A Novel Knowledge Discovery based approach for Supplier Risk Scoring with application in the HVAC Industry

Bilal Akbar Chuddher

Thesis submitted for the Degree of
Doctor of philosophy (PhD)

Brunel University London

February, 2015

Abstract

This research has led to a novel methodology for assessment and quantification of supply risks in the supply chain. The research has built on advanced Knowledge Discovery techniques and has resulted to a software implementation to be able to do so. The methodology developed and presented here resembles the well-known consumer credit scoring methods as it leads to a similar metric, or score, for assessing a supplier's reliability and risk of conducting business with that supplier. However, the focus is on a wide range of operational metrics rather than just financial, which credit scoring techniques typically focus on.

The core of the methodology comprises the application of Knowledge Discovery techniques to extract the likelihood of possible risks from within a range of available datasets. In combination with cross-impact analysis, those datasets are examined for establish the inter-relationships and mutual connections among several factors that are likely contribute to risks associated with particular suppliers. This approach is called conjugation analysis. The resulting parameters become the inputs into a logistic regression which leads to a risk scoring model the outcome of the process is the standardized risk score which is analogous to the well-known consumer risk scoring model, better known as FICO score.

The proposed methodology has been applied to an Air Conditioning manufacturing company. Two models have been developed. The first identifies the supply risks based on the data about purchase orders and selected risk factors. With this model the likelihoods of delivery failures, quality failures and cost failures are obtained. The second model built on the first one but also used the actual data about the performance of supplier to identify risks of conducting business with particular suppliers. Its target was to provide quantitative measures of an individual supplier's risk level.

The supplier risk scoring model is tested on the data acquired from the company for its performance analysis. The supplier risk scoring model achieved 86.2% accuracy, while the area under curve (AUC) was 0.863. The AUC curve is much higher than required model's validity threshold value of 0.5. It represents developed model's validity and reliability for future data. The numerical studies conducted with real-life datasets have demonstrated the effectiveness of the proposed methodology and system as well as its future potential for industrial adoption.

Dedicated
to
My Parents and Family

Acknowledgements

My utmost gratitude goes to my supervisor, Dr Harris (Charalampos) Makatsoris, for his direction, encouragement and support throughout the research and the writing process. Without his continuous support and guideline, it was not possible for me to conduct this research.

Special thanks to my father for his continuous encouragement, support, and guidance. In addition, I would like to thank my mother for her patience and prayers because I could not have achieved anything, whatever I have, without it.

Finally I would like to thank all family members and my friends, who helped me according to best of their abilities.

Table of Content

List of Tables

List of Figures

1 INTRODUCTION

1.1 Background

Today's competitive business environment which is characterised by short product lifecycles, increased complexity of products and services, globalised nature of markets, rapidly changing consumers' behaviour and dynamic demand have forced the firms to focus on their core competencies (Weele and Rozemeijer, 1996). The philosophy of focusing on company's core competencies resulted in reduction of internal value added activities, increased reliance on outsourcing and networking for required competencies that shifted competition from single company vs single company to supply chain vs supply chain (Lambert and Cooper, 2000). The increased dependency on outsourcing and supply network has exposed the companies to supplier's risk profile (Zsidisin and Ellram, 2003). In recent years, companies have also implemented the cost cutting methodology such as lean management, centralized production and distribution and single source etc. These efforts result in reduction of redundant stock and operational slacks that may worsen the risk's consequences (Tang and Musa, 2011). The situation of companies' exposure to risk becomes even terrible, because in current globalized business environment companies' supply chain structures are very complex and are very vulnerable to political crises, natural disasters and the dynamics of the market place (Braunscheidel and Suresh, 2009). Regardless of risk drivers (as some mentioned above), the occurrence of risk in supply chain has significant negative impact on company's operational and financial performance (Hendricks and Singhal 2005a,b).

A few examples of risk cases affecting supply chains are given in Table 1.1, clearly showing that companies' increasing reliance on supply networks leads to more exposure to risks in supply chain. Over the last 10 years, the increasing numbers of research studies on supply chain risks have proven that the issue of risk in supply chain operations has become a focal research agenda for supply chain operation researchers (Kleindorfer and Saad, 2005; Narasimhan and Talluri, 2009). Gartner Inc., 2011 report ("predicting supply chains for 2012") indicated that companies are now more concerned about risk in supply chain and emphasize the need of quantitative risk assessment and management. Companies are also showing increased interest in utilizing advanced technology for efficient risk management in supply chains.

Table 1.1: Examples of different supply chain risks and their impact

Year	Description	Source
1997	Boeing faced the estimated loss of $2.6 billion due to supplier's failure to deliver two crucial parts	Radjou, 2002
2000	Ericsson has single-sourcing policy; a fire accident at location of its sole chips' supplier disrupted the supply that cost about 400 million US dollars loss to Ericsson.	Norrman and Jansson, 2004
2001	Land Rover experienced difficulties in production of new model because one of main supplier for chassis filed for bankruptcy.	Sheffi, 2005
2002	Union strike at west coast port disrupted the normal operational and some container are delivered six months late than their schedule delivery date	Cavinato, 2004
2007	World leanest car manufacturer Toyota halted entire production in Japan due to an earthquake that severely damaged its major supplier for piston and seal rings. Other companies such as Mitsubishi Motor Corporation, Suzuki Motor Corporation and Honda Motor Corporation also suspended their production as they also had same supplier.	Hayashi et al., 2007
2008	Volvo Cars received 28% less revenue as compared to last year in same period because of the weak dollar that reduced the revenue and it also reduced further opportunities for R&D.	Tang and Musa, 2011
2010	The eruption of volcano in Iceland disrupted the Airline flights' schedule resulting disruption of global supply chains	Field, 2013
2013	Horsemeat found in beef burgers on sale in UK and Ireland forced the Tesco immediately withdrew sale of all products from the supplier in question	BBC (15 January 2013 Last updated at 22:24)

The examples of risks in supply chain context shown above along with many others available in previous literatures raised the question that "*how can risk be efficiently assessed in the supply chain context to achieve high operational and financial performance*", intrigue the current research. It is common notion that efficient risk assessment is entirely dependent on excellent prior risk identification, and that these activities should be performed in sequence to deliver real benefits for companies (Kern et al. 2012, Berg et al., 2008; Craighead et al., 2007; Zsidisin et al., 2000), build the focus of this research thesis.

For effective risk assessment the accurate knowledge about the underlying factors affecting the performance of a company at a given time is extremely important. The supply risk identification process provides the required knowledge (i.e. factors affecting the performance) for risk assessment. Despite the recognised importance of

supply risk identification in the ways it affects risk assessment, it is currently under-represented in existing risk assessment models (Matook, et. al., 2009). This is mainly due to the dependency on the quality and the accuracy of the available knowledge (expert view) about the factors affecting the performance of supply chain (Behdani et al, 2012; Adhitya et al, 2009). Furthermore data modelling difficulties related to the uncertainties associated with wide variety of supply chain factors and their interactions that may or may not be affecting the supply performance at a given time is another reason (Thun et. al., 2011; Dani, 2009). In the majority of the cases the incorporation of supply risk identification in supplier risk assessment is based on an expert's knowledge and experience (Ghadge et al, 2013; Behdani et al, 2012). This highlights the need for new more efficient and accurate approaches for risk identification that can overcome the highlighted drawbacks. An alternative approach for supply risk identification can be entirely data-driven where a priori assumptions about the role of supply chain factors are not necessary. Knowledge discovery approach is data driven and designed to identify unknown patterns and relationships in the data that may exist, is used in this research thesis for supply risk identification.

Knowledge discovery is the non-trivial process of discovering of valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et. al., 1996 a). Although knowledge discovery is quite a well-established area for risk identification in financial and other sectors (see section 3.3.1), its application for risk identification in supply management context is a new but very promising area of research. The complexity of uncertainties associated with supply chain characteristics (risk factors), along with the large size of factors affecting the supply performance not only justifies the application of knowledge discovery in a supply risk identification but also make it highly attractive.

In risk assessment process the overall impact of identified supply risks on supply performance is assessed to aid decisions making for risk mitigation and monitoring. The supply risk assessment procedure that is focused on individual suppliers in order to calculate the overall impact of supply risks is termed as supplier risk assessment. In previous studies both qualitative/semi-quantitative and quantitative approaches have been employed for supplier risk assessment. Although these approaches contribute toward supplier risk assessment, they have some limitation. Such as the qualitative/semi-quantitative approaches depends on the expert knowledge, which can be biased (Tazelaar and Snijders, 2013). Furthermore, the quantitative approaches

such as simulations and modelling used prior assumed descriptive criteria about supply risk for assessing suppliers and provide the result according to described criteria. The lack of inclusion of changing knowledge initiative makes these approaches reactive in nature to predict the probability of supplier risk and reduce the effect of supply risk in supplier risk assessment (Thun et. al., 2011). However, current supply chain environment is very dynamic in nature and require inclusion of changing knowledge initiative for supplier assessment (Zsidisin et. al., 2004; Chopra and Sodhi, 2004; Kleindorfer and Saad, 2005; Manuj and Mentzer, 2008). This highlighted that there is still a need for new, more efficient and accurate knowledge discovery based risk assessment models (Matook, et. al., 2009). Those can have the ability to include the changing knowledge initiative about supply risk and proactively predict the probability of supplier risk. Risk scoring approach in combination with data-driven supply risk identification approach can be an appropriate solution for supplier risk assessment, that will have the ability to predict the supplier risk by include the changing knowledge initiative about supply risk.

Risk scoring is well known risk assessment approach used by lending organization to assess overall risk of individual customer. Risk scoring requires selection of appropriate factors (i.e. strong predictor of defined risk), historical data and data modelling techniques for building a risk scoring model. Adoption of such an approach into supplier risk assessment presents a number of challenges. First, the biggest challenge in the risk scoring model is the selection of the right parameters, which have strong relationship with the dependent variable i.e. defined risk. Second the transformation of numerical variables into categorical variables, which can reduce the overall redundancy of available data. These issues are handled by proposing a simple attribute and discretization method based on the knowledge discovery about supply risk (see detail in section 4.1.3.1).

1.2 Research Aim and Objectives

Current research thesis is aiming at **"the *development of an integrated Knowledge discovery based approach and system for supplier risk assessment that provide the supplier risk score analogy to the famous credit score used for assessing lending consumers creditworthiness*".** First, a snapshot of supply chain risk management (SCRM) is taken to position the current research and define the risk in this research context. Then, methods for effective risk identification and assessment are identified

and implemented to crop the desired supplier risk assessment approach. In order to do this, the supplier risk assessment system is divided into subsystems based on the supply risk assessment processes of defining the risk, identifying risk and assessing the risk. The development of long-term relationship between suppliers and buyers depend upon the actual performance (Mohanty and Gahan, 2012). Therefore, we believe that effective supply risk identification based on actual supply performance can be a comprehensive way of providing feedback for the development of supplier risk scoring model.

A research framework is developed according to these subsystems, as shown in Figure 1.1. This figure shows that the supply performance and supply chain characteristics (structure, culture, and process) can be used to define the risk. A solid risk identification process could identify the impact of supply chain characteristics on supply performance. This could be established by identifying the hidden knowledge about relationships between supply performance and supply chain characteristics. For example, relationship between production process and performance can reveal how the production process in a given supply chain is impacting the required performance. With a proper implementation of risk scoring process, the final supplier risk score model can be cropped for supplier risk assessment.

Based on the research framework, the following research objectives and Questions (research questions are written as RQ) are developed for the current research thesis.

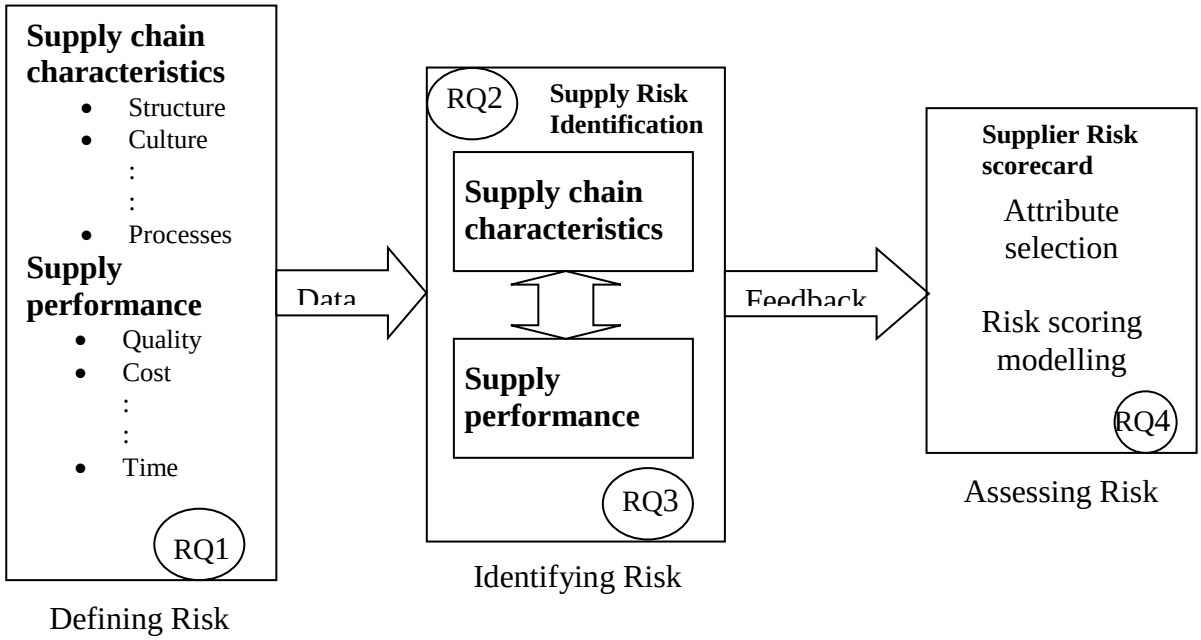


Figure 1.1: A research framework for supplier risk assessment approach

Objective I: The determination of risk factors that have direct impact to supply and defining the supplier risk that account for supply risk explicitly

It is essential to understand the current agenda in field of supply risk management to position the current research thesis in the field of supply chain risk management. The discovery of gaps in literature will identify the research opportunities in this field. The exploration of different definitions, terminologies and processes involved in this field, further help to clarify the scope of current research and provide the ground for defining the risk in the current research perspective. Risk has been examined in various management fields and has justified its significance in decision making for different business management functions. Different definitions of risk normally inherent three dimensions: uncertainty of outcome, expected outcome, and potential outcome (Sitkin and Pablo, 1992). Risk may be seen in an analogous manner within context of supply management. For example, there is uncertainty of outcome associated with a supplier's ability to make the product according to the desired design and specification changes within the specified time (Bidault et al.,1998). Scholars have addressed the issue of risk within supply management context i.e. supply risk, and concluded that supply risk has multi-dimensional construct depending upon the purchasing organizations' perception (Zsidisin, 2003a).The current research is dealing with risk in context of supply management for developing supplier risk assessment approach, therefore, first question raised is:

RQ1: What are the different dimensions of supply risk and how can it be used for supplier risk assessment, when measured on actual supply performance?

Objective II: The development of a Supply Risk Identification approach based on Pattern Discovery

The second research objective focuses on finding how supply risk can be identified effectively by an approach that can be a viable alternative to traditional approaches employed for supply risk identification. Most of the traditional supply risk identification approaches exclusively depended on experts' perception and experience such as expert interviews, survey or brainstorming etc. or partially depended on experts' perception and experience such as AHP, fish bone diagram etc (Behdani et. al., 2012). However this study is focused on pattern discovery in the available data for supply risk identification rather than experts' perception and experience. To achieve

this objective, Knowledge discovery approach for pattern discovery is selected and analysed for its competency and robustness in identifying supply risk. The implementation of Knowledge discovery approach has two main requirements: appropriate data and identification of a suitable data mining algorithm for available data.

Current study uses the data about supply chain characteristics and supply performance. The data about supply performance is purely obtained from specific company; however data about supply chain characteristics is available from both publically available data sources and company specific data. For example, the data regarding company's purchasing strategy (one of supply chain characteristic) is obtained from specific company, however the data about supply chain operating environmental characteristics such as occurrence of natural disaster during the purchase cycle time is obtain from publically available data source. Further, different data mining algorithms are available; these can be used for pattern discovery in available data, singling out the most suitable data mining algorithm for available data is tricky problem (Pechenizkiy et. al., 2005). This raised the following research question

RQ2: Could the publically available data be combined with company specific data into a dataset to be used for supply risk identification and which algorithm is the most suitable data mining algorithm for available data in supply risk identification?

RQ3: What knowledge can be extracted from the available data and which combinations of supply chain characteristics best identify the supply risk?

Objective III: The design of a Supply Risk Scoring Methodology based on Knowledge discovery

This research thesis is focused on the development of a supplier risk assessment approach that can provide the supplier risk score. To obtain the supplier risk score, a supplier risk scoring model is developed that is entirely driven by feedback obtained from supply risk identification. Therefore the natural question arises:

RQ4: Could such a supply risk-aware methodology for supplier risk scoring model development add any value against the state of art variable selection and discretization approaches?

Objective IV: The development and implementation of a prototype software system for the evaluation of the risk scoring methodology

The proposed supplier risk assessment approach is used to develop a prototype system. This prototype system will assist in the demonstration of the proposed approach.

Objective V: The validation of the methodology through a case study from the HVAC industry

The proposed supplier risk assessment methodology and system is tested on real data of HVAC industry. This application will allow the evaluation of the proposed methodology's practicality from an industrial perspective.

1.3 Structure of the Thesis

This section provides a guide through this thesis, which consist of seven chapters and is organized as follows.

Chapter 2 reviews previous studies related to supply chain risk management and in particular supplier risk assessment. First, some general concepts about risk in supply chain are presented, followed by a review of the current risk assessment process and techniques. This chapter provide the understanding about the field of study and identified the relevant research gap.

Chapter 3 presents a review of the knowledge discovery and risk scoring modelling. First part of the chapter descries the detail of knowledge discovery process, while the second part explains the risk scoring process in detail. Third part of the chapter described the evaluation methods used for assessing the validity of the knowledge discovery process and risk scoring. This chapter provided: (1) the understanding of knowledge discovery process implementation, and (2) the understanding of risk scoring model process implementation.

Chapter 4 presents a novel approach for supplier risk assessment, which consists of three parts. First part deals with defining the supply risk measurements and supplier risk, the second part, explain about proposed rule-based method for supply risk identification. Finally, the third part provides risk scoring modelling approach for developing supplier risk score.

Chapter 5 focuses on providing the foundation for evaluation of the proposed methodology. In the first section the system design is explained. The section two explains the background of the case study company is given. In the final section of this chapter, the initial experimental results are provided to progress toward answering the research questions.

Chapter 6 provides the description and analysis of the results. First section provides results and discussion about the supply risk identification model and second part deal with the supplier risk scoring model.

Chapter 7 outlines the summary and conclusion together with implication and limitation of the study.

2 SUPPLIER RISK ASSESSMENT

Supply chain management and risk management are two complex but broadly separate fields of study. Supply chain risk management combines both these management elements into a much wider and more diverse field of study. It is important, therefore, to understand the premises of this subject by examining the theory, paradigms, methodology and policies that have been adopted in this field.
The objective of this chapter is to provide the theoretical background of supply chain risk. Furthermore, the premises of supply chain risk are explained to identify areas of interest for research. Finally, the paradigm of supply risk management is examined and by reviewing various methodologies presented in the literature that address the identification and assessment of risk in supply chain context.

The first question to ask in risk management is "what is risk?". Therefore, in the following initial section relevant definitions, terms and perceptions of risk are presented to provide an identity to risk according to current study.

2.1 Risk and Uncertainty

From the beginning of the last century, risk has been investigated in diverse fields of literature, from economics to finance, engineering, strategic management, international management, operational research and, more recently in supply chain (Jüttner et al., 2003; Tang, 2006). Despite this, there is no general consensus about the nature, perception or definition of risk. Instead there exist several definitions and concepts of risk, according to the field of study and focus of research. This diversity further presents considerable ambiguity between risk and uncertainty.
Samson et al. (2009) stated that there is no general definition for these two terms but rather many definitions dependent on discipline and context. Different definitions and relationships between risk and uncertainty have been identified in economics, finance, operations management and engineering. Some authors, especially in the areas of economics and finance, consider uncertainty and risk as synonymous. In contrast, there are many scholars, especially in the fields of engineering and operations management, who believe that uncertainty and risk are different concepts, but they are not agreed on their relationship. Some state that they are independent concepts while

others believe they are dependent. Furthermore, those that believe uncertainty and risk are related to each other also divide into two groups. Some authors consider that risk depends on uncertainty and others that uncertainty depends on risk. As the current research thesis is in the field of engineering and operational management, In this case, the viewpoint of engineering and operations management seems more appropriate in terms of current research.

2.1.1 Uncertainty

Uncertainty has been defined as the indefiniteness or variance of an event. It reflects the potential of a favourable or unfavourable occurrence falling to the left or right of a mean or median value. Uncertainty allows the possibility of listing potential future events, but without offering conclusions about which will actually happen. Zimmermann (2000) provides an exhaustive classification of the causes or causal pathways of uncertainty, as indicated below:

- Lack of information or knowledge: this is possibly the commonest cause for uncertainty;
- Excess information (complexity): this produces uncertainty due to the limitations in our capacity to observe and process large amounts of data simultaneously;
- Conflicting evidence: this produces uncertainty due to information available being incorrect but not identifiably so, or information is received from non-relevant sources;
- Ambiguity: this causes uncertainty when information has different meanings, depending on the context;
- Measurement: this uncertainty occurs when there is some uncertainty about a measure when the only measure known is indicated by the measurement tool.

2.1.2 Risk

Despite the lack of consensus in definition, risk is generally thought of as reflecting uncertainty in the scope of potential outcomes, the chances of these outcomes occurring, and a subjective valuation of their impact (March and Shapira, 1987). Risk in literature of engineering and operations management is defined as the result of an event's probability and an estimate of the expected consequences if the event occurs. Based on this definition (equation 2.1), risk has two components: the probability or likelihood of an event outcome, and the consequences of that outcome.

$$\text{Risk} = \text{An event's probability} \times \text{Consequence} \qquad (2.1)$$

However, study of the corresponding literature reveals disagreement on the nature of risk. In particular, the consequences or output of risk give rise to two dimensions (Mitchell, 1995). First, risk is conceived as a random instability (variability) associated with an expected value (mean) of an outcome (e.g. Arrow, 1965). In this view, risk is synonymous with variance and has the potential for both a downside (loss) and an upside (profit, opportunity). Second, in contrast to the above, dictionaries generally define risk simply as a threat of injury, damage or loss (McKechnie, 1983). Furthermore, in some fields, a distinction is made between risk and threat. A threat is a low probability event with high negative outcomes, but its probability is difficult to assess. A risk, in contrast, is a higher probability event, but assessment is possible of both the probability and the consequences.

Based on the above distinctions, this research work considers risk and uncertainty as two distinct but related concepts where risk depends on uncertainty as shown in Figure 2.1. Furthermore, the concept that risk carries mainly negative consequences is accepted as being more commonly held than the negative/positive viewpoint (March and Shapira, 1987). Therefore, this study considers the negative perception of risk, where both probability and the negative outcomes can be captured. As the objective of this study is to provide a quantitative measure of supplier risk, this consideration seems appropriate.

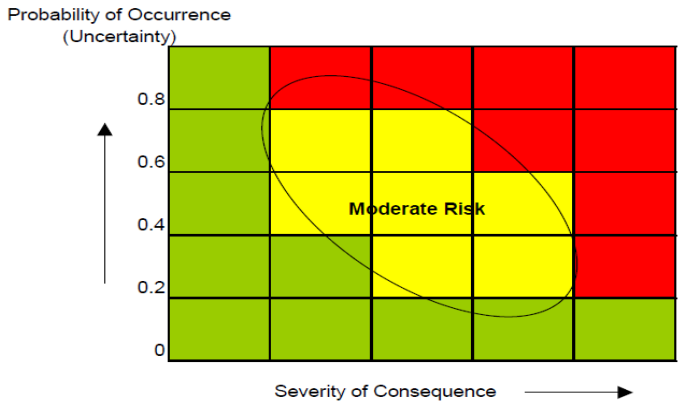


Figure 2.1: The Risk Matrix

The previous section has explained the nature of risk in different fields of study and the relationship between uncertainty and risk. The following section will explain risk in the field of supply chain management and in particularly supply risk.

2.2 Supply Risk

A supply chain focuses on the material that flows between supply chain members, but as a broader concept can also include other flows, such as financial and information flows within the supply chain as shown in figure 2.2. As in other fields of study, risk also inherent in supply chain management, further, risk in supply chain context can be divided into different classifications (Tang, 2006). Christopher and Peck (2004) provided three categories of risks in supply chain: "internal to the firm", "external to the firm but internal to the supply chain network" and "external to the network". Risks that were internal to the firm were further divided into "process risks" and "control risks". Risks external to the firm but internal to the supply chain network were further classified as "supply side" (upstream) and "demand side" (downstream). Similarly, Thun and Hoenig (2009) distinguish between "internal to company" and "cross-company" risks. The cross-company risks consist of "purchasing risks" (upstream) and "demand risks" (downstream). Many other authors has also divided the supply chain risk into "supply" and "demand" risks perspective (Wagner and Bode, 2006; Sodhi and Lee, 2007; Tang and Tomlin, 2008). According to supply chain structure that normally consists of supplier networks-main company-customers (see figure 2.2), afore mention division of supply chain risks into supply" and "demand" risks seems very appropriate. It clearly distinguishes the role of different agents in the supply chain, their interaction and the location of relevant risk in supply chain. Supply chain risks associated with supply side (upstream) activities and members of a supply chain termed as supply risk, while Supply chain risks associated with downstream activities and members of the supply chain are termed as demand risk.

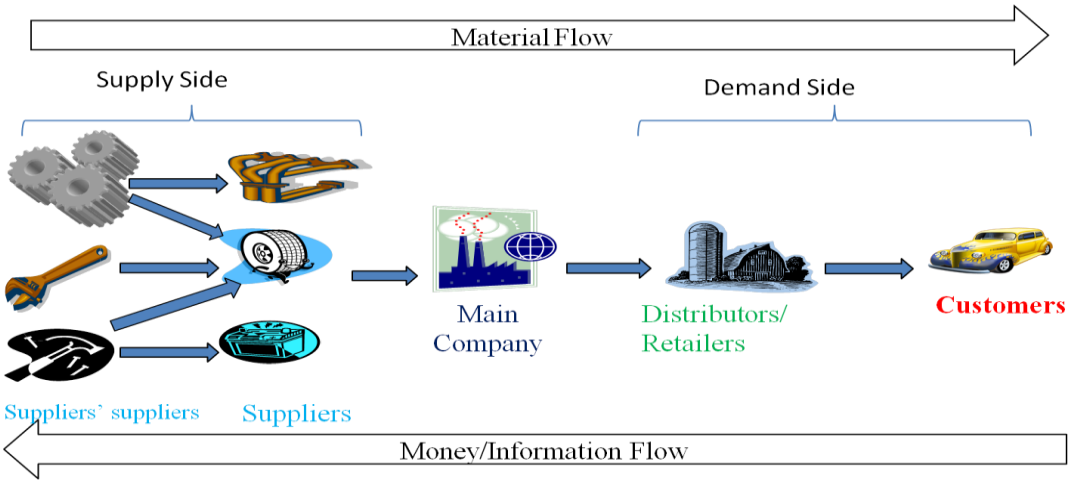


Figure 2.2: A general supply chain structure

Effective management of supply chain risks (either supply risk or demand risk) is highly desirable for company high operational and financial performance (Hendricks and Singhal 2005 b). Research to date has proposed various, frameworks, models and tools for the management of supply chain risk. However, supply and demand risk management initiatives have their own requirements and procedures, which can differ from each other. Tang's (2006) study about the risk management initiatives in supply chain also distinguishes between supply side and demand side management initiatives. Previous research has shown that there is a clear distinction between supply side and demand side risk management processes, with each requiring different forms of measurement and control (Kouvelis et al., 2006; Manuj and Mentzer, 2008; Wagner and Bode, 2008; Tang, 2006). This study focuses on the supply side of supply chain risk management, which may also be called supply risk management or upstream supply chain risk management (Wagner and Bode, 2006; Zsidisin et al., 2004), in order to ensure a clear focus of the study and its measurement items. It is believed that analysing the supply chain system by parts i.e. analysing upstream supply chain separately and downstream supply chain separately can be a comprehensive way of dealing with the complicated issue of risk in supply chains.

Typically, a supply system comprises of many tiers of suppliers, and involved suppliers at each level providing goods or services to suppliers at the next level of the supply chain. In such networks, there is rarely a linear flow of goods due to multiple parts or suppliers interlinking at each level, which results in uncertainty (Riddalls et. al., 2000). Uncertainty can also exist due to changes in upstream supply network

operating environment such as markets, technologies, competition, politics and statutory regulations (Wagner and Bode, 2006). Failure to manage supply risk can be very costly (see example in Table 1.1) and cause extreme delays in delivery to customers (Martha and Subbakrishna, 2002). Therefore, an organization's success relies on understanding of supply risk and managing them effectively (Harland et al., 2003; Roth et al., 2008; Tang and Tomlin, 2008; Wagner and Bode, 2008).

2.2.1 Supply Risk Definition

Similar to risk in other fields of study, risk in supply chain context is also vague in concept. Many other terms, such as vulnerability or disruption, are often used synonymously to describe risk in supply chain. Previous literature includes several different definitions of risk and the related concepts of supply chain disruption and vulnerability.

Peck (2005) refers to risk in supply chain context as anything that could disrupt the flow of information, material or product in the supply chain. This disruption may create a mismatch between supply and demand, affecting either cost or quality, that is, a divergence from target values. Wagner and Bode (2008) view risk as a negative change to an expected performance measure, which results in failure to meet necessary demand, delivery deadlines or pre-determined costs, etc. However, they consider supply chain disruptions to be low probability but high impact events, that is, anomalous events occurring in the supply chain that significantly threaten normal supply chain operations. In this description supply risk and disruption are equated with risk and threat. Hou et al. (2010) agree and define supply disruption as the unexpected non-availability of supply due to an unexpected event that makes one or more source of supply unavailable. Chopra and Sodhi (2004) and Craighead et al. (2007) consider supply risk and disruption to be the same in upstream supply chain context. Vulnerability is also an important and closely related concept. Christopher and Peck (2005) describe supply chain vulnerability as exposure to a significant disturbance arising from risks to the supply chain. Svensson (2000) defines vulnerability as a condition that affects a company in achieving its goals.

The supply risk also has different dimensions either positive or negative similar to risk in different field of study (Wagner and Bode, 2008). For example, Jüttner et al. (2003) considered risk as changes in the distribution of potential supply chains outcomes, their likelihood and their subjective values. However, contrasting with the mean-

variance approach, several other scholarly definitions focus solely on the downside of supply risk. Harland et al. (2003) state that supply risk is associated with the undesired consequences of alterations in danger, damage, loss, injury or other such negative impacts. Table 2.1 below reports some definitions provided in literature concerning risk in upstream supply chain context along with the risk perception approach they follow.

Table 2.1: Supply risk definition and their dimensions

Author	**Risk dimension**	**Definition**
Zsidisin et al. (1999)	Negative	Significant and/or disappointing failures with inbound goods and services.
Norrman and Lindroth (2002)	Negative and positive	Anything that can affect the normal flow of material and information between supplier and customer.
Jüttner et al. (2003)	Negative and positive	Variation in the distribution of possible supply chain outcomes, their likelihood's, and their subjective values.
Harland et al. (2003)	Negative	Adverse effects on inward flow of any type of resource that enables operations to take place; also termed 'input risk'.
Zsidisin, (2003)	Negative	The probability of an incident associated with inbound supply from individual suppliers or the supply market, in which the outcome results in the inability of the purchasing firm to meet customer demand or causes threats to customer life and safety.
Bogataj and Bogataj (2007)	Negative	Potential variation of outcomes that decrease the value added at any activity cell in a chain, where the outcome is described by the volume and quality of goods in any location and time in a supply chain flow.
Manuj and Mentzer (2008)	Negative	The distribution of outcomes related to adverse events in upstream supply that affect the ability of the focal firm to meet customer demand (in terms of both quantity and quality) within anticipated costs and time, or causes threats to customer life and safety.

As it can been seen, most previous definitions portrait the negative dimension of risks in upstream supply chain context, therefore, this work focuses on the negative outcomes of a possible risk event. Further, almost all the above definitions of supply risk have two constructs: source of supply risk and outcomes of supply risk. The source of supply risk is uncertainty associated with the upstream supply chain characteristics either individual supplier characteristics, the supply network environment or market factors. The outcome of supply risk is the effect of uncertainty

associated with the upstream supply chain. Nevertheless, the sources of supply risk and the outcomes are varied according to context and industry (Zsidisin, 2003a). Furthermore, most of the previous research is qualitative largely relying on perception of the mangers. Wacker (2004) stated that a good definition is "concise, clear verbal expression of a unique concept that can be used for strict empirical testing" (p. 631). Therefore, current study defines the supply risk in context of supply chain management that can offer a clear description letting alone the analytical measures.

Based on the findings about supply risk the **proposition of current research about supply risk** is build that: *"An uncertainty associated with upstream supply chain characteristics, the results of which affect desired outcome (supply performance) negatively is considered as supply risk".*

In the previous sections an overview of supply risk and definition of supply risk has been provided. The following section will expand on the supply risk management procedure and the role of risk assessment within the supply risk management process.

2.3 Supply Risk Management

This study focuses on the supply side of supply chain risk management, generally referred as supply risk management (Wagner and Bode, 2006; Zsidisin et al., 2004), by adapting definitions of supply chain risk management from, Manuj and Mentzer (2008) and Tang (2006). Supply risk management can be defined as:

> *The recognition and assessment of supply risks and resultant losses in the supply base, and execution of suitable strategies through a synchronized and combined approach with the intent of mitigating one or more of the following – financial or product losses, speed of losses, probability of events, speed of events, detection time of events and frequency of events to ensure continuity and profitability of supply chain.*

From the above definition it can be seen that supply risk management involves supply risk identification (i.e. recognition), assessment, and the introduction of a strategy to reduce the effects of supply risk (i.e. mitigation) in order to maintain profitability and continuity of supply chain. The supply chain risk management literature shows that

there are different frameworks to support the supply risk management process. Although differences exist in each framework concerning methodology and the number of stages involved, in general the supply risk management process can be divided into four phases (Hallikas et al., 2004; Kleindorfer and Saad, 2005; Wagner and Bode, 2009). These four common phases are: (i) risk identification: to provide an organization with recognition of the factors affect negative effect on supply chain, (ii) risk assessment: to evaluate the probability of a risk event and its potential impact, (iii) risk mitigation: to determine and implement actions directed at minimising the probability of disruptive events and/or their impact, and (iv) risk monitoring: to confirm that risks are properly identified and assessed and that suitable controls and responses are established. In Table 2.2 the steps constituting different supply risk management frameworks are described and classified according to the above four phases. Further general definitions of risk identification, assessment, management and monitoring are provided.

The literature on supply risk management procedure reveals a variety of frameworks underpinning the supply risk management process. While differences in procedure exist, however, there is agreement that risk assessment is at the centre of the entire supply risk management process. There is further agreement that each stage of the supply risk management process is interdependent and requires active feedback from previous stage of the supply risk management framework (Kern et al., 2012, Manuj and Mentzer, 2008). This implies that decisions taken in one phase of the process will directly or indirectly influence other phases. Thus, accurate supply risk assessment requires early and precise risk identification, and actions need to be taken in sequence by companies for satisfactory benefits to be derived (Kern et al., 2012, Berg et al., 2008; Craighead et al., 2007; Zsidisin et al., 2000).

Table 2.2: Supply risk management frameworks

Stage	Harland et al. (2003)	Hallikas et al. (2004)	Matook et al. (2008)	Manuj and Mentzer (2008)	Description
Risk Identification	1-Map supply network; 2-Identify risk and its current location	Risk identification	Supplier risk identification	Risk identification	Identification of all characteristics of supply chain; Definition and categorization of all the risk factors; Identification of all associated volatility related to these supply chain characteristics.
Risk Assessment	Assess risk	Risk assessment	Supplier risk assessment reporting and decision of supplier risks	Risk assessment and evaluation	Estimation of the probability of occurrence and the possible consequences (quantitative, semi-quantitative or qualitative) displayed of identified risks in a probability and consequences matrix after assessment.
Risk Mitigation	Manage risk; Form collaborative supply network risk strategy	Risk management action	Supplier risk management responses	Risk management strategies	Selection and implementation of strategies to control risk.
Risk Monitoring	Implement supply network risk strategy	Risk monitoring	Supplier risk performance outcomes	Implementation of strategies mitigation	Review of performance to identify opportunities for improvement; Execution of regular audits of policy and standards compliance and feed-back to whole system.

Supply risk assessment can be focused on individual product (Kaljic, 1983) individual supplier (Wu et.al. 2006, Blackhurst, et. al., 2008) or supply network (Hallikas, 2003). A survey conducted by AMR Research (O'Marah, 2009) showed that perception of supply risks in most companies had been changing in a sensible direction. Previously the main concerns had concentrated particularly on transportation costs (due to rising fuel expenses) and increases in commodity prices. However, the survey found that now supplier failure is the most apparent issue for companies. The results showed that supplier failure due to various supply risks was rated by 38% of respondents as the main issue that affects a company performance. Furthermore, an extensive analysis by Thun and Hoenig (2009) of supply risk management in 67 German car manufacturing plants, in terms of probability and impact of risks on the supply chain, showed that supplier quality problems were acknowledged as the most critical risk. Analogously, the most severe problem carrying the highest impact on the supply performance was perceived as supplier failure due different supply risks is the focus of current research thesis. The application of supply risk assessment procedures to determine the individual suppliers' overall risk is termed as supplier risk assessment.

2.4 Supplier Risk Assessment

Supplier risk assessment is an ordered process (Sinha et al., 2004) and the most comprehensive element of a supply risk management system (Wu et al., 2006). Based on a review of the literature, the supplier risk assessment process can be classified into four components: recording and categorising the risk factor, identifying the supply risk, supplier risks assessment, and implementation (Sinha et al., 2004; Wu, et. al., 2006; Costantino and Pellegrino, 2010; Blackhurst, et. al., 2008). A literature review in supply risk management concerning the supplier risk assessment stages is conducted to identifying the research gaps required to be filled is presented in following sections. The supplier risk assessment procedure can be divided into three stages, i.e. listing and categorising the risk factor, risk identification and risks assessment.

2.4.1 Listing and Categorising of Risk Factors

The main purpose of listing and categorising risk factors is to identify and classify a comprehensive range of factors that could be risk sources and then to target them in a risk identification process. Craighead et al. (2007) argue that the time involved in predicting or discovering a supply risk can affect the severity of supply risk. This

implies that companies need to ensure that they have supply risk sources (i.e. risk factors) register by careful prior scanning of supply chain and its environment. That can enable companies to identify the supply risks early and corrective measures promptly begin (Craighead et al., 2007; Tomlin, 2006; Zsidisin et al., 2004). However, normal constraints on company resources means that supply risks need to be identified in the most efficient way possible. A haphazard search for supply risks is an inefficient use of resources, thus observation fields also need to be defined to most efficiently locate vulnerabilities and potential sources of risk. It has been noted that listing and categorising risk factors through established risk awareness protocols enables an efficient risk identification process within restricted resources (Hallikas et al., 2002; Kleindorfer and Saad, 2005; Stecke and Kumar 2009). Listing and categorization enables the introduction of key protective measures to be clearly understood and helps the focus on collecting suitable data for future risk identification and assessment processes (Harland et al., 2003). Systems of risk factor classification can address a particular supply chain (Harland et al., 2003) or can be chain independent by focusing on a single supplier or product (Sinha et al., 2004, Wu et.al. 2006).

Previous literature has identified many sources of supply risk, i.e. uncertainty related to particular characteristics of the upstream supply chain. Kraljic (1983) was an early proponent of recognising that the supply portfolio can be a source of supply risk and therefore requires proper management. Zsidisin (2003a) provided a comprehensive list of characteristics that affect supply risk perception of supply managers. In a study of supply risk perceptions, he began by identifying risk factors (supply risk sources) through a literature review and subsequently produced a classification of risk factors based on feedback from purchasing organizations. Zsidisin (2003a)'s study suggested that product and supplier characteristics were not the only categories that supply managers perceived with supply risk, but they also included wider characteristics of the supply market. Based on previous studies, Ho et al.(2005) listed many factors within the purchasing process that represent supply uncertainties related to critical material supplies, such as frequency of changing suppliers, complexity of materials, complexity of procurement technology, time specificity of materials procurement, delivery frequency, delay of delivery and fluctuations in the selling price.

Rao and Goldsby (2009) argue that environmental factors are also key sources of supply risk in the upstream supply chain, together with both individual organization

and market characteristics. Trkman and McCormack (2009) pointed out that knowledge of the political climate is important if working with overseas suppliers, as conducting business where there is political unrest is impractical. Regulatory, statutory and bureaucratic policies are important factors in supply risk, as well as their frequency of change (Wagner and Bode, 2008). Uncertainties in nature (earthquake, floods, etc.) and the social sphere (terrorism, strikes, etc.) are also significant issues for managers to consider when developing sourcing strategies (Sheffi, 2001; Norman and Jansson, 2004).

Further, market and environmental characteristics are considered to increase supply risks when suppliers are located in natural disaster regions or in markets susceptible to fluctuating currency rates, volume constraints, or price inflation; or where availability of qualified suppliers is limited. Supplier characteristics create perceptions of higher supply risk when they offer limited capacity, cost inflexibility, incompatible information systems, lower quality, unpredictable cycle times that result in delivery delay, high delivery frequency and inflexibility of response to volume or mix demands. Finally, the internal procurement process raises supply uncertainty in cases where procurement technologies are highly complex and a company is dependent on its business protocols for maintaining critical material supply. Table 2.3 lists examples of the risk factors identified from the literature review; however, this table is by no mean exhaustive.

The sources of supply risk are widespread, therefore, it is important for a company to execute a deep and well-organized review of uncertainties in the upstream supply chain, and then produce a risk register or risk portfolio. Clearly, it is impossible to list every risk source, however, using approaches such as a literature review, group discussions, brain storming or surveys a satisfactory risk register can be compiled and these factors can be categorised. Some authors have attempted to compile general categorisations of these risk factors. However, there are no rules for compiling such listings and categorisations, therefore the format will depend on specific contexts, type of industry, available resources and the company management's judgment. This study will adopt the taxonomy and brainstorming approach, to identify possible risk factors and their categorization.

Table 2.3: Risk factors involved in supply risk

Author	Risk Factors
Vlajic, et. al. (2012)	Price fluctuation, exchange rate, regional economics, market capacity, legal requirements, infrastructure, political unrest, criminal acts, public reputation, natural disasters, man-made disasters, biological disasters, product quality, supply network complexity, sourcing type, weak alternatives, production capacity constraints, equipment reliability, sourcing reliability, management control, forecasting accuracy, information infrastructure, technology support, information visibility, information reliability, trust, coordination and labour skill and training.
Lockam and McCormack (2012)	Misalignment of interest, supplier financial stress, financial risk, supplier leadership change, tier 2 stoppage, service problems, supplier HR problems, supplier locked, merger/divestiture and production quality.
Thun and Hoenig (2011)	Supplier failure, supplier quality problems, oil crisis, terrorist attack, strike, malfunction of IT-system, accident (e.g. fire), natural disaster, machine breakdowns, import or export restrictions, transportation failure, delivery chain disruptions, increasing customs duty, change in customer demand, technological change, increasing raw material prices, number of nodes, stringent security and customs regulations, port/vessel capacity restrictions, product complexity, stringent storage quality requirements, volatility of supplier's location, supplier capacity and labour restrictions.
Cooke (2002)	Risks of supplying market from, or over-dependence on, a foreign plant, especially during times of political tension and risk stemming from increased regulation.
Dickson (1996)	Quality, delivery, price, production facilities and capacity, technical capability, financial position, management and organization, performance history, warranties and claim policies, procedural compliance, communication system, operating controls and labour relations record.
Zsidisin (2003)b	Capacity constraints, cycle time, disasters, financial health of suppliers, legal liabilities, currency fluctuations, management vision, market price increases, incompatible information systems and product design changes.
Zsidisin and Ellram (2003)	Unanticipated changes in the volume requirements and mix of items needed, production or technological changes, price increases, product unavailability and product quality problems.

2.4.2 Risk Identification Approaches

Risk identification aims to discover all relevant supply risks. According to definition, supply risk has two constructs: the source of uncertainty and its negative outcome. This infers that an initial judgement is required about factors that can be possible supply risks i.e. listing risk factors. Then, to identify supply risk a holistic approach is

required to determining if considered factor is actually a supply risk on the basis of its negative output (Buhman et al., 2005). Risk identification literature addresses this important issue by adopting different risk identification methods.

Turning to the "expert view" is a natural and most common approach for identifying supply risk, methods such as survey (Thun and Hoenig, 2009) or brainstorming (Norrman and Jansson, 2004) are used for this purpose. Referring to historical data, obtaining reports from similar companies or reviewing literature can assist experts in the risk identification process. A further step can be to involve a cross-functional team of employees together with a diverse group of experts (Hallikas et al., 2004; Norrman and Jansson, 2004). The benefits of this latter approach are the wider perspectives it can provide and the broadening of understanding and commitment to the risk management process within the company. Although a variety of selective methods exist to identify risks, the choice of method is ultimately case specific. Factors influencing this choice include the complexity of the supply chain and availability of time and appropriate experience. The expert-based methods for risk identification (such as brainstorming or risk questionnaire) can be rapid to deploy, but require a competence that may not be available inside a company. The alternative, of hiring external consultants to perform the risk identification, may be expensive and time intensive, and also undermines the commitment from staff involvement in risk management process. A systematic and disciplined approach, however, can enable a more comprehensive risk identification process.

One such systematic method is Analytic Hierarchy Process (AHP), AHP was utilised by Schoenherr et al., (2008) to identify the supply risk involved in making the decision about outsourcing for a US manufacturing company. First, they defined three sourcing characteristic's categories as the primary decision objectives; namely, the product, the foreign supplier and the foreign environment. Next, these were subdivided into sub-objectives and finally into 17 risk factors, that are identified as supply risk. Wiendahl et al., (2008) adopted a method called Ishikawa Diagrams for a case study to determine logistic risks for a forging company. Beginning with objectives and their potential negative effects (such as "low output rate"), supply risk was identified through complied list of risk factors that can be possible events causing adverse effects using the input from the expert. Chapman et al., (2011) have applied quality management techniques such as process chart/flow chart and histogram to identify key processes. Cause and effect analysis is then used to isolate the origin of

the delivery variance supply risk. These approaches such as AHP, Ishikawa Diagrams, chart/flow chart and histogram utilized the expert judgment in structural manner to identify the supply risk. However, semi-quantitative, expert-based methods have limitations in for their implementation in complex and many tiered supply chains (Neiger et al., 2009). This highlight the need of approaches those can identifying supply risk with thorough illustration of the supply chain structure.

Adhitya et al., (2009) discussed the application of Hazard and Operability (HAZOP) for supply risk identification thorough illustration of the supply chain structure. The HAZOP method is widely-used to identify hazard risk in chemical process plants by methodically identifying every significant activity within an organization. Adhitya et al., (2009) considered that supply chains and process plants display similar structures, and suggested that implementation of the HAZOP risk management method for supply risk identification can be an effective approach. A HAZOP study for a process plant uses process flow diagrams (PFDs), therefore a supply chain flow diagram (SCFD) and work-flow diagram (WFD) were similarly defined to represent the supply chain structure and operations. Identification of risk can be performed by progressively identifying deviations occur in different supply chain characteristics' values and their consequences. Neiger et al., (2009) applied the value-focused process engineering (VFPE) concept to develop a risk identification framework for identifying supply risk according to the design of the supply chain. Goal modelling and process modelling are integrated to implement the VFPE concept in a systematic way for risk identification in five steps. In Step 1, risks are identified as functional objectives and linked to the structure of supply chain objectives. Step 2 involves identifying the functional risk objective components. In Step 3, synchronized decomposition is conducted following the VFPE guidelines to combine the two previous structures into a single objectives structure. In Step 4, possible risk outcomes are linked to possible sources of uncertainty by applying e-EPC taxonomy of risk sources. Finally, in Step 5, the risk objectives structure and the risk sources/adverse events structure are joined to identify supply risk. Although approaches such as HAZOP and VFPE provide frameworks to identifying supply risk in a structural way, however their application is very complicated and requires up-front resources and knowledge to establish the inter-relationships between risk factors for supply risk identification (Oehmen et al. 2009).

Oehmen et al. (2009) adapted a system oriented modelling approach to identify the significant relationships between different risk factors. Their approach involved the integration of both a structure model and a dynamic model of risk. The supply chain risk structure model details the connections that produce the causes and effects of supply chain risks, that is, the risk factors and their potential impacting relationships. The supply chain risk dynamics model is used to predict the potential dynamics, or directional spread, of risk development. This type of model attempts to represent the complexity of supply chain relationships by addressing their network character. However, although the supply chain risk dynamics model represents probabilities, it does not fully integrate the charting of transitional events. Pfohl et al. (2011), on the other hand, employed interpretive structural modelling (ISM) to establish and examine the interrelationships among potential risks. ISM is an interpretive, qualitative procedure that delivers solutions to complex problems through the questioning and mapping of the intricate interconnections between elements. In this case, the potential risks are diagnosed in both supplier and manufacturer supply chains and then categorised by location – either inside or outside the supply chain. After identifying the directional impact of one risk upon another, these are then classified into four impact groups dependent on the force of impact and the measure of dependency of each risk upon another. Fuzzy MICMAC analysis is then applied to gauge the strength of influence of the risk factors upon each other. Finally, to readily appreciate the risk levels implied by this impact structure, a hierarchy model is produced. ISM is a qualitative and interpretive method which generates solutions for complex problems through discourses based on the structural mapping of complex interconnections of elements (Malone, 1975; Watson, 1978). The application of this technique is suitable for identifying the interrelationship between the supply risk factors; however it depends upon the accuracy of the input given for the final modelling. The self-interaction matrix (SSIM) by pair wise comparison is constructed for providing the input. This process which is very tedious and demanding depends upon participants to decide upon the pair wise relationship between the elements. Furthermore, the ISM model shows only that there is a connection between two risks if the impact of this connection is significant. There is need for methodology to identify the significant as well as minor relationship among risk factors and their respective impact. Guertler and Spinler (2014) have implemented the cross-impact analysis to identify the supply risk interaction. Based on the literature review a list of

supply risk is identified for four levels of supply chain operation, then based on the expert judgement values of relationship between two supply risks are used to develop cross-impact matrix. The cross-impact analyses reveals calculation of interrelationships among the supply risks in given system and finally, a short lists of supply risk are identified for supply risk monitoring.

These techniques seem very suitable for identifying complex interrelationships between supply risks. However, these approaches were only applied qualitatively, in accordance with requirements. These have not yet been used for quantitative simulations and do not provide predictive output about supply risk, which is more desirable in for effective supply risk management. Table 2.4 summarised the methods used for supply risk in previous literature. All the methods have some contribution for supply risk identification along with their limitation. That highlights the need for more research for supply risk identification to overcome the some mentioned limitation of previous approaches.

Table 2.4: Methods for risk identification

Approach	**Methods**	**Reference**
Qualitative approaches	Personnel brainstorming	Norrman and Jansson (2004)
	Expert interviews	Tuncel and Alpan (2010), Kumar et. al. (2014)
	Expert view (survey)	Thun and Hoenig (2009), Yang and Yang (2010), Markmann, et.al. (2013), Hoffmann et. al. (2013)
	Literature review	Wu et al. (2006), Canbolat et al. (2008), Yang and Yang (2010), Badurdeen et. al. (2014)
Semi-quantitative approaches	Action Research method and AHP	Schoenherr et al. (2008)
	Interpretive structural modelling	Pfohl et al. (2011)
	Quality management	Wiendahl et al. (2008),
	Hazard and Operability (HAZOP)	Adhitya et al. (2009)
	Value focused process engineering	Neiger et al. (2009)
	Cross impact matrix	Guertler and Spinler (2014)

2.4.3 Risk Assessment Approaches

Risk assessment is next step of overall supplier risk assessment procedure followed by risk identification. This step involves an analysis of the probability and/or frequency of occurrence of a supplier risk based on the impact of supply risk. Previous literature has suggested various approaches to assessing the supplier risk.

Similar to the supply risk identification, expert opinion based methods are also very common in risk assessment. Norrman and Jansson (2004) suggest a judgemental approach to risk assessment within a supply chain risk management framework. In their study, probability is rated using a qualifying scale that can term events as rare, unlikely, likely or almost certain. Similarly, the scale for severity of impact also includes several levels, which can be represented in different forms, for example, severe or light; high, medium or low; and negligible, minor, major or severe. Once combined, these levels of probability and impact can be used to estimate the supplier risk with categorisations such as low, medium, high and very high supplier risk.

Matook et al. (2009) presented a supplier risk management framework based on the Association of Insurance and Risk Managers (AIRM) approach together with a rating method for supplier risk assessment. This method uses two-sided prospective: internal firm ratings and external supplier ratings to identify supply risk. Firms explore the difference of opinion obtained through two different perspectives about supply risk to create the supplier risk profile. Trkman and McCormack (2009) created a conceptual framework for categorising a supplier network into four groups: Rock, Star, Millstone and Bouncer. In this framework, two factors determine each supplier rating: the performance of the supplier within the chain and the degree of disturbance in its operating environment. Rock group performs high in a stable environment, while Millstone performs low in a stable environment. Star group suppliers have high performance in a high turbulence environment, while Bouncer suppliers have low performance in a turbulent environment. Individual supplier assessment was achieved using both personal interviews and online surveys, which can allow an organisation to assess each supplier in their supply base. The online surveys were conducted with key supplier personnel to measure characteristics and supply chain structure (location, transport routes, etc.). Risk analysts were used to rate market and technology turbulence and external uncertainties. The resulting ratings based on expert judgement were then used to position each supplier into a categorized turbulence index. This study provided important insights into the supplier assessment process, determining

that both the supplier environment and company's strategy should be considered when assessing the potential of supplier's non-performance. All the above mentioned experts' judgement based risk assessment studies provide the basic foundation for supplier risk assessment and mostly method for categorization of supplier into groups. However, these methods require the proficiency that may not be available, or experts perception that is focus on certain aspect of supply risk and ignoring other can cause the biasness of obtained results (Wei et al. 2010).

Semi-quantitative method such as weighting methods and the Analytic Hierarchy Process (AHP) are commonly used for risk assessment. In these methods the 'expert judgment' is replaced with numerical values for providing a numerical rating of supplier risk. Blackhurst et al. (2008) used factor weighting in a supplier risk assessment and monitoring tool that produces risk indices for production parts and suppliers. The weighting is given to each factor in the category indicates their effect on the company. The weights relate to the probability of each category of supply risk causing disruption, and the degree of impact each category would have on supply performance is calculated. Each risk category is further divided into subcategories, which are also given relative weights. The final supplier score is a result of summing up the category weighting and its subcategory weightings. Although this method can be implemented easily, its drawbacks arise in common with the complexity of subcategories. The more subcategories that are introduced the less relative weight each will carry. As a result, this method can be less responsive to large risk ratings arising from any individual factor.

Gaudenzi and Borghesi (2006) employed AHP in a two phase method to identify risk factors in the supply chain and evaluate the strength of each factor. The first phase involved prioritizing the supply chain objectives, such as: prompt delivery, order completeness/correctness and damage/defect-free arrival. Secondly, the relative significance of the risk factors was assessed against each of these objectives. Similarly, Wu et al. (2006) have suggested the use of AHP to analyse risk factors in the supply base. This involves classifying supplier-oriented risk factors into categories, then applying AHP to measure an overall risk index. Levary (2008) conducted a case study by using AHP to compare an existing supplier against two potential suppliers. Each supplier was evaluated according to supply reliability criteria: supplier reliability, country risk, transportation reliability, and reliability of the supplier's suppliers. An expert panel evaluated the criteria and the suppliers'

priorities for each criterion. The ranking of each supplier was determined by summing the result of over all the criteria with multiplying a criterion's priority for each supplier. Other applications of AHP for assessing risk factors in the supply chain can be found in Schoenherr et al. (2008) and Enyinda et al. (2010).

Aside from AHP, other semi-quantitative methods, which rely particularly on experts' opinion and experience, have been proposed. Such as Sinha et al. (2004) suggested a methodology to reduce supply chain risks that used failure mode and effect analysis (FMEA) for supplier risk assessment. In the proposal, each potential failure mode is accorded a risk potential number (RPN) that is a product of the probability of a failure occurring (P) and impact severity (S) if failure occurs. P and S were subjectively judged using a scale of 1 to 10. Although these methods can enable a wide and inclusive range of supply risks, however effectiveness diminishes as the number of elements being evaluated increases. In addition, these methods depend on the contribution of personal values and judgment and may ignore the reality of dependence between risk events.

In addition to qualitative and semi-qualitative procedures, quantitative methods for risk assessment are also described in the literature. Wu et al. (2007) created the Disruption Analysis Network (DA_NET), based on a Petri Net (PN) modelling approach, to represent the multiple effects of supply risks through a supply chain system and to assess the impact of supply risks on system performance. Tuncel and Alpan (2010) also used PN-based simulation to indicate the impact of several supply risk situations (e.g. risks in demand, transportation and quality) together with possible remedial measures for the supply chain network. Risks to transportation supply and their impact on supply chain performance were studied by Wilson (2007) who proposed a system dynamics model for a supply chain comprising five levels: retailer, warehouse, tier-1 supplier, tier-2 supplier and source material supplier. Different scenarios of transportation risk – typically between adjacent stages – were modelled and their effect on fulfilment of customer orders and changes to stock were calculated. The conclusion was that transport disruption between the tier-1 supplier and the warehouse produced the largest impact.

Kull and Closs (2008) drew on resource based theory to develop an Arena simulation model to analyse impacts on system performance from the interaction between inventory levels and supply chain risk. It is proposed that inventory level has negative impact on supply risk. Both first tier supplier and buyer's inventory level has negative

impact on system performance in second tier supply uncertainty. A simulation model is developed for single retailer, main supplier and second supplier with prior selected failure probability. The simulation model provides the system performance (customer order filled) as output by considering days on hand for inventory as input under different settings of second supplier failure probability.

More contemporary quantitative methods for supply chain risk assessment include the proposal by Wei et al. (2010) for use of Inoperability Input–Output Modelling (IIM) to determine the expanding impact of disruption to a particular node of supply chain. IIM uses two metrics (inoperability and economic losses) to assess disruption impacts on supply chain networks. This can provide a comprehensive model of the resilience of each node in the network and enable appropriate choice of nodes for mitigating actions following disruptive events. This was illustrated by the case of an alcohol manufacturer in which a potential disruption in one supplier was addressed by increasing the number of suppliers. IIM has also been applied by Bogataj and Bogataj (2007) in assessing how lead-time disruptions expand through a supply network. Sawik (2011) used mixed Integer programming for optimizing the supply portfolio considering single period supplier selection under disruption risk (disaster) and delay risk. The expected cost and worst case cost were calculated under different delivery scenarios with deliver plenty cost. The portfolio approach helps determine order allocation for the minimum cost. Simulation based method for assessing the risk provide the empirical evidence however these methods both too numerically complex to apply and too subjective to be effective (Gereffi et.al. 2005; Knemeyer et al. 2009). Furthermore, these methods depend upon the assumption of supply risk occurrence for risk assessment. Making assessment on the assumption supply risks will occur is wasteful, if they do not (Zsidisin and Smith, 2005; Jung et al., 2011). The use of risk prediction methods, such as data mining, can provide satisfactory alternative routes to supplier risk assessment (Dani, 2009).

Mohtadi and Murshid (2009) used extreme value theory to develop a method of estimating the probability of catastrophic events with a dataset of terrorist attacks that involved nerve agent weapons. However, while such methods could help identify probability and impact of key supplier supply risk they focus only on location and low probability–high impact risk, whereas thorough supply chain risk assessment requires a study of all types of risk. A risk assessment model from Jung et al. (2011) enabled buyers to assess supplier risk using operational capability indicators and financial risk

indicators. The proposed approach used logistic model using data on five variables: switching cost, profit margin, asset to sales ratio, quality capability and technology capability. This data driven risk assessment model can be used to assess supplier risk using supplier operational and financial capabilities, however, the model ignores both supplier environment and the focal company's buying strategy. Lockamy and McCormack (2012) and later Lockamy (2014) adopted the Bayesian networking modelling approach to evaluate the impact of supplier risk on company revenue. A set of measure and scale is used to generate supplier risk profile by calculating the probabilities of network, operation and external risk factors. The impact of supplier risk is measured in term of value at risk for each supplier profile on company revenue. A risk profile reduction analysis is done to determine, how different risk categories can impact the value at risk for particular supplier. These data mining technique can be used to develop supplier risk profiles to determine the risk exposure of a company's revenue stream for its supplier base. Limitation of these models is that the identification of supply risks is not properly modelled. The identification of supply risk solely depends upon judgmental approach, which can be biased and can lead toward biased estimation of supplier risk. Furthermore, supplier risk defined on perceptual index and companies overall financial calculation such as revenue, however supplier risk should be measured in terms of supply performance.

Table 2.5: Methods for risk assessment

Risk Assessment Method		Reference
Qualitative/ semi-quantitative	AHP and Fuzzy AHP	Wu et al. (2006), Gaudenzi and Borghesi (2006), Levary (2007), Levary (2008), Schoenherr et al. (2008), Enyinda et al. (2010), Kull and Talluri (2008), Ganguly and Guin (2013), Badea et. al. (2014),
	Expert group rating	Norrman and Jansson (2004), Blackhurst et al. (2008), Matook et al. (2009), Punniyamoorthy et.al. (2013)
	Expert opinion and Delphi techniques	Thun and Hoenig (2009), Blos et al. (2009), Yang (2010), Sanchez-Rodrigues et al. (2010), Markmann et al. (2013)
	Failure mode and effect analysis (FMEA)	Sinha et al. (2004), Pujawan and Geraldin (2009)
	Conjoint Analysis	Atwater et.al. (2014)
Quantitative (modelling and simulation)	Petri Net	Wu et al. (2007), Tuncel and Alpan (2010)
	System dynamics	Wilson (2007), Ghadge et.al. (2013)
	Discrete event simulation	Munoz and Clements (2008), Kull and Closs (2008),
	Monte Carlo Simulation	Finke et al., (2010), Mangla, et.al. (2014)
	Statistical approaches	Lockamy and McCormack (2012), Jung et al. (2011), Tse and Tan (2012), Lockamy (2014),

2.5 Research Gaps and New Opportunities

The literature review for this study has revealed that the majority of approaches to risk identification and assessment are based either, completely or partially on judgments from experts about probability of risk event occurrences and their impacts. Further risk identification and assessment that use the quantitative modelling approaches. Such as Monte Carlo technique, Petri Nets and Fault and Event Trees (Kleindorfer and Saad, 2005; Wu and Olson, 2008; Tuncel and Alpan, 2010) for quantitative risk analysis are too complex mathematically or in their implementation. Further these simulation based approaches require priory assumptions about the different functions of supply risk. To build these assumptions for risk identification and assessment a considerable amount of knowledge and specific data is required, which companies may not monitor or may fail to record. Therefore, there is a need for a comprehensive methodology that provides companies with accurate knowledge about supply risk and

how to assess it from a quantitative perspective by utilizing past data that is readily available within the organisation (Dani 2009).

Khan and Burnes (2007) highlighted that there is need for exploring the already well-known risk approaches in other fields for application in supply chain risk management. These approaches should focus on how risk factors influence the key indicators of supply performance that companies already monitor. Furthermore, the methodologies should uncover the knowledge concerning supply risk that is hidden in available data by using validated models (Khan and Burnes 2007). In this way, it is possible to stimulate an easy understanding and communication of the causes and effects of supply risk. In the views of Macgillivray et al. (2007) and Smillie and Blissett (2010), communication has an intricate but essential role to play in improving standards in risk management, by supporting its institutionalisation and thus improving risk management controls.

In order to address the gap as discussed in supply risk management, this work proposes a framework that integrates both risk identification and analysis by using well-known data driven risk practices in other fields, such as knowledge discovery and risk scoring modelling. This proposed methodology is based on data currently recorded by companies for purposes other than risk investigation such purchasing, quality control or inventory control etc.

2.6 Summary

There is no common definition of supply chain risk, but there are multiple perceptions coming from multiple domains and type of risk dealing with supply chain risk. Current study selected the engineering and operation domain and supply risk type of supply chain risk. Although it is acknowledged that there may be further definitions and types of supply chain risk coming from other disciplines such as finance, emergency management, utility theory, health and safety. Most conceptual research is focused on categorizations of supply risk sources; those are often taken as synonymous to supply chain risk that is understood starting point for supply risk identification. There are different approaches adopted by previous literature for risk identification and assessment. Qualitative and semi-qualitative techniques are mostly used for risk identification and risk analysis such as: failure mode and effect analysis (FMEA) (Sinha et. al., 2004), empirical analysis (Thun and Hoenig, 2011; Wagner and Bode, 2006), and analytic hierarchy process (AHP) (Schoenherr et al., 2008). Quantitative techniques include analytical and simulation models such as analytical optimization models (Sawik, 2011), simulation modelling such as Monte Carlo simulation (Finke et al., 2010), and Petri nets (Tuncel and Alpan, 2010). Although many studies discussed the risk identification and assessment in context of supply chains, however there is need for research to analyse the application of established risk practices in other field of studies in supply risk assessment. The current thesis combines the established risk practices i.e. knowledge discovery and risk scoring techniques for supply risk identification and assessment to provide a supplier risk assessment approach and system.

The following chapter will provide the overview of two selected approaches i.e. knowledge discovery and risk scoring.

3 KNOWLEDGE DISCOVERY AND RISK SCORING

3.1 Knowledge Discovery (KD)

The term knowledge discovery (KD) was introduced at the first knowledge discovery in databases (KDD) workshop held in 1989 (Piatetsky-Shapiro, 1991). A widely used definition of the term was proposed by Fayyad et al. (1996a), which describes knowledge discovery as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (p.6). Data, in this expression, refers to facts or cases in a database, and pattern describes a subset of the data that can be extracted to fit a model, form a structure, or be descriptively classified in some way. By calling KD a process, reference is made to the necessary stages of data preparation, pattern search, evaluation and repeated refinement. Non-trivial means that the process is not pre-defined but involves the flexibility of search, discovery and inference.

Since the introduction of KD, it has experienced continuous development in new knowledge discovery techniques applicable to many areas of research involving both industry and academia. Independent scientific progress in fields such as biology and e-commerce has also compelled knowledge discovery methods to undergo considerable transformation (Piatetsky-Shapiro, 2007). There are notable examples of successful KD applications deployed on large-scale, real-world problems, as well as many KD-based systems now employed in daily business use. SKICAT is a KD system applied in astronomy to undertake image analysis, cataloguing and classifying of vast amounts of data received from a sky surveying observatory (Fayyad et al., 1996). The KD application in business areas include finance and investment, detection of fraud, marketing, industrial manufacturing, Internet information systems and telecommunications. Other successful areas applying KD methodologies include web-mining (Kolari and Joshi, 2004), biological research (Page and Hawley, 2004) and genomics (Lee et al., 2008).

KD techniques are of particular interest to a variety of academic communities and are a popular framework for problem solving in many research fields. This has led to the development of a number of methodologies that differ according to the field of study and the perspective of the problem-solving developers. The next section provides an overview of the basic KD process framework.

3.1.1 The KD process

In the literature, a number of methodologies have been proposed for knowledge discovery. Nevertheless, each follows the essential points of the scheme: data preparation, data mining and interpretation of the knowledge extracted. KD is both an interactive and an iterative process that involves many steps and many decisions during model construction. Fayyad et al. (1996b) describe the KD process in practical terms, with an emphasis on its interactive nature. Attention is drawn to iterative procedures, which may require returning to earlier steps to improve the quality of the process. The process is also noted for requiring "artistic" as well as scientific skills, because it requires gradual building of the best choice of elements rather than simply applying a fixed formula. A full understanding is therefore needed for each process and the options available at each step. The following is a simple graphic illustration of the nine-steps in the KD process those describe the whole KD process:

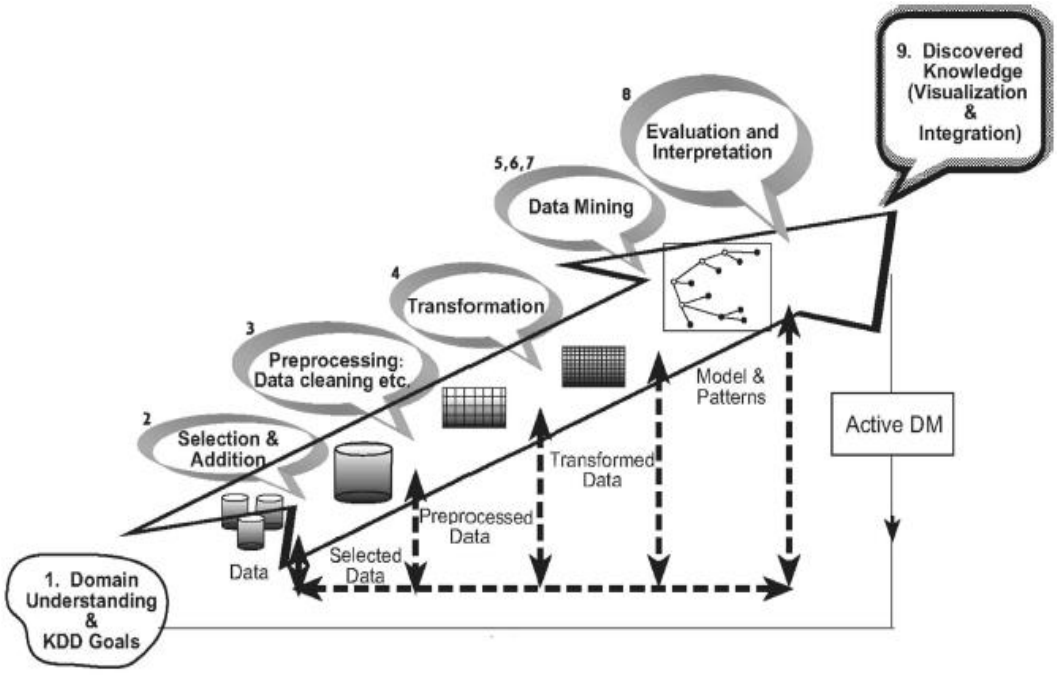


Figure 3.1: The Knowledge Discovery (KD) process (Maimon and Rokach, 2005)

Step 1: Understanding the application domain and KD goals

The first step before implementation of the KD process is to fully understand the objectives and define the goals of the process in a given application domain (i.e., supply risk management in the current thesis). This understanding should include the nature of the problem for which the KD process will be conducted (i.e., supply risk

identification for the current thesis) and the environment in which the output may be put to use (i.e. supplier risk scoring in the current thesis). Understanding the application domain and defining the KD goals will aid decision making concerning the type of data required, its transformation, the algorithms to be used, how results are to be represented, etc. Due to the iterative nature of the process, as described, these decisions may still be revised.

Step 2: Selecting data and creating a dataset

Once the goals of the process have been defined, the data will be used for KD, needs to be determined. This involves finding the data that is available, adding further data as necessary and combining all data into a single data set for the knowledge discovery. The importance of this step is that the chosen data mining algorithm learns and discovers from the data provided, which is the evidence base for the construction of models. Absence of any significant attributes could result in failure of the study, therefore more attributes is preferable to fewer. However, inclusion of more attributes can increase the complexity of data, therefore this need to be balanced. The interactive and iterative aspects of KD can aid in identifying the required balance that will result in the most suitable data set.

Step 3: Pre-processing and cleansing

This is the stage at which the reliability of data can be enhanced. Data pre-processing and cleansing involves dealing with missing values and removing noise and extraneous data. The time devoted to this task can vary from nil to being the most intensive part of a KD project. There are data mining controlled algorithms and simple statistical methods, those can be used for data pre-processing and cleansing. The degree to which such cleansing is undertaken can depend on many factors, such as the level of data pre-processing required for the data mining algorithm and resultant knowledge, etc.

Step 4: Data transformation

This stage involves preparation and development of the data so that it is in the best condition for the data mining. The methods used include reducing dimensions (e.g., attribute selection and extraction; and sampling records) and attribute transformation (e.g., discretization of numerical attributes and functional transformations). This

procedure is generally very project specific and can be critical for project success. However, it is not necessarily essential to choose the right transformations at the start of the project. The nature of the KD process is to reflect upon itself and be able to indicate the transformation that needs to follow in the project.

Step 5: Deciding the type of data mining task

This step requires a decision on the appropriate data mining task to be used, for example regression, classification, or clustering according to the KD objectives and previous steps in the KD process. Data mining techniques generally use the principle of inductive learning, in which an operational model is constructed based on training data. This can either be used to deliver an explicit predictive (or supervised) response or to deliver an implicit descriptive (or unsupervised) response. Whether the model learns under explicit supervised conditions or develops its own unsupervised learning, it is assumed that the trained model can be applied to future data cases.

Step 6: Choosing the data mining algorithm

At this stage an algorithm is chosen that can enable the desired pattern search in the given data. The choice of algorithm selection depends upon the goals of the project such as a requirement for high precision or for understand-ability, or both. Each algorithm has its own parameters and methods of learning from the given data.

Step 7: Employing the data mining algorithm

At this stage, the chosen data mining algorithm is employed to train the dataset. The algorithm may need to be employed several times until satisfactory results are produced that meet objectives.

Step 8: Evaluation

In this step, evaluation and interpretation of the mined data patterns (i.e., rule's reliability and model's accuracy) is undertaken according to defined goals. An assessment is made of whether the model induced has produced comprehensible and useful results and whether the goals of the data mining – defined at the outset – have been met.

Step 9: Using the discovered knowledge

The knowledge gained from the data mining process is now available for introduction into another system. However it is very important to understand the conditions under which the knowledge was obtained, because changing conditions may change the data structure and require a different implementation of the KD process. For example, the knowledge derived from the system's data was static in nature, but now the system has becomes dynamic in nature and therefore may require an implementation of the KD process that suits the dynamic nature of the data. Therefore, knowledge should be incorporated into another system that requires the same conditions from which the knowledge is derived.

Data mining is the algorithmic step of the whole KD process that enables meaningful knowledge to be produced from available data. The previous section provided some understanding of the data mining process and the importance it has in the KD process. The next section will offer an overview of data mining and its relative tasks and techniques.

3.2 Data Mining

Extracting useful patterns from data has been given many descriptive names, including data mining, information discovery, knowledge extraction, data pattern, data processing and data harvesting or archaeology. However the statistical data analyst and management information system communities have generally use the term data mining.

"Knowledge discovery" is a term that was first coined at the 1989 KDD workshop to stress that "useful knowledge" is the desired outcome of data-driven discovery. Data mining is an important step in this process and involves applying particular algorithms to uncover useful patterns within data. Fayyad et al. (1996b) defined data mining as "the application of specific algorithms for extracting patterns from data" (p. 39). Data mining techniques are applied to data to uncover unseen patterns and relationships that can help in decision making (Baradwaj and Pal, 2011). Various data mining algorithms and techniques are available to achieve knowledge discovery from data. A brief taxonomy of data mining methods is given in the next section to provide better understanding.

3.2.1 Taxonomy of Data Mining Tasks and Techniques

Taxonomy is useful to gain an understanding of the different data mining methods, and how they are grouped and inter-relate. Overall, data mining methods are divided into two main types related to knowledge discovery goals, namely verification-oriented and discovery-oriented. Verification-oriented methods concern with the evaluation of a hypothesis often proposed by an external expert. They usually involve common traditional statistics methods of hypothesis testing, such as goodness-of-fit, hypotheses tests and analysis of variance. Discovery-oriented data mining methods generally involve the discovery of a hypothesis from a dataset and validating it, rather than verifying a known hypothesis. Despite some overlaps with verification methods, the main objective of discovery-oriented methods is identification and construction of models based on statistical evidence obtained from available data. Discovery methods identify patterns automatically in data. This research thesis is primarily concerned with the identification of patterns in data about supply risk. Therefore, the main focus of this research thesis is discovery-oriented data mining methods. Figure 3.2 presents taxonomy of data mining methods.

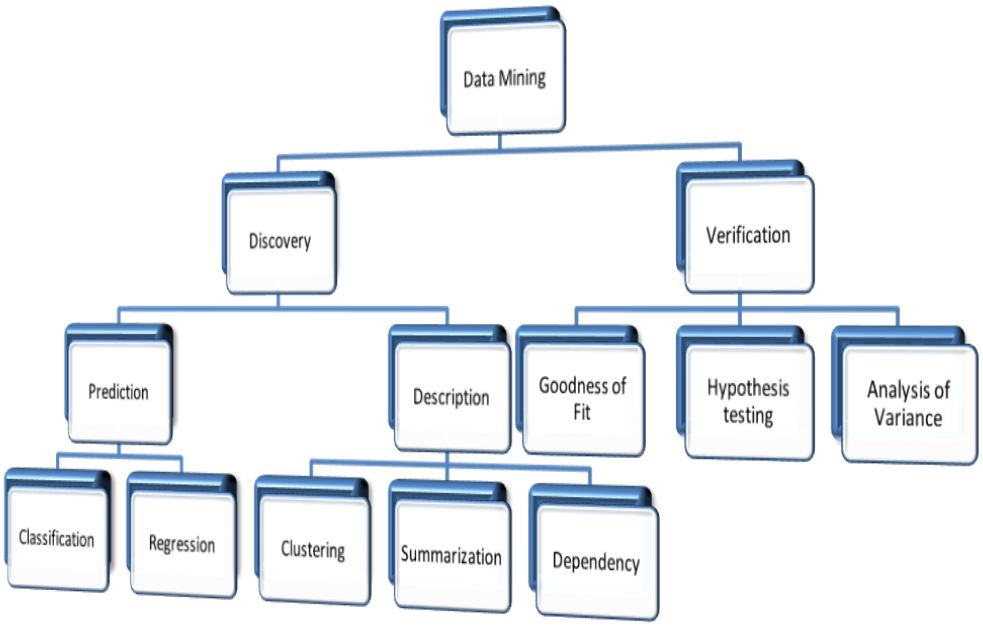


Figure 3.2: Data mining task taxonomy

Discovery-oriented methods can be further subdivided into description-oriented and prediction-oriented methods. Descriptive methods discover patterns within data and attempt to understand (for example, by visualization) how underlying data

interrelates. Predictive methods seek to create a behavioural model, by identifying new, unseen patterns, then predicting the values of target variable in relation to the identified patterns. Most discovery-oriented data mining techniques require inductive learning, which involves explicit or implicit construction of a model on training sample. The prediction-oriented and description-oriented methods can be divided into different data mining tasks (Miller and Han, 2001) as given below.

3.2.1.1 Clustering tasks and techniques

The Clustering task is a process of identification of different clusters or groups within a given dataset of objects. Objects are clustered so that intra-cluster similarities are maximized, and inter-cluster similarities are minimized. Once the new clusters are determined, the corresponding objects are labelled according to their clusters. Clustering techniques include both the statistical approaches and machine-learning approaches, all clustering techniques can be categorised as either hierarchical clustering or data partitioning techniques.

3.2.1.2 Association task and techniques

Association task is a process of identifying patterns for objects in a given dataset, by revealing objects co-occurrences with other objects, or by identifying significant dependencies between objects. Association task can be conducted using techniques based on graphical theory approach such as Bayesian network, the machine-learning approach such as association rule mining, or statistical functions such as principal component analysis and correlation analysis techniques.

3.2.1.3 Summarization task and techniques

Summarization task is the abstraction or generalization of data. A data set that is task relevant can be summarised or abstracted to provide an overview of the data, generally in aggregate form. Most of the summarization task techniques are statistical in nature, such as mean calculation or standard deviation.

3.2.1.4 Regression task and techniques

Regression task focuses on predicting a value for a given continuous dependent variable based on its relationship with independent variable values. Regression analysis techniques can be divided into linear functional and nonlinear functional techniques. Linear functional techniques mainly use statistical approaches such as linear regression, multi linear regression, generalized linear model techniques and time series analysis. Nonlinear regression techniques can employ both statistical and

machine learning approaches, but most involve machine learning approaches such as neural networks, k-nearest techniques, regression techniques, and so on.

Another type of regression task is Trend Analysis, which discovers interesting patterns in the evolutionary history of data objects. A model or function is constructed, based on identified patterns in an object's history, that simulates the object's behaviours, such as up, down, peak, valley, etc. The model reflecting the object's past can then be used to predict future behaviours.

3.2.1.5 Classification task and techniques

Classification task is the creation of a model which uses an object's attributes to determine its class. The classification model is built by analysing the objects in the training set to establish the relationship between each object's attributes and the class. This constructed model can be used to classify future objects.

Classification techniques can be categorised into five main categories (Witten, et. al., 2011). Bayesian classification techniques use the Bayesian function to build the model and deliver results in probabilistic values such as naive Bayesian classification technique, Bayesian net classifier, etc. Functional techniques use a statistical or mathematical function to develop the classification model, such as neural network techniques, logistic regression, support vector machine, etc. Instance-based learning classification techniques compare new problem instances with those encountered in training and stored in memory. Instance-based learning is also known as lazy learning, for example, k-nearest neighbour classifier is an instance-based learning technique. Rule-based classification techniques create classification models in the form of a readable rule set; these techniques include RIPPER algorithm, PART algorithm, Zero rule, one rule, etc. Decision tree techniques represent the model as a tree, where every leaf is a class, and nodes represent the attributes of independent variables.

All the above mentioned data mining tasks can be applied separately or in a combined way depending upon the requirement of concerned problem. Based on the objectives of current research thesis and acquired domain knowledge (i.e. nature of problem formulation for supplier risk assessment approach development), the classification task of supervised learning is mostly concerned with current research thesis.

All data mining task and techniques are aimed at addressing real world issues and uses real world data; however for given issue data may not be available from single data source and further data source is not designed by keeping data mining task

implementation in mind. Therefore, a database needs to build by keeping data mining task implementation in mind. Further, issue such as noise, absent values or attributes, unwanted information, overlarge data sets or sparse data should be resolved to have appropriate dataset for data miming algorithm implementation (Frawley et.al., 1992; Matheus et. al. 1993). The current study also faces these challenges as available data is not stored with the purpose of data mining implementation, nor is it available from a single data source. Therefore, it is important to analyse whether the obtained data from different data sources is suitable to meet the current study objectives.

Further as the different data mining techniques are available for given task, those can represent the considerably varying evaluation performance on given sample (Michie et.al. 1994, King et. al. 2002). However on all methods NO Lunch Theorem (NLT) applied, which states that if model A performs better than model B on one evaluation metric, then B can perform better than A on other metric. It is not clear that which technique can create the best result for given objective and training sample (Todorovski and Dzeroski 1999), Which naturally create a question: *which data mining technique can be most suitable for available data to meet the objectives of KD process?*. The selection of appropriate algorithm is one of the key challenges in knowledge discovery systems implementation (Pechenizkiy et. al. 2005; Lindner and Studer, 1999). Furthermore, selection of model depends upon the knowledge of analyst about algorithm and problem domain (Brachman and Anada 1996). The utility of more than one quantitative measure for model's performance according to problem domain and ranking of models on these measurements can be good solution to identify the single best model (Brazdil and Soares 2000). The issue is resolved through the adoption of appropriate methodology in current research thesis.

Table 3.1: Different data mining tasks and their techniques with a brief description

Task	Techniques	Description
Clustering	Hierarchical	Provides different clusters having homogeneity within a cluster and some dissimilarity among different clusters.
	Partitioning	Provides a single partition of the data instead of a clustering structure.
Association	Association rules	Provides interpretable rules about association among different variables using heuristics search methods.
	Bayesian networks	Provides graphical interpretation of causal relationships among variables together with conditional probabilities using Bayesian methods.
	Statistical	Identifies the co-relation between variables using statistical methods; there can be different methods for numerical and categorical variables.
Summarization	Summery rule	Generalization and characterisation about the data is performed either by mining or summary rules such as characterisation rules.
	Attribute-oriented induction	Using the hierarchical aggregation of data attributes by compressing data into generalised relations based on background knowledge.
Regression	Linear	Provides a model which identifies the linear relation between the dependent and independent variables.
	Non-linear	Provides a regression model which identifies the linear and non-linear relationship between the dependent and independent variables.
Classification	Bayesian	Provides a model using the Bayesian function to develop a classification model; these techniques can be used for both regression and classification.
	Instance base	Classifies a new instance based on some similarity function using the instance-based learning approach.
	Functional	Provides the discriminate functions to distinguish between predefined classes that can be linearly or non-linearly separable.
	Rule base	Provides a set of classification rules using a heuristics search method that can be used to classify predefined classes.
	Decision Tree	Provides graphical representation of a tree to classify given data according to predefined classes; the graphical representation can be easily converted into rules.

3.3 Application of Knowledge Discovery in Risk Identification

KD can be used for risk identification in a variety of risk management systems in order to achieve organizational goals. Common KD based risk identification systems in different fields are given below.

3.3.1 Fraud Risk or Non-Compliance Risk Identification Systems

The process of compliance monitoring for risk identification uses the knowledge discovery process to develop a monitoring system, which compares predetermined acceptance conditions with actual data. When the knowledge discovery based monitoring system detects variance (risk) from the predetermined conditions an exception report is produced to identify the variance (risk). Examples of these systems include credit card protection, monitoring privacy compliance, auditing checks, etc. (Goldschmidt, 2009).

3.3.2 Intrusion detection systems

In this case, KD is used to develop an intrusion detection system that monitors information systems, identifies security breaches and raise an alarm. This system monitors and analyse events in computer through the implementation of various data mining techniques to detect any intrusion or security risk (Singhal and Jajodia 2005). Projects such as MADAM-ID, ADAM, and clustering project used knowledge discovery processes to the construction of operational IDs and clustering audit log records for risk identification.

3.3.3 Lie Detection Systems

Many lie detection systems are available in the marketplace, such as Clementine, SGI's Mine-set, IBM's intelligent miner or SAS's text miner. These systems use KD process to automatically detect the lies or misinformation in email or website data. Different data mining techniques are applied in KD process to identify risk in business deals, communications with angry customers, and many other similar situations.

3.3.4 Risk Identification Systems in Manufacturing

Knowledge discovery has been successfully applied in the improvement of manufacturing processes, with many organizations using the KD process to discover useful informative rules within manufacturing data to improve risk identification, such as quality failures or time delay, etc. Boeing has successfully applied knowledge discovery processes to its manufacturing data to identify rules that can help predict

part quality inspection failure or delay at individual tooling machines. Printing company R.R. Donnelly has used knowledge discovery to reduce quality (banding) problems, and also to create rules that establish process parameters (e.g., viscosity of ink) to reduce quality risk.

Furthermore, knowledge discovery have been also applied successfully in many other fields for risk identification such as part failure detection, web site personalization, disease (risk) detection in healthcare, diagnosis error (risk) identification in healthcare, project failure risk identification, etc. The variety of these successful applications of the knowledge discovery process shows its suitability for risk identification and the relevance to interest for current research thesis, i.e., supply risk identification.

3.4 Risk Scoring

Risk scoring is a process of finding an empirically valid estimate of risk probability that represents the population under consideration, i.e., the given data (Barman, 2005). Risk scoring models can be produced using data mining techniques to provide a probability of risk and to make predictions for new data (Schreiner, 2004). Historical data provide the foundations for model building. Applying data mining techniques to historical data enables the building of a risk scoring model, which can then be applied to new data to predict future behaviour. However, the procedure of using the model for prediction is different from the process of model creation. Once a model is created, it may be used many times to provide risk scores on newly inputted data.

3.4.1 Definition of Risk Scoring

There are a number of different definitions of risk scoring, which can vary according to environment and perspective.

- A method of quantitatively predicting the probability of risk that a company/person may be unable or unwilling to honour an obligation under terms of a business contract (required performance) and thus cause a loss (Mester, 1997).
- A formula based on known data that assigns points to attributes of data for predicting future outcomes (Perrine, 2007).

- The usage of the data about the performance and characteristics of historical loans to predict the performance of future loans (Schreiner, 2004).

A major objective of risk scoring is to assist the organizations in quantifying and managing risk when conducting business. The United States Circuit Court is held to consider that actuarial evidence indicates risk scores are a good predictor of risk of loss (Johnson-Speck, 2005). Another actuarial study concludes risk scores are powerful predictors of risk and also one of the most accurate predictors of loss (Miller, 2003).

Risk scoring was initially successfully applied as a method of evaluating the risk of lending customers such as credit card applicants i.e. credit scoring (Anderson, 2007). Credit scoring aid decision makers to predict customer default probability on given loan is the most widely used risk scoring model. The insurance industry also uses risk scoring models to aids decisions on applicants for new insurance policies and renewal of existing polices. GE Capital Mortgage Corporation, for example, applies risk scoring for screening mortgage insurance applications (Prakash, 1995). Setting and adjusting insurance premiums also uses risk scoring, since clients with poor risk scores can be identified as more likely to file insurance claims and thus suitable to be charged higher premiums. Such risk information is also used to assess performance and accountability according to insurance policy conditions. Landlords can use risk scoring to determine the likelihood of potential tenants making timely rent payments.

Suppliers of utilities (such as electricity, gas) in the United States have used risk scores to decide if they should offer services to potential customers. Scorecards are also use to predict the efficacy of patients receiving certain medical treatments. Finally, employers can make use of risk scorecards before hiring potential employees, especially if positions involve handling large amounts of money (Consumer Federation of America, 2002). All these examples are evidence that the risk scoring process can be used in different fields of study, if aligned with the study goal. The current study is also concerned with estimating the probability of loss a supplier can inflict on a buyer firm. Therefore, implementation of a risk scoring model to assess supplier risk seems eminently suitable. However, understanding the application of risk scoring requires an explanation of the full process in detail, which the following section will provide.

3.4.2 Risk Scoring Development Process

The aim of a risk score model is to build a single aggregated risk indicator for the given risk factors. To develop risk score model a number of steps need to be followed (Siddiqi 2006):

Step 1: Understanding the business problem

The aim of the model should be determined clearly, as this will affect other decision in scoring model building process such as which technique to use and which independent variables (data elements) will be appropriate to include. It will also influence the choice of the dependent variable, or outcome to be presented. In current research thesis chapter two provided an understanding of relevant business problems and the foundation for building risk score models for supplier risk assessment.

Step 2: Defining the dependent variable

For building the risk scoring model, the dependent variable (also known as target variable) need to be defined, where it has binary value either risk vs. no-risk. Most of the risk scoring model focus on quantifying the probability of risk (PR), tradition known as probability of default.

PR reflects the probabilistic assessment of an obligor or counterparty defaulting on contractual obligations within a particular time period. Therefore, a dependent variable in risk scoring is defined in two dimensions: a loss definition and a period in which the level of loss can occur, usually called the outcome period. The outcome period is thus the time over which obligators' performance in the sample is observed to classify them as no-risk (good) or risk (bad).

In the current study, the proposed methodology will define the supplier risk by taking the above mentioned consideration in accordance with purchasing and supply base context.

Step 3: Data, segmentation and sampling

Since it is unlikely that many business situations will present a perfect scenario or data for modelling, some cases may be inevitable. Therefore, decisions on data selection must comply with some basic requirements:

- *Past business experience*

  Since development of a scoring system requires the analysis of past decisions, the organization must have offered some business to other parties in the past, such as purchasing orders to different suppliers. So it can provide the data for modelling, because, no historical data availability equals no scoring system.

- *Data retention*

  Information used to support past decision must have retained in a usable form in order to build a model. For example, in purchasing decision the existing supplier survey and purchased order receiving report data would be relevant for supplier risk model development.

- *Known outcome of past decision*

  The outcome of past decision must be available in quantifiable form. Suppliers past performance histories can be used to classify outcomes as good or bad suppliers. The level of detail of historical performance records must be examined, and data archiving and purging procedures are important. For instance, when suppliers purge performance accounts from the records, efforts must be made to recover information on these accounts.

- *Age of decision*

  The age of trading decisions must be sufficient to provide a practical framework for measurement and classification. Suppliers who have only recently received their first purchasing order will not offer sufficient performance history to be accurately classified. Equally, suppliers engaged many years earlier with supplier survey and order receiving reports available only from earlier years will not be reflective of current trading conditions or relationships. An appropriate time frame for including data should therefore be selected, which will depend upon problem objective and the risk decision being tested.

- *Sample size*

  To obtain an appropriate sample size the number of business decisions included must be sufficient to capture essential outcomes. Negative outcomes, such as defaults, are infrequent in business performance; therefore the ability to include such performance may influence both sample time frame and sample size. A small business, conducting fewer transactions, may require a

longer sample time frame, whereas a large business may provide sufficient data from recent history. Ensuring that data selection delivers clean, accurate and appropriate data, that is the most important aspect of model development and typically requires the most time and effort. The availability of data will typically be influenced by the type of model required.

- *Development and testing sets preparation*

  After defining 'bad' events and the relevant outcome period, relevant data can be collected for inclusion in the data set from which the scorecard can be developed. A training sample should be created to build the scorecards and a testing sample created for accurate testing of the scorecard. This testing should be conducted on separate performance data from that used to compile scores.

Step 4: Model building

A risk scoring model can be built using a number of different data mining techniques (see detail in section 3.4.3). Data mining technique examines the relationships between a dependent variable and a set of independent variables. The resulting output from a data mining model are coefficients, which indicate the correlation between the dependent variable and the independent variables.

Regardless of the technique used to build the model, the predictability of built model should be evaluated. Such evaluation will attest the scorecard's readiness to perform its intended tasks.

Step 5: Generalization

A risk scoring model is ultimately intended for its application on samples other than the development sample. Therefore, to be useful, the model should not be too specific to the training data. For this reason, a test sample for the same time period is used in model testing. A separate sample from a different time period can also be used on the completed model to validate predictability. This can ensure robustness of scoring across data from different periods of time.

Step 6: Ongoing monitoring

After the model has been developed and is in implementation, monitoring its functionality at regular intervals is also important. The business environment is subject to constant change; therefore model predictions will periodically need to be

re-tested. Where the population has changed from the source data, reliable model predictability may only require small changes. However, monitoring can also indicate when predictive capacity is below requirements and redevelopment of the model needs to be considered.

3.4.3 Data Mining Techniques for Risk Scoring Model Building

In previous literature, different data mining techniques are used for risk scoring model building, such as: discriminant analysis, linear regression, logistic regression, mathematical programming, probit analysis, Markov chain models, nonparametric smoothing methods, recursive partitioning, expert systems, conditional independent models, genetic algorithms and neural networks. Most techniques used for risk scoring model building are related to classification task. The following section gives a short literature review of data mining techniques used for risk scoring model building.

Galindo and Tamayo (2000) investigated several classification techniques for their efficacy in credit risk assessment. Those considered were probit, neural network, decision tree, k-NN models. The k-NN method was found to deliver the best results.

Doumpos et al. (2002) compared the Multi-group Hierarchical DIScrimination (M.H.DIS) method with certain traditional methods, such as discriminant, logit and probit analyses, for classification of credit risk. Their conclusions pointed to M.H.DIS as having greater classification accuracy.

Xiao et al. (2006) assessed a variety of classification methods for credit scoring models. Their study gave approval to SVM, MARS, logistic regression and neural network in terms of classification results, but LDA and CART were also considered particularly user-friendly for credit scoring tasks.

Satchidanand and Jay (2007) examined five types of classifying techniques: machine learning methods, Bayesian theory, statistical tools, neural networks and kernel-based models. The study sought the best predictor of default probability and concluded that the kernel-based RBF neural network was superior in identifying true positives.

Atish and Huimin (2008) made a comparison on cost effectiveness metric for seven classification data mining techniques; those utilize the domain knowledge in their implementation. Using area under the curve and misclassification cost for analysis, the study concluded that inclusion of domain knowledge enhanced effectiveness with certain data mining techniques. Ince and Aktan (2009) evaluated performance on prediction metric using discriminant analysis, neural network, logistic regression, and

classification and regression trees. In this study, the best performers were CART and neural network.

Table 3.2: Data mining approaches in risk scoring (Keramati and Yousefi, 2011)

Techniques	**Reference**
Neural network	West (2000), Malhotra and Malhotra (2003), Hsieh (2004), Angelini et al. (2008), Yu (2008), Abdou (2008), Tsai (2009) and Khashman (2010)
Bayesian classifier	Baesens et al. (2002), Li and Guo (2006), McNeil and Wendin (2007), Kadam and Lenk (2008), Panigrahi et al. (2009), Stefanescu (2009) and Antonakis and Sfakianakis (2009)
Discriminant analysis	Altman (1968), Eisenbeis (1977), Taffler and Abassi (1984), Yobas (2000), Kumar and Bhattacharya (2006), Abdou (2009)
Logistic regression	Steenackers and Goovaerts (1989), Laitinen (1999), Alfo and Trovato (2005), Tang and Chi (2005), Ma and Tang (2007), Sohn and Kim (2007), Luo and Lei (2008) and Liang and Xin (2009)
K-nearest neighbour	Paredes and Vidal (2000), Hand and Vinciotti (2003), Islam et al. (2007), Marinakis et al. (2008) and Li (2009)
Decision tree	Mues et al. (2004), Lee et al. (2006), Xiao et al. (2006), Zhao (2007), Lopez (2007), Yeh et al. (2007), Bastos (2008) and Li et al. (2010)
Survival analysis	Thomas et al. (2001), Stepanova and Thomas (2002), Baesens et al. (2005), Noh et al. (2005), Sohn and Shin (2006), Carling et al. (2007), Beran and Djaidja (2007), Andreeva et al. (2007), Bellotti and Crook (2009), Cao et al. (2009) and Sarlija et al. (2009)
Fuzzy rule-based system	Baetge and Heitmann (2000), Tung et al. (2004), Tang and Chi (2005), Laha (2007), Hoffmann et al. (2007), Jiao et al. (2007), Lahsasna et al. (2008), Liu et al. (2009) and Xinhui and Zhong (2009)
Support vector machine	Huang et al. (2004), Gestel (2006), Chen and Shih (2006), Gestel et al. (2006), Yang (2007), Martens et al. (2007), Huang et al. (2007), Xu et al. (2009), Zhou et al. (2009), Chen et al. (2009), Luo et al. (2009), Bellotti and Crook (2009), Hardle (2009), Zhou et al. (2010), Yu et al. (2010) and Kim and Sohn (2010)
Hybrid models	Lee et al. (2002), Malhotra and Malhotra (2002), Wang et al. (2005), Lee and Chen (2005), Hsieh (2005), Huang et al. (2006), Zhou and Bai (2008), Zhang et al. (2008), Yu et al. (2008), Lin (2009), Chen and Li (2010) and Hsieh and Hung (2010)

Li, F.C. (2009) examined support vector machine, k-nearest neighbour and neural network for classification accuracy when used without features selection methods. Greater classification accuracy was found when these methods were integrated with effective features selection. Twala (2010) analysed credit risk prediction with the options of using individual classifiers or classifier pair combinations. The test results indicated that predictive model accuracy was improved by pairing individual classifiers. Paleologo et al. (2010) constructed a composite credit scoring model capable of coping with unbalanced data, missing information and non-data points.

The literature review above has listed many techniques for risk scoring. However it cannot be determined that there is any outstanding technique that can be stated as universally best for risk scoring model building.

Currently the neural network and logistic regression are most widely used techniques in banking and other sectors for risk scoring model building. The neural network is complex concept and black box technique; however logistics regression technique can be easily understood and implemented. Therefore, current research thesis uses logistic regression technique for building the risk scoring model.

It is shown that utilization of knowledge domain (Atish and Huimin 2008) and features selection methods (Li, F.C. 2009) can enhance the risk scoring model performance regardless of the data mining algorithm. The current thesis also proposed knowledge driven features selection method for developing a risk scoring model.

3.4.4 Rules for Selection of Appropriate Variables and Discretization

The choice of variables for inclusion in a risk scoring model is a major consideration. There can be numerous numbers of independent variables that can be included into supplier risk scoring model being potential sources of supply risk and ultimately for supplier risk. Such independent variables can have both numerical and categorical type data. Independent variable having numerical type data can also add considerably redundancy to the available data for model building. As the numbers of independent variables and numerical type data increase the efficiency of the model decrease especially for logistic regression algorithm.

The higher the number of variables in the model the greater the estimated standard errors become and the less applicable the model will be to other data sets. Minimizing the inclusion of variables will result in more stable model that will be appropriate to

use in wider spectrum of future application situations. However too few variables can result in loss of information, therefore, the objective must be to choose variables that satisfactorily explain the data and produce the best model to address the particular risk scoring task.

In order to determine the selection of variables, both statistical and machine learning techniques can be applied. Commonly used statistical techniques involve bivariate analysis such as likelihood ratio test, Pearson chi-square test, Weight of Evidence (WOE), Spearman rank order correlation, Gini coefficient, information value, and principal component analysis. Any of method can be used to reduce model dimensions provided the choice is appropriate for the required variable selection and for transformation of numerical to categorical variables. Nevertheless, there are certain unbreakable rules on selecting variables for inclusion in risk scoring model building. The following is a summary of the main essential principles to be considered for selecting variables (Bolton 2009).

- *Logical and predictive*

  A constant principle to apply is that the simpler the model the better and more robust it is likely to be. The variables chosen should be logical, have significant predictive power and be easily explained to business managers. These principles should help extend the model's useful life.
- *Multi-co-linearity*

  Highly correlated variable in data sample can create a model that will over-fit on testing sample. Co-related variable can calculated using the same input value. Therefore, the selected variable should not be co-linear and make logic.
- *Available in Future*

  The choice of variables should be limited to those that will remain available in future. Variables should be excluded if they are not likely to be representative in future, are new and few in number.
- *Compliant*

  When risk assessment models are used to make decisions regarding third parties, the developer must make sure that variables are compliant with any legal, policy or ethical conditions on their use.

- *Minimum information loss*

  Reducing excess variables is desirable but this should be done with the minimum impact on vital information. Although some variables may appear weak, if they are closely correlated with significant variables their exclusion could limit the effectiveness of the model.

Implementing data mining in the KD process enables construction of a model based on the given data. The constructed model now needs to be tested for the goodness-of-fit and predictive ability of the model. The following section will provide details of different model evaluation methods and measurement metrics.

3.5 Evaluating the Classification Performance of a Model

The constructed model needs to be tested for its ability to predict the class labels of previously unseen records. Measuring performance enables an unbiased appraisal of its generalization error. In addition, the accuracy or error rate determined from the test can be used to compare the relative performance of different classifiers on the same domain. The following section reviews some of the methods and metrics commonly used to evaluate the performance of a classifier.

3.5.1 Evaluation Methods

The two main evaluation methods are: Holdout and Cross-Validation. These methods vary from each other on basis of dividing the available dataset for data mining into subsets and the way these subsets are used for testing the constructed model.

- *Holdout Method*

  In this method the available dataset is divided into two subsets: training subset and testing subset. Training subset is used to build the model and testing subset is used to validate the built model. This method is very suitable for validating the final developed model.

- Cross-Validation

  In this method the available dataset is divided into k subsets having same number of record. In this method $k-1$ subsets are used for training and one subset is used for testing. The testing process is repeated for k iteration. Let acc_i be the model accuracy during the i^{th} iteration. The overall accuracy is

given by $acc_{sub} = \sum_{i=1}^{k} acc_i / k$. This method is very useful for model building stage.

3.5.2 Evaluation Metrics

3.5.2.1 Predictive performance metrics

A classifier is, typically, evaluated by a confusion matrix as illustrated in Figure 3.3 (Chawla et al., 2002). The columns are the Predicted class and the rows are the Actual class. In the confusion matrix, true negatives (TN) is the number of negative examples correctly classified, false positives (FP) is the number of negative examples incorrectly classified as positive, false negatives (FN) is the number of positive examples incorrectly classified as negative and true positives (TP) is the number of positive examples correctly classified. There are several metrics that have been developed from the confusion matrix.

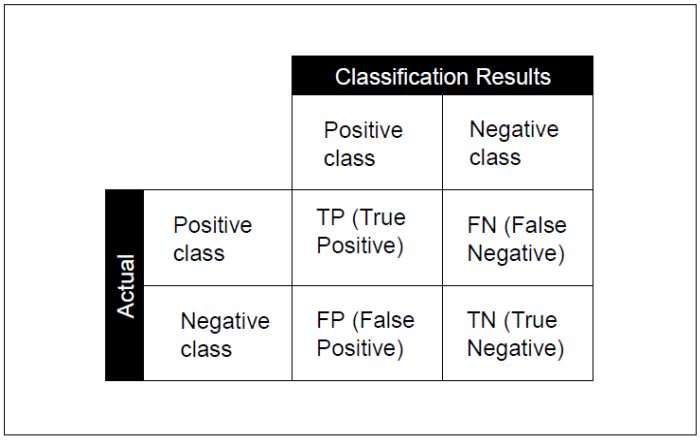


Figure 3.3: A confusion matrix for a two-class problem

- *Accuracy*

  It determines the percentage of correctly classified examples. Based on the confusion matrix accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \qquad (3.2)$$

- *Recall*

  It determines the percentage the percentage of correctly classified positive examples. Recall is also termed as: true positive rate, hit rate or sensitivity. It calculated as

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3.3)$$

- *Precision*

  The precision determine the percentage of correctly classified examples of given class either positive or negative. For the positive class, precision is calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (3.4)$$

- *F-measure*

  The main goal for model is to improve the recall without losing the precision. However, generally as recall increase the precision can decrease. Therefore a trade-off between these two metric is required. F-value combines the precision and recall values to provide a single metric that represent the trade-off between recall and measure. F-measure represents the trade-off among different values of TP, FP, and FN (Buckland and Gey, 1994). The F-value is give as

$$\text{F} - \text{measure} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}} \qquad (3.5)$$

- *Receiver Operating Characteristic (ROC) curve*

  ROC curve is a standard technique through identifying the trade-offs between true positive and false positive error rates for summarizing classifier performance as shown in figure 3.4. The Area under the Curve (AUC) is an accepted performance metric for a model validation (Bradley, 1997). For random guessing, the AUC coefficient = 0.5, a model will be valid if it has the AUC >0.5. Further higher the AUC represent the better model performance on new unknown data. Weiss (2004) indicate that AUC is more suitable than accuracy as it is not biased to minority class.

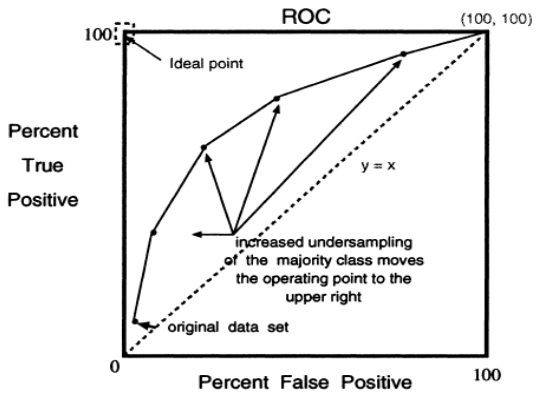


Figure 3.4: Illustration of a ROC Curve (Crook et al. 2007)

3.5.2.2 Rule's quality performance metrics

A classification rule is a knowledge representation, Let R_i be a rule that predicts classC_i, |A| is the set of instances which belong to classC_i, |B| the set of instances to which rule R_i is applicable and N is the total number of instances in the dataset. When evaluating the quality of a rule, the three common factors to be taken into account are the coverage, the completeness and the confidence factor of the rule, defined as follows:

- C*overage*
  It is determined by the number of instances satisfied by the rule antecedent and is given by |B|.

$$\text{Coverage} = \frac{|\text{B}|}{N} \qquad (3.6)$$

- *Rule's completeness*

  Rule's completeness or Support is the ratio between the number of instances in which the rule is applied and predicts correctly over the total number of instances in the class corresponding to that rule. Support is calculated according to following equation:

$$\text{Support}\ (R_i) = \frac{|\text{A} \cap \text{B}|}{|\text{A}|} \qquad (3.7)$$

  Where

  $|A \cap B|$ *is the number of instances correctly covered by the rule for given class* C_i

- *Rule's confidence (precision)*

  Rule's confidence (precision) is the probability that the rule classifies correctly the instances to which it is applied. Confidence value is calculated according to the following equation:

$$\text{Confidence}(R_i) = \frac{|\text{A} \cap \text{B}|}{|\text{B}|} \qquad (3.8)$$

- *PS measure of rule interestingness:*

  Piatetsky-Shapiro (1991) proposed rule interestingness measure, which considers the three most important metrics for evaluating rule's quality: confidence, support and coverage. This measure can be termed as PS (Piatetsky-Shapiro's) measure and is defined as:

$$\text{PS}(R_i) = |\text{A} \cap \text{B}| - \frac{|\text{A}||\text{B}|}{\text{N}} \qquad (3.9)$$

- *Comprehensibility*

  It is another metric used for rule set evaluation. It measures the complexity of rule set and rule. For a given rule set, the less number of rules is considered more comprehensible. For a single rule, the shortest rule with less complexity is considered as more comprehensible.

3.6 Summary

In this chapter the general concepts of knowledge discovery and especially their application for risk identification have been discussed and identified as an appropriate research area. The implementation of Knowledge discovery in different field of studies such as finance and banking, information system, economics, medicine etc. is a well-documented and recognised area. Although knowledge discovery in supply risk management is a relatively new area, however this emerging research domain can potentially lead to compelling results (Dani 2009, Ghadge 2013).

A typical knowledge discovery process involves nine basic steps: (1) Understanding the application domain and KD goals, (2) Selecting data and creating a set on which to perform discovery, (3) Pre-processing and cleansing, (4) Data transformation, (5) Deciding the type of data mining task, (6) Choosing the Data Mining Algorithm, (7) Employing the Data Mining Algorithm, (8) Evaluation, (9) Using the discovered knowledge. Data mining is a key component in KD process and consists of steps 5, 6, and of KD process. Data mining algorithms application enable pattern discovery. Common data mining tasks include: Clustering tasks, Association task, Summarization task, Regression task and Classification task. Any of the data mining task can be implemented in a KD process depending upon the goals and problem formulation. Classification task has been selected as being most appropriate in the context of research in this thesis. There are various techniques available for given data mining task according to anticipated outcome and given problem perspective. The selection of most suitable techniques for anticipated outcome and given problem perspective is critical question to be answered, current thesis propose a method in methodology chapter.

Risk scoring modelling is a method of quantitatively predicting the probability of risk that a company/person may be unable or unwilling to honour an obligation under terms of a business contract (required performance) and thus cause a loss (Mester, 1997). A typical risk scoring process consists of six steps (Siddiqui 2006); (1) Understanding the business problem, (2) Defining the dependent variable (3) Data, segmentation and sampling, (4) Model building, (5) Generalization, (6) Ongoing monitoring. Different techniques have been used previously for model building stage of risk scoring process, that involve simple discriminant analysis to advance neural network techniques. The current research thesis chose the logistic regression

technique depending upon its suitability with desire goal and its easy implementation. Each risk scoring model require the selection of appropriate variables, those can be best predictor of target variable. However there is no universal algorithm, but some important rules are there for selection of appropriate variables. Current research thesis will propose a method for selection of appropriate variable that can satisfy the rules explained in this chapter. Finally different classification performance evaluating methods and metrics are described.

4 KNOWLEDGE DISCOVERY BASE SUPPLIER RISK SCORING

The previous two chapters identified the research gap (chapter 2) and appropriate approaches and techniques (chapter 3) that can be used to deal with the identified issue in field of interest respectively. This chapter explains the proposed methodology that is designed to develop a supplier risk scoring model that can be used to assess the supplier risk based on knowledge discovery about supply risk.

4.1. A Novel Approach for Supplier Risk Scoring

Supply base of a company is aimed to provide competitive advantage through purchasing or outsourcing decisions. Besides providing competitive advantage, it also exposes a company to potential operating and financial losses (Chopra and Sodhi, 2004). Consequently, the purchasing decisions involve balancing the trade-off between the expected rewards (supply performance) from a supplier against its risk of loss. Additionally, it is also very important to consider the buyer's attitude toward risk in purchasing and procurement context as it significantly affect his(her) final decision about supplier selection and the order quantity (Chen et al., 2007). Previous studies considered both the risk neutral and risk aversion attitude. Although the risk aversion attitude is identified as most dominating risk attitude in previous studies for risk management in different field of subjects (Harrison and Rutström, 2008). Many prior studies in purchasing and procurement (upstream supply chain) have implicitly or explicitly assumed a risk-neutral attitude since (Harrison, et. al., 2009): the risk-neutrality assumption is suitable as it improves the systematic tractability of decisions. Furthermore, it does not require additional experiments or other analyses to assess the decision maker's attitude toward risk. The current thesis also considered the risk neutral behaviour as this study is more focused on overall expected performance gain or loss. Secondly the case study company also have risk neutral behaviour for purchasing and procurement.

The ultimate objective of the risk neutral buyer is to maximize its supply performance (reward) through purchasing decisions. In risk neutral attitude, if the supply performance is measured in total purchase value then opportunity cost of purchase value "V" at riskless value added rate "r_i" can be calculated as,

$$Opportunity\ cost = P \times V \times r_i + (1 - P) \times V_r \times r_i \qquad (4.1)$$

$P =$ probability that desired performance will be delivered
$(1 - P) =$ probability that desired performance will not be delivered
$r_i =$ riskless value add rate for purchase value V such as 100% insurance
$V =$ actual purchase value
$V_r =$ actual loss value in supply base portfolio due to suppliers' failures

However it is almost impossible to have riskless value add rate in practical purchasing decisions. The risk neutral buyers have a linear utility function, meaning that maximizing expected utility performance maximizes the overall performance. Therefore, before giving a purchase order to supplier, an expected performance is calculated to maximize the utility performance by considering the value added rate offered by supplier and values of loss for given purchase value. The expected performance values for purchase values "V" and value add rate "r" is given as

$$Expected\ performance = P \times V \times r - (1 - P) \times V_L \qquad (4.2)$$

Where, V_L is value of loss cause by supplier failure, it can be calculated as

$$V_L = q \times V_r \qquad (4.3)$$

Where, q is rate of supplier failure. The rate of supplier failure can be calculated as

$$q = \frac{number\ of\ suppliers\ failed\ to\ fulfilled\ contract}{total\ number\ of\ suppliers} \qquad (4.4)$$

Adopting Boyes et. al. (1989)'s concept for risk assessment in lending decision with risk neutral attitude. The risk neutral buyer (lender) will place order (loan the business) to a supplier (creditor), who offer an expected performance (expected return) at a value added rate "r" higher than the opportunity cost (4.1) for purchase value V,. i.e.

$$P \times V \times r - (1 - P) \times V_L > P \times V \times r_i + (1 - P) \times V_r \times r_i \quad (4.5)$$

Putting the value of V_L in inequality equation (4.5) then,

$$P \times V \times r - (1 - P) \times V_r \times q > P \times V \times r_i + (1 - P) \times V_r \times r_i \qquad (4.6)$$

By simplifying the above inequality equation (4.6), we can get following inequality

$$P > \frac{V_r(q+r_i)}{[V\times(r-r_i)+V_r\times(q+r_i)]} \quad (4.7)$$

According to equation (4.7) the supplier risk evaluation depends upon the value of "P", the probability of contracted performance will be fulfilled. If a supplier has higher probability than the critical level given at right hand side of equation (4.7), it will be given a purchase order, otherwise it will be rejected. Accordingly, if a buyer knows all the parameters in equation (4.7), it is simple to evaluate the supplier risk. The critical level can be approximated by the buyer through the analysis of purchase spending according to right hand side of equation (4.7); however the probability of supplier risk evaluation given at left hand side of equation (4.7), hinges on the buyers perception or judgement. This perception or judgement can be biased, which can lead to biased supplier risk assessment. Therefore, a tool is required that provides quantitative value for "P" that can be used in supplier risk evaluation. This proposed methodology provides straight forward quantitative value for supplier risk evaluation in term of supplier risk score. To develop a supplier risk scoring model, a dependent variable needs to be defined and independent variables need to be identified. In the context of supply risk management, supplier risk is a dependent variable and supply risks are independent variables. Therefore, efficient supply risk identification is very crucial for supplier risk assessment.

Supply risk identification will be efficient if it is done on the bases of empirical evidence rather than explicitly and mainly based on decision maker's knowledge, perception or experience. The proposed methodology for supplier risk assessment approach attempts to provide data driven supply risk identification and also explain the role of supply risk in the supplier risk assessment. It does not require a priori assumptions about the supply risk; the whole process relies on pattern discovery from available data about supply risk (i.e. empirical evidence). Figure 4-1 shows a diagram of the proposed methodology. As shown, the process has been designed so that supply risk can be taken into account implicitly during the supplier risk assessment process.

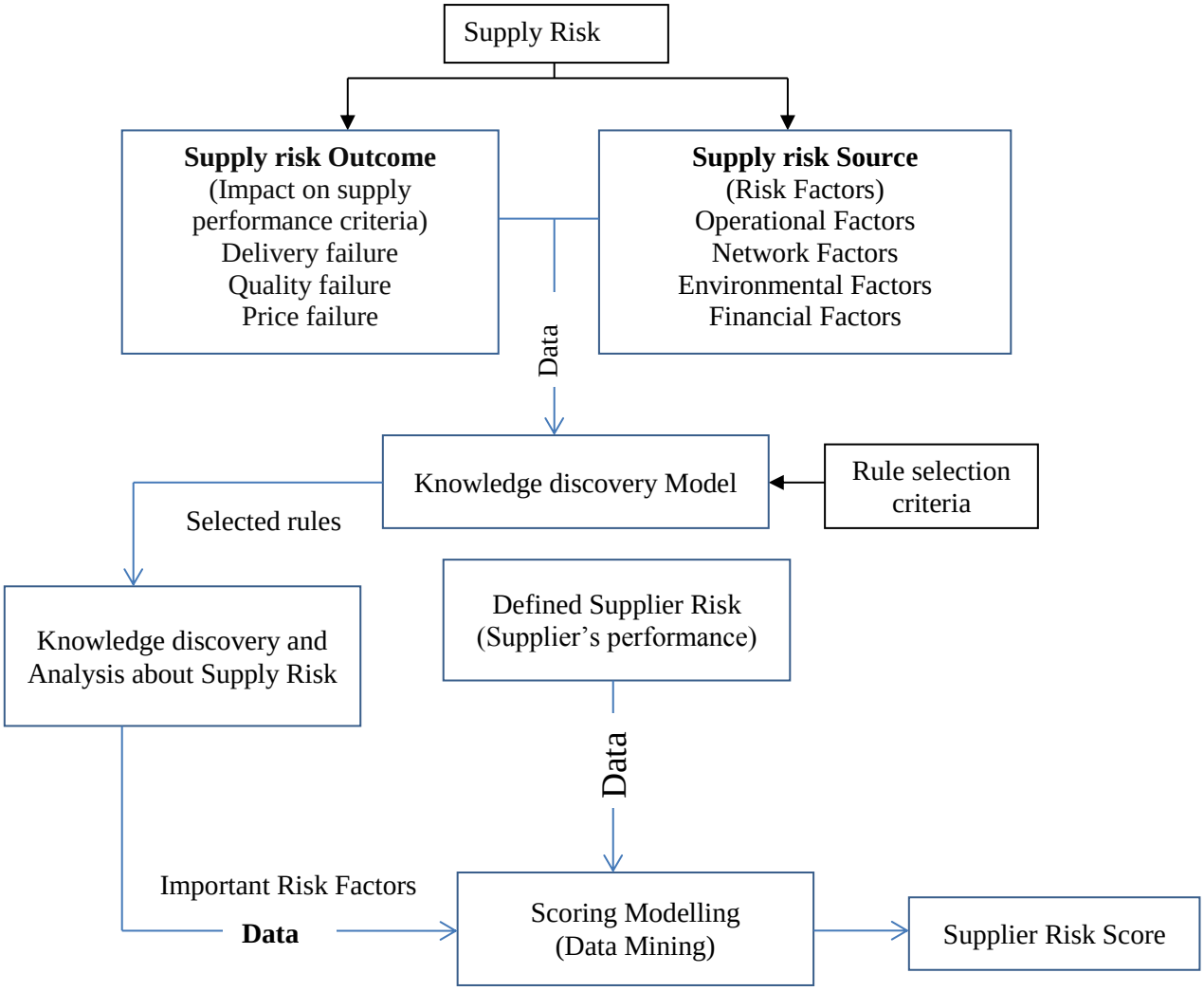


Figure 4-1: Structure of the proposed methodology

Based on definition of supply risk (chapter 2 section 2.2.1), *uncertainty associated with upstream supply chain characteristics (factor), that cause failure to meet the desired outcome is considered as supply risk.* Supply risk has two construct the source of supply risk (i.e. uncertainty associated with upstream supply chain characteristics) and outcome of the supply risk (i.e. failure to meet the desired outcome). Therefore, any supply risk identification effort should consider both constructs. These consideration are taken into account in current research thesis, first, supply risk measurement i.e. risk factors (source of supply risk) are selected according to available data and desire outcome failure (supply risk outcome) is measured on supply performance criteria for which both parties (supplier and buyer) are agreed. Furthermore, supplier risk is defined in such a way that explicitly explains the role of supply risk.

Second, the uncertainty associated with supply chain characteristics (risk factors) is the source of the supply risk, for which the decision maker may have insufficient information, understanding, knowledge and experience to recognize all the uncertainties that cause the desired outcome failure. In current study, a rule-based knowledge discovery approach used to tackles this issue as it provides information about risk factor uncertainty and desired outcome failure. Furthermore, the desired outcome (supply performance criteria) may not affect by risk factor in an isolated way. Especially within complex supply chain environments where multiple factors are affecting a supply chain operation, therefore, impact on performance cannot be determined based on one isolated risk factor alone and its contribution toward the outcome in presence of other factors. In that case, it is more valid to consider the inter-relationships (conjugations) among different risk factors with respect to their impact to desired outcome failure. A knowledge driven cross impact analysis techniques is used to analyse risk factor interconnectivity and their impact toward specific desired outcome failure. Considering the complexities associated with supply risk the application of knowledge discovery approach seems ideal. This is different from previous studies, those lies on expert opinion rather than the knowledge discovery in available data about the two constructs of supply risk. Previous studies proved that the application of knowledge discovery approach is very ideal solution to complex problem for revealing previously unknown information hidden in data.

Finally, the supplier risk scoring model is developed to identify the probability of risk as a supplier risk score. For risk scoring model, the data about identified supply risks and supplier's performance will be labelled into two classes based on the defined supplier risk. The logistic regression algorithm is applied on available data to build the data mining model that will be used for supplier risk score calculation. This make the current approach different from previous studies, which mostly focused on the aggregating the effect of supply risks rather than considering the impact of supply risk on supplier's overall risk profile. The proposed methodology consists of three main sections which are explained followed.

The purpose of supplier risk assessment is to develop a structured way of defining, identifying and assessing the supply risks with respect to given supplier. Defining the supply risk and supplier risk facilitates the application and logic of integrating the supply risk and supply performance in supplier risk scoring model. The following section will also provide listing and risk factor categorization of risk sources.

4.1.1 Definition of Supply Risk Measurement and Supplier Risk

Assessing the exposure to supply base risks requires understanding of the conditions that increase the probability of risk (loss). To clarify the conditions a hierarchy structure is presented in Figure 4.2, to understand the structure of supply risk and supplier risk that can be used to analyse the probability of risk. This will help to define the measurement and provide the inputs for data driven approach. It also provides the understating about the two constructs of supply risk (i.e. sources and outcome), and their role in defined supplier risk. This process provides a list of risk factors that can be possible a cause of supply risk and their categories. In this hierarchy, supply risks for a firm are primarily associated with failures in supply performance metric such as delivery, cost, and quality. We later utilize this information in operationalizing supply risk, and defining the supplier risk for scoring model. The following two sub-sections will explain the supply risk and supplier risk respectively.

The risk factors given Figure 4.2 are not specified name of any risk factors, just the representation of any first operation risk factor (O_1) and Nth number of operational; risk factor (O_n). Similar representation is shown for financial risk factors and other risk factors categories.

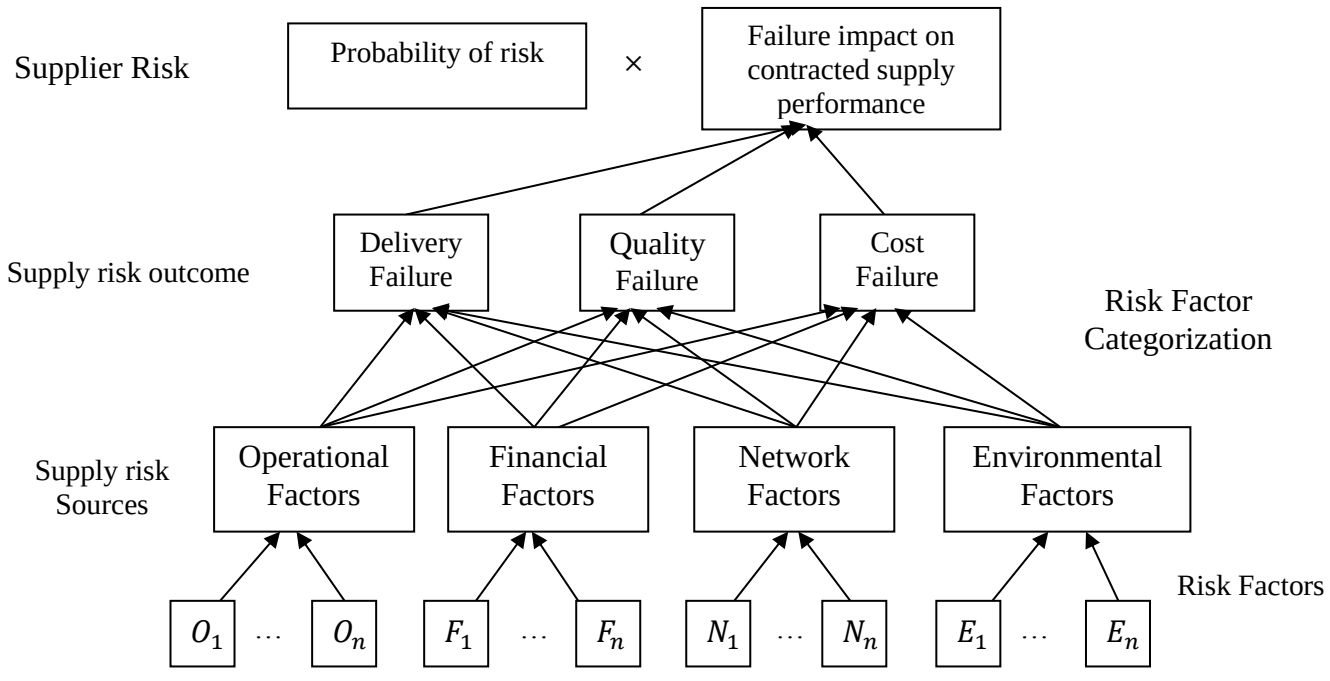


Figure 4.2: The hierarchy structural of supply risks and supplier risk relationship

4.1.1.1 Supply risk measurements

The defined supply risk has two constructs (1) sources of supply risk (uncertainty associated with risk factors) and (2) outcomes of supply risk (impact on desired outcome). Ritchie and Brindley (2007) argued that performance and supply risks are interconnected to each other and changes sought in the performance needed to be recognised as the consequences of supply risk. This argument is further supported by the Wagner and Bode (2008) empirical study that supply risk derived from different sources undermines the supply performance. Based on these arguments, in the current thesis the supply risk outcome (impact on desired outcome) is measured on supply performance metrics, those are used to calculate overall supply performance. The overall supply performance that the supply chains partners (suppliers) pursue have three target areas of quality, cost, and delivery (Sinha et al. 2004). Therefore, supply risk outcome or impact on desired outcome can be measured in term quality; cost and time as used in given case study to categories the purchased orders into different classes (see section 5.5.1.2).

- Delay risk is that the purchased order will not deliver on time as agreed also known as delivery failure
- Quality risk is that the supplied goods or services do not meet the quality requirements as specified also termed as quality failure
- Cost risk is that the goods or services will not be delivered at the price that was agreed upon when the order was placed also termed as price failure

Supply risk sources (uncertainties associated with supply chain characteristics) termed as risk factors in the current thesis, can be internal and external to the upstream supply chain. They can be located anywhere in the upstream supply chain, for example the quality standards of a supplier are internal to supplier and natural disasters occurred at the supplier location, are external to both supplier and supply chain (see chapter 2 section 2.4.1). However it is not possible to consider all the risk factors due to limitation of available resources. Therefore, to use the limited resources of an organization in the most efficient way in risk assessment process, the current thesis adopted literature review method in combination with brainstorming method. After the in-depth literature review, brainstorming session is conducted with the cross-

functional team of managers at case study Company to list and categories the risk factors.

A wide range of risk factors has been identified from the literature review; some examples of risk factors identified from literature review are given in Table 2.3. This initial list of possible risk factors provides foundation for discussion at the brainstorming sessions. In the brainstorming session it had been discussed with managers that what can be the main risk factors and how these risk factors affect the supply performance. Furthermore it has been discussed the most common factors for which the data can be readily available in industry. Each of the members of the brainstorming team from the company provided their inputs according to their knowledge, experience and resources they have in their industry. The initial risk factors list was reviewed and changed in an iterative manner until it did not provide any addition to previous list after the brainstorming session. *Note this list of risk factors is not considered as supply risk but only used to obtain the data about these factors.* **These risk factors are quantifiable, having categorical or numerical values.** *Further those factors can be exempted for which data is not readily available in organization.*

Based on the in-depth literature review and the brainstorming sessions, the risk factors have been listed and categorised into four types:

- *Financial risk factors* represent the financial state of the supplier and other financial factors such as exchange rate etc.

- *Operational risk factors* are more focused on the supplier manufacturing capabilities and processes.

- *Network risk factors* represent buyer's purchasing policy, purchasing market and purchasing network characteristics. For example does the company have sole supplier purchasing policy or dual?

- *Environmental risk factors* address the upstream supply chain's operating environment that could cause variability in entire supply chain performance.

The complete list of risk factors and their categorization based on literature review is given in appendix I.

4.1.1.2 Supplier risk

Suppose that a purchase order of purchase value "V" is placed to any supplier at value added rate "r" (percent) of purchase value then the value added can be calculated as

$$\text{Value added} = \text{r} \times V \tag{4.8}$$

If the performance is measured as total purchase value "V_p", then

$$V_p = V + (\text{r} \times V) = (1 + \text{r})V \tag{4.9}$$

If a supply risk occurred due to one of the performance metric failure such as quality or delivery failure then the total purchase value (performance) for a given order will decrease. Supposing value "Y" of supply risks, then the actual performance will be

$$\text{Actual performance} = \begin{cases} V_p, if\ y = 0 \\ (V_p - Y),\ if\ 0 < Y < V_p \\ -V_p\ , if\ Y > V_p \end{cases} \tag{4.10}$$

Equation (4.10) shows that the actual performance depends upon the value of realised supply risk. If the realised supply risk is higher than expected purchase value then buyer will loss the entire purchase value.

The buying firm develops a supply performance criteria by taking into account the supply risk associated with purchase (i.e. left hand side of equation 4.5). This performance criterion is then finalized in-form of a contract with supplier after negotiations on rewards and risks, which is an obligation for supplier to fulfil. After the contract, buyers engage in monitoring the supplier adherence to the contracted performance after end of contract; however performance metrics such as the quality, time and cost are the main focus of buyer and supplier over the given period of time.

Suppose that purchased orders "N" are given to supplier for time period "t" for contracted performance "V_c". It is considered that suppliers will fulfil the contracted performance, however due to high realized supply risk during the period; there is risk that supplier may fail to fulfil the contract and buyer face losses (i.e. the Actual performance depends upon realised supply risk: see equation 4.10). A contracted performance can have two outcomes either it will be fulfilled or not. Therefore the

probability distribution of contracted performance using Bernoulli trial can be presented as,

$$\text{Contract performance} = \begin{cases} V_c \; with \; probability \; \text{P} \; that \; contracted \; fulfilled \\ V_L \; with \; probability \; (1-P) that \; supplier \; failed \end{cases} \quad (4.11)$$

Taking in account the supply risk and requirement for defining the dependent variable for risk scoring (see chapter 3 section 3.4.2), supplier risk is defined as "*failure to fulfil the obligatory contracted performance due to realised supply risk during given time period, which cause loss to buyer*".

The above section provides definitions of the supply risk measurements and supplier risk and also provides the foundation for the data modelling process to discover the hidden knowledge about supply risk in the available data. The following section will provide the detail about the knowledge discovery model.

4.1.2. Supply Risk Identification Model (SRIM)

Since the purpose of this methodology is to perform supplier risk assessment based on the information about supply risks. Supply risk identification model (SRIM) is aimed at providing the data driven input (knowledge about supply risk) to supplier risk scoring model. Furthermore, efficient supply risk management demands proactive actions for the management of identified supply risks. However, proactive actions are possible if the system has the ability to predict supply risk before it's realized. Therefore, this study is also focused on building the Supply risk identification model (SRIM) that can also predict the supply risk before it occurs.

To build a supply risk identification model (SRIM), data about risk factors and desired outcome failure is required. In the current study the desired outcomes are measured on supply performance metrics, therefore the data about the supply performance metric is required. The purchased order receiving reports of a company show which purchase order is failed to meet the desire outcome according to required supply performance metric. Consequently, the data about the desire outcome failure is obtained from the order receiving reports. The data about most of the risk factors is already available in supply base database in the form of supplier selection parameters, network characterises, and purchasing policy etc. Additional data about the risk

factors is obtained from selected data sources. All the available data is used in conjunction with order receiving histories to build a supply risk identification model (SRIM).

KDD modelling process (see detail in section 3.1.1) is adopted to build the supply risk identification model (SRIM) using data on hand. The current research thesis is focused on achieving a knowledge driven data model which can provide the useful knowledge in-form understandable rules and able to predict unknown supply risk. To achieve the desired goal a four stage data modelling framework is implemented. These stages are (1) data processing (2) model building (3) model testing and selection (4) knowledge extraction and analysis. Figure (4.3) provides the overview of the data modelling framework for supply risk identification model (SRIM) and sequence of the processes. By implementing the data modelling framework the knowledge discovery will crop up in form of patterns (rules), to provide the input for supplier risk scoring. Additionally, at the same time this approach will provide the ability to system for predicting the unseen supply risk.

Data pre-processing stage involves data pre-processing of selected datasets from available data sources to remove the noise from the data. Model building involves the selection of appropriate data mining task and techniques according to required data pattern type and implementation of selected techniques on final dataset. Model testing and selection involves the selection of best performing algorithm for final model on evaluation criteria. Further Knowledge extraction involves the removal of unwanted rules discovered in the knowledge discovery process and further impact base analysis provide the important inputs for supply risk identification and supplier risk scoring model. The detail of each stage and appropriate techniques implemented in each stage is given in the next sections.

4.1.2.1 Data pre-processing

The objective of data processing is to obtain a training dataset for knowledge discovery, which can result in better prediction accuracy and representation. The data about the risk factors and performance metrics is obtained from different data sources. Therefore, a database is developed in the current research to collect the information from different data sources and keeping in a format which can be used for data modelling. In the database a table is created which combines the information about all the risk factors and received order history according to single purchase order and

against specific ID. This process will aid in combining the information related to one specific purchase order in one row in data the table. However the collected data has issues regarding data noises and missing values. To overcome these data problems, appropriate data cleaning and handling the missing value are applied.

- *Data cleaning*

  It is a process of removing the outliner from available data. Data pre-processing starts with data cleaning. There are different approaches for the data cleaning process but the most common approach called "dealing variable-by-variable" is applied in the current study. It is a statistical approach, where the suspicious attribute is removed due to its relationship with probability distribution of variable. For example, for a given variable the mean value in the data is 8 and standard deviation is 4, then in this process the attribute having value 15 for given variable is removed.

- *Handling of Missing values:*

  TO handle missing values, the missing values are replaced by implementing the statistical technique called "Mean substitution". For given variable, the mean value is calculated from available cases belonging to one specific class and then this value is used to replace all the missing values in the same class.

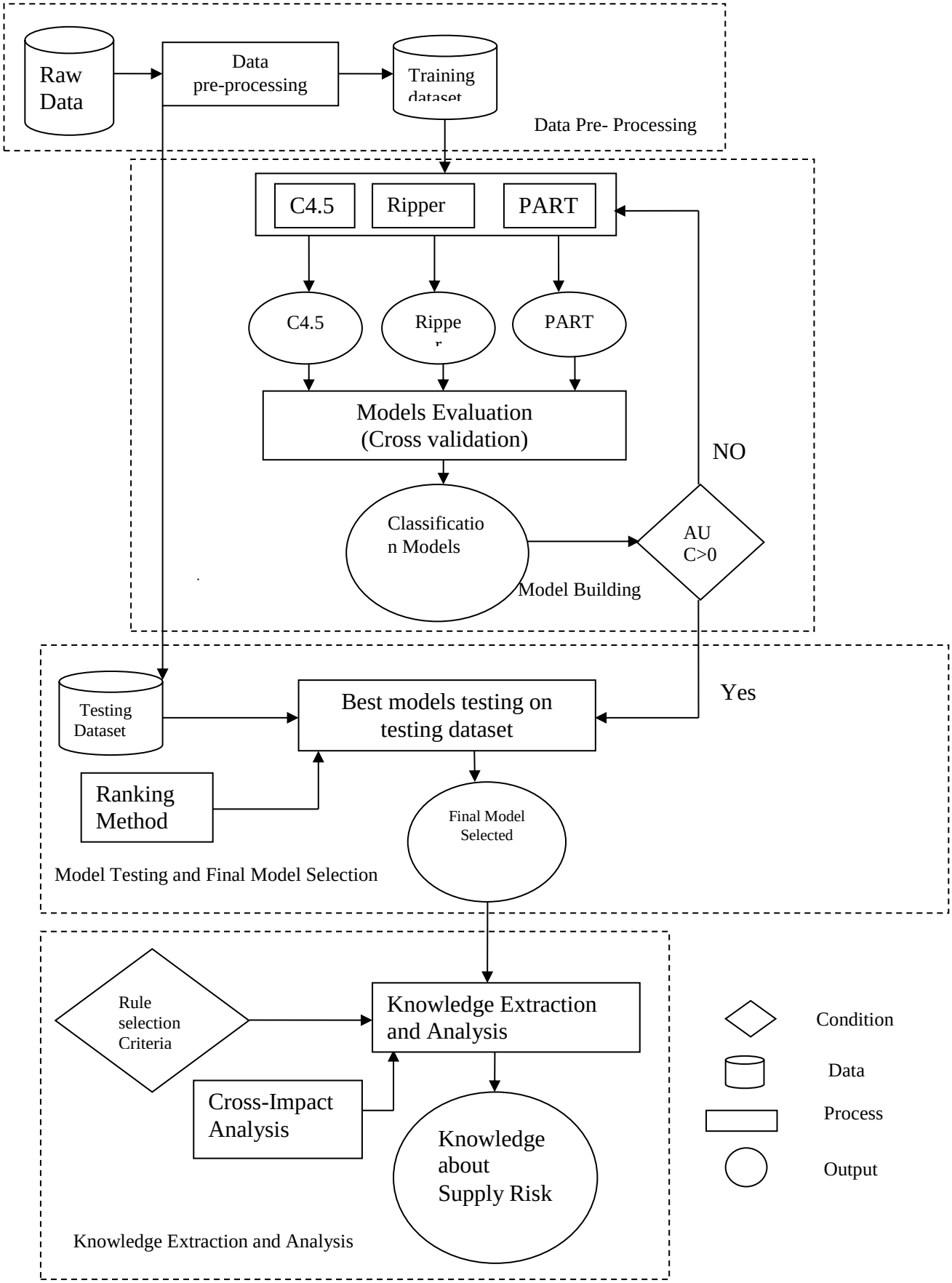


Figure 4.3: A data modelling framework for Supply risk identification model

4.1.2.2 Model building

Model building is the core component of the data modelling framework. The current thesis is focused on predicting the unknown outcome and knowledge discovery about of supply risk. Therefore, problem formation is a rule-based classification data mining task. The rule-based classification task provides a classification model that has the ability to predict unknown outcome and provide the knowledge in form of readable rules. For model building, risk factors are considered as independent variables and performance metric are considered as dependent variables (classes).

The given problem formulation is rule base classification, therefore three rule base classification algorithms: decision tree C4.5, Ripper and PART are selected. The C4.5 is selected as it most used algorithm for rule base classification and normally considered as the state of the art algorithm for rule base classification. RIPPER is selected as previous experiments reported that the classification accuracy of RIPPER algorithm is comparable to state of the art C4.5 algorithm (Yang et. al., 1999). Two main Approaches: Divide and Conquer and Separate and Conquer are used for developing the most of rule base algorithms. Both these approaches have their own procedure for developing the rules and providing the comparative result to each other in-term of classification accuracy. C4.5 algorithm belongs to Divide and Conquer approach, while RIPPER belongs to separate and Conquer approach. The PART algorithm is developing by the combination both Divide and Conquer and Separate and Conquer approaches, which also provide the competitive result to other algorithm based on its parent rule base approaches (Frank and Witten 1998). Therefore, PART algorithm is also selected for given rule base classification problem.

These three algorithms are implemented on processed dataset using Java based data mining workbench called WEKA with its default parameter setting to build classification models. However, if the condition in model building stage is not met then these parameters are changed to meet the desire condition. The cross validation method is best to use in model evaluation during the model building stage as it provide the multiple test results (Frank et al. 2004). The detail functionality of selected algorithms implementation is given below.

$C_{4.5}$ Algorithm

$C_{4.5}$ Algorithm is an extension of the ID3 algorithm for developing the optimized (minimum generalization error) decision tree model (Quanlin 1993). The algorithm consists of two phases: tree growth and tree pruning. First in tree growth phase, a

complete tree is grown by recursively partitioning the training dataset into subsets having the same class. Then in tree pruning phase, the size of grown tree is reduced to minimum generalization error of grown tree. The framework for decision tree growth and pruning is given in Table 4.1.

In the current study training set "S" consists of the independent variables set "X" which are risk factors and different classes C_i which are supply performance metric outcome. In the tree growing phase, the algorithm calculates test criteria for independent variables. The variable with highest test criteria value is selected as decision tree node to partition the training dataset "S" into subsets ((s1, s2, s3,…, s_i).

Table 4.1: The Decision Tree $C_{4.5}$ Algorithm (Rokach and Maimon 2002)

Algorithm: C4.5
Input: training set =S, variable set in S = X, target class=C
Output: decision tree =T

T={}
If
All training set S belong to same class C_i OR stopping criteria met ***then***
labelled Tree note "t" with class C_i and terminate
end if
for all *variable $X \epsilon S$* ***do***
calculate information criteria i.e. Gain ratio if split on X
end for
X_{best}= variable have the highest gain ratio
T = create tree node 't' that tests X_{best}at tree root
S_i= induced sub- dataset from S based on X_{best}
for all *S_i* ***do***
$T_{sub-tree}$= repeat above for each S_i
attach $T_{sub-tree}$to corresponding branch of T
end for
return *T*
pruning tree T
training set =S , decision tree =T ,target class=C, node=t
confidence threshold = ϕ
do
Select a node t in T such that is maximally decrease the error rate
If
$t \neq \phi$ ***then*** *pruned (T,t)*
Until t= ϕ
Return *T*

If the selected (node) variable "X_i" has numerical type data then the threshold value "h" for given values of $X_i < h > X_i$ is identified to partition dataset into subsets which maximize test criteria. If the variable values are categorical then the data is

partitioned into subset according to each value of the selected variable"X_i". After the partition of subsets the node is attached with a leaf having most common class in subset. Otherwise the node is attached with its parent node and process is continued for next variable until the leaf node is obtained. The test criteria used for selecting variable at node in $\boldsymbol{C_{4.5}}$ algorithm is normalized information gain called gain ratio. This gain ratio can be measured as

$$Gain\ ratio = \frac{\Delta_{info}}{info\ split\ (entropy)} \tag{4.12}$$

$where\ \Delta_{info}\ is\ the\ information\ gain\ after\ split\ at\ variable\ X_i$

$$\Delta_{info} = E(S) - \sum_{i=1}^{m} P_r(X_i)\ E\left(S_{x_i}\right) \tag{4.13}$$

$where\ E(S)\ is\ the\ entropy\ of\ set$ "S" or partent node "t"
$m\ is\ the\ number\ of\ variable\ values\ of\ attribute\ "X_i"$

$P_r(X_i)\ is\ the\ fraction\ of\ records\ have\ variable\ value\ x_i\ in\ set\ S\ or\ node\ "t"$

$E\left(S_{x_i}\right)\ is\ the\ info\ split\ (entropy) of\ subset\ S_{x_i} of\ training\ set\ S$

$$E(S) = -\sum_{i=0}^{n} P(C_i) \times log_2\ P(C_i) \tag{4.14}$$

$where\ n\ is\ the\ number\ of\ class\ "C_i"\ in\ set\ S$

$and\ P(C_i) is\ the\ fraction\ of\ records\ belong\ to\ class\ C_i\ in\ given\ set\ "S"$

$$E\left(S_{x_i}\right) = -\sum_{i=1}^{k} P(x_i)\ log_2\ P(x_i) \tag{4.15}$$

$where\ k\ is\ the\ total\ number\ of\ split\ made\ by\ variable\ value\ x_i$

$and\ P(x_i)\ is\ the\ fraction\ of\ records\ having\ variable\ value\ x_i\ in\ given\ subset\ S_{x_i}\ or\ child\ node$

This growing phase is continued in a recursive manner to partition the dataset into subsets until the stopping criteria met. The stopping criterion is that all the instances are labelled.

After growth phase, the non-predictive nodes of decision tree are removed in pruning stage using the error base estimation. The pruning process conducted after the growth phase is called post pruning. The basic strategy is to compare the error estimation for decision tree before and after the removal of node and then decide accordingly to minimum error. The error estimation metric can be calculated as,

$$E = \frac{\mathrm{e+1}}{\mathrm{N+m}} \tag{4.16}$$

where "e" is misclassification instances at node t
"N" is total number of instances at node t
"m" is total number of instances in training set

In the pruning process at each node, the weighted error of each child node is compared to the misclassification error rate of the new node when it is labelled with majority class after the removal of child node. The node connected connect to above node is called child node in decision tree. In the punning phase, the training data used to build the decision tree is also used to calculate the misclassification error rate at each node instead of new data. A statistical significance threshold for pruning is determined in Weka implementation termed as confidence values. Confidence value close to zero will conduct the more pruning and can result in an over pruning problem. On the other hand, confidence value close to 1 can conduct less pruning resulting in under pruning. The default values for confidence factor set at 0.25 in Weka implementation. The current thesis uses the default setting of Weka implementation to avoid any complexity in implementation stage.

RIPPER Algorithm

The rule learning algorithm Repeated Incremental Pruning to Produce the Error Reduction (RIPPER) is an extension of Incremental reduce error pruning algorithm (Furnkranz and Widmer 1994) proposed by Cohen (1995). The RIPPER algorithm consists of two phases: rule set building and rule set optimization. The framework of algorithm process is given in Table 4.2.

In the current study, we have multiple classes in our training set, therefore the algorithm order the classes according to their dominance. First, the least dominant class is selected to build the rule set. The selected class is considered as the positive class and rest of the classes are considered as negative class in a given dataset. The built rule set separate the selected class from other classes, this rule set building process continued until most dominate class is left.

The rule set building phase consists of two process, rule growth and rule pruning. In the rule set building stage, the training dataset "S "is partitioned into growing subset S_G and pruning subsetS_P. The subset S_G is used for rule growing and subset S_P is used for pruning to grow new rule.

In the growing process: Initially the rule is null, conditions (variable values) are repeatedly added to rule that maximize theoretical test criteria of rule. The condition for the numerical variable can be $X_i \leq x_i \geq X_i$, where x_i any value in variable X_i, the condition for the categorical type variable X_i will be X_i=x_i. The rule growth process

is continued by adding a condition in the rule, one by one until the new addition do not increase the test criteria value of the rule from the previous test criteria value of the rule or it is 100% correct for the corresponding class on growing datasetS_G. The theoretical test criteria called FOIL information gain can be measured as

$$G = p[\log\frac{p}{t} - log\frac{P}{T}] \quad (4.17)$$

Where p is the number of instances covered correctly by rule after adding condition, t is the total number of instances covered by rule after adding the condition to initial rule, P is the number of instance measured correctly by rule before adding the condition and T is the total number of instances covered by rule before adding the condition.

In the pruning process, a newly grown rule is pruned immediately. During the pruning process, newly grown rule is modified by removing any final sequence from the rule that results in the maximum pruning metric value of rule on the pruning subsetS_P. The pruning metric value is a ratio between the numbers of instance correctly covered by the rule to total number of instance covered by the rule in pruning subsetS_P. This pruning metric value in current implementation can be measured as

$$W = \frac{p}{p+n} \quad (4.18)$$

Where
"p" is the number of instances covered correctly by rule
"n" is the number of instances incorrectly covered by rule

The pruned rule is added into rule set, and the instances covered by this rule are removed from both growing subset S_G and pruning subsetS_P. The remaining data is combined to generate the new rule using same growing and pruning process (as described above) to add new rules in rule set. The rule set building continues in an iterative manner until the description length of the new rule set is not "64" bit greater than smallest description length of previous rule sets. The rule building process stopped, if description length of new rule set is more than 64 bit larger than smallest description length of rule set so far after the addition of the rule in rule set. After the stoppage of the rule building stage an initial rule set is obtained.

Table 4.2: The RIPPER Algorithm (Cohen 1995)

```
procedure Build Rule-Set (P,N)
P = positive examples
N = negative examples
Rule-Set = {}
DL = Description Length(Rule-Se, P,N)
while P ≠ {}
        // Grow and prune a new rule
        split (P,N) into (GrowPos, GrowNeg) and (PrunePos, PruneNeg)
        Rule = Grow Rule(GrowPos, GrowNeg)
        Rule = Prune Rule (Rule, PrunePos, PruneNeg)
        add Rule to Rule-Set
        if Description Length(Rule-Set, P,N) > DL+ 64 then
                // Prune the whole rule set and exit
                for each rule R in Rule-Set (considered in reverse order)
                        if Description Length (Rule-Set - {R}, P,N) < DL then
                                delete R from Rule-Set
                                DL = Description Length (Rule-Set, P,N)
                                end if
                        end for
                return (Rule-Set)
                end if
        DL = Description Length (Rule-Set, P,N)
        delete from P and N all examples covered by Rule
        end while
end Build Rule-Set

procedure Optimize Rule-Set (Rule-Set, P,N)
for each rule R in Rule-Set
        delete R from Rule-Set
        UPos = examples in P not covered by Rule-Set
        UNeg = examples in N not covered by Rule-Set
        split (UPos, UNeg) into (GrowPos, GrowNeg) and (PrunePos, PruneNeg)
        Rep. Rule = Grow Rule(GrowPos, GrowNeg)
        Rep. Rule = Prune Rule (Rep. Rule, PrunePos, PruneNeg)
        Rev. Rule = Grow Rule(GrowPos, GrowNeg, R)
        Rev. Rule = Prune Rule(Rev. Rule, PrunePos, PruneNeg)
        choose better of Rep. Rule and Rev. Rule and add to Rule Set
        end for
end Optimize Rule-Set

procedure Ripper(P,N, k)
Rule-Set = Build Rule-Set(P,N)
repeat k times Rule-Set = Optimize Rule-Set (Rule-Set, P,N)
return (Rule-Set)
end Ripper
```

In the optimization stage, the overall error rate of rule set obtained in the rule building stage is minimized using the minimum description length principle technique. During

the optimization stage for each rule R_i in initial rule set R, two alternative rules R_{i1} and R_{i2} are constructed from randomized data using the rule growth and rule pruning process. The alternative rule R_{i1} is generated is generated using the same way as initial rule R_i in initial rule set, however the pruning metric is aimed at reducing error rate of rule set instead of rule itself. This pruning metric can be calculated as

$$\frac{TP+TN}{P+N} \tag{4.19}$$

Where

TP is the number of instances covered by the rule correctly for positive class
TN is the number of instances covered by the rule correctly for negative class
N the total number of instances belongs to other classes
P the total number of instances belonging to class under consideration

The alternative rule R_{i2} is generated by greedily adding the condition in the original rule R_i instead of using the empty rule. The algorithm calculated the description length of the original rule R_i and alternative rules $R_{i1}$and R_{i2}. The rule with the minimum description length is kept in a final rule set and the other two rules are discarded. This optimization process is conducted for each rule in the initial rule set according to their sequence of generation. The basic strategy of the optimization is to improve the error rate of initial model by decreasing the size of rule set.

PART Algorithm

PART algorithm (Frank and Witten 1998) combines a divide and conquer strategy used for decision tree ($\boldsymbol{C_{4.5}}$) and a separate and conquer strategy used for rule learning (RIPPER). The algorithm converts the largest covering leaf of a partial decision tree into a single rule. The instances covered by this rule are removed from the training set. The remaining training set is used to build a new partial decision tree for extraction of a new rule. This process recursively continues until all the instances are covered by the rules. The method for generation of partial tree is given in Table 4.3.

Table 4.3: Method for constructing partial decision tree (Frank and Witten 1998)

Process to construct subset
Select the split of given set into subsets **While** subset that are not expanded **and** all the subset provide leaf are divided so far Choose the next subset to divide and divide it If all the subsets provide the leaf Try to replace node by leaf

A partial decision tree is an ordinary decision tree that contains the branches to an undefined sub-tree (Witten and frank 2005). The training set S containing the attributes $\{X_i\}$ and class $\{C_i\}$ is divided in to subsets in the same way as the in algorithm $\boldsymbol{C_{4.5}}$ (see above). However it differs from normal algorithm $\boldsymbol{C_{4.5}}$ that the subset having the lowest entropy value is expanded further until the leaf node appears instead of expanding all the subsets to develop a complete tree as in normal $\boldsymbol{C_{4.5}}$ algorithm. As the leaf node appeared in the partial decision tree, the pruning process is conducted on the leaf node. The pruning process is same as in the $\boldsymbol{C_{4.5}}$algorithm except no sub tree rising occur. After completion of the partial tree, the leaf with the highest instance coverage in the partial tree is selected as a single rule.

After the selection of the rule, the partial decision tree is discarded and instances covered by the rule are removed and the selected rule is added into rule set same as in the RIPPER algorithm. However, this algorithm does not conduct any optimization on the rule set as in RIPPER algorithm. The algorithm proceeds recursively to add the rules in the rule set until all the instances in training set are covered.

4.1.2.3 Model testing and selection

To selected single best model for knowledge discovery in current study, a simple ranking method is proposed. In this simple method, each classification model obtained through three different selected algorithms implementation in model building stage are tested on testing dataset. These models are ranked according to their average performance on the selected evaluation metric and assigned a ranking accordingly ranging from 3 to 1. The highest average performance model will be ranked 3; the runner up will be ranked 2 and so on, however this ranking is inverted for model's complexity metric. The average performance of each metric is calculated by equation (4.20). For example, recall is selected as evaluation metric and there are 3 total classes in data set and algorithm is PART. According to equation (4.20) the recall values of these three classes given by PART algorithm's model is summed up and divided by 3 to calculate the average performance of evaluation metric "recall" for PART algorithm's model. Then based on this average performance value PART algorithm's model is ranked ranging from 3 to 1.

$$V_a = \frac{(\sum_{i=1}^{n} v_i)}{n} \qquad (4.20)$$

v_a = the metric′s average performace value for given algorithm
v_i = the metric′s performance value for ith class
n = the total number of classes in dataset

In the model selection stage, three of the best models obtained in the model building stage, are tested on a dataset using the handout evaluation method. The average ranking of each algorithm's model is calculated for selected evaluation metrics and the highest average ranking valued algorithm's model will be considered as the single best algorithm's model. The objective of Knowledge discovery model is to provide the prediction for new data and disclose the valuable knowledge about the supply risks. Therefore, the evaluation criteria involved both the quantitative classification performance metrics and knowledge discovery metrics. The model evaluation performance metrics includes precision, recall, F-measure, area under the curve (AUC) and comprehensibility (detail is given in chapter 3 section 3.5). The final average ranking of each algorithm is calculated as

$$R_a = \frac{(\sum_{i=1}^{m} r_i)}{m} \qquad (4.21)$$

R_a = the algorithm′s average ranking value
r_i = the algorithm′s ranking for ith evalaution metric
m = the total number of evaluation metrics

The best models developed using three selected algorithms are compared based on their average ranking value. The highest average ranking value algorithm's model is selected as final model. The final selected model is used to make the prediction about the supply risk called as supply risk prediction model and selected for knowledge extraction stages.

4.1.2.4 Knowledge extraction and analysis

Rules generated by the final selected model are considered to be highly usable and readable for discovering hidden knowledge from the data. However there can be large number of the rules and all the rules may not be useful in rule set. Thus, it can be difficult for human expert to find out the valuable knowledge. The value knowledge of knowledge can measured on quantitative and qualitative measures. It is very difficult to measure valuable knowledge qualitatively, because it is influenced by importance of purpose and amount of knowledge human expert has (Wang et. al., 2002). Therefore, in current research thesis only quantitative measures are used to

extraction the valuable knowledge (i.e. select the rules) based on rule's quality metrics. The rule's quality can be measured using the support, coverage and confidence metrics (see section 3.5.2.2). Considering only one metric for rule's quality can produce the biased results. Such as considering only confidence value as important factor make the knowledge extraction very susceptible to over fitting (Furnkranz and Flach, 2005). For example, the rules with 100% confidence value but so much less support can lead to wrong results as it cannot be a truly representative of the data. To overcome the above issue, the PS measure is selected for rule's quality measure, PS measure considered both the support and confidence factor of a rule.

Furthermore to select a rule, the current research thesis adopted the principle for rule's quality (rule interestingness) proposed by Freitas (1999), which states that at a given constant coverage value, rule's quality increase with high number of correctly classified instance by each rule. In other words, at constant coverage value the higher the PS measure value provides better quality rules for knowledge extraction. Therefore, in the current thesis, a rule is selected according to following principle,

$$PS_M \leq PS_i \text{and } \mathrm{Cr_i} \geq C_a \tag{4.22}$$

Where,

$$PS_M = \text{the model average rule PS value}$$
$$PS_i = \text{the PS value of ith rule}$$
$$C_a = \text{the model average coverage value}$$
$$Cr_i = \text{the coverage value of ith rule}$$

A rule is selected if it has both its PS value and Coverage value above than or equal to model's average PS and coverage values. The Model's average PS value and average coverage value is measure as equation (4.23) and (4.24) respectively.

$$PS_M = \frac{\sum_{i=1}^{n} \mathrm{PS_i}}{n} \tag{4.23}$$

$$C_a = \frac{\sum_{i=1}^{n} \mathrm{Cr_i}}{n} \tag{4.24}$$

$$n = \text{the total number of rules in rule set}$$

In this research thesis, to analyse the inter-relationships (conjugation) among risk factors with respect to specific output (class), an approach is proposed. This approach is based on the Cross-impact analysis method (Godet et al. 1979) and termed as conjugation analysis. Cross-impact analysis is modelling approach use to analyse systematically interrelationships and mutual connections between variables based on

pair-wise expert judgments. Cross-impact analysis is suitable for explorative modelling whenever the use of theory-based computational models is not possible due to a lack of theoretical advancement. As this study is not tied to theories about supply risk so the adoption of such method is very suitable.

The conjugation analysis is slightly different in term of its implementation and objective than the basic cross impact matrix. This study uses the conjugation analysis evaluating the inter-relationship (conjugations) among the factors with respect to specific outcome (class) and their contribution toward specific output rather than factors' impact on each other as in basic cross impact matrix. In the conjugation analysis pair-wise input in conjugation matrix is based on the knowledge discovery about supply risk, however in cross impact method, the pair-wise input in cross impact matrix is based expert judgments. To complete a conjugation matrix, selected rules are used rather than using the expert judgement method.

The n*n conjugation matrix (n= number of risk factors appeared in selected rule set), showed the pair wise inter-relationship (conjugation) of risk factors in the first row with risk factors in the left column. The conjugation of a risk factor with other factors is counted by calculating the number of time considered factors appeared pair-wise in rule set, with same sequence for specific class. To complete the n*n conjugation matrix, a simple procedure is proposed based on the developed rule. A rule is set of conditions for given factors that required for classifying the given class (in decision tree algorithm rule is path from root of tree to leaf node). All the factors with given condition appear in a rule has same importance toward output classification. However for development of conjugation matrix later appearing factor is given in first row and first appearing factor is given in left column.

Tabel4.4: Example of the selected rule set

X_1= 1.83 AND X_2 <= 4.03 AND X_3 > 80.06 AND X_4 = Low THEN G-Risk
X_1 <= 2.68 AND X_2 <= 3 AND X4= Med then G-Risk
X_1 <= 1.84 AND X_4 = Low AND X_3 <= 80.08 then G-Risk

To understand the conjugation matrix filling procedure a simple example is given. For example the table 4.4 shows selected rule set, the conjugation matrix is consist of 4*4 matrixes as total number of factors (i.e. X_1, X_2, $X_3$and X_4) are 4 in selected rules set. To fill the conjugation matrixes, first pair wise conjugation among the factors is

calculated. The conjugation values of X_1 with X_2, $X_3$and X_4 are 2, 2 and 3 respectively. As the X_1 appear first in rule, so other factors are in conjugation with X_1. According to proposed simple method, X_1 is shown in left column and factors in conjugation with X_1 are shown in first row of conjugation matrixes. As $X_1$and X_2 appeared pair-wise two times in rule set, where X_1 appeared first than X_2, therefore, the conjugation value of X_1 with X_2 is shown in second row and third column under X_2 in conjugation matrixes (Table 4.5). In a similar manner the conjugation value of X_1 is calculated with X_3 andX_4. Using this proposed method conjugation matrixes is completed by calculating the conjugation values of X_2, X_3 and X_4 with other factors.

Table 4.5: The conjugation matrix

Interrelationship		OF			
		X_1	X_2	X_3	X_4
With	X_1	0	2	2	3
	X_2	0	0	1	2
	X_3	0	0	0	1
	X_4	0	0	1	0

The simple conjugation matrixes show the direct pair-wise interrelationship among the factors, however factors in a system can also be interrelated indirectly, therefore, both direct and indirect interrelationship need to be considered in order to capture true system behaviour (Fried and Linss, 2005). In order to capture the true behaviour of system, Advance impact analysis technique (Linss & Fried 2009, 2010) is used.

In the current study, according matrix filling process the sum of each row called active sum represent the strength of a risk factor toward specific class and its interrelationship with other factors. The sum of each column called Passive sum shows the strength of a risk factor toward specific class and its conjugation with other factors.

The higher active or passive sum value of a factor represent its high predictive probability toward classification and its higher interrelation with other factors. The direct active sum and passive sum of risk factor "i" is calculated as,

$$d\text{AS}\,(i) = \sum_{a=1}^{n}\left(R_{i,a}\right) \qquad (4.25)$$

$$d\text{PS}\,(i) = \sum_{i=1}^{n}\left(R_{i,a}\right) \qquad (4.26)$$

Where

$d\text{AS}\,(i)$ is direct active sum
$d\text{PS}\,(i)$ is direct Passive sum

$R_{i,a}$ *is* preceding risk factor i and succeeding risk factor a *in a rule*
n is *total number of risk factors apeared in rule set*

In order to quantify indirect interrelationships, the direct active and passive sum are extended to the order k = n − 1 for a conjugation matrix with n risks factors (Linss & Fried, 2009), where 2nd order is obtained by multiplying the conjugation matrix with itself further orders are obtained by multiplying the resultant matrix with initial conjugation matrix by n-2 time until order k is reached. The active sum $d\mathrm{AS}_k\ (i)$ and passive sum $d\mathrm{PS}_k\ (i)$ of risk factor "i" for order k is given by equation (4.27), (4.28) respectively.

$$d\mathrm{AS}_k\ (i) = \sum_{a=1}^{n}\left(R_{i,a} \times dAS_{K-1}(a)\right) \qquad (4.27)$$

$$d\mathrm{PS}_k\ (i) = \sum_{i=1}^{n}\left(R_{i,a} \times dPS_{K-1}(i)\right) \qquad (4.28)$$

Indirect active sum and passive sum of a risk factor "i" is calculated adding up all the direct active and passive sums from the first order to order k respectively, given as,

$$I\mathrm{AS}\ (i) = \sum_{K=1}^{n-1}(d\mathrm{AS}_k\ (i)) \qquad (4.29)$$

$$I\mathrm{PS}\ (i) = \sum_{K=1}^{n-1}(d\mathrm{PS}_k\ (i)) \qquad (4.30)$$

Where

IAS (i) is *in*direct active sum
IPS (i)is *in*direct passive sum
$d\mathrm{AS}_k\ (i)$is direct active sum of K order
$d\mathrm{PS}_k\ (i)$is direct Passive sum of K order

The higher indirect active and passive sum of risk factor "i" shows the high interrelation with other risk factors and impact toward specific class. Consequently, high level of "integration" of a risk factor indicates that strong interrelations with other variables in a system and its predictive power toward specific class. The "integration" I (i) of risk factor "i" is represented by the arithmetic mean of relative indirect active sums and relative indirect passive sum in equation below,

$$\mathrm{I}\ (i) = \frac{I\mathrm{AS}_r(i) + IPS_r(i)}{2} \qquad (4.31)$$

To calculate "integration," of risk factors, indirect active and passive sums needed to be converted in relative values Linss & Fried (2010). The indirect active sums and

passive sums are put in relation to their common maximum value in-order to calculate the relative indirect active sums $IAS_r\ (i)$and relative indirect passive sum $IPS_r\ (i)$ as,

$$IAS_r(i) = \frac{IAS\ (i)}{Max_{i=1}\ \{IAS\ (i); dPS\ (i)\}} \times 100 \qquad (4.32)$$

$$IPS_r(i) = \frac{IPS\ (i)}{Max_{i=1}\ \{IAS\ (i); IPS\ (i)\}} \times 100 \qquad (4.33)$$

The previous section of proposed methodology i.e. Supplier risk identification Model (SRIM) was dealing with the risk identification stage, now the next section of the proposed methodology i.e. risk scoring model deal with the risk assessment stage.

4.1.3 Risk Scoring Model

The estimated risk score Pi will be used as a proxy to the contracted performance probability P required according to equation (4.7), which provide a condition

$$P = P_i + e \qquad (4.34)$$

Where

e is error associated with the statistical estimation.

Risk scoring model is a function, f(X, β), where X is a vector of independent variables that are the significant predictor of dependent variable and "β" is the vector of weight for these independent variables i.e. the regression analysis co-efficient values (see section 4.1.3.2). In the context of current study, supplier risk is a dependent variable (defined in section 4.1.1.2). The formulation of supplier risk scoring requires the selection of appropriate variables that are the significant predictor of dependent variable (supplier risk). A method is proposed for selection of appropriate independent variables and further discretization process is proposed to reduce the redundancy of selected variables.

The estimated weight "β" of independent variables will be used to calculate the risk score. A data mining technique is required to estimate the weight "β"of independent variables. The following section will provide details of appropriate variables selection method and technique to estimate the weight of selected variables.

4.1.3.1 Variable selection and discretization

The goal is to select the variables that results in a "best" model within the context of supplier risk scoring model development. Therefore, a simple method is proposed, to

select appropriate independent variables will most likely in result more stable and easily generalized model. Further selected numerical type data variables are converted into categorical type data as the numerical type data can increase also increase the redundancy problem.

According to definition, supplier risk is *failure to fulfil the obligatory contracted performance due to realised supply risk during given time period, which cause loss to buyer*". Therefore, supplier risk is caused due to supply risk; hence the variables that are significant predictor of supply risk can be the significant predictor of supplier risk. The active sum and passive sum provide the information of how much a variable has predictive power toward specific class i.e. supply risk outcome. Therefore, in the current thesis, the variables selection principle is given as

$$X_i selected\ if$$
$$IAS_r(i) \geq IAS_r\ (averge)\ \ Or\ \ I\ (i) \geq I\ (averge)\ Or\ IPS_r(i) \geq IPS_r\ (averge) \quad (4.35)$$

Variables have higher active or passive sum or integration value than average value will be selected. According to principles explained for variable selection (see section 3.4.4), selected variable should be logical and predictive, furthermore excluded variable should cause less information loss. The variable with high active or passive sum or integration value represents the more predictive contribution toward specific supply risk outcome. The supply risk outcome (see Figure 4.2) is linear relationship with the actual performance of supplier (equation 4.10). Consequently, it is assumed that variable has high predictive power for supply risk outcome also have high predictive power toward supplier performance (failure to fulfil the obligatory contracted performance). Further the lower values will not cause unacceptable levels of information loss.

The selected numerical type variables data are discretised based on the knowledge discovery. Selected numerical type variable is converted into categorical type (groups or bins) through the cut off values of selected variable shown in selected rule set. Numerical value of selected variable is replaced by categorical values or bins by considering following steps:

1. Round the cut of values to whole number if the numerical type variable has maximum values more than 10 in data set, otherwise round to one digit point. After rounding apply step 2.
2. Select only two most repeating cut off values of selected variable in selected rule set and convert numerical type variable into categorical type variable with three groups. The repetition of cut-off value should be more than or equal to 20 percent of total number of the cut-off values appeared for given variable in rule set. Otherwise consider only one most repeating cut off value with minimum repetition threshold value (20%) and convert the selected variable into binary type.
3. If no cut value meets the required minimum threshold value of repetition i.e. (20%), then round the whole number to nearest zeros and round the digital value to whole number and then apply step2
4. If more than two cut-off values have same repetition, then selected those two values those increase the rage between two selected cut-off values.

Table 4.6: Example of discretization based proposed method

For example $X_1$is selected variable for risk scoring, which has numerical type data ranging from 0 to 9. Selected rule set in table 4.1 shows the cut off values forX_1. According to interesting rule set, two most repeating cut off values of $X_1$are 2.7 and 1.8 (round to one digit point) also have repetition more than required threshold. According to proposed discretization method, after discretization X_1 will have 3 groups that are

$$X_1 > 2.7$$
$$2.7 \geq X_1 \geq 1.8$$
$$1.8 > X_1$$

Similar X_2 and X_3respectively,

$$X_2 > 4, \qquad 4 \geq X_2 \geq 3, \qquad 3 > X_2$$

$$80 > X_3 \text{ or } X_3 \geq 80,$$

4.1.3.2 Model building

In the current thesis, Logistic regression technique is selected for model development to estimate the weight of independent variables i.e. "β_j". Estimation and validation of supplier risk scoring model requires data on supplier contracted performance and its potential predictors (risk factors). In the current study we have "n" number of observation for (X_i, C_i) in given dataset S. where $X_i$are independent d-dimension vector of risk factors and C_i is target variable with binary value (1, 0). It holds value C_i =1 if supplier is good (no risk) and C_i=0 otherwise. A probability function p for the given situation can be expressed through logistics model as

$$P(X_i) = \frac{e^{\sum_{j=1}^{i} \beta_j X_{ij}}}{1+e^{\sum_{j=1}^{i} \beta_j X_{ij}}} \quad (4.36)$$

Where,
P (X_i) is the probability that ith supplier is good means C_i=1
(1- P (X_i)) is the probability that ith supplier is bad means C_i=0
X_{ij} is j^{th} variable of i^{th} supplier
β_j is weighted estimator or co-efficient parameters for j^{th} variable in logistic regression.

The logarithm likelihood for this model can be expressed as

$$\log(\beta) = \sum[C_i logP(X_i) + (1 - C_i)(1 - logP(X_i)] \quad (4.37)$$

In current study the WEKA data mining bench tool for building the logistic regression model that use the ridged estimation technique (Cessie and Houwelingen 1992) for estimating the weighing estimators "β_j"of variables. In this technique the difference between two successive estimated parameters is restricted to$(\beta_{j+1} - \beta_j)^2$. The ridged parameter controls the values of "β_j". When the ridged parameter is equal to zero, then solution is same as ordinary MLE, however when ridged parameter tends toward infinity the values of "β_j" tends toward zero. Therefore, the default setting of ridged parameter is used when using Weka base logistic regression implementation for obtaining a good estimation of "β_j" for model building.

After the development of regression model for risk scoring, it is needed to be evaluated for its predictive power. An appropriate benchmark rate is required to compare the predictive power of risk scoring model for evaluation. In general, the benchmark for the dichotomous model is 50 percent because the dependent variable is binary, however in most cases the portion of target classes in a given population are not same. Consequently, the Neter (1996)'s method is used for calculating the benchmark rate in condition of unbalance data about target class population. This method assumes that the observation can be classified correctly at the same rate as their population portion. Such as, suppose there is 80 percent of good supplier and 20 percent of bad supplier in known population then benchmark can be calculated as,

$$0.8 \times 0.8 + 0.2 \times 0.2 = 68\% \quad (4.38)$$

So the developed model predictive accuracy should be higher than the calculated benchmark.

4.1.3.3 Standardized risk scores

The aim of current research is to develop a standardized risk scoring measure, similar to well-known credit scoring model with fixed score limit ranging from minimum score to maximum score such as from 300 to 900 (FICO Score range).

To generate such standardized risk score, first, a raw risk score for each supplier in available dataset is calculated by adding up the β_j values of independent variables estimated though logistic regression model. The β_j values depend upon the specific profile of each supplier, based upon categorical values for each independent variable, in which supplier falls. The raw score is calculated as

$$Raw\ Score_i = \alpha + \beta_1 + \beta_2 + \cdots \ldots \ldots \ldots \ldots + \beta_j \quad (4.39)$$

Where

$$\alpha = intercept\ of\ logistic\ regression\ model$$

$$\beta_j = \text{weighted estimator or co} - \text{efficient parameter for } j^{th} \text{ variable}$$

Standardized risk score is calculated based on the raw scores, as

$$\Delta_i = Score_{max} - Score_{mini} \quad (4.40)$$

$$\omega_i = Raw\ Score_i - Score_{mini} \quad (4.41)$$

Where

$$Score_{mini} = the\ minimum\ value\ of\ Raw\ Score_i\ for\ all\ suppleirs\ in\ dataset$$

$$Score_{max} = the\ maximum\ value\ of\ Raw\ Score_i\ for\ all\ suppleirs\ in\ dataset$$

$$Score_i = Range_{mini} + \{(\frac{\omega_i}{\Delta_i}) \times (Range_{max} - Range_{mini})\} \quad (4.42)$$

To use the scorecard after its development, a minimum score value is determined according to objectives of scorecard implementation. The minimum score value is termed as cut-off value and can represent the threshold for risk, profit depending upon scorecard user's objectives. A detailed analysis was conducted to determine best cut off level that produced the best, most reliable and useful results.

4.2 Summary

This chapter covered the aspects of design and modelling of the proposed approach that accommodates the knowledge discovery into supplier risk assessment. This approach is not based on the development of a fixed, enormous and hard coded mathematical model where variables are fixed, rather based on knowledge discovered about supply risk in available data of supply chain characteristics and supply performance. To meet this, an appropriate supply risk identification model that will be the base for the supply identification and input of risk scoring model is proposed. It is a knowledge discovery model that captures the relationship between supply sources (i.e. uncertainty of supply chain characteristics) and supply risk outcome (i.e. their impact on the supply performance). This relationship is expressed in terms of simple readable and easy to understandable classification rules. These classification rules enable the investigation of higher order inter-relationships and mutual connection among risk factors through conjugation analysis.

For the classification rules, three different rule base data mining algorithms are implemented on available data, through proposed method best model is singled out for rule selection. Rules obtained from single-out model are scrutinized based on rule's quality. These selected rules are used to develop a conjugation matrix. The conjugation matrix provides the inter-relationships among risk factors and calculates the importance of each factor within supply risk identification system.

These values are used to select the appropriate variable for developing the supplier risk scoring model. The logistic regression technique is used to build the model utilizing the data about the selected variables and supplier's performance. The outcome of the built model is used to develop standard supplier risk score with fixed maximum and minimum range. All these processes are conducted in an integrated manner to get the required output i.e. supplier risk score.

5 SYSTEM DESIGN AND IMPLEMENTATION

The previous chapter provided the detail description of the proposed methodology to meet the objectives of this research and the rationale behind its formulation. In this chapter, issues related to the design of the developed system that facilitates the proposed methodology are presented. The content of this chapter falls into two parts. The first part provides a procedure for the design of the system that covers all the aspects of the design phase of the whole system. The second part is more focused on the implementation of the designed system for its validity. First a background of the case study company is provides and then the identification of the datasets used for validating the proposed system for supplier risk assessment is discussed.

5.1 Design Methodology

The Unified Modelling Language (UML) methodology is used for the description of the detailed design of the system. UML is a general-purpose visual modelling language that includes notation and general guidelines for required specification, semantic concepts, visualisation, construction and documentation of the software system (Rumbaugh et. al., 2005). It describes the information about the static and dynamic behaviour of a system through inclusion of static, dynamic, environmental, and organisational parts of system. All these parts are categorised into views that are expressed in the form of different diagrams. Views can be further organised into four major areas: structural, dynamic, physical and model management (Rumbaugh et al.,2005).

Each of these four major areas can be viewed with different view diagrams, such as structural classification can be represented by the class diagram, internal structure, collaboration diagram, component diagram and use case diagram. Actor and class are the key elements of the structural classification, where (e.g. actors) represent behavioural concepts and class represent the objects. Dynamic behaviour can be describes by the use of state machine diagrams, activity diagrams, sequence diagrams and communication diagrams. Dynamic behaviour shows the activities of the system over the time. Finally, the physical layout is represented by conceptual architecture diagram and the model management is represented by technology package containing views that explain the computational resources and the organisation of the models.

Diagrams that were used to fully explain the proposed system are: the conceptual architecture diagram, use case diagram, technology package diagram, class diagram, and the sequence diagram. These are further analysed in the following sections.

5.2 Conceptual Architecture and Requirements of the System

The design of an integrated system that incorporates knowledge discovery functions for supplier risk assessment is consisting of four main tiers: application layer, operational layer, database layer and data collection layer.

The *Application Layer* contains User Interface components for service provided by supplier risk assessment system. The User Interface components are consists of the variables selection, user input data and visualisation of required results. The *Application Layer* communicates with implemented workflows and user actions are performed in the operational Layer. *Application Layer* is the main access point for users of the system.

The *Operational Layer* integrates all required functionality of system that can realise the concepts of proposed methodology. The *Operational Layer* is consists of supply risk identification engine and risk scoring engine. Supply risk identification engine is the implementation of knowledge discovery base data mining algorithm and proposed conjugation analysis. The main functionality of the knowledge discovery engine is to implement the data mining algorithm for the extraction of the knowledge from the data and give it to knowledge extraction and analysis engine. The knowledge extraction and analysis engine develop the conjugation matrix and perform the analysis according to proposed methodology (section 4.1.2.4) for developing the knowledge about supply risk. The knowledge about supply risk is feedback to risk scoring engine that provide the supplier risk score. The risk scoring engine is consists of the data mining engine and standardized score engine. The functionality of the data mining algorithm is to identify weight-age of selected variables and placed it in standards score engine that perform supplier risk evaluation.

The *Database Layer* is a construct of system that maintains the required data. *Data Layer* constitutes a distributed infrastructure, where, each engine is responsible for their data. Figure 5.1 illustrates the complete conceptual architecture of system.

The *Data collection Layer* is responsible for collected the data from different identified data sources according to required business needs. The data obtained is then transformed into a file that is used to load the data into main database inhabit

Database Layer. However this layer is not integrated part of the system. Current all these functional are done separately and transformed file is loaded into database.

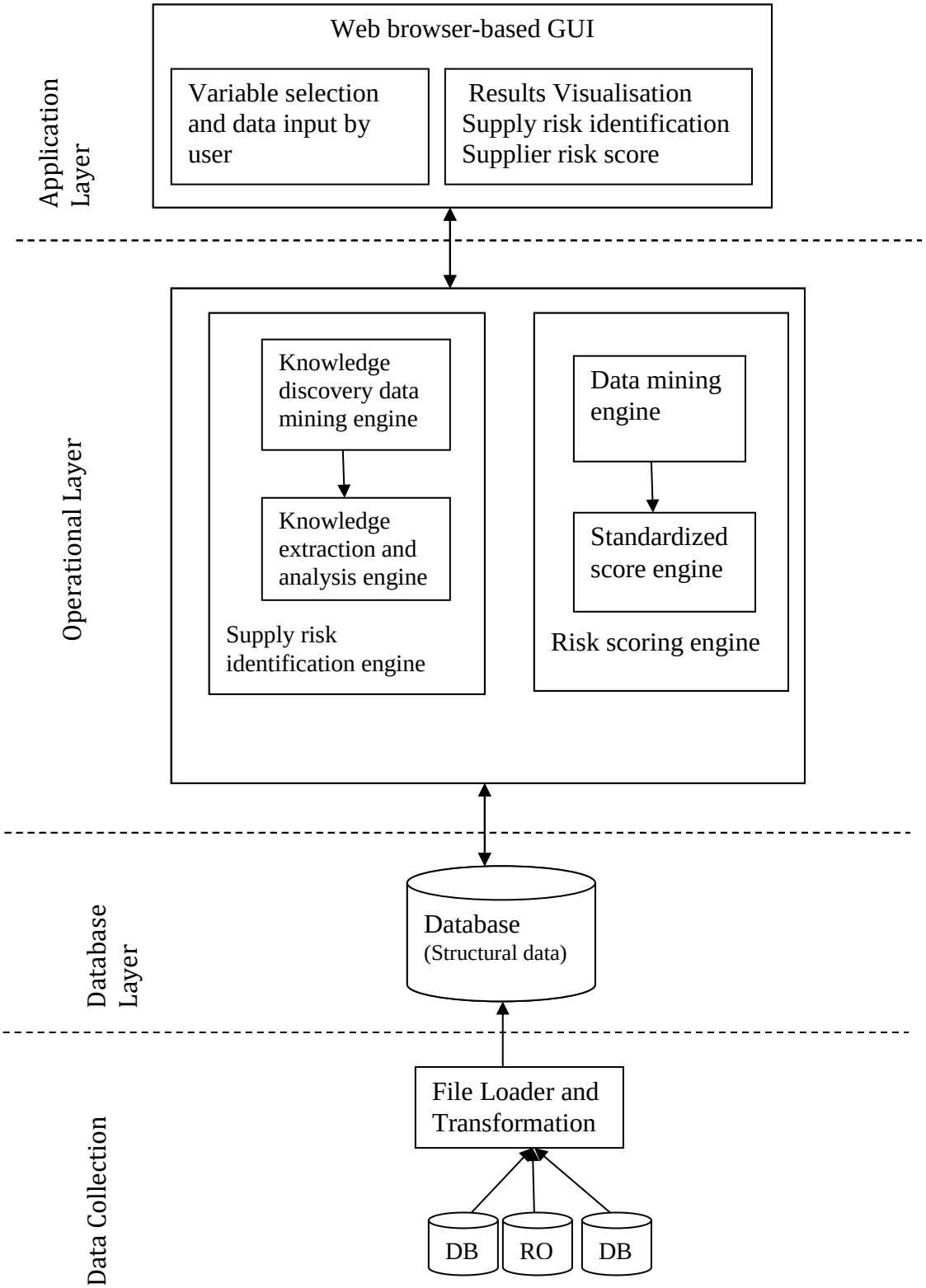


Figure5.1: The conceptual architecture of supplier risk assessment system

The focus of this research was on the design and development of the three first components: *Application Layer*, *Operational Layer* and *Data Layer*. A number of requirements for the proposed system have been set. Such a system:

5.2.1 Support for Rules Selection

The input is a flat file from database of structural data containing the characteristics and class that is given to knowledge discovery data mining algorithm. The data mining algorithm provide the classification rules for given class. These rules needed to be filtered to select the specific rule that meet the required criteria for rule selection. The data mining algorithm were used from the Weka repository that does not provide the function for specific rule selection. Therefore, the designed system should support the calculation of each rule's quality and then according to defined criteria selection of the rules.

5.2.2 Support for Automatic Development of Conjugation Matrix and Analysis Operations

The system must support development of conjugation matrix and analysis operations or provide the required tools for the development of these operations. Additionally, it must facilitate the efficient storage of the matrix and their calculation. The whole matrix can be considered relatively stable with the selected rules set being the only possible source of frequent updates. In the case of new model development, the update of the selected rule set is quite straight forward. It first involves the calculation of the rule's quality of the new model then rules are selected based on the required rule's quality criteria. Then system should be able to develop the conjugation matrix and calculations based on selected rule set.

5.2.3 Support for the Integration of Knowledge Discovery for Risk Scoring

According to proposed methodology, the knowledge discovery about supply risk gives feedback to supplier risk scoring model, by identifying the important variables and discretization of the numerical type variables. The need to integrate the knowledge discovery with the risk scoring is required for system development. This integration will provide unified approach for developing the input flat file from database for data mining component of risk scoring engine.

5.2.4 Usability

The portal should be simple and should give explanations about the entities to user, wherever it is necessary. It should ensure similar visual experience for users, for different screen resolutions.

5.2.5 Performance

The system should repose as fast as possible for given actions triggered by users. This can result good user experience.

5.2.6 Software as a Service (SAAS)

It should support at least the most common used browsers such as

- Google Chrome
- Mozilla Firefox
- Microsoft Internet Explorer

5.3 System Design

System design section will explain the structure and dynamic behaviour of the whole system with the utilization of the use case diagram, resources utilization, class diagram and sequence diagram.

5.3.1 Use Case Diagram

Use case diagram is used to portray the external behaviour of the system as this can be viewed from outside users. Use case diagram presents a logical explanation of the required functionality. Therefore, it can be used as a first presentation of the usage requirements of the system.

This use case diagram demonstrates the system's functionality in relation to its users. This functionality is provided by the classifiers and expressed in terms of their interactions. Classifiers shown in this diagram are actors and use cases. Relationship types such as generalization, usage and association, model the interactions between the classifiers. It should be noted that these actors denote roles that do not necessarily coincide with real persons but they can represent processes or other systems.

The functionality provided by each classifier should be in accordance with the main objective of the system, which is to support development of supplier risk score by utilising knowledge discovered about supply risk in available data. Hence, each of the

use cases represents a piece of functionality that can either be autonomous or can be mixed with that of other use cases.

Figure 5.2 presents a high-level overview of the usage requirements of such a system. In the Figure 5.2 use cases are drawn as ellipses while different types of links denote the type of relationships between actors and use cases. Two main actors can be distinguished, modeller and user. Modeller associates with Import data use cases. User associates with the supplier risk evaluation.

A generalisation relationship holds between *the import data* use case and the supply chain characteristics specific to purchase order, purchase order performance/outcome, supply chain characteristics specific to supplier and supplier's performance use cases denoting a parent-child relationship. Classification rule base classifier is a generalisation of the classification rules, determined supply risk identification classifier denoting again a parent-child relationship. Use cases that are linked with dashed arrows denote a usage dependency.

Use case *Import data* is a key use case. Its purpose is to deal with the construction of the input dataset. This case is invoked when the actor Modeller initialises the application. The Modeller defines a number of parameters that relate to the size and dimensions of the input dataset. On completion, the input data is prepared to be further processed.

Use case *Classification rule base classifiers training* is also a key use case. Its purpose is to develop the best model for given algorithm and then select the best performing model. The best performing model discovers the knowledge in-form of classification rules in the input data. The best performing model act as the supply risk identification model and predict the unknown output based on user specified inputs. This use case is conducted when a valid input dataset based on user specified criteria has been generated. Classification rules and prediction value for given specified input stored and displayed on the screen. Use case *knowledge extraction and analysis* is responsible selected the rules based the rule's quality. Further based on the selected rules the conjugation matrix are developed and analysed. It is invoked when the *Classification rule base classifiers training* use case is completed.

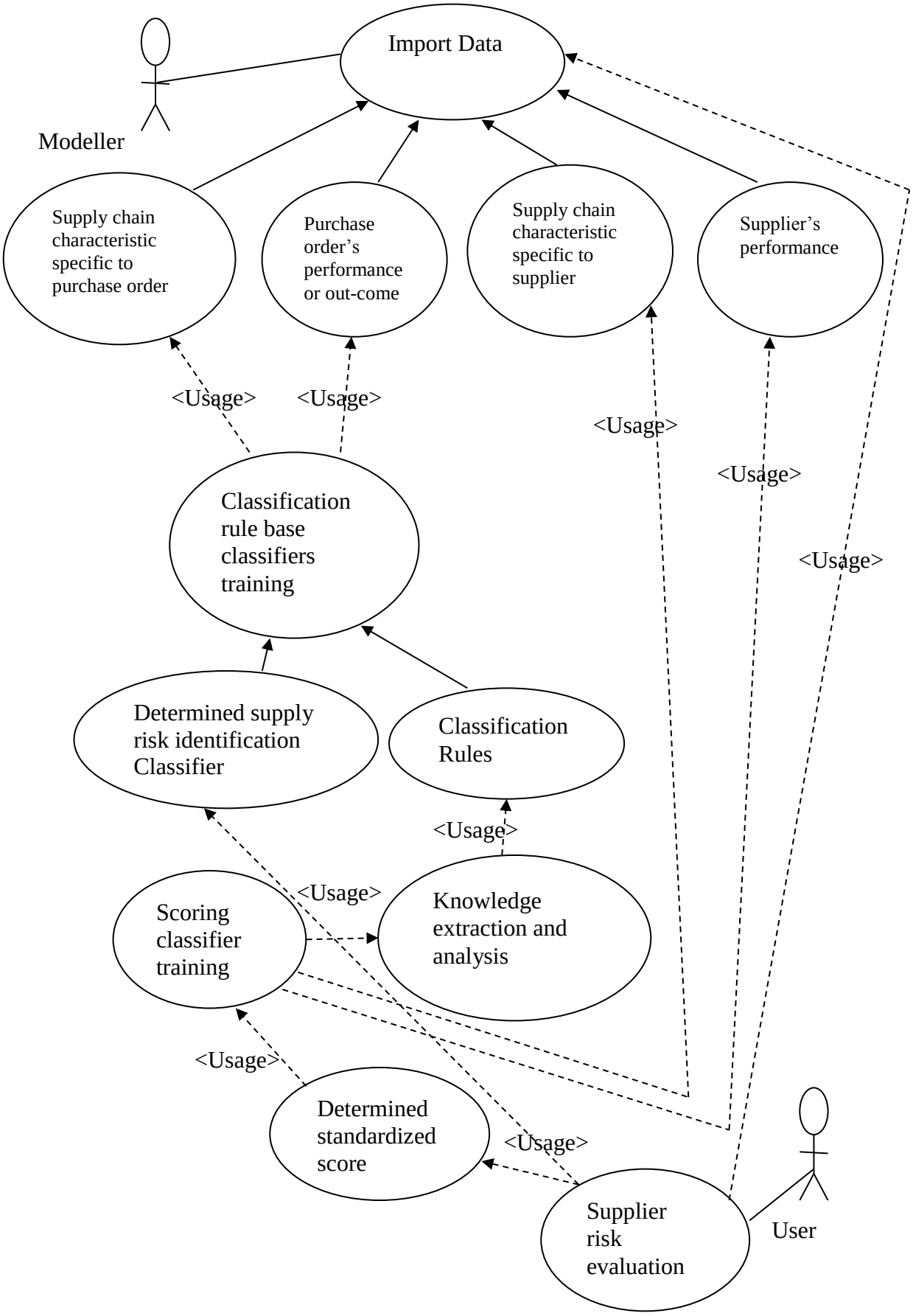


Figure 5.2: The Use case diagram of supplier risk assessment system

Use case *scoring classifier training* is responsible for the determination of the classifier for providing the estimation weight of the selected variables with respect to dependent variable. It is invoked after the completion of the *knowledge extraction and analysis* use case. It uses Supply chain characteristic specific to supplier, Supplier's performance and *knowledge extraction and analysis* to determine dataset for the scoring classifier. On completion it provides the input to the *determined standardized score* use case. The *determined standardized score* use case provide the standardized score for the generated dataset.

The final use case is *the supplier risk evaluation* use case. It is invoked by the user and is responsible for the supplier risk assessment function. It uses the *determined standardized score* in conjunction with a test dataset or a test case. It calculates the supplier risk score and finally displays.

5.3.2 Technology Packages

This section provided an overview of the technologies used to build the supplier risk assessment system. Figure 5.3 shows the technologies used for the system in the form of a package diagram. These were used to develop the whole system. A brief description of them follows.

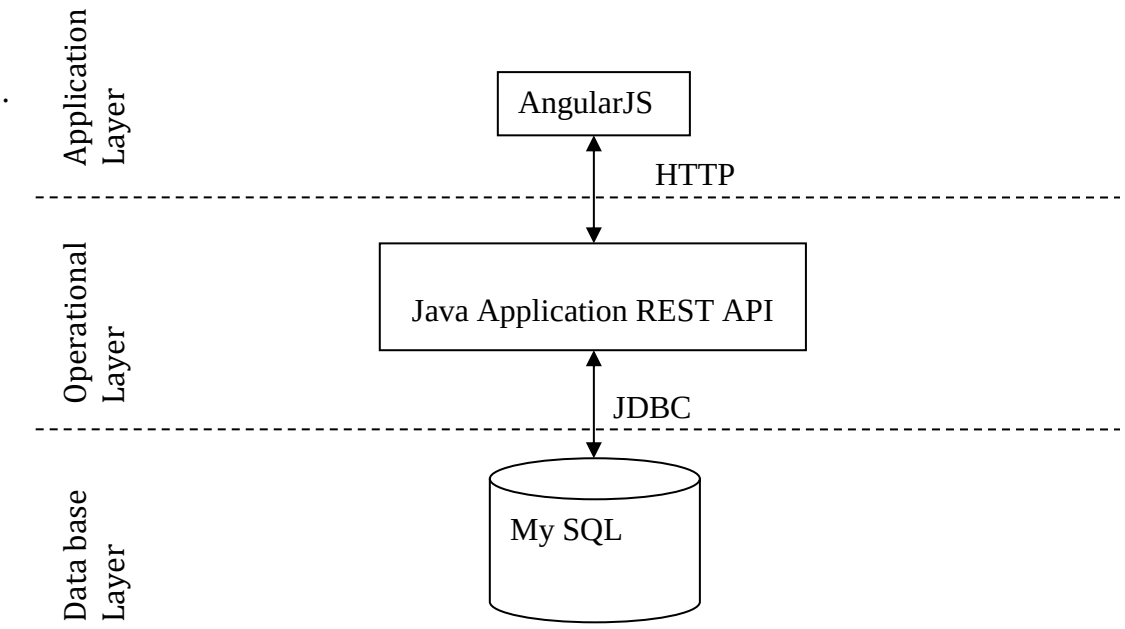


Figure 5.3: The Package Diagram

5.3.2.1 AngularJS

The Application Layer of the supplier risk assessment system is implemented using AngularJS. The framework introduces HTML tags (Directives), which dynamically bind data to the corresponding HTML structure. It is very easy to fetch data itself from a RESTful Web service; AngularJS works together with any server-side technology (AngularJS Developer Guide 2014).

To fulfil the requirements of the supplier risk assessment system, AngularJS seems an ideal choice. It enables seamlessly separation of the view and user experience from the server-side implementation.

5.3.2.2 Java application RESTful API

The Application Layer communicates with the Operational Layer through API calls using the Representational State Transfer (REST) pattern.

Supplier risk assessment system is implemented and use a straightforward, modern RESTful API based on the HTTP protocol and JSON as the representation format. REST is not fixed to a pre-definite technology, but is a bundle of requirements for designing an API. Following this requirements leads to a state-of-the-art, robust API architecture. Some core constraints are (Sandoval, 2009):

- It should have client-server architecture
- The requests should be independent from each other
- Uniformly Accessible – resources have to have a unique address

Supplier risk assessment system is implemented such an API in order to fulfil the requirements for the proposed architecture. The main conditions of a RESTful API regarding Supplier risk assessment system are given as:

- Resources: A resource is everything within the platform, which is addressable. In current system this can be Supplier risk assessment model.
- Representation: The representation is the kind of data, which is communicated between user and server. For supplier risk assessment system the common representation in JSON is used to keep the communication simple between the JavaScript based Application Layer and the operational layer.

5.2.3.3 Database

Each engine in operational layer of system is required data for its functionality and therefore a database is required. The data structures required for supplier risk assessment is tabular, where there are relations among factors, a relational database management system (RDBMS) seems ideal. The open source software MySQL has been selected for this purpose. MySQL is licensed under the GNU General Public License and is an established, reliable and second most used database system in world (What is MySQL, 2012).

5.3.3 Class Diagram

In the class diagrams a graphical representation of the model's static elements (classes, relationships) is provided. The algorithms described are implemented in the classes discussed below. Classes are drawn as rectangles while their inter-relationships as arcs using the Enterprise Architect 12. Enterprise Architect is a high performance modelling, visualization and design platform based on the UML 2.5 standard.

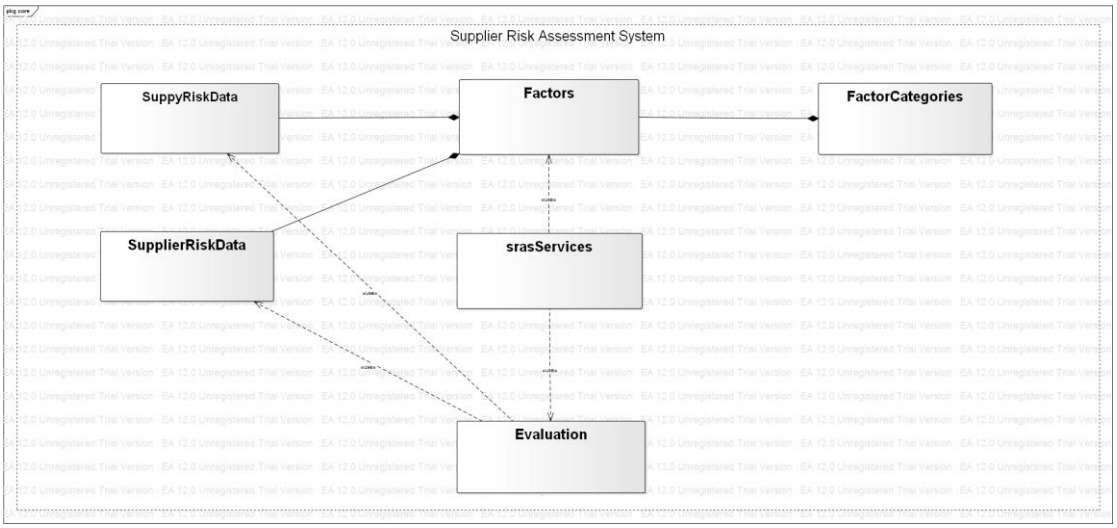


Figure 5.4: The class diagram of supplier risk assessment system

Figure 5.4 is the class diagram for the supplier risk assessment system. The supplier risk assessment system contains six classes: the SRASServices, Evaluation, Factors, Factor Categories, Supply risk data and supplier risk data. SRASService is the main service (server) of the system. It include the methods that use the class Factors (e.g. retrieveFactors (int.), retrieveFactorsvalues ()) and Evaluation (e.g. Evaluation (JSONObject)).

The filled diamond symbol shows composition relationship between classes which denotes the class opposite side of the diamond has life dependency on the other class. Factors class has the dependency on the Factorcategories class. It includes method that deal with the factors (e.g. getFactorsID()). Factors class is responsible for interfacing to the database. Two classes: supply risk data and supplier risk data are in composite relationship with the factors class, having life dependency on Factors class. Factors class includes methods that manage the data (e.g. set values type (), getUnit (), setUnit () etc.).

The evaluation class mainly manages the knowledge discovery, and whole algorithmic part of the system. The methods in this class are used to initialise, configure and execute the data mining algorithms. It contains methods that handle the input data (e.g. GenerateDataFile ()), implement the Weka (e.g.RunWeka ()), and also methods that set up the training and test datasets such as (CrossValidationSplit ()).

5.3.4 Sequence Diagram

The sequence diagram (Figure 5.5) shows the behaviour of the system with respected to timeframe. It has two dimensions: horizontal and vertical. The horizontal dimension corresponds to the objects, while the vertical dimension corresponds to time. The vertical dashed lines represent the lifeline of given object and filled box on dash line shows the execution of specific procedure (activity) for given object. The arrows denote message (calls) between the objects.

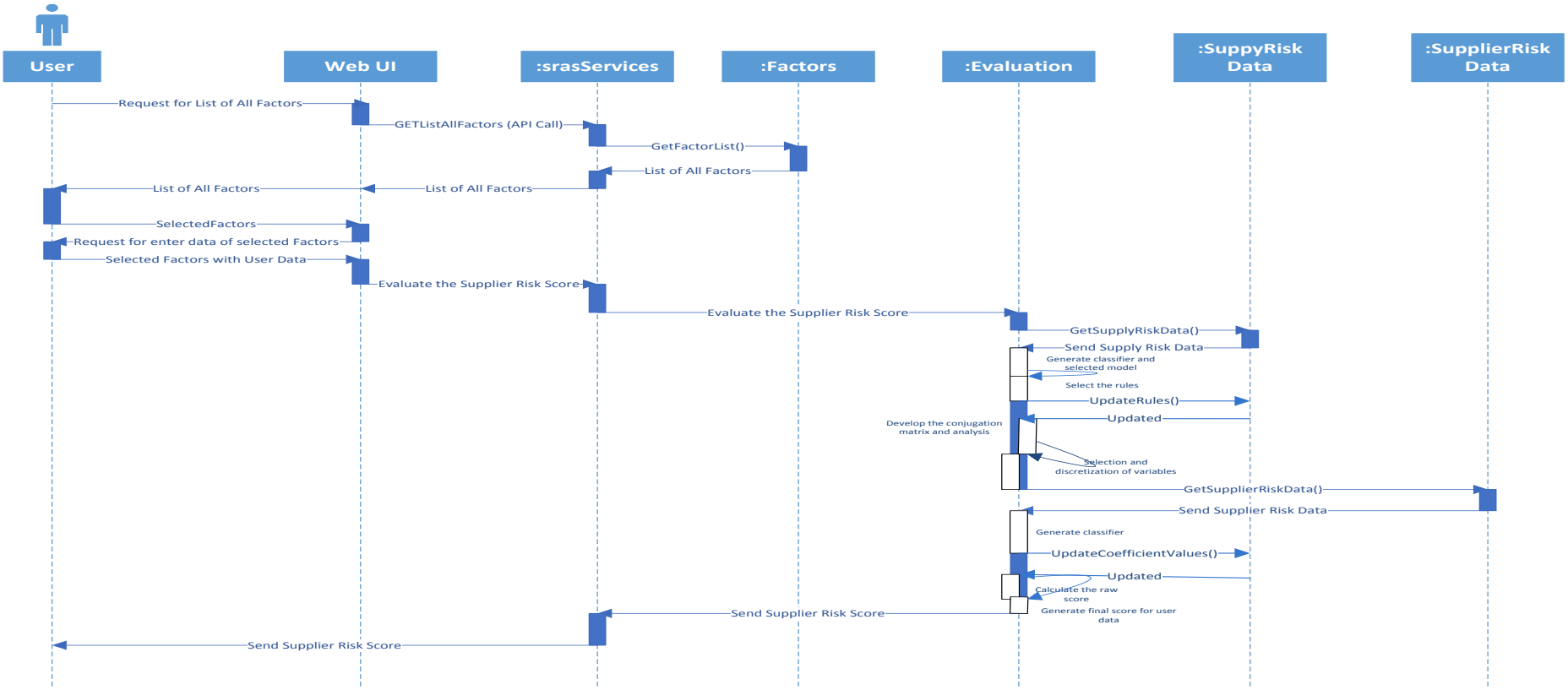


Figure 5.5: The sequence diagram of supplier risk assessment system

5.4 Database Design

5.4.1 Data Sources

The proposed approach for supplier risk assessment is a data-driven; therefore it is complete dependent on data that contain direct or indirect information that relates to supply risk and supply performance (see Figure 4.2). A number of factors that could be source of supply risk (see appendix I) have been identified and classified as financial, operational, network and environmental. The data related to these factors may be found in different data source depending upon the completeness and accessibility to data source. Therefore, the objective is to select the factors from given categories and performance metric for those data is already available either within Case Study Company or from publically available data sources (see appendix II). The collection of the required data and converting it into a format that can be used for the required system functionality is another task of this research thesis. The list of the data sources used acquire the data in this research given as

Supplier evaluation reports: The purchasing and procurement department normally contain supplier evaluation forms when requesting supplier quotations. These provide details of supplier facilities and operational capabilities. It can be a useful data source already available within the company for conducting supplier risk analysis.

Order delivery reports: The Stock Control department records all items received from suppliers. This internal data source tracks supplier delivery metrics according to delivery requirements. This data source is again useful in performing analysis of current and historic supplier delivery performance.

Quality control reports: The Quality Control department records supplier product quality. The data related quality assessments records for all products received from the company's suppliers are useful in identifying previous quality failure.

Procurement audits: Audits are performed by the purchasing and procurement management department as a matter of policy. These cover both the products supplied and the policies applied in purchase, and they are produced both for the company and as requirements for government and other agencies. Data available from these audits

can contribute to details about supplier contracts and include commodity pricing, product availability, sourcing strategy, cost comparisons, licensing requirements, etc.

Publically available online resources: Considerable supplier's country information is available freely. This information can cover factors such as occurrence for natural disasters, political and economic situation, and logistics indices. Data from such open source databases is incorporated into the data for the current study (see Appendix II for source details).

5.4.2 Logical Design

The developed database is required to store the data obtained from different data source to be used for system implementation. Further different activities of the system such as the rules selection, classifier model required to store their data in database for the further use. The database is designed according to proposed categories of the risk factors and system requirements. The conceptual data model with its attributes is given (see appendix III). The proposed data model includes nine main components that consist of two types: components that related to the input data storage and activity data storage. The components those related to input data are: FactorCategories, Factors, SupplierRiskData, SupplyRiskData and UserInputData. The components those are related to system's activity data storage are: RulesOnSupplyRiskData, SelectedRulesDetail, FactorCoefficeint and ConjugationMatrix.

5.5 Implementation

To analyse the practicality and functionality of the proposed approach and design system, the system is implemented for one of leading air condition and refrigeration manufacturing company situated in Pakistan. The name of this company is kept confidential and called Case study Company in current study. First, the Purchasing process and Issues of the case study company is provided than the design of experiment is explained to identify of the dataset required to upload the system.

5.5.1 Case Study Company Background

Case study Company is heating ventilation and air conditioning system (HVAC) company. The history dates back to four decades of continuous achievements. With around four decades of experience, the company is primarily engaged in the design, manufacture, supply, installation, commissioning and operation and maintenance of all kinds of sophisticated refrigeration and air-conditioning systems. The company's thrust for innovate tool and technology is one of main reason that company agreed to implement this proposed methodology and provided the required resources and support (as some of companies approached by researcher denied to provide the required resources). With a state of the art manufacturing unit and a workforce of highly qualified engineers and technicians, the company has been manufacturing premium products which are a hallmark of engineering excellence and precision.

Since last decades due to the change in current market dynamics and new opportunities to expand business in global market, company is changing its strategies to deal with market dynamics and expanding its business to sell its products and service beyond their domestic market. To pursue new business opportunities and deal with market dynamic, company now faced with situations where its business specifically requires growth in the supply network. The supply network growth and new requirement demands a larger focus on the issue of risk in the supply chain and selecting the "right" suppliers. Therefore, Case study Company require a tool that will consolidate all of the supply base management resources at disposal and provide easy methods for monitoring and assessing the risk in the escalating number of supplier contracts.

In these respects the proposed methodology and design system in this research is implemented at Case study Company to provide help to start the process of managing the risk in their supply base.

To implement the designed system it is very important to understand the company's purchasing process, current methodologies and the process being used to monitor the different aspects of supply performance. The following section will describe the purchasing process and different processes currently in-place at case study Company for supply base management.

5.5.1.1 Purchasing process and Issues

At Case study Company, four departments are directly involved in procedures concerning purchase orders from suppliers: Purchasing and procurement department, Production department, Stock control department and Quality control department.

During normal production, the production department issues a bill for materials to the stock control department. The stock control department provides the production department with available materials and parts from stock, but for unavailable materials issues a purchase request to the purchasing department. The purchasing department consults a directory of suppliers, creates a shortlist and sends requests for quotations to selected suppliers together with a supplier evaluation form. The supplier evaluation form contains information such as supplier's capability, capacity, operation, etc. The purchasing department makes an initial supplier evaluation based on received evaluation forms and price quotations. The department's current buying policy is to favour lower bidders able to provide 90 days credit after product delivery. Consequently, the department negotiates with selected supplier(s), which are finalised by a deed of agreement. The binding deed of agreement details product specifications, quality standards, volumes, delivery terms, price and payment details, and legal terms. Each contracted supplier is given a unique identification number, against which all future contract and performance data for the supplier are stored.

After the contract is agreed, the purchasing department issues a purchase order to the supplier for the required material(s) or part(s). Every purchase order carries a unique identification and includes specific information about the order such as quantity and delivery date. The stock control department issues a delivery report upon receipt of the order from the supplier. The quality control department issues a quality report on the products received. Both of these reports are sent to the purchasing department, which (if all is in order) issues a supplier payment request to the finance department.

Data collected during the above process represents valuable monitoring information on the supply performance. Delivery times and product quality are both monitored

and recorded. However, this information is only used by the purchasing department in order to issue an agreed payment to the supplier. Failure on delivery or quality would affect such a payment. In order to conduct the evaluation of supplier risk, it would be advantageous for all such monitoring records to be available and examined together. Furthermore, the growth in scale of the company's business increases the number of materials purchase orders, which places additional strain on the purchasing department and the time available to effectively evaluate each supplier.

Currently, PAEC is facing increasing problems with its purchase orders, particularly in terms of delivery and quality. The purchasing department is focused on cost, but failure to properly assess supplier risk results in increased costs for the company as well as damage to the company's brand name in market. By integrating the all the information available within the company into the supplier risk, these problems such as delivery failure, quality failure etc could be more effectively and cost-efficiently addressed.

In the current case study, the supply performance metrics that are the focus of concern for the company are; *on-time delivery specified quality* and *agreed price*. The present purchasing and procurement procedures of the company lack knowledge of supply risk and fail to take advantage of known performance metrics in making risk assessments. Further, there is a requirement for a faster and more efficient method of supplier risk assessment. Therefore, there is a need for a methodology or tool that can rapidly make a baseline analysis of the potential risks that exist in the supply base. Consequently, this company represents an ideal case study for implementing the proposed methodology and examining its practicality in a real world environment.

The proposed methodology is a data driven approach for identifying supply risk and overall supplier performance by developing a supplier risk score model. This requires data concerning risk factors and supply performance metrics. The following section provides information about the selected data and its parameters.

5.5.1.2 Data sample

The development of a knowledge discovery based supplier risk scoring model requires the creation of two data samples. The purpose of the first data sample is to develop supply risk identification (SRIM) model to discover knowledge about

potential supply risk. The purpose of the second data sample is for development of a supplier risk scoring model using supplier overall performance data.

To create the SRIM, a data sample related to case study company's purchase orders and their output performance was collected covering a period of six years. Selection of the purchase orders included in the study was determined by three factors. The first factor was availability of the suitable data to provide the necessary information to conduct supply risk identification (e.g., supplier evaluation reports). The second factor was the ability to compare purchase order performance against the company's required performance metric for each order. The third factor was the need to ensure that the selection was representative of the full range of purchase data, thus attention was paid to product type, supplier location (including international suppliers).

Data about purchase orders and their outcomes was available on a yearly basis. Thus, a total of 696 purchase orders were available for the year 2008 and, similarly, totals of 726, 753, 724, 770 and 761 were available for the following years 2009 to 2013, respectively. In each of these years different numbers of purchase orders failed to meet the required performance metric due to causes of quality failure, delivery failure or cost failure. In total, data for 4430 purchase orders over the last six years with their respective performance output was collected. For example, out of the 4430 orders, 880 orders did not meet the delivery performance metric and were labelled as delay risk, 750 orders were labelled as quality risk based on the quality reports of received orders, 800 purchase orders failed to meet the price requirement and were labelled as cost risk and 2000 orders were labelled as no risk as these orders fulfilled all the performance metric requirements.

The data sample consisted of 25 input variables and a target variable. The input variables contained numerical, categorical and binary type data, as given in Table 5.1, while the target variable contained the categorical data with four outputs (no risk, cost risk, delay risk and quality risk). These four categories are obtained based on the defined supply risk measurement (see section 4.1.1.1). Each order included in the data sample was arranged according to its date to ensure the accuracy of data. This becomes important in relation to such incidental environmental factors as occurrence of a natural disaster.

Table 5.1: The variables used in the case study and their data type

Category	Risk Factor		Data type	Unit/ Value
Financial	Z-Score	RF1	Numerical	Score
	Commodity Price	RF2	Numerical	Percent
	Price comparison	RF3	Categorical	Low, Average (Avg), High
	Exchange Rate	RF4	Numerical	Percent
Operational	ISO-certification	RF5	Binary	Yes , NO
	Quality Award	RF6	Binary	Yes , NO
	Warranty	RF7	Binary	Yes , NO
	Quality Record	RF8	Binary	Yes , NO
	Quality Improvement	RF9	Binary	Yes , NO
	Quality inspection	RF10	Categorical	SPC, BI, Judg.
	Technical capabilities	RF11	Numerical	Score (0-5)
	Manufacturing Yield	RF12	Numerical	Percent
	Production Facility	RF13	Numerical	Score (0-5)
	Cycle Time	RF14	Numerical	Week
	Capacity utilization	RF15	Numerical	Percent
Network	Availability	RF16	Categorical	Low, Medium (Med), High
	Relationship	RF17	Numerical	years
	Supplier lock	RF18	Categorical	Sole, Duel, Multiple(Multi.)
	Information Sharing	RF19	Categorical	Low, Medium (Med), High
Environmental	Natural Disasters	RF20	Categorical	Green, Yellow, Orange, Red
	Manmade Disaster	RF21	Categorical	Green, Yellow, Orange, Red
	Political Stability	RF22	Categorical	Green, Yellow, Orange, Red
	Infrastructure	RF23	Numerical	Score (0-7)
	Economic Freedom	RF24	Numerical	Score (0-100)
	Logistics Performance Index	RF25	Numerical	Score (0-5)
Class	Supply risk		Categorical	Delay risk, Quality risk, Cost risk, No-risk
Note the to measure the level of man-made disaster on single value following rating criteria is used based on economic cost and fatalities (economic cost is used if fatalities are not involved) No. of dead =0 (green), 1-10(yellow), 11-25(orange), above 25 (red) Economic cost = unknown (green), less than 200 thousand (yellow), 200 thousand to 1 Million (orange), Above 1 million (Red)				

***SPC=**statistical process control, **BI=** Batch inspection, **Judg. =** judgemental

In order to construct the data sample for the supplier risk score model (SRCM), data related to Case study Company suppliers' annual purchase performance and the risk factors involved in performance failure was collected. To calculate the overall performance of suppliers, a financial calculation was made of each supplier's annual performance. Firstly, the total value of purchase orders placed with each individual supplier in a year was calculated. Secondly, the total value of losses cause by individual supplier for the same years due to failure in meeting a performance metric such as quality failure, delivery failure or cost failure was calculated. Then, the annual

performance was calculated for each supplier by subtracting the value of losses from the total value of purchases. Finally, using the supplier risk evaluation equation (4.7), the threshold value (required contracted performance) was calculated. The evaluation of suppliers was made, as either good or bad, based on the threshold value of required contracted performance. Any supplier failing to meet the required threshold value was labelled as bad otherwise good. The objective of the risk scoring model was to provide the probability of good or bad in form of risk score. For accessibility, each supplier's annual performance output and information on their risk factors in each specific year was stored in a single row with specific ID in the database.

The data sample related Case study Company suppliers' performance was collected over a period of six years. The suppliers included in the study were selected according to the availability of necessary risk assessment data over the period of six year. It was very difficult to obtain satisfactory data prior to 2008. Therefore 2008 data was selected as starting year for data. In total, data for 136 suppliers was collected over the six year period from 2008–2013. Attention was paid particularly to the distribution of suppliers, based on geographic location, company size, product type, and other factors to ensure the data accurately represented a full population of suppliers in company's supply base. The number of 136 suppliers and the six year time horizon were selected purely on the availability of data. The number of suppliers in the study was intended to be greater than 136, but this was again a function of what data was available and accessible within the timeframe of the project. However, as each supplier's performance was measured on a yearly basis, it was possible to consider each year's performance as an observation, or test case. In total, this provided a sample of 820 observations (test cases) available for data analysis. The data size, consisting of 820 instances to build the risk model, can be regarded as suitable when considering the average number of suppliers to a medium sized firm and the difficulties in data availability. Out of the 820 observations, 475 supplier performances were labelled as Good and 345 supplier performances were labelled as Bad. The initial input parameters were the same as the first data model, however, in risk score model building only selected parameters based on the knowledge discovered in the previous model are used. Furthermore, numerical type data was discretized (converted into a categorical data type) based on the knowledge discovery of supply risk in the supply risk identification model.

5.5.2 Experiment Calibrations for Input Data Selection

Some initial experiments based on obtained data sample have been carried out to assist in initial decisions about the input data upload to design system. The first step in the testing process involved the identification of the input data; input data is consisting of risk factors and number of classes in target variable. These input data parameters affect the dimensions and size of dataset as summarized in Table 5.2.

Table5.2: The set of parameters that need to be calibrated

Data parameters	Description
Number of input variables	Impact the dimension of the data
Number of class	Impact the data size
Time period	Impact the data size

The first parts of input data parameters are related to target variable, which have the different number of classes' base on performance metric. The inclusion of all supply performance metric is very important as these metric are used to calculate the supplier's overall annual performance. The number of classes has direct impact on both the size of data and the success of the output for proposed approach. In the current study for the supply risk identification model, there will be four different classes labelled as no-risk, cost-risk, delay risk and quality risk based on the selected performance metric.

The second part of input data parameters contains the risk factors, which impact the dimension of the data size. It is apparent that there are an extremely large number of input parameters that can be used to test proposed study, further it more desirable to have more input variables at knowledge discovery. The data related to these factors is obtained from different data sources, therefore some initial experiments have been conducted to analyse the impact of these parameters integration in single dataset. Number of initial tests had to be performed to analyse the impact different data sources integration in to single dataset in term in-term of classifier's accuracy and area under curve (AUC) that can be use for knowledge discovery. To conduct the initial test, fifteen data segmentation are made through combination of risk factors related to four parameters categories (see Figure 4.2).

First four segmentations contain related to individual category such as only risk factors related financial category or operation. In next six segmentations data related to two different categories is combined such as risk factors related to financial and operational category, financial and network category and similarly for other

categories. In four segmentations are data related to three different categories is combined such as data related to financial, operational and market category is combined, data related to financial, network and environmental category is combine and similarly two more segmentation are made. Final segmentation is consisting of all the data related to all the risk factors given in case study belong to defined categories. These segmentations are used to analyse the impact of these parameters integration in single dataset.

To conduct the tests, the year constraint has been used, which result in all the available purchase order related to year 2008, this dataset consists of 697 records. All types of products supplied from suppliers are measured on same supply performance metric and have the same input parameters for supplier selection. Therefore, inclusion of all type of product does not affect the homogeneousness of data.

To analyse the impact of these parameters integration in single dataset, in total 45 models were built, each model is named according to its categorical segmentation. For example F-model consist of input parameter which are related to financial parameter category and FON-model consist of input parameter which are related to financial, operational and network parameter categories.

The outcome of this investigation varied with the different segmentations. The results are summarized in Table 5.3. In conclusion, the overall results illustrate that as the data obtained from different data sources is combine the overall accuracy of classifier increase. The knowledge discovery process prefers the inclusion of more and more factors to deliver to better results (Maimon and Rokach, 2005).

The overall results shows that, the uses of different parameters related to all four categories provide the much better result. We have further analysed the impact of these parameters categories on class coverage, average confidence and support value. As the number of parameter categories increase in dataset, the class coverage for each class increases and become more evenly distributed. The FONE model has the highest average confidence value for the rule set than the other model and comparatively good average support value than rest of the models. Based on results of initial experiment for input parameter selection it is assumed that inclusion all parameters (for which the data is available) related to all the categories in model building provide good result. Therefore, in further study will use all parameters in given case study related to all categories.

Table 5.3: The impact of combining data obtained from different data source on model performance

Model Name	Input data description		Model Performance						Class Coverage
	No. of Parameters	Data size	C4.5 Algorithm		Ripper Algorithm		PART Algorithm		
			Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	
F-Model	4	696	73.28	0.814	71.7	0.805	71.41	0.83	Very low for quality risk class
O-Model	11	696	65.23	0.783	67.96	0.805	64.22	0.795	Very Low for cost risk class
N-Model	4	696	54.45	0.741	51.58	0.606	55.32	0.763	Very low for cost and quality risk
E-Model	6	696	66.38	0.817	65.8	0.783	66.38	0.83	Almost evenly distributed
FO-Model	14	696	76.44	0.884	84.33	0.906	77.59	0.877	Better than model F and O
FN-Model	8	696	73.99	0.835	72.99	0.819	72.27	0.858	Much better than F and N model
FE-Model	10	696	77.3	0.852	75.14	0.853	75.86	0.871	Better than model F
ON-Model	15	696	72.7	0.864	74.86	0.85	74.57	0.871	Much better than model O and N
OE-Model	17	696	68.96	0.823	76.15	0.865	69.971	0.846	evenly distributed class coverage
NE-Model	10	696	74.43	0.86	74.71	0.834	73.85	0.858	Low for quality risk class better than model N
FON-Model	19	696	78.74	0.873	82.76	0.903	80.03	0.885	Almost evenly distributed
FOE-Model	21	696	81.17	0.873	85.77	0.903	79.74	0.885	Almost evenly distributed
FNE-Model	14	696	76.01	0.844	78.88	0.876	78.3	0.88	Almost evenly distributed
ONE-Model	21	696	76.21	0.875	80.6	0.882	79.02	0.886	Almost evenly distributed
FONE-Model	25	696	80.14	0.881	85.06	0.923	80.17	0.89	Almost evenly distributed

In current study all the metric used for calculating the supplier performance and all type of products purchase were included in data sample. Therefore, the issue related to sample size used in design process complete depend upon the data time period. The study involves purchase orders that spread over a period of 6 years. Supply chain operating environment at the case study company is very dynamic in nature, hence assuming stability over such a long period may not be suitable. So, this creates a question that "Which is best suitable period for sample to produce meaningful results, in such a dynamic environment"? To answer this question some initial tests are carried on the given data. To conduct initial tests, first five year data is used to make ten samples with different data size. First five samples contained data related to individual year. Then using this five year data, 5 sample periods were created. The period one contains the data about the first year only. Period two contains the data from previous year and second year. Similarly the period three contain the data related to previous two year and third year. Similar manner last period contain the data of all the previous years and last year. This division of the data in the periods and individual years will help in answering the above question. All the initial tests are conducted using the Weka machine learning bench-work with cross validation evaluation method.

To identify the best sample size, in total 30 different models were built for these samples using three algorithms (C4.5, RIPPER, and PART). The results of these models are summarized in Table 5.4. The period_1 models have same result as yearly_ 2008 models because they used same input sample, data of year 2008. All the periodic models outperformed their respective yearly models, for example the period 2 contains the data of year 2008 and 2009. Period_2 models outperformed these yearly 2008 and 2009 models on the accuracy, AUC and PS measure. In the similar way the period_3, period_4 and period_5 models outperformed their respective yearly models. So, these results show that performing the study on the completed sample irrespective of their date of purchase can provide more meaningful results; the Knowledge discovery process also prefers more amount of data as much as possible (Frawley et.al., 1992; Matheus et. al. 1993). If a case study has three year data, then the study should be performed at all available samples by keeping the holdout testing sample used for model selection. In current study the sixth year data is kept as hold-out test sample to be used in model selection method.

Table 5.4: The impact of different sample size on Models' performance

Algorithm	Performance	Yearly Model					Periodic Model				
		2008	2009	2010	2011	2012	Period 1	Period 2	Period 3	Period 4	Period 5
	Data size	696	726	753	724	770	696	1422	2175	2899	3669
C4.5	Accuracy	80.14	79.55	77.35	79.35	78.10	80.14	84.53	87.22	88.93	89.23
	AUC	0.88	0.86	0.86	0.88	0.86	0.88	0.92	0.94	0.94	0.95
	RI	426.97	445.05	463.05	430.65	501.82	426.97	935.13	1421.23	1896.62	2387.34
RIPPER	Accuracy	86.36	84.99	84.94	84.68	83.77	85.06	90.08	90.25	92.41	93.13
	AUC	0.92	0.91	0.90	0.91	0.91	0.92	0.95	0.95	0.96	0.97
	RI	436.77	470.31	476.91	440.74	514.76	436.77	926.78	1420.64	1928.27	2474.12
PART	Accuracy	83.61	82.87	80.39	83.64	80.74	80.17	88.54	89.06	90.07	92.23
	AUC	0.91	0.91	0.88	0.91	0.90	0.89	0.94	0.94	0.95	0.97
	RI	453.94	457.73	495.16	446.89	506.34	453.94	934.30	1449.09	1933.22	2447.51

5.6 Summary

The system is designed in such a way that accommodates the proposed methodology for supplier risk assessment. The UML is used to design and presented the design with various UML diagrams. The proposed methodology is data-driven therefore, data requirements are set as: the appropriate data related to the factors that affect the supply performance, quantity of the data; finally the data should be readily available within company or publically available data sources. The design of the database was presented that is used within the system.

For the validation and practicality of the proposed approach, design system is implemented to a case study company. For the purpose of implementation and testing, actual data was collected from an air conditioning company's supply-base (procurement) and from data-sources that are available publically. The collected data was used to create two databases that formed the test environment for design system. The first data sample contained the information related to specific purchase orders and their outcomes and the second data sample contained the information related to specific suppliers and their performances. The data requirements set in the previous section guided the process towards the dataset identification.

After the data was sourced and imported in a required database, some initial experiments are conducted to obtain the best sample for data model development. The final selected data is uploaded in database to be used for the development of three classification models, each for the selected algorithm. These models have been tested on the testing dataset for selection of final model.

6 RESULTS AND DISCUSSIONS

In the previous two chapters, the detail of the proposed methodology and the data used for this research has been given. Here, the focus is on presenting the case study results. The chapter can be divided into three main sections, first section discuss the performance of three classification models and selection of final model as supply risk identification model (SRIM). The results of the knowledge discovery form supply risk identification model are provided in second section as it is used in the supplier risk model. In the third section, the results about knowledge discovery base supplier risk scoring model are presented and discussed.

6.1 Models Testing and Selection

Three classification models are constructed on the best identified set of input data parameters and sample size using three selected algorithm (see chapter 4 section 4.1.2.2). However there is need for the selection of one model which performs the best from these models, which can be used as the supply risk identification model (SRIM). The supply risk identification model (SRIM) is aimed at providing the hidden knowledge about supply risk in the available data and to predict the supply risk of new input data. Therefore, these three classification models are tested on the new data using hold-out validation method to single out one model according to the proposed method (section 4.1.2.3). The testing dataset is not used in the model building stage i.e. last year data of available sample in current case study. The models' classification performance on the new dataset is measured on five different evaluation metrics consisting of precision, Recall, F-measure, area under the curve (AUC) and comprehensibility. The selection for these evaluation metrics is based on the fact that the available data is unbalanced for the given target classes. The evaluation of any data mining model on these metrics is very suitable for an unbalanced data situation. The discussion section is divided into six sub-sections, according to the considered evaluation metric and final model selection.

6.1.1 Supply Risk Precision

The classification accuracy is the predominant measurement in machine learning used for evaluating classification models. It represents the ability of a model to correctly classify the instance, however due to data imbalance for different classes in the

dataset (see chapter 5 section 5.5.1.2); it may not under-represent the result of minority classes. In this context, it is important to analyse the precision of each class. For each supply risk class, precision is defined as fraction of the examples classified as positive that are truly positive (see section 3.5.2.1). Figure 6.1 represents the precision of each class for selected three algorithms.

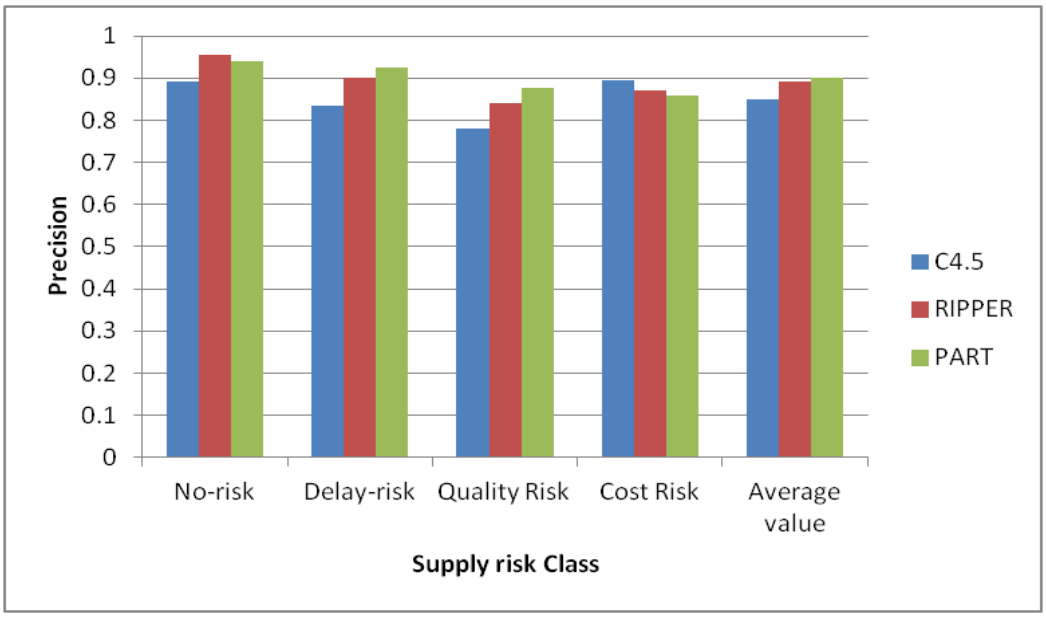


Figure 6.1: Comparison of C4.5, Ripper and PART on precision

All algorithms performed well for all the classes, however for all algorithms, the supply risk precision is decreasing from majority to minority class. C4.5 performs better than both the Ripper and PART algorithm on cost risk class; however it performs worse than both the RIPPER and PART algorithm for all other supply risk classes. Different algorithms perform differently for each class, however these models does not reflect a significant difference on the average value of precision metric.

Furthermore, a comparison of algorithms is conducted to identify the significant difference among algorithms for supply risk precision through one way within subject ANOVA test. ANOVA test with 95% level confidence means α=0.05, represent that there is no significant difference between the algorithms on supply risk precision metric.

6.1.2 Supply Risk Recall

Since one of the objectives of supply risk identification model is to predict supply risks, it is important to analyse the number of each supply risk class that are correctly predicted. Recall is an evaluation metric that takes into account this consideration.

Recall is also appropriate for imbalanced data problems as it can be used to measure the performance of either the majority or minority class. Supply risk recall corresponds to the fraction of positive examples that are correctly labelled. Supply risk recall performance for each supply risk class according to selected three algorithms is shown in Figure 6.2.

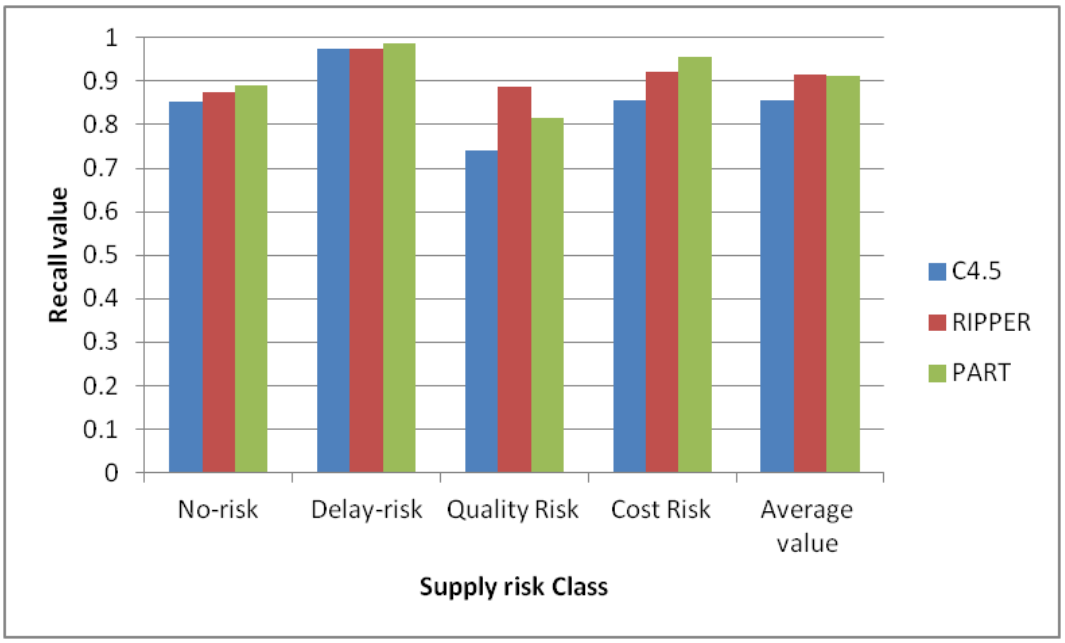


Figure 6.2: Comparison of C4.5, Ripper and PART on recall value

All algorithms show good performance on recall across all supply risk classes. The minimum recall value is 0.74 obtained from C4.5 for quality risk type, however all other algorithms provide more than 0.8 (80%) recall value across all supply risk classes. C4.5 algorithm also provide above .8 recall value for all supply risk classes except quality risk. PART algorithm exhibits better performance than other algorithms for almost all supply risk classes, except quality risks type, Where the RIPPER algorithm has the highest performance, than both C4.5 and PART algorithm. Quality risk class is a minority class in the available data set and RIPPER algorithm performed better than both other algorithms. This can be due to RIPPER natural tendency toward minority class than other algorithms, especially C4.5 more focused on the majority class; however it is desirable to have better recall value for minority class in unbalance data situation.

On the average value of recall, there is not a significant difference among the algorithms. A comparison of algorithm is conducted using one way within subject

ANOVA test. It does not provide any significant difference among all algorithms' performance on recall metric at ANOVA significance "α" value of 0.05(95%).

6.1.3 Supply Risk F-measure

The F-measure enables to observe the simultaneous effects of recall and precision. Precision focuses on the number of instances correctly classified relating to one specific class. Recall focuses on overall performance considering both correctly classified and incorrectly classified. However there is a need for a trade-off between these two values. F-measure is a trade-off between the precision and recall; it is weighted harmonic mean of precision and recall. Figure 6.3 shows F-measure of each supply risk class for three selected algorithms.

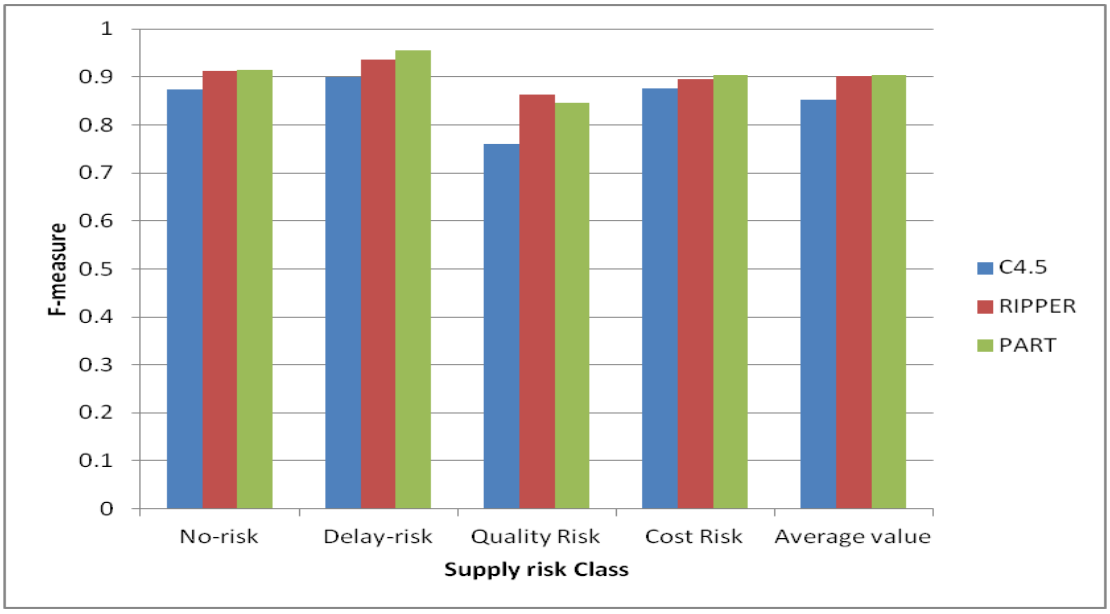


Figure 6.3: Comparison of C4.5, Ripper and PART on F- measure

All the algorithms performed well with respected to F-measure. The minimum performance is give by C4.5 algorithm for quality risk class, which is 0.76. The maximum performance is found at delay risk, where all algorithms almost touch the 0.9 value for F-measure. Both RIPPER and PART algorithms comparatively perform better than C4.5 on all types of risks. PART and RIPPER performed comparatively to each other for almost all types of risks; however PART provides a lower value for F-measure at quality risk than RIPPER algorithm. The PART algorithm performs better than the C4.5 algorithm and it clearly outperform the C4.5 with regard to quality risk and delay risk type with a difference of performance of 10% and 7% respectively. On

the average value of F-measure, it can be seen that there is a small difference between C4.5 and PART algorithm.

One way within subject ANOVA test with α=0.05 is conducted to identify the significant difference among algorithms performances on F-measure metric. ANOVA test results show that there is significant difference between C4.5 and PART algorithm with alpha value α=.015. However there is no significant difference between the C4.5 and RIPPER and RIPPER and PART algorithm.

6.1.4 Area under Curve (AUC)

ROC graphs are two-dimensional graphs in which true positive is plotted on the Y axis and FP rate is plotted on the X axis (see chapter 3 section 3.5.2.1). The area under the ROC curve (AUC) is used to portray the model behaviour on new data. A classification model is realistic if it has AUC >0.5, otherwise it is worse than random guessing. Figure 6.4 shows the area under curve (AUC) of supply risk classes for selected algorithms.

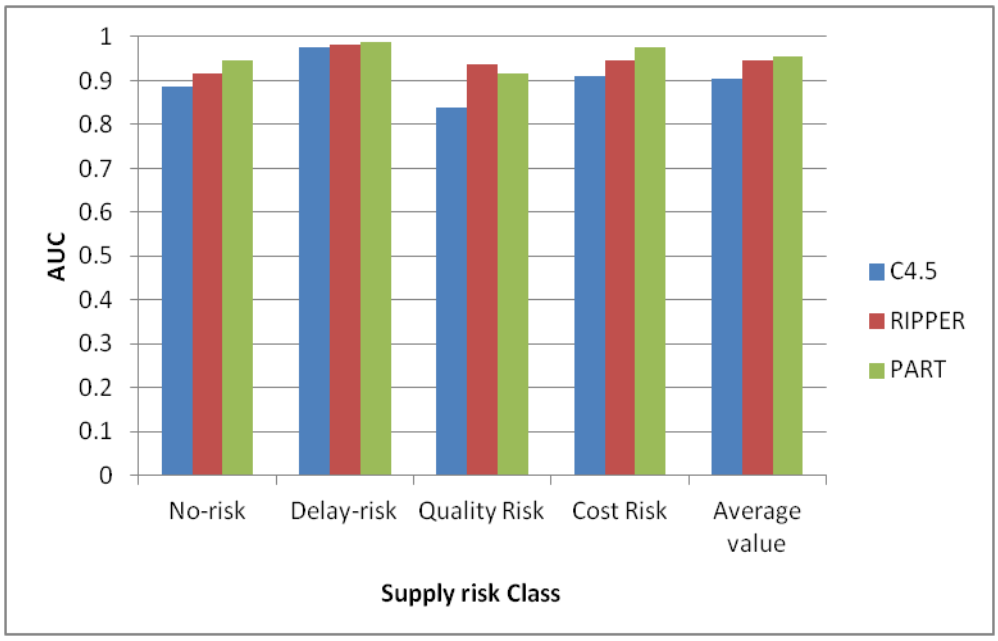


Figure 6.4: Comparison of C4.5, Ripper and PART on AUC

All algorithms exhibit the realistic performance for all supply risk classes, as all the algorithms have higher value than required threshold (AUC=0.5) for all supply risk classes. Different algorithm have different AUC for different classes, C4.5 has the minimum value of 0.84 for quality risk, however much higher than required threshold (AUC=0.5). C4.5 algorithm performs much worse than both RIPPER and PART

algorithms on all supply risk classes. This is a direct consequence of C4.5's tendency to blindly classify positive class as negative class. PART performs better than RIPPER algorithm on all classes except quality risk class. On the average value of AUC, there is higher difference between PART and C4.5 than between PART and RIPPER and between C4.5 and RIPPER.

To analyse the significant difference between the algorithms performance on AUC, one way within subject ANOVA test is conducted with significance value (α=.05). It represent that in pair-wise comparison there is a significant difference between the C4.5 algorithm and PART algorithm with significance alpha value (α=.02). RIPPER algorithm does not represent any significant difference with both C4.5 and PART algorithm.

6.1.5 Model Comprehensibility

For testing a model from a comprehensibility view point, the number of rules generated by model is used. The comprehensibility of a classification model is high if the number of rules for specific class type is low and vice versa (Verbeke et al., 2011). Figure 6.5 shows model comprehensibility of different supply risk classes for selected algorithms.

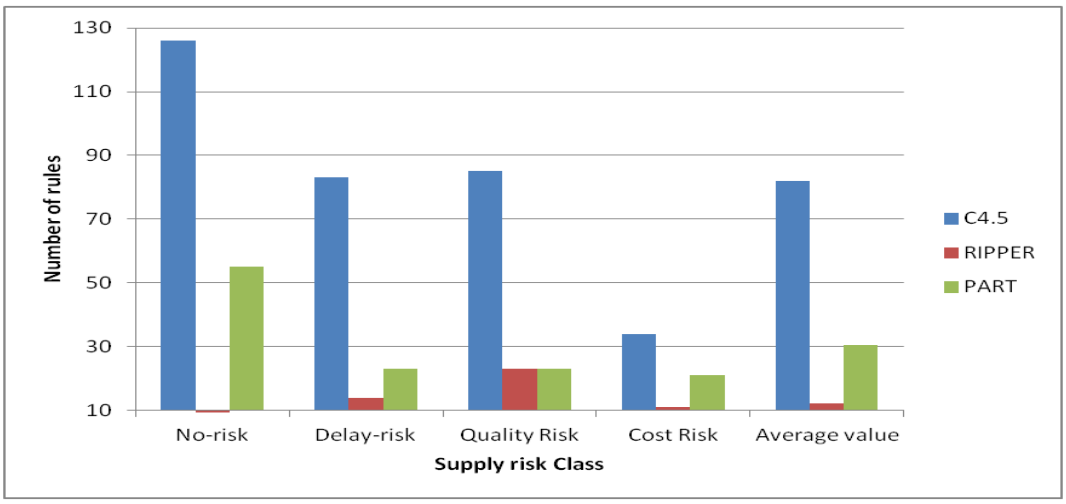


Figure 6.5: Comparison of C4.5, Ripper and PART on comprehensibility

All the algorithms performed different from each other with respected to comprehensibility. According to the comprehensibility criteria the RIPPER algorithm outperformed both the C4.5 and PART algorithms for all the supply risk classes. This can be a result of the ability of C4.5 algorithm to over expand for a given data. Furthermore, the Ripper algorithm does not create the rules for the most dominating

class; just provides one default rule for the dominating class. On the other hand, RIPPER algorithm tries to create more rules for the least minority class as PART and C4.5 generate more rules for majority class. This can be seen from the results that RIPPER algorithm has also outperformed the PART algorithm on all classes except quality risk. Since, the nature of the RIPPER algorithm functionality and its ability to optimize rule set provides a more comprehensible model. One the average value of comprehensibility there is significant difference between the three models. The difference between the C4.5 algorithm with both RIPPER and PART is much higher than the difference between RIPPER and PART.

To analyse the significant difference between the models' comprehensibility, one way within subject ANOVA test is conducted with significance value (α=.05). The results of one way within subject ANOVA test reject the null hypothesis and represent that there is significant difference between models' comprehensibility. It represent that in pair-wise comparison the C4.5 algorithm has a significant with both the RIPPER and PART algorithm with significance alpha value (α=.038) and α=.021 respectively. However there is not a significant difference in pair-wise comparison between the RIPPER and PART model's comprehensibility with α=.378. This can be due to the closer number of rules between RIPPER and PART model for most of supply risk class except no-risk class.

6.1.6 Supply Risk Identification Model (SRIM) Selection

The selection of most suitable model is an important issue in the knowledge discovery process. In the current study initially three different algorithms are applied to given problem formulation. A method is proposed in current the research to single out one model that performs the best, which can be used as supplier risk identification model.

According to the proposed method, algorithms were ranked according to their performance according to the evaluation criteria (see chapter 3 section 4.1.2.3). The average values of evaluation metrics i.e. precision, recall, F-measure, AUC and comprehensibility is calculated for three classification models. The ranking of the three classification models on selected evaluation metrics and average ranking is given in Figure 6.6.

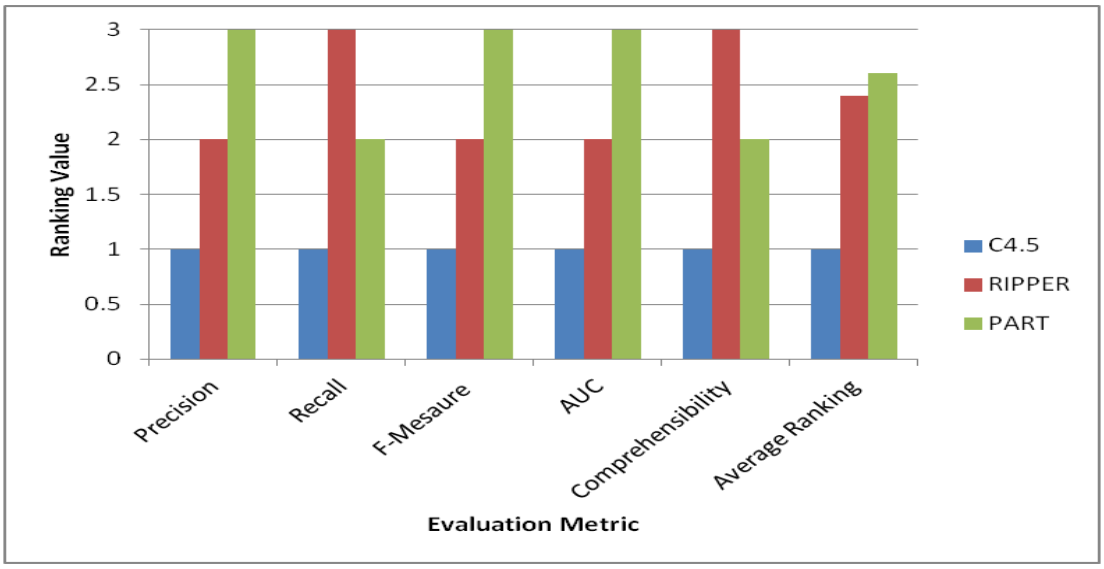


Figure 6.6: Comparison of classification models on their ranking value

Both PART and RIPPER algorithm's models outperformed the C4.5 algorithm's model. PART algorithm outperformed RIPPER algorithm on precision, F-measure and AUC, however RIPPER outperformed PART on recall and comprehensibility.

This can be two bases; first the RIPPER algorithm does not produce the more than one general rules for majority class, which increase its comprehensibility. Second it has rule set optimisation function that further reduces the number of rules for different classes.

The PART algorithm outperformed both C4.5 and RIPPER algorithms on average ranking value. Brazdil and Soares (2000) argue that using only one evaluation metric may be meaningless for comparing the models' performance because there may not be significant different between two algorithms on one evaluation metric. The detail analysis of classification models performance in this section showed that no algorithm consistently outperformed the other on all the evaluation metrics (No free lunch theorem) and also does not have showed significant differences mostly on evaluation performance metrics. It makes selection of single best algorithm difficult; the proposed selection method in methodology seemed to be useful. The above results support the selection of more than one evaluation metric and ranking method in algorithm selection. Based on the above result the PART algorithm model is selected as single best model for current case study and selected as supply risk identification model (SRIM). This supply risk identification model will be used for knowledge extraction and making the prediction about supply risk for new input data.

6.2 Knowledge about Supply Risk

The supply risk identification model (SRIM) is aimed to predict accurately the unknown supply risk and produce knowledge about the supply risk from the available data. To evaluate the results of its ability to predict accurately the unknown supply risk, a number of evaluation metrics and proposed approach is used to identify the best model as final supply risk identification model (SRIM), which provides the best result in terms of prediction accuracy. To get the knowledge about supply risk, the classification rules from supply risk identification model's rule set have been analysed in this section.

The supply risk identification model (SRIM) provided in total 102 rules. There is high number of rules generated for no-risk class, which account for 34.31% of all the rules; however rules generated for delay risk, quality risk and cost risk class counted 20.59%, 17.65% and 27.45% of total rules respectively. These rules discover the data pattern and provide the hidden knowledge in data; however all the rules may not be valuable. Therefore, the rule PS value that measures the rule value-ability or interestingness and coverage value of each rule is calculated. Rules related to four given supply risk classes has been selected from the rule set based on their PS value and coverage values. The rule is selected, if meet the criteria given in equation 4.22 as,

$$RI_a \leq \mathrm{RI_i} \text{and } \mathrm{Cr_i} \geq C_a$$

Where,

$$RI_a = \text{the model average rule interestingness value (PS)}$$
$$RI_i = \text{the rule interestingness value of ith rule}$$
$$C_a = \text{the model average coverage value}$$
$$Cr_i = \text{the coverage value of ith rule}$$

The selected rules about supply risk shows that they have risk factors related to each category of supply risk factors. It is prominent that supply risk outcome is a result of the combined effect of different risk factors as depicted in the selected rules related to each supply risk class. Hidden knowledge obtained from the data related different supply risk class is discussed in the following sub sections.

6.2.1 Delay Risk

According to defined measurement (see chapter 4 section 4.1.1.1), if a purchase order is not delivered on time it will be labelled as a delay risk. Only 38.1 percent of overall delay risk rules are selected according to defined criteria of rules selection. These rules are given and arranged according to their highest PS value in Table 6.1.

Table 6.1: The selected rules about delay risk

1. Z-Score <= 1.83 AND Technical capabilities <= 4.03 AND Price comparison = Avg. AND Logistics Performance Index <= 4 AND Infrastructure <= 4.12 AND Availability = Low: Delay-Risk

2. Z-Score <= 2.68 AND Logistics Performance Index <= 3 AND Exchange rate <= -4.94 AND Infrastructure <= 5.21 AND Z-Score <= 1.86 AND Manmade Disaster = Red: Delay-Risk

3. Z-Score <= 1.84 AND Warranty = NO AND Availability = Low AND Technical capabilities <= 4.08 AND Supplier lock = Dule AND Logistics Performance Index <= 2.98 AND Manmade Disaster = Red: Delay-Risk

4. Z-Score <= 1.84 AND Technical capabilities <= 4 AND Warranty = YES AND Infrastructure <= 3.98 AND Logistics Performance Index <= 2.98 AND Production Facility <= 3.1 AND Cycle Time <= 5.23: Delay-Risk

5. Z-Score <= 1.84 AND Technical capabilities <= 3.03 AND Quality record = NO AND Warranty = YES AND Production Facility <= 2.98 AND Capacity utilization > 75.06 AND Manmade Disaster = Red: Delay-Risk

6. Z-Score <= 1.84 AND Warranty = NO AND Infrastructure <= 4.07 AND Availability = Low AND Supplier lock = Sole AND Natural Disasters = Red AND Price comparison = Low: Delay-Risk

7. Z-Score <= 2.65 AND Logistics Performance Index <= 2.99 AND Z-Score <= 1.8 AND Capacity utilization > 79.3 AND Supplier lock = Sole AND Manmade Disaster = Red AND Information Sharing = Low: Delay-Risk

8. Z-Score <= 2.73 AND Technical capabilities <= 3.97 AND Quality record = NO AND Logistics Performance Index <= 3.01 AND Manufacturing Yield <= 76.29 AND Exchange rate <= -4.98: Delay-Risk

According to extracted rule set about delay risk, it is observed that risk factor "Z-score" having value in financial default zone (<1.8) have the higher probability of causing the delivery failure (delay risk). However, Z-score having value in safe financial zone (<2.7) cause the delay risk, when exchange rate tends towards a decrease between the dollar and the considered currency. Some other interesting results obtained about environmental factors such as logistic performance index,

infrastructure and disaster. Data pattern reveals that medium (below 3) logistic performance index with medium infrastructure and higher manmade disaster probability cause the delay risk.

According to the pattern discovered, low availability of product especially with sole or duel supplier lock caused the delay risk. These finding are very consistent with previous literature such as duel or sole sourcing policy can increase probability the delivery failure, especially when the availability of product is low in market (Quayle 2002). According to discovered knowledge, the supplier with highest technical capability (>4) is more desirable in situation of medium production facility (<3), high capacity utilization (>75). These results seems interesting as high capacity utilization left less opportunity for manufacturer to cope with change in design and demand, low technical capability has left less room for the supplier to find quick alternatives with medium production facility requiring the high maintenance time, which can cause the delivery failure. These data patterns can be very useful to identify the delay risk for a given purchase order and can aid risk mitigation and monitoring decision.

To analyse the inter-relationships among risk factors with respect to delay risk, a conjugation matrix shown in Table 6.2 is developed according to selected rules about delay risk (see Table 6.1). In total 17 risk factors out of 25 risk factors appeared in selected rule set, these risk factors are used for further analysis to identify their interrelationship and their importance toward delay risk. conjugation matrix shows that the factors from different categories are inter-connected to each other with respected to delay risk, such as financial factor (Z-score) is connected with operational factors (technical capability, capacity utilization etc.), network factors (availability etc) and environmental factors (disaster, infrastructure etc.) and vice versa. These results are very consistent with initial experiments for input data selection (see chapter 5 table 5.2), that represents as the integration of data from different categories increase class coverage for supply risk classes become much better.

Table 6.2: The inter-relationship among risk factors with respect to delay risk

	RF_1	RF_3	RF_4	RF_7	RF_8	RF_{11}	RF_{12}	RF_{13}	RF_{14}	RF_{15}	RF_{16}	RF_{18}	RF_{19}	RF_{20}	RF_{21}	RF_{23}	RF_{25}
RF_1	2	2	2	4	2	5	1	2	1	3	3	4	2	1	6	4	6
RF_3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
RF_4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
RF_7	0	1	0	0	0	1	0	2	1	1	2	2	0	1	2	2	2
RF_8	0	0	1	1	0	0	1	1	0	1	0	0	0	0	1	0	1
RF_{11}	0	1	1	2	2	0	1	2	1	1	1	1	0	0	2	2	4
RF_{12}	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{13}	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0
RF_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{15}	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2	0	0
RF_{16}	0	1	0	0	0	1	0	0	0	0	0	2	0	1	1	0	1
RF_{18}	0	1	0	0	0	0	0	0	0	0	0	0	1	1	2	0	1
RF_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{20}	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{21}	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
RF_{23}	1	1	0	0	0	0	0	1	1	0	2	1	0	1	1	0	1
RF_{25}	2	0	2	0	0	0	1	1	1	1	1	1	1	0	3	2	0

RF_1= Z-Score, RF_3= Price comparison, RF_4= Exchange rate, RF_7= Warranty, RF_8= Quality record, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{13}= Production Facility, RF_{14}= Cycle Time, RF_{15}= Capacity utilization, RF_{16}= Availability, RF_{18}= Supplier lock, RF_{19}= Information Sharing, RF_{20}= Natural Disasters, RF_{21}= Manmade Disaster, RF_{23}= Infrastructure, RF_{25}= Logistics Performance Index

The conjugation matrix presents that 95 pairs of risk factors out 289 pairs for 17 risk factors has interrelation, where 35% of risk factors have strong interrelation above the average value of 1.57. The strongest interrelationships are obtained for Z-Score (RF_1) and Manmade disaster (RF_{21}), Z-Score (RF_1) and Logistic performance index (RF_{25}), Z-Score (RF_1) and Technical capability (RF_{11}), and Technical capability (RF_{11}) and logistic performance index (RF_{25}) risk factors pairs. The proposition of strong inter-relationships between risk factors is confirmed by these results. These inter-relationships might be triggered as financial problems can be the cause of operational issue (including network and environment) and operational problems can be cause of financial issues (Carter and Giunipero 2010). For example, financial distress may happen due to failure in obtaining financing at a crucial point in the production season. And manmade disasters such as a strike at a factory or port can decrease production capacity or entire interruption of supply.

The results of risk factors behaviour with respect to delay risk is summarized in Table 6.3. Factors related to financial, operational categories have higher active sum, network factors have higher passive sum, while the environmental factors has both

higher active and passive sum. It appears that the inter-relationship among the risk factors expand across all characteristics of upstream supply chain.

Table 6.3: The behaviour of risk factors with respect to delay risk

Risk Factor		AS (i)	PS (i)	IAS$_r$ (i)	IPS$_r$ (i)	I(i)
Z-Score	**RF1**	50	6	100	12.97	56.48
Price comparison	**RF3**	3	8	7.54	13.51	10.53
Exchange rate	**RF4**	3	7	14.20	11.91	13.05
Warranty	**RF7**	17	7	21.20	9.39	15.29
Quality record	**RF8**	7	4	8.48	5.65	7.06
Technical capabilities	**RF11**	21	7	32.40	10.87	21.63
Manufacturing Yield	**RF12**	1	4	1.68	6.55	4.11
Production Facility	**RF13**	3	9	0.07	13.97	7.02
Cycle Time	**RF14**	0	6	0	11.02	5.51
Capacity utilization	**RF15**	4	8	0.59	12.39	6.49
Availability	**RF16**	7	10	10.03	17.53	13.78
Supplier lock	**RF18**	6	12	5.01	20.71	12.86
Information Sharing	**RF19**	0	6	0	14.69	7.35
Natural Disasters	**RF20**	1	5	0.89	9.56	5.22
Manmade Disaster	**RF21**	1	22	0	39.21	19.60
Infrastructure	**RF23**	10	12	19.83	20.06	19.95
Logistics Performance Index	**RF25**	16	17	33.82	25.77	29.79

AS (i)= Direct Active Sum, PS (i)= Direct Passive Sum, IAS$_r$(i)= Relative Indirect Active Sum, IPS$_r$(i)= Relative Indirect Passive Sum, I(i)=Integration,

Further analysis of indirect interrelations of risk factors will provide in-depth analysis of risk factors behaviour within system. High IAS$_r$(i) is identified for Z-score and Logistic performance index represent that the probability of delay risk highly depends upon these factors. Risk factors such as technical capabilities and warranty also shows high IAS$_r$(i) with respect to delay risk. It is very interesting as normally technical capabilities and warranty represent the supplier ability to fulfil the quality requirement. According to results, technical capability is strongly interconnect with logistic performance index that has a direct effect on normal supply operation or logistics of the products.

The highest IPS$_r$(i) is represented by a manmade disaster, it has a strong interrelationship with z-score and logistic performance index. According to interaction results, logistic performance index, infrastructure and supplier lock also have comparatively high IPS$_r$(i). Both infrastructure and supplier lock are strongly interconnect with Z-score.

Integration I(i) values represent the level of interrelationship factors have within the system. The results show the high I(i) value for z-score and logistic performance index. The existence of mutual connections is further investigated for these variables. Both of these factors are strongly interconnected to each other. Furthermore they have an indirect connection with each other through manmade disaster, technical capabilities and infrastructure. As Both z-score and logistic performance index is strongly interconnected with the manmade disaster, technical capabilities and infrastructure. These results confirmed that mutual connections and even feedback loops between risk factors exist with respect to delay risk. Therefore, it is very important to have such data driven methodology as the current thesis for supply risk identification, those can identify these hidden inter-relationships which may not be identified by expert's base or fixed mathematical functional base approaches.

6.2.2 Quality Risk

Quality risk is a minority class in the available data sample; therefore, it has the least number of rules in the model, where about 22.22% of overall quality risk rules have been selected. A purchased order that failed to meet the quality requirements is labelled as quality risk. Table 6.4 shows selected rules about quality risk and arranged according to their PS values.

Table 6.4: The selected rules about quality risk

1. Technical capabilities <= 3.01 AND Quality improvement = NO AND Manufacturing Yield <= 70 AND Quality Award = NO AND ISO-certification = NO AND Quality record = NO AND Relationship <= 5.11 AND Infrastructure <= 6.23 AND Manmade Disaster = Orange: Quality-Risk

2. Quality improvement = NO AND Quality record = NO AND Natural Disasters = Orange AND Quality Award = NO AND Technical capabilities <= 4.04 AND Capacity utilization <= 80.13 AND ISO certification = NO AND Manufacturing Yield <= 60.09: Quality-Risk

3. Availability = High AND Quality record = NO AND ISO certification = NO AND Manufacturing Yield <= 69.71: Quality-Risk

4. Capacity utilization > 79.93 AND Warranty = NO AND Quality record = NO AND Production Facility <= 2.02 AND Cycle Time > 0.87 AND Quality Award = NO: Quality-Risk

Knowledge discovered about quality risk, states that a supplier with no quality record and manufacturing yield less than 70% cause quality risk. Furthermore, the lack of

quality award either local or international, such as ISO certification and no continuous quality improvement philosophy such as Six-Sigma or Total Quality Management are also identified as the main cause of quality risk. These results are confirmed by Adanur and Allen, (1995) who stated continuous quality improvement and ISO certifications decrease the quality risk from supplies. Rlue#2 states capacity utilization below 80% with other factors and conversely in rule#4 states capacity utilization above 80% with some other factors cause quality risk. However, the discovered knowledge is not contradictory for specific class, as capacity utilization below 80% without highest technical capability (>4) and ISO certification cause quality risk. Conversely, capacity utilization above 80% with bad production facility and no warranty cause quality risk. The warranty is closely related to quality and the contractual aspect of the supply chain. If a company provides warranty it reflects maturity level toward best technical capability for quality and service (Diaz et al 2012). Therefore, unavailability of a warranty reflects a lack of technical capabilities even if capacity utilization is higher. Furthermore, the bad production facility with high capacity utilization reflects the supplier's focus is on quantity rather than quality. These results are further supported by the knowledge discovered about delay risk, where a lack of high technical capabilities and bad production facilities cause delay risk in presence of no quality record. Multiple factors are affecting a supply chain operation; therefore, impact on performance cannot be determined based on one isolated risk factor alone. Further interrelation analysis among the factors is conducted according to discovered knowledge.

The conjugation matrix use for interrelationship analysis has only 12 risk factors out of 25 risk factors with respect to quality risk as shown in table 6.4. The conjugation matrix consists of 12×12 pairs of risk factors, where 51.4% pairs show risk factors has interaction with each other. The strong interconnection is observed for risk factors such as Quality record (RF_8) show the strong interrelationships with ISO certification (RF_5), Quality Award (RF_6), and Manufacturing Yield (RF_{12}). Further all these factors show the strong relationship with each other such as ISO certification (RF_5) has strong relationship with Manufacturing Yield (RF_{12}) and Quality Award (RF_6) has with ISO certification (RF_5). The Quality improvement (RF_9) also show the strong interrelationship with above four risk factors. Further, Technical capabilities (RF_{11}) exhibits strong relationship with ISO certification (RF_5) and Manufacturing Yield

(RF_{12}). All these factors are related to the operational factors category, which reveals that the quality failure is more related to supplier internal operational capabilities to produce the product for obligatory quality requirements.

Table 6.5: The inter-relationship among risk factors with respect to quality risk

	RF_5	RF_6	RF_7	RF_8	RF_9	RF_{11}	RF_{12}	RF_{13}	RF_{14}	RF_{15}	RF_{16}	RF_{17}	RF_{20}	RF_{21}	RF_{23}
RF_5	0	0	0	1	0	0	2	0	0	0	0	1	0	1	1
RF_6	2	0	0	1	0	1	1	0	0	1	0	1	0	1	1
RF_7	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0
RF_8	2	2	0	0	0	1	2	1	1	1	0	1	1	1	1
RF_9	2	2	0	2	0	1	2	0	0	1	0	1	1	1	1
RF_{11}	2	1	0	1	1	0	2	0	0	1	0	1	0	1	1
RF_{12}	1	1	0	1	0	0	0	0	0	0	0	1	0	1	1
RF_{13}	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
RF_{14}	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{15}	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0
RF_{16}	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
RF_{17}	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
RF_{20}	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0
RF_{21}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{23}	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

RF_5= ISO certification, RF_6= Quality Award, RF_7= Warranty, RF_8= Quality record, RF_9= Quality improvement, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{13}= Production Facility, RF_{14}= Cycle Time, RF_{15}= Capacity utilization, RF_{16}= Availability, RF_{17}= Relationship, RF_{20}= Natural Disasters, RF_{21}= Manmade Disaster, RF_{23}= Infrastructure

The results of risk factors behaviour within system for quality risk are given in Table 6.6. The risk factors accounted 58.33% of total factors appeared in conjugation matrix have the higher active and passive sum than average value (calculated 5.67), this output supports that there may be interrelationship among risk factors to cause the desire outcome failure. The maximum active sum value is shown for quality record and quality improvement and maximum passive sum value obtained for ISO certification and Manufacturing Yield. That represent that the no ISO certification and low manufacturing yield increase the probability of quality risk in the presence of other factors such as no quality records and quality improvement. There are 21 mandatory records and panning documents required for ISO certification (ISO 9001:2008). Furthermore, the quality improvement is not possible without keeping proper previous record and planning documents. The results obtained in this study are clearly corresponding to such evident requirements.

Table 6.6: The behaviour of risk factors with respect to quality risk

Risk Factor		AS (i)	PS (i)	$I\text{AS}_r$ (i)	$I\text{PS}_r$ (i)	I (i)
ISO certification	**RF5**	6	12	25.52	67.78	46.65
Quality Award	**RF6**	9	11	52.04	51.27	51.66
Warranty	**RF7**	4	1	28.47	4.90	16.68
Quality record	**RF8**	14	9	75.38	47.67	61.53
Quality improvement	**RF9**	14	1	100	4.11	52.06
Technical capabilities	**RF11**	11	4	72.74	21.55	47.14
Manufacturing Yield	**RF12**	6	12	29.19	70.42	49.80
Production Facility	**RF13**	2	3	11.82	14.93	13.38
Cycle Time	**RF14**	1	4	9.93	17.78	13.85
Capacity utilization	**RF15**	7	5	44.34	25.66	35
Availability	**RF16**	3	0	24.82	0	12.41
Relationship	**RF17**	2	6	0	50.15	25.08
Natural Disasters	**RF20**	5	2	42.71	9.88	26.30
Manmade Disaster	**RF21**	0	8	0	71.12	35.56
Infrastructure	**RF23**	1	7	0	59.72	29.86

AS (i)= Direct Active Sum, PS (i)= Direct Passive Sum, $I\text{AS}_r(i)$= Relative Indirect Active Sum, $I\text{PS}_r(i)$= Relative Indirect Passive Sum, I(i)=Integration

The in-depth risk factors behaviour is analysed to understand the indirect interrelation of risk factors. The highest $I\text{AS}_r(i)$ is obtained by the quality improvement followed by the quality record and technical capability risk factors, which are 100, 75.38 and 72.74 respectively. Quality improvement is directly related to quality record, quality award, manufacturing yield and ISO certification, where quality record is directly related to quality award, manufacturing yield and ISO certification and technical capability is directly related to manufacturing yield and ISO certification. Therefore, these factors are reinforcing each other and increase their indirect interaction within the system for quality risk occurrence. Consequently, supplier's deprived abilities regarding these risk factors increase the probability of quality risk.

Further the manufacturing yield and ISO certification has high $I\text{PS}_r$ (i) value i.e. 70.42 and 67.78 respectively. These two factors are strongly related to highest $I\text{AS}_r$ (i) value factors. Therefore the lack of ISO certification and low manufacturing can effectively increase the probability of quality risk in conjugation with supplier's deprived technical capabilities, quality records and quality improvement abilities.

To analyse the feedback loop of system the Integration I(i) values are estimated for each risk factor. The result exhibit the quality record has the highest integration value, it has feedback loop with quality award, manufacturing yield and ISO certification. Further it has an interconnection with quality improvement factor, which has second

highest Integration I(i) values, quality improvement is also interconnected with quality award, manufacturing yield and ISO certification. These results show high value of mutual connection between the factors for quality risk.

6.2.3 Cost Risk

A purchase order that is not delivered at the price that was agreed at the time of order placement is termed as cost risk. The implementations of proposed criteria for knowledge extraction provide 5 interesting rule about cost risk. This accounts about 17.86% of all cost risk rules available in supply risk identification model's rule set. Table 6.7 provide selected interesting rules about cost risk class.

Table 6.7: The selected rule set about cost risk

1. Z-Score > 2.63 AND Commodity price > 5 AND Availability = Low AND Economic Freedom > 70.12 AND Manufacturing Yield > 60.14 AND Quality record = YES: Cost-Risk

2. Z-Score > 2.69 AND Commodity price > 5 AND Availability = Low AND Economic Freedom > 70.12 AND Manufacturing Yield > 60.14 AND Relationship > 0.43 AND ISO certification = YES AND Political Stability = Orange: Cost-Risk

3. Z-Score > 2.69 AND Commodity price > 5 AND Availability = Low AND Economic Freedom > 70.12 AND Capacity utilization <= 80.82 AND Commodity price <= 15.76 AND ISO certification = YES AND Natural Disasters = Yellow: Cost-Risk

4. Warranty = YES AND Availability = Low AND Capacity utilization <= 80.46 AND Economic Freedom > 69.53 AND Supplier lock = Sole AND Logistics Performance Index <= 4.76: Cost-Risk

5. Technical capabilities > 3.92 AND Commodity price > 4.65 AND Exchange rate > -5.07 AND Availability = Low AND Capacity utilization <= 80.82 AND Supplier lock = Sole AND Exchange rate <= 6.96 AND Price comparison = Avg.: Cost-Risk

According to discovered knowledge about cost risk in available data, the low product availability and increase in commodity price cause cost risk. Further the supplier has good technical capabilities, with good manufacturing yield, capacity utilization less than 80% and has ISO certification cause the cost risk. The further analysis of the case study company shows that the cost risk is mostly related to certain group of products. The discussion with the case study manager revealed that this group of product is high technology products, which are available from specific suppliers. Above mention factors such as good technical capabilities, with good manufacturing yield, capacity utilization less than 80% and has ISO certification shows the strong technical and

operational capability, generally those are required to produce the high technology products.

Further, suppliers with technical competence would be less likely to confront operational problems and, because of its stronger price-bargaining power and semi-structured order, would be less likely to have financial difficulties. This is very consistent with the discovered knowledge as the supplier in safe financial zone (Z-Score >2.7) with high technical capabilities and sole purchasing policy and located to high economic freedom location cause cost risk. Supplier manufacturing products using a valuable and rare operational capability that cannot be easily copied by competitors can enjoy a certain monopolistic position and can sell their products as sole supplier (Jung et al 2011).

The conjugation matrix given in Table 6.8 presents the mutual inter-connection values for 17*17 pairs of risk factors with respected to cost risk. All the factors related to financial category in this case study appeared along with other different factors from the rest of three factor categories. The strongest interrelationships are observed for pairs: Commodity price (RF_2) and Availability (RF_{16}), Z-Score (RF_1) and Commodity price (RF_2), Availability (RF_{16}) and Economic Freedom Index (RF_{24}). This illustrates that the factors located outside wall of supplier such as commodity price, availability, economic freedom index can be the main cause of the cost risk.

Table 6.8: The inter-relationship among risk factors with respect to cost risk

	RF_1	RF_2	RF_3	RF_4	RF_5	RF_7	RF_8	RF_{11}	RF_{12}	RF_{15}	RF_{16}	RF_{17}	RF_{18}	RF_{20}	RF_{22}	RF_{24}	RF_{25}
RF_1	0	4	0	0	2	0	1	0	2	1	3	1	0	1	1	3	0
RF_2	0	1	1	2	3	0	1	0	2	2	4	1	1	2	1	3	0
RF_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_4	0	0	2	1	0	0	0	0	0	1	1	0	1	0	0	0	0
RF_5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
RF_7	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1
RF_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{11}	0	1	1	2	0	0	0	0	0	1	1	0	1	0	0	0	0
RF_{12}	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0
RF_{15}	0	1	1	1	1	0	0	0	0	0	0	0	2	1	0	1	1
RF_{16}	0	1	1	1	2	0	1	0	2	3	0	1	2	1	1	4	1
RF_{17}	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
RF_{18}	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
RF_{20}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{22}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RF_{24}	0	1	0	0	2	0	1	0	2	1	0	1	1	1	1	0	1
RF_{25}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

RF_1= Z-Score, RF_2= Commodity price, RF_3= Price comparison, RF_4= Exchange rate, RF_5= ISO certification, RF_7= Warranty, RF_8= Quality record, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{15}= Capacity utilization, RF_{16}= Availability, RF_{17}= Relationship, RF_{18}= Supplier lock, RF_{20}= Natural Disasters, RF_{22}= Political Stability, RF_{24}= Economic Freedom Index, RF_{25}= Logistics Performance Index

The results of the in-depth behaviour analysis are summarized in Table 6.9. The results reveal a maximum AS (i) of 24 and a maximum PS (i) of 12 with 82% of factors either high AS (i) or high PS (i) than average value. These results support the presence of strong interrelationships between the risk factors. Commodity price and availability show the highest active sum values of 24 and 21 respectively, while economic freedom and ISO certification show the highest passive sum value of 12. These outputs strengthen the implication obtained through knowledge discovery from the available data, since less availability of purchased product in market can help suppliers to have high bargaining power. Therefore they can be in a position of transferring the increase in commodity price to their customer and keeping themselves in a safe financial position. By taking a closer look at the indirect interrelationships among risk factors, it is revealed that z-score show the maximum relative indirect active $IAS_r(i)$ of 100. Z-score is strongly interconnected with the commodity price, availability and economic freedom. These factors have high direct active sum means strongly interconnect with other factors, that results in the z-score having a strong interaction within the system. Commodity price has second highest $IAS_r(i)$ accounted

76.60, it is strongly interconnected with availability, ISO certification and economic freedom.

Table 6.9: The behaviour of risk factors with respect to cost risk

Risk Factor		AS (i)	PS (i)	IAS_r (i)	IPS_r (i)	I (i)
Z-Score	**RF1**	19	0	100	0	50
Commodity price	**RF2**	24	9	76.60	16.72	46.66
Price comparison	**RF3**	0	7	0	26.59	13.30
Exchange rate	**RF4**	6	8	16.97	25.02	20.99
ISO certification	**RF5**	2	12	0	39.14	19.57
Warranty	**RF7**	5	0	17.44	0	8.72
Quality record	**RF8**	0	5	0	16.19	8.09
Technical capabilities	**RF11**	7	0	34.72	0	17.36
Manufacturing Yield	**RF12**	4	8	0	23.47	11.73
Capacity utilization	**RF15**	9	10	22.66	26.26	24.46
Availability	**RF16**	21	10	46.61	17.43	32.02
Relationship	**RF17**	2	5	0	16.19	8.09
Supplier lock	**RF18**	3	9	3.22	29.75	16.48
Natural Disasters	**RF20**	0	7	0	27.31	13.66
Political Stability	**RF22**	0	7	0	26.68	13.34
Economic Freedom	**RF24**	12	12	19.44	27.72	23.58
Logistics Performance Index	**RF25**	0	5	0	19.19	9.59

AS (i)= Direct Active Sum, PS (i)= Direct Passive Sum, $IAS_r(i)$= Relative Indirect Active Sum, $IPS_r(i)$= Relative Indirect Passive Sum, I(i)=Integration,

The maximum relative indirect passive sum $IPS_r(i)$ is shown for ISO certification (RF_5) and supplier lock (RF_{18}). It shows that these factors are also highly contributing toward cost risk. ISO certification is connected with commodity price, z-score, availability and economic freedom. Also the supplier lock is connected with availability and capacity utilization. Both ISO certification and supplier lock relevant with supplier position in network to provide the superior product. In general there is high risk of supply disruption due to sole supplier (Odette 2013), disruption risk can be avoided by imposing the sole supplier to maintain the excess capacity by especial contract, however imposing such requirement would add to cost especially when it is providing the superior product (Pereira 2005).

The integration vale is estimated for risk factors to identify the feedback looping factors. The highest value is obtained for z-score, which has the highest $IAS_r(i)$, however $IPS_r(i)$ is zero. The feedback loop of z-score is due to its strong interaction with three highly direct interconnected factors (Commodity price, availability and economic freedom) with the system. These results further support the proposition that

any risk factors in a dynamic environment such as a supply chain may not affect the output in an isolated way. Furthermore, the factors showing interconnection belong to different categories of risk factors, which indicate that the reasons of risk can be exist anywhere in upstream supply chain. Therefore, for conducting a supply risk identification a 360 degree overview operating supply chain is necessary, where data drive approach such as the proposed one can be very suitable.

The current study is aimed at developing a risk scoring model, which can be used as tool to assess the supplier risk. The previous sections discussed the results of the supply risk identification model that is designed to provide feedback to develop the risk scoring model. Now, the next section will provide the results about the supplier risk scoring model.

6.3 Risk Scoring Model

Risk scoring model is a function, f(X, β), that requires the classification of a data mining task. The classification task is aimed at building a data mining model that estimates the β values to predict the dependent variable (supplier risk) based on its relationship with the independent variables. Before building the data mining model for risk scoring, first appropriate variables are selected that have an adequate relationship with the supplier risk and used as independent variables "X" for developing the supplier risk scoring model. To develop the best model for risk scoring, the results of appropriate variables selection and discretization are provided in the following section.

6.3.1 Selection of Appropriate Variables and Discretization

The selection of appropriate variables is aimed at providing the independent variables that has a relationship with dependent variable (supplier risk). Supplier risk is the dependent variable, according to supplier risk definition; a supplier is classified as risky if its "supply performance" is lower than the contracted obligatory supply performance due to realised supply risk. A rule-based knowledge discovery approach is adapted that convey as much additional information as possible relating to the impact of risk factors toward supply performance. Based on this information (discovered knowledge about supply risk), appropriate variables those fulfil the define criteria as given in equation 4.26 are selected for risk scoring from initial dataset of variables given in Table 5.1.

$RF(X)$ *is selected if*

$$IAS_r\ (i)\ \geq IAS_r\ (averge)\ \ Or\ \ \mathrm{I}\ (i) \geq \mathrm{I}\ (averge) Or\ IPS_r\ (i)\ \geq IPS_r\ (averge)$$

Active and passive sum represents each factors overall contribution toward one specific supply performance criteria (supply risk outcome). While integration represents each factors overall interaction with other factors in the system with respect to one specific supply performance criteria (supply risk outcome). These supply performance criteria (supply risk outcome) are used to calculated the overall supply performance that is used to define the dependent variable (supplier risk) in this current study (see Figure 4.2). Therefore, the value of active and passive sum and integration of risk factor is considered as representative of its predictive power toward the dependent variable (supplier risk). The result is presented in the form of bar chart in Figure 6.7, where each risk factor shows its active sum $IAS_r\ (i), IPS_r\ (i)$ and integration $\mathrm{I}\ (i)$ with respect to three selected supply risk outcomes with different colour.

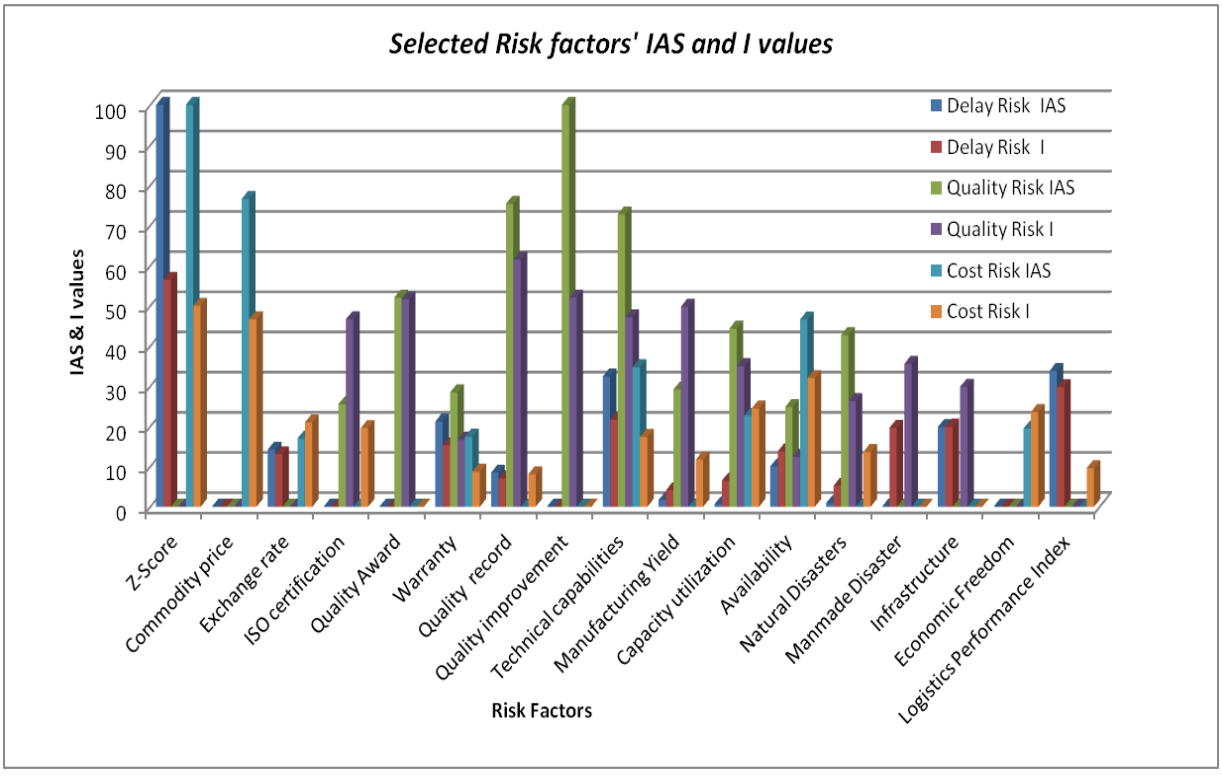


Figure 6.7: The selected variables for risk scoring model

After the selection of variables, there is a need for converting the numerical type variables in to categorical type to have "best" risk scoring model in terms of its stability and generalization. Based on extracted knowledge about supply risk, a method is proposed (see chapter 4 section 4.1.3.1) for converting the selected numerical type variables into categorical type. Table 6.10 shows the results of discretization process.

Table 6.10: Discretization of numerical type variables

Variable	Group	Value (Range)	Variable	Group	Value (Range)
Z-Score	1	$RF_1 > 2.7$	Commodity Price	1	$RF_2 > 5$
	2	$2.7 \leq RF_1 \geq 1.8$		2	$RF_2 \leq 5$
	3	$RF_1 < 1.8$	Technical capabilities	1	$RF_{11} > 4$
Exchange rate	1	$RF_4 > -5$		2	$4 \leq RF_{11} \geq 3$
	2	$RF_4 \leq -5$		3	$RF_{11} < 3$
Manufacturing Yield	1	$RF_{12} > 70$	Capacity utilization	1	$RF_{15} > 81$
	2	$70 \leq RF_{12} \geq 60$		2	$RF_{15} \leq 81$
	3	$RF_{12} < 60$	Infrastructure	1	$RF_{23} > 4.1$
Logistics Performance Index	1	$RF_{25} > 3$		2	$RF_{23} \leq 4.1$
	2	$RF_{25} \leq 3$	Economic Freedom	1	$RF_{24} > 70$
				2	$RF_{24} \leq 70$

RF_1= Z-Score, RF_2= Commodity price, RF_4= Exchange rate, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{15}= Capacity utilization, RF_{23}= Infrastructure, RF_{24}= Economic Freedom Index, RF_{25}= Logistics Performance Index

After selection of X is a vector of independent variables (section 6.5.1), now there is need to estimate "β" values. In the current study, data mining model is built using logistic regression technique to estimate the "β" values used for risk scoring. The following section will provide the result of model build for risk scoring.

6.3.2 Risk Scoring Model Building

The risk scoring model is built using the data about selected independent variables and dependent variable (supplier risk) to estimate the weight ""β_j" of independent variables. Initially two data samples was collected, first sample was containing the data related to individual purchase orders performance and second data sample was containing the data related to suppliers' annual supply performance and their related supply chain characteristics (see chapter 5 section 5.5.1.2). The second data sample is used to build and test data mining model for supplier risk scoring. The dataset consists of total 820 observations covering six year time period (2008-2013), out of which 684 observation covering five year time period (2008-2012) are used to build the model termed as training data. Statistical description of training data is given in Figure 6.8.

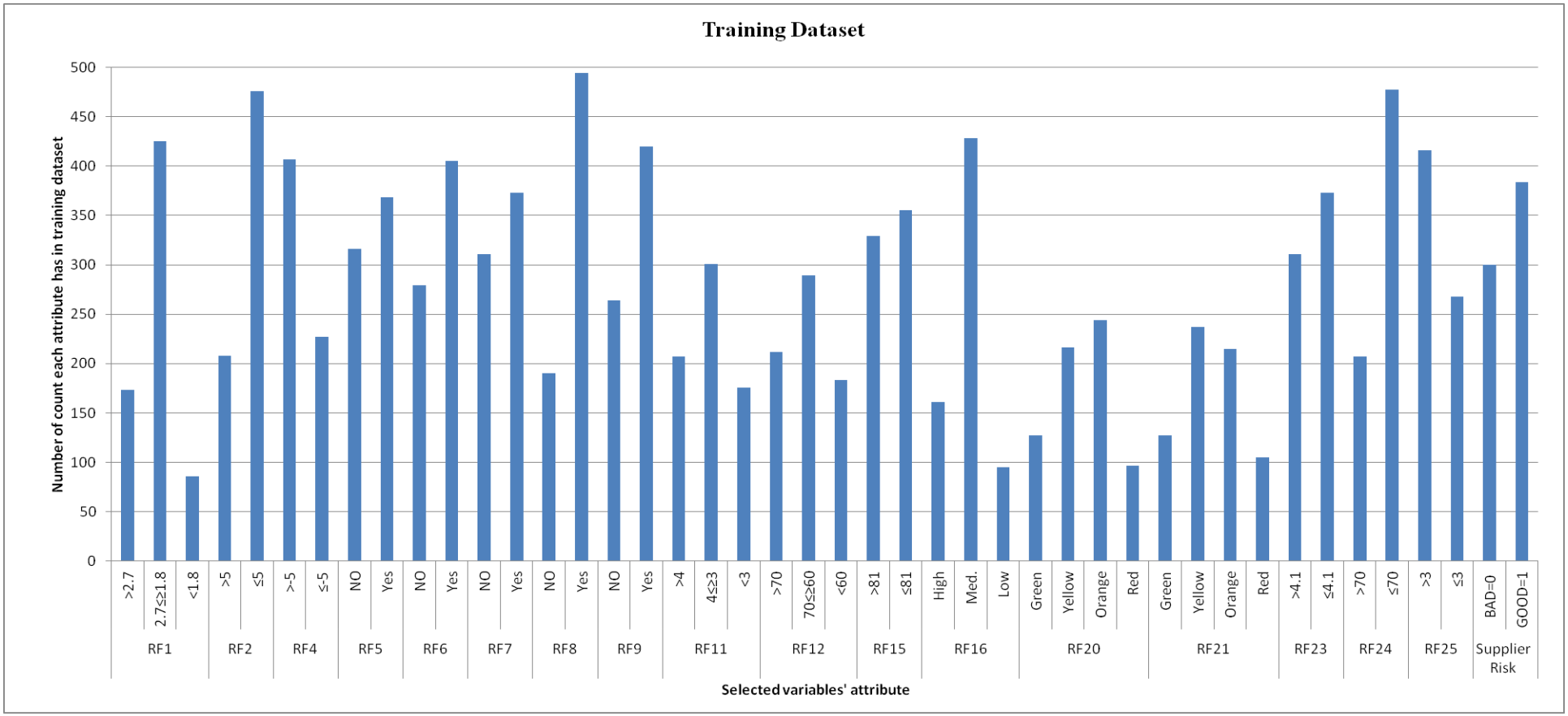


Figure 6.8: The general statistics of training dataset

The model is built for predicting the supplier risk class (labelled "0" for "Bad/risky" or "1" for "Good/non-risky" supplier) using the logistic regression algorithm. The logistic regression model is built to estimate "β_j" for selected independent variables is given in Table 6.11.

Table 6.11: The estimated weighing "β_j" values of independent variables

Variable	Attribute	β_j	Variable	Attribute	β_j
Z-Score	$RF_1 > 2.7$	-0.14	Capacity utilization	$RF_{15} \leq 81$	-0.16
	$2.7 \leq RF_1 \geq 1.8$	-0.41	Availability	$RF_{16} = High$	-0.48
	$RF_1 < 1.8$	1.11		$RF_{16} = Med.$	-0.87
Commodity price	$RF_2 \leq 5$	-1.82		$RF_{16} = Low$	2.42
Exchange rate	$RF_4 \leq -5$	1.04	Natural Disasters	$RF_{20} = Green$	-0.17
ISO certification	$RF_5 = YES$	-0.61		$RF_{20} = Yellow$	-0.07
Quality Award	$RF_6 = YES$	-0.89		$RF_{20} = oragnge$	0.18
Warranty	$RF_7 = YES$	-0.41		$RF_{20} = Red$	-0.01
Quality record	$RF_8 = YES$	-2.38	Manmade Disaster	$RF_{21} = Green$	-0.04
Quality improvement	$RF_9 = YES$	-1.38		$RF_{21} = Yellow$	-0.1
Technical capabilities	$RF_{11} > 4$	-0.28		$RF_{21} = oragnge$	0.08
	$4 \leq RF_{11} \geq 3$	0.45		$RF_{21} = Red$	0.07
	$RF_{11} < 3$	-0.27	Infrastructure	$RF_{23} \leq 4.1$	-0.73
Manufacturing Yield	$RF_{12} > 70$	-0.45	Economic Freedom	$RF_{24} \leq 70$	-1.4
	$70 \leq RF_{12} \geq 60$	0.23	Logistics Performance Index	$RF_{25} \leq 3$	0.87
	$RF_{12} < 60$	0.2	Intercept	α	5.99

RF_1= Z-Score, RF_3= Price comparison, RF_4= Exchange rate, RF_5= ISO certification, RF_6= Quality Award, RF_7= Warranty, RF_8= Quality record, RF_9= Quality improvement, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{15}= Capacity utilization, RF_{16}= Availability, RF_{20}= Natural Disasters, RF_{21}= Manmade Disaster, RF_{23}= Infrastructure RF_{24}= Economic Freedom Index, RF_{25}= Logistics Performance Index

The results indicates that availability ($RF_{16} = Low$) has the highest value of "β_j"to calculated supplier risk class labelled 0 i.e. Bad, followed by z-score ($RF_1 < 1.8$) both accounted 2.42 and 1.11 respectively. According to discovered knowledge about supply risk low availability cause the delay risk and cost risk, while low z-score cause the delay risk. Interrelationship analyse of factors represent that "availability" has reasonable values of active sum and integration with respect all the supply risk type i.e. delay risk, quality risk and cost risk. While Z-score showed the highest active sum and integration values for delay risk and cost risk. Furthermore, quality records ($RF_8 = YES$) and Commodity price ($RF_2 \leq 5$) shows the lowest value of "β_j"to calculated supplier risk class labelled 0 i.e. Bad, both account "-2.38" and "-1.82". According to discovered knowledge about supply risk, quality records ($RF_8 = NO$) is

the main cause for quality and delay risk, while commodity price ($RF_2 > 5$) is the main cause of cost risk. Quality records' active sum and integration values are very high for quality risk and moderate for delay risk. While the commodity price shows the high active sum and integration values with respect to cost risk. In the current case study, the highest type of supply risk is delay risk followed by the cost risk. The above indicated result seems very appropriate under these conditions. The value of "β_j" of each attribute of independent variable will be used to calculate the raw score.

6.3.3 Risk Scoring Model Testing and Validation

The model is built for predicting the supplier risk in future; therefore, validity of the model should be based on its predictive accuracy for unseen new data (testing data). Data set about 136 suppliers consists of 820 observations over the period of six years out of which 684 observations were used for model building and 136 observations last year i.e. 2013 were left out as testing data. The testing data was not used for model building, so it is unseen data. Of the 136 suppliers in testing data sample, 58 have failed to deliver contracted performance labelled as bad (0) and 78 have fulfilled the required contracted performance labelled as Good (1). General statistic about the selected variables after discretization for given testing dataset is shown in Figure 6.9.
The training model is tested on testing sample using hold-out evaluation method. The output result about supplier risk class labelled as (1or 0) is compared with its actual known performance and results are presented in confusion matrix Table 6.12.

Table6.12: The classification confusion matrix for testing dataset

Actual classification	Predictive Classification	
	BAD =0	GOOD =1
BAD =0	46	12
GOOD =1	6	72

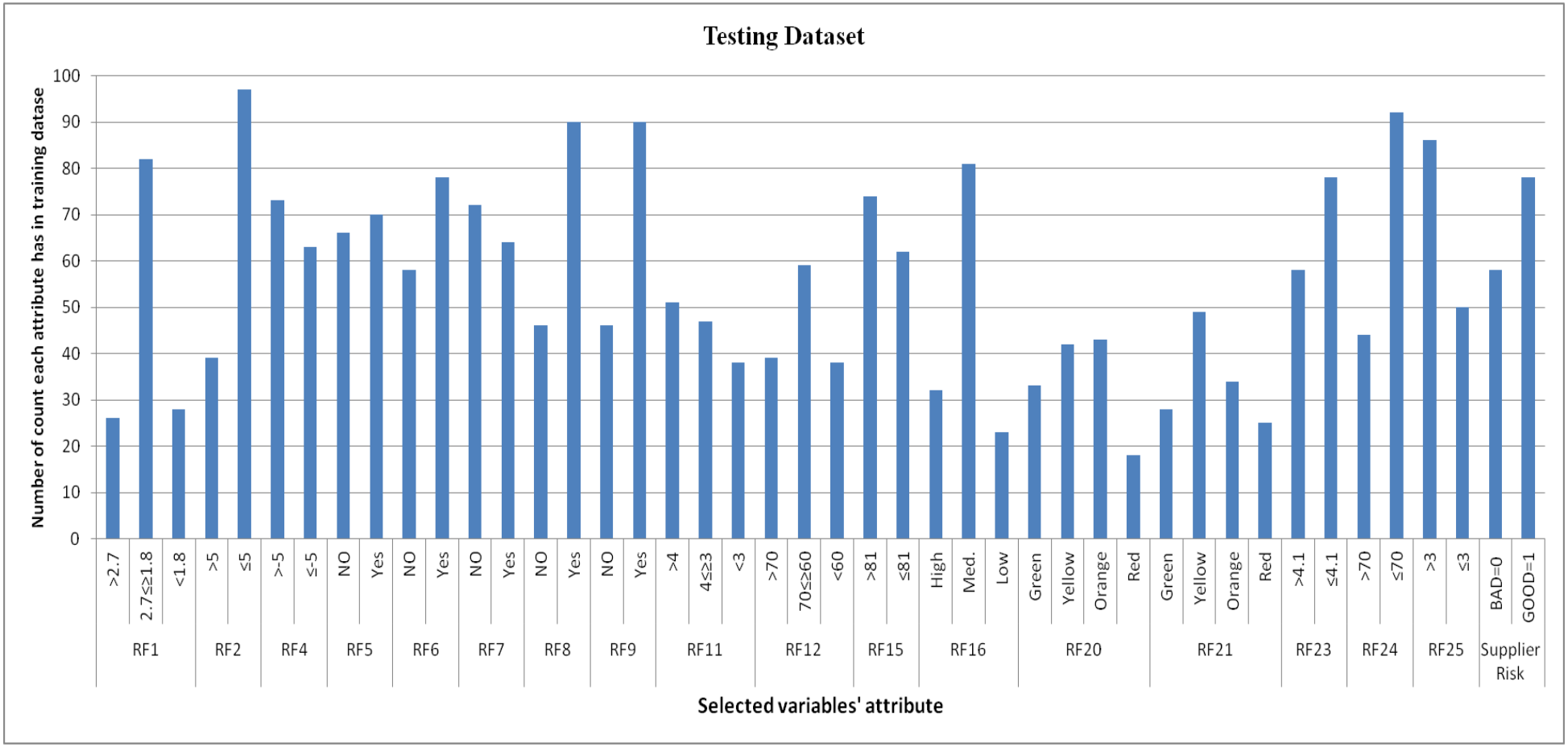


Figure 6.9: The general statistics of testing dataset

The model exhibit very high accuracy with 86.76% for correctly predicting the instances in test data set. The model has mean absolute error rate of 0.2282. Area under the receiver operating curve (AUC) is higher than the required threshold value of 0.5 and account about 0.863. This shows that the model has less over fitting probability on new data. A summary of model's performance on classification evaluation metric is given in Table 6.13.

Table6.13: The classification performance of model on testing dataset

Class/model	Accuracy	AUC	Precision	Recall	F-measure
BAD=0	0.793	0.863	0.885	0.793	0.836
GOOD=1	0.923	0.863	0.857	0.923	0.889
Overall	0.868	0.863	0.869	0.868	0.866

Although the model shows the high predictive accuracy, however there is a need to check its predictive power against an appropriate benchmark. The available sample for model building and testing was unbalanced. Therefore, this study used the Neter (1966) method for calculating the benchmark. Out of 820 total observations for overall sample, 42% observations are bad and 58 percent observations are good. However for tainting dataset and testing data set 44% and 43% observations are labelled as bad and 56% and 57% observations are labelled as good respectively. Using the Neter (1966)'s method for overall sample, training and testing sample the benchmark accuracy value is calculated as respectively

$$(.42 \times .42 + .58 \times .58) = .5128$$
$$(.44 \times .44 + .56 \times .56) = .5072$$
$$(.43 \times .43 + .57 \times .57) = .5098$$

The benchmark values for overall, training and testing sample is about 51.28 %, 50.72% and 50.98% respectively. Further, Hair et al., (2006) states that a model should be considered valid if it has classification accuracy greater than at least one quarter of what would be achieved by chance. In the current study, as the class output is binary so by chance prediction accuracy is 50%. Therefore, a valid model must have accuracy of 62.5% or better for both training and holdout dataset. Current model has predictive accuracy of 88.16% and 86.76% for both training and holdout dataset respectively. That is much higher than the required benchmarks; therefore build model is valid and can be used to calculate the standardized score for supplier risk assessment.

In the current study, the proposed approach (methodology) for supplier risk assessment is tested against two requirements.

- The first is its ability to predict accurately the unknown supply risk and produce meaningful knowledge about the supply risk in available data.
- The second is the viability of such an approach in building model for supplier risk score

To evaluate the results of the first, a number of evaluation metrics and proposed approach is used to identify the best model, which provides the best result in term of high prediction accuracy and interesting rules for supply risk.

A build model for supplier risk score is aimed at providing high prediction accuracy supplier risk class. Therefore, to evaluate the second requirement, the impact of extracted knowledge on classification accuracy of build model for risk scoring is analysed in following section.

6.4 Performance Comparison

To calculate risk score, a data mining model is developed for available data about the independent variables and dependent variable. In the current case study, initially there are 25 independent variables available for building a model for risk score (see Table 5.1). However, the proposed approach adopted a different view point and selected only variables (risk factors) those are significant predictors of supply performance that is used to define the dependent variable (supplier risk). In the current case study we have selected a list of factors according to purchasing managers' perception and previous literature about supply risk. However instead of using all these factors directly into supplier risk assessment, a rule based knowledge discovery approach is used to reduce the number of independent variables by identifying the factors actually contributing toward realised supply risk or impacting on supply performance. Furthermore, numerical type data is dsicretized into bins (categorization).

To see the impact of the proposed approach on classification performance it is compared with other approaches,

- A model is developed without using knowledge discovery approach. This means that a model is developed for initial listed risk factors (25 independent variables) and actual supplier performance without using knowledge discovery approach. Then the result of this model (given in detail in appendix III) is compared with knowledge driven model build for risk scoring.
- A model is developed without knowledge discovery and using state of the art discretization approach such as equal-width binning method (results are shown in appendix IV). Equal width binning approach convert the numerical type data into categorical type data by converting numerical values into equal width of bins (results are shown in appendix V).
- A model is developed using state of the art variable selection approach such as Correlation-based Feature selection method (Hall 1998). Correlation-based Feature selection method identifies the subset of important variables by considering individual predictive capability of each feature along with the degree of redundancy between them. The result of this model (shown in Appendix VI) is compared with knowledge driven model build for risk scoring.
- A model is developed using both state of the art variable selection and discretization approaches. The result of this model (shown in Appendix VII) is compared with knowledge driven model build for risk scoring.

Sample about 136 supplier over of period of six year with 25 risk factors and their supplier performance consists of total 820 observations, out of which 684 observation were used to build different models (statistics are given in appendix VII). A sample of 136 observations is used to test these build models (statistics are given in appendix IX). Output of resultant models in terms of classification performance is compared with knowledge discovery model in term of classification performance is given in Figure 6.10.

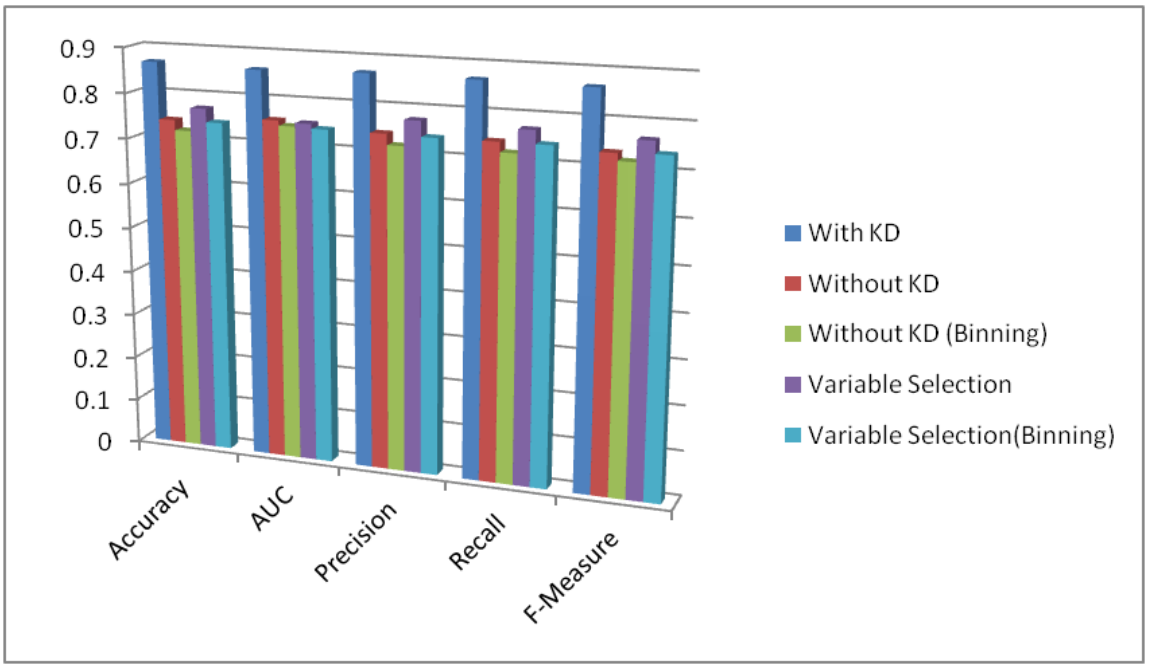


Figure 6.10: The comparison of KD risk scoring with other approaches

All the models without knowledge discovery approach have acceptable overall accuracy with respected to both Neter (1966)'s method and Hair et al. (2006)'s benchmark, however all model behaves poorly for the minority class. All models except without knowledge discovery and with discretization approach model have not achieved required benchmark stated by Hair et al. (2006)'s i.e. 61.5% for minority class. In the case of unbalance data, the minority class higher accuracy is more desirable especially when minority class is the main target. Such as in current study, minority class labelled as "bad" is main target and it is desirable to have high accuracy. The proposed knowledge discovery model provides much higher accuracy for both minority and majority class than the required benchmarks (both Neter (1996)'s method and Hair et al. (2006) benchmarks). Furthermore, the knowledge discovery base supplier risk scoring model outperformed all other models without knowledge discovery approach on all metrics.

The build model without Knowledge discovery has 74.3% accuracy, while the proposed approach conduct the variable selection and provide 86.8% classification accuracy. According to Occam's Razor's principle the simplest is best and furthermore unnecessary predictors will add noise to the estimation of desire output i.e. supplier risk.

Automatic variable selection method such one used in this study is aimed to construct a model that predicts well or explains the relationships in the data; however automatic

variable selections does not guarantee the consistency for these goals. Piramuthu, (2004) compared different feature selection techniques in his study but did not find a real winner. Automatic variable selection is a means to an end and not an end itself. In the current study, proposed method for variable selection provided higher prediction than automatic variable selection method in the current thesis. These results highlight the validity of knowledge driven risk scoring model building method that enhanced the classification performance.

Further, in the proposed approach, a knowledge driven discretization method is proposed to reduce the number of possible values of numerical type variable. The problem of choosing the interval borders and the correct artily for the discretization of a numerical value range remains an open problem in numerical feature handling (Kotsiantis and Kanellopoulos, 2006). The models built using most common discretization technique "equal width binning" are outperformed by the knowledge driven model on all performance evaluation metrics (see Figure 6.10). There is common harmony in data mining literature that there is no universal approach for building best data miming model, however different methods or techniques can be applied that perform better for a given problem in available resources. The current proposed approach performed better in the current problem domain and available data that underline its validity for the stated problem with the given resources, excluding the claim of its "comprehensiveness" in such problem domain and given resources.

6.5 Standardized Risk Scores

The final objective of study is to develop standardized risk score for supplier risk assessment similar to famous credit score (FICO) with fixed range. To develop a standardized risk score first the raw risk score is calculated using the estimated ""β_j"" values for independent variables according to raw score equation 4.39 (see chapter 4 section 4.1.3.3). Based on the result, supplier raw risk score is calculated by putting the binary values (0 *or* 1) according to supplier profile in equation below

$$supplier\ raw\ rsik\ score$$
$$= 5.99 - 0.14 \times (RF_1 > 2.7) - 0.41 \times (2.7 \leq RF_1 \geq 1.8) + 1.11 \times (RF_1 < 1.8) - 1.82 \times (RF_2 \leq 5) + 1.04 \times (RF_4 \leq -5) - 0.61 \times (RF_5 = YES) - 0.89 \times (RF_6 = YES) - 0.41 \times (RF_7 = YES) - 2.38 \times (RF_8 = YES) - 1.38 \times (RF_9 = YES) - 0.28 \times (RF_{11} > 4) + 0.45 \times (4 \leq RF_{11} \geq 3) - 0.27 \times (RF_{11} < 3) - 0.45 \times (RF_{12} > 70) + 0.23 \times (70 \leq RF_{12} > 60) + 0.2 \times (RF_{12} < 60) - 0.16 \times (RF_{15} \leq 81) - 0.48 \times (RF_{16} = High) - 0.87 \times (RF_{16} = Med.) + 2.42 \times (RF_{16} = Low) - 0.17 \times (RF_{20} = Green) - 0.07 \times (RF_{20} = Yellow) + 0.18 \times (RF_{20} = oragnge) - 0.01 \times (RF_{20} = Red) - 0.04 \times (RF_{21} = Green) - 0.1 \times (RF_{21} = Yellow) + 0.08 \times (RF_{21} = oragnge\ ) + 0.07 \times (RF_{21} = Red) - 0.73 \times (RF_{23} \leq 4.1) - 1.4 \times (\ RF_{24} \leq 70\ ) + 0.87 \times (RF_{25} \leq 3)$$

Table 6.14: Example to calculate the supplier's raw risk score

Suppose that given supplier has profile as given in table below

RF_1	RF_2	RF_4	RF_5	RF_6	RF_7	RF_8	RF_9	RF_{11}	RF_{12}	RF_{15}	RF_{16}	RF_{20}	RF_{21}	RF_{23}	RF_{24}	RF_{25}
2.8	7.2	2	Y*	N*	Y	Y*	N*	3	72	75	low	red	red	5	60	3.5

*Y=Yes & N=No

To calculate raw score for supplier's profile the binary values are placed in supplier raw risk score equation. As the supplier has$RF_1 = 2.8$, this values is greater than 2.7 therefore, "1" value will be placed at $(RF_1 > 2.7)$, however "0" value is placed at $(2.7 \leq RF_1 \geq 1.8)$ & $(RF_1 < 1.8)$ in raw risk score equation. In current supplier profile$RF_2 = 7.2$, this value is greater than 5, therefore, "0" value is placed at $(RF_2 \leq 5)$ in supplier raw score equation. $RF_5 = yes$ For given supplier therefore, "1" is placed at$(RF_5 = YES)$. Similarly the binary (0 *or* 1) values are placed in supplier raw score equation according to given supplier's profile and score is calculated. For current supplier profile the final raw score is obtained as

$supplier\ raw\ score\ = 5.99 - 0.14 - 0.61 - 0.41 - 2.38 + 0.45 - 0.45 - 0.16 + 2.42 - 0.01 + 0.07\ - 1.4 = 3.37$

This raw score will be used to calculate the standardized risk score.

Raw score of all the suppliers' profiles available in sample is calculated using the same method. In total 684 suppliers' profiles are available training sample, for which raw score is calculate the results are summarized in Table6.15.

Table 6.15: Standardized risk score for training data sample

	Max.	Mini.	Mean	St.dev.
Raw Score	10.2	-5.74	0.07	2.74
ω_i	15.94	0	5.8	2.74
Δ_i	$Raw\ Score_{max} - Raw\ Score_{mini} = 15.94$			
Range	200	800		
Standardized Risk Score	800	200	418.54	103.2918

"ω_i" is calculated by subtracting minimum raw score from each supplier's raw score (see equation 4.41). Δ_i is calculated by subtracting minimum raw score from maximum raw score (see equation 4.40). Finally the standardized risk score for each supplier profile is calculated using equation 4.42 (see chapter 4 section 4.1.3.3) by setting $Range_{mini} = 200$ and $Range_{max} = 800$. Higher score represents the high level of supplier risk.

Further analysis is conducted to determine best cut off level that produced the best, more reliable and useful results for the current case study. As the objective of the supplier risk score is aimed to assess the risk level of new supplier or current supplier in change circumstances. Therefore, testing data (unseen for trained model) is used to determine the optimal cut-off value. Standardized risk score is calculated for each supplier profile in testing sample. To determine the cut-off value available standardized risk score is divided into ten groups (range). For each group number of bad and good supplier is calculated. The results are given in the Table 6.16.

Table 6.16: The analysis for cut-off value

SCORE RANGE	**#GOOD**	**#BAD**	**CU.GOOD**	**CU.BAD**	**CU.GOOD %**	**CU.BAD %**	**CU.BAD Avoided %**
200-294	12	2	12	2	15%	3%	97%
295-335	11	3	23	5	29%	9%	91%
336-358	13	1	36	6	46%	10%	90%
359-373	12	2	48	8	62%	14%	86%
374-390	13	1	61	9	78%	16%	84%
391-451	11	3	72	12	92%	21%	79%
452-534	1	13	73	25	94%	43%	57%
535-618	3	11	76	36	97%	62%	38%
619-724	1	13	77	49	99%	84%	16%
725-800	1	9	78	58	100%	100%	0%

$\#\mathbf{GOOD} = number\ of\ good\ supplier, \#\mathbf{BAD} = number\ of\ bad\ supplier,$
$\mathbf{U.GOOD} = cumulative\ good\ supplier, \mathbf{CU.BAD} = cumulative\ bad\ supplier$

Based on the current analysis, the cut-off value can be determined depending on company's objective. For example, for current case study the data sample is stretched on six year period (2008-2013), at that time company was expanding its supply base. Therefore, in that situation the score range 391-451 can give cut-off value, as it provides the 92% good supplier. According to expanding objective, the supplier having lower risk score than 391 can be given purchase order and supplier with higher risk score than 451 can be denied purchase order, however between one can be

referred to manger. ROC is curve is developed using cumulative bad and cumulative bad to analyse the validity of risk score ranges.

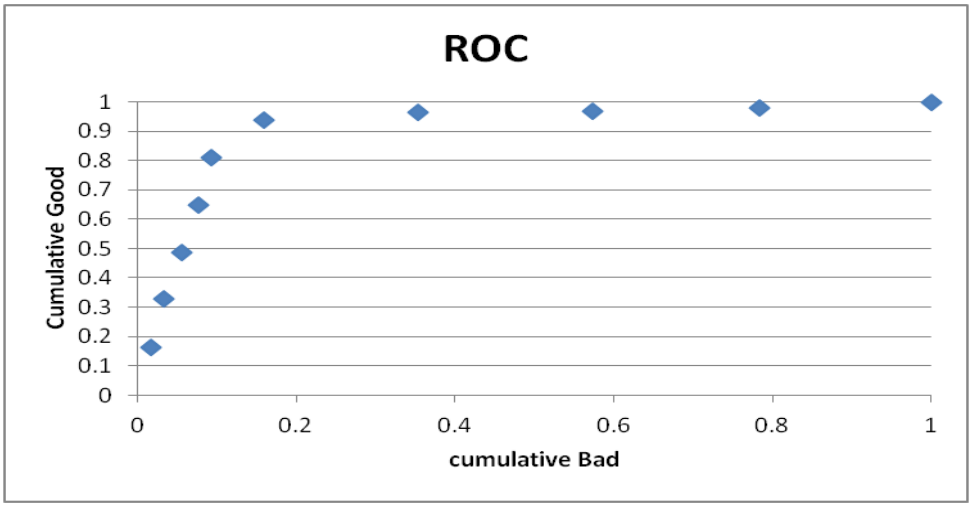


Figure 6.11: ROC curve based on cumulative Good and bad

Furthermore, a cost base profit analysis can also be conducted to determine the cut-off value. If profit loss ratio is known for the good and bad supplier then using number of good and bad supplier in each range, the total profit can be calculated to determine cut-off value. For example if profit loss ratio is 1:2 for the good and bad supplier then using this ratio profit base cut-off range 391-451 is sleeted. As it provides highest profit according to results summarized in Table 6.17.

Table 6.17: The analysis for cut-off values based on profit and loss

SCORE RANGE	**#GOOD**	**#BAD**	**Profit** (thousands)
200-294	12	2	8
295-335	11	3	13
336-358	13	1	24
359-373	12	2	32
374-390	13	1	43
391-451	11	3	48
452-534	1	13	23
535-618	3	11	4
619-724	1	13	-21
725-800	1	9	-38

The developed score card can help the purchasing manager to make fast, reliable decision about new supplier and continuing business with current supplier.

6.6 Summary

In order to test the methodology, a case study has been carried out that consisted of two parts. The first part involved the investigation of the way risk factors affects required outcomes i.e. supply performance metric quality, time and cost. For this, three classification models are developed based using three selected algorithm. These models are tested on testing dataset and according to proposed model selection method one model is selected for knowledge discovery. The best model uses PART algorithm as it performs better than the other two algorithms.

The knowledge discovery resulted in the selection of classification rules for the labelled classes: quality risk, delay risk and cost risk. The rules where further processed in order to investigate the interrelationships and mutual connection among different risk factors for specific labelled class. This analysis produced valid results and proved the ability of the method to identify the impact of supply chain characteristics (risk factors) on the desired outcome and relationships among factors. The analysis highlighted the fact that the effect varies upon the different conditions of supply chain characteristics such as Z-score >2.7 cause the cost risk, while Z-score <1.8 cause the delivery risk, however no validate impact of Z-score is identified for quality risk. Further it showed that different risk factors have strong mutual interaction with respect to specific supply risk outcome.

The second part of the case study involved the selection of appropriate variables and their preparation (discretization) for risk scoring model development that resulted in the supplier risk score. Based on the knowledge discovery the appropriate variables are selected and categories (binning process) using the proposed approach. The data sample related suppliers' performance is used to develop a risk scoring model by implementing the logistic regression algorithm. The results demonstrated high accuracies and outperformed the traditional methods used for similar purposes. Finally a standardized supplier risk score is generated that represents the suppliers' reliability for doing business in a similar manner to how the credit score represents the creditworthiness of a customer.

7 CONCLUSIONS

This chapter provides the conclusion of current research thesis that consists of five sections. In section 7.1, a brief overview of the scope of the study, importance and differentiation of this research thesis from previous work is given. The section 7.2 explains how the given research questions have been addressed followed by a list of the research outcomes that are given in section 7.3. Implementation and Limitations and are provided in section 7.4 and 7.5.

7.1 Thesis Overview

Nowadays, in current business operational structure, outsourcing and purchasing constitutes the large portion of companies' annual cost. Therefore, companies' profitability and the growth highly depend upon the performance of sourcing and purchasing decisions. However the purchasing performance is a future event and cannot be predicted with certainty. Even a supplier who meets all the required supplier selection criteria (high technical capability, production facility, financial position etc.) may end up in causing loss to company because of undesired output caused by future event happen i.e. supply risk in supply network. Hence, purchasing companies should have the techniques to identify suppliers who are more likely to perform according to desire performance in the presence of supply risk.

As mentioned in Chapter 2, previous most of the suppliers risk assessment approaches are based on deductive learning (expert-centred). These deductive learning approaches for supplier risk assessment are designed in such a way that capture the appropriately knowledge of the experts about supply risk for supplier risk assessment. These approaches face problems related to subjectivity, knowledge explication and updating due to their high dependency on expert's knowledge, experience and perception about supply risk. Current research thesis proposes the idea of implementing the inductive learning approach for supplier risk assessment. In this case, learning about supply risk is achieved through implementation of knowledge discovery approach on the data related to supply chain characteristics and performance. Further knowledge about supply risk is integrated into supplier is risk scoring model that predicts the probability of supplier risk in terms of risk score.

This study was designed to investigate the impact of supply chain characteristics on supply performance for supply risk identification and their incorporation within supplier risk scoring model development assessment. This methodology is closely associated with nature of the input data related to supply chain characteristics and defined purchasing performance. Hence the availability and good understanding of data are of great importance and strongly related to the success of the methodology. In this study, a database was developed for collecting the relative data from different data-sources. This is a primary activity to have an input for implementation of knowledge discovery and risk scoring approaches.

Knowledge discovery approach provides the supply risk identification model and risk scoring approach provides the supplier risk scoring model. However the implementation of Knowledge discovery approach that enables the uncovering of the previously unknown desired information about supply risk and the implementation of risk scoring approach for the development of supplier risk scoring model is articulated in an integrated manner for supplier risk assessment.

Knowledge discovery methodologies have been widely used for risk identification in different fields of studies such as such as banking, environment, and customer service etc. However, its implementation in field of the supply risk management is a relatively new area of research. Further, this approach is different from previous studies, as it focuses on actual available data related to supply chain characteristics and performance and the way adopted methodology is implemented produce optimal results for supply risk identification.

In majority of previous supplier risk assessment studies, the norm is to employ an approach that aggregates the impact of identify supply risk for supplier risk assessment. In contrast, in this research an algorithmic approach was followed to obtain a supplier risk score based on the supplier's actual performance and realised supply risk rather than aggregated the impact of identify supply risk. A system is design on the proposed approach that offers an integrated supplier risk assessment platform. It also supports to model and mine the data that is available in both public database and standard supply-base data available within a company for knowledge development about supply risk.

To validate applicability of the proposed approach for real-world problem, a case study is conducted. Two datasets were built using the standard supply-base data of AC manufacturing company and publically available data about supply chain

characteristics and supply performance. First dataset consisted of data about actual purchased order performance and upstream supply chain characteristics (environmental, operational, network and financial). Second dataset consisted of suppliers' annual supply performance and relative upstream supply chain characteristics.

First dataset was used to discover the hidden knowledge about supply risks. Three rule base algorithms, i.e. C4.5, RIPPER, PART were implemented to the data for developing knowledge discover models. To select the optimum model for knowledge discovery, a model selection is proposed and implemented. The optimum model is selected as supply risk identification model that is used to predict the supply risk of new purchase order and also used for knowledge discovery. The hidden knowledge about different supply risk types is presented in form of easy readable statesmen i.e. classification rules. Further analysis is conducted for identify the interrelationships among different risk factors and their significance towards specific type of supply risk identification using cross impact based proposed approach.

Based on the results of risk factors interrelationships and their significance towards specific type of supply risk, input parameters to develop the supplier risk scoring model are selected. Further the numerical type data are converted into categorical type data based on the rules discovered about the supply risks. This approach efficiently integrated the knowledge about supply risks in supplier risk scoring. The data about these selected parameters supplier's annual performance is obtained from second dataset. The logistic regression technique is applied to the estimated weight of the theses parameters toward target variable i.e. supplier risk. Based on the results of logistic regression model, a risk scoring model is developed. This risk scoring model calculates the risk score of existing or new suppliers. Higher score of supplier represent higher risk level. The model based on knowledge discovery approach does not only enhance the accuracy of supplier risk assessment, but can also be served as a core element of supply chain management tool. However it should be noted that risk score are for advisory purpose to assess supplier risk for take final decision by decision maker.

7.2 Revisiting Research Questions

In Chapter 1 the research questions that set the framework for the current research were presented. In this section these questions are revisited in a way how they were addressed throughout the research.

RQ1: What are the different dimensions of supply risk and how it can be used for supplier risk assessment, when measured on actual supply performance?

The literature review in the field of supply chain risk management clearly distinguished the background, borders and settings of the supply risk and demand risk both constitute the supply chain risk. The supply risk is concerned with flow material, finance and information that transpire in upstream supply chain. Supply risk is no differ than general risk in other field, having impact on desire target that can be two-dimensional, upward (positive) or downward (negative). However, most of the previous work considered the supply risk as negative directional action, so as the current research thesis.

Based on the different definition of supply risk, it is understood that it composes of two components: (1) source of supply risk and (2) the outcome of the supply risk. Where these two components depend upon the perception of concerning company, thus in current research thesis these two components are precede according to proposition of the current research thesis about supply risk: *"An uncertainty associated with upstream supply chain characteristics, the results of which affect desired outcome (supply performance) negatively is considered as supply risk".* Based on this proposition uncertainty associated with upstream supply chain characteristics is source of supply risk and negative effect on desired outcome (supply performance) is the outcome of the supply risk.

To clearly demonstrate the role of the supply risk in supplier risk assessment; both its constructs are used. Supplier risk assessment provides the probability of supplier risk (i.e. risk score) based on the supplier performance and factors affect supplier performance. First the negative outcome of the supply risk is use to calculate the supplier performance to define the supplier risk. Second the sources of the supply risk are used as the factors affect supply performance for calculating the probability of supplier risk (risk score). This consideration is logical as the defined supplier risk is based on supply risk outcome.

RQ2: Could the publically available data be combined with company specific data into a dataset to be used for supply risk identification and which algorithm can be most suitable data mining algorithm for available data in supply risk identification?

The study uses the data from both company specific databases and publically available databases for supply risk identification. This data is combined to implement the data mining algorithm for developing a supply risk identification model. The data collected from different data sources is related to supply chain characteristics. Initial experiments are conducted to analyse the impact of combining the data obtained from different data sources on knowledge discovery algorithms' performance.

The obtained results clearly showed that as the combining of the data obtained from different data sources have showed positive impact on knowledge discovery process. For example FONE model contain the data related all the supply chain characteristics used for current case study, is obtained from all the available data sources both company specific and publically available databases. This (FONE) model showed the better evaluation performance metric than almost all the models, those have less data combination. Further the Model "O" that contains only company specific data and model "E" that contain the publically available data are outperformed by the model "OE" that combine the data obtained from both publically available and company specific database on all the evaluation performance metric.

The current thesis adopted the knowledge discover approach to identify the supply risk, the problem formulation is a classification task. As there are different data mining algorithm are available to solve the same problem formulation in knowledge discovery process. Therefore, there is need for justification of selected algorithm. An algorithm selection method is proposed in the current thesis for justifying the algorithm selection in classification task base problem formulation. In the proposed selection method different algorithm are ranked according to their performance on classification performance evaluation metrics. The highest ranking algorithm is selected to develop the supply risk identification model. Several evaluation metrics are utilized rather than any one single metric, which ensure the selection of best available algorithm. However, this did not exclude No Free Lunch Theorem. Based on the results, PART algorithm is standout as most appropriate algorithm for available data sample to achieve the desired goals and objectives of knowledge discovery.

RQ3: What knowledge can be extracted from the available data and which combinations of supply chain characteristics best identify the supply risk?

A number of standard datasets, commonly used within the areas of supply base management and other field of studies have formed the dataset related to supply chain characteristics (risk factors) and supply performance. In accordance with the main requirement of the study that no a-priori assumption is made about the way supply chain characteristics interact with the supply performance. Rule base classification method is used to extract the knowledge from data. This enabled the discovery of the knowledge in form of simple statement (rules) to understand the way supply chain characteristics interact with supply performance.

Further analysis on base of discovered knowledge; provide the in-depth understanding about supply chain characteristics interrelation among them to interact with supply performance. The results highlighted that different combination of supply chain characteristics interact with supply performance in different way. This information supports the identification of supply chain characteristics combinations that best is to identify the supply risk as discussed in section 6.2.

RQ4: Could such a supply risk-aware methodology for supplier risk scoring model development add any value?

Since the interest was also to investigate whether is it possible to base a supplier risk assessment process entirely on knowledge extracted about supply risk, second part of the case study was performed to develop supplier risk scoring model. The supplier risk scoring model used the knowledge about supply risk to form its input data. The supplier risk scoring model building is a classification task. Therefore, the value of supply risk-aware methodology for supplier risk scoring model is measured in-term of classification performance. Supply risk-aware methodology for supplier risk scoring model outperforms the other methodologies those are not supply risk-aware on the classification performance evaluation metrics as discussed in section 6.4.

Further the value addition of implementing standardized risk score that is final product of supply risk-aware methodology for supplier risk scoring is measured in-term profit loss. It clearly demonstrates the implementations of standardized risk score can avoid the losses as discussed in section 6.5.

7.3 Research Contributions

The main contributions of this research are listed below:

- The development of a holistic supply risk aware approach for supplier evaluation by integrating knowledge discovery techniques. The risk scoring method developed in this work resulted into a meta-model to provide supplier risk scoring metric. The methodology developed and implemented in this work is novel and has not been reported elsewhere in the supply chain risk management literature.

- Use of new metrics for supplier risk assessment combining economic freedom and logistics performance indices. Most variables used in this study have been previously employed for either supplier selection or supply chain risk assessment, however their combination has not been reported elsewhere. Therefore, the use of such quantitative metrics as independent variables to measures the risk of an individual supplier along with other supply chain characteristics is one of the key contributions to the supplier risk assessment body of knowledge.

- Real world application: An additional key contribution of this work is the evaluation of the methodology based on real life datasets from the HVAC industry. Although the methodology is not limited to this particular industry, the method has been validated using those real life datasets demonstrating that the proposed method cannot only produce meaningful results that can be acted upon but that they are also easy to interpret.

- A novel approach for modelling the interactions among risk factors and the selection of key variables for classification data model building by integrating cross impact analysis and knowledge discovery approaches.

- Design and development of a discretisation method that discovers all cut-off values in a derived rules set of numerical type variables to convert them into categorical type variables is another contribution to the body of knowledge.

7.4 Managerial Implementation and Recommendation to Case Study Company

- The developed knowledge discovery base supplier risk assessment model can be used in two ways. It can be used as a pattern recognition tool that aims to uncover patterns in data about supply risk. As demonstrated in the first part of the case study, the identification of supply risk depends upon the nature of risk factors output and their combinations. This information can then be used to improve decision making in supply risk mitigation and purchasing or supplier selection.

- The second use of the model is as an automated risk assessment system. Data mining techniques are designed to deal efficiently with large amount of data. Therefore, in case of mass appraisal such as purchasing data of a company, this model can be used as automated system to conduct supplier risk assessment. However, the use of general performance measurement limits the applicability of this method for strategic supplier or equity joint venture suppliers, where the other requirement are more important than only general performance measurements.

- According to current purchasing policy at case study Company, the purchasing department prefer low bidding supplier which can provide the products on financial credit of 90 days after products received. Based on the supplier risk scoring model and optimum cut-off value (see section 6.6) the company has been suggested to select the supplier which has the score range below 391 or 451 along with manger's consideration.

7.5 Limitation

Despite its differentiation from other works and its contribution, there are also some limitations associated with it due to the exploratory nature of this work.

- The whole approach is based on classification that heavily relies on data for the revealing of the possible relation makes this approach vulnerable to data and its size. Therefore, the size of dataset should be sufficient as classification

is prone to errors if the size of the dataset is small. Further, the results are entirely dependent on the level of data availability and the quality of the data. Conductively, dense and uniformly distributed sample such as having the balance data for risky/non risky supplier contribute to better results, especially in case of large dataset.

- Another limitation is that the sample of this study was limited to a single purchasing firm and its suppliers. Thus, the supplier risk assessment model needs to be adjusted to the characteristics of the industry in question. These characteristics include the degree of parts diversity, the complexity of the upstream supply network of the first-tier suppliers and a shortened product life cycle.

- Final limitation of the study is related to the perceptual index of the information about the risk factor. The information about some risk factors such as logistic performance index, economic freedoms are obtained from publically available databases, which are results of different surveys. Although the survey teams try to conduct surveys as objectively as possible, and these results were created through strict auditing, the perceptual index could be subject to measurement errors.

7.6 Future work:

Based on the limitation of this research, some future work opportunities can be as

- The data set used to belong to one Case Study company; therefore, there is an opportunity for validating the Supplier risk assessment framework on data sample related other case study companies.
- Another opportunity of this work can be the implementation of the proposed Supplier risk assessment framework on the big data.
- This study proposed a ranking method for selecting the optimal algorithm for rule base classification; the proposed method can be tested for its usability in other rule base classification problem formulation using the evaluation metric such as F-measure and comprehensibility. As these factors has provided the significant difference between different algorithms.

7.7 Conclusion

This thesis has demonstrated validity of a novel framework for supplier risk assessment that combines two complementary approaches: Knowledge discovery and Risk scoring. The motivation for current research thesis is to develop a methodology focusing on quantitative risk assessment to bridge the research gap in the field of supply risk management, especially within globalized supply chain operating environment. Where, it is very difficult obtain the expert opinion for decision making due to difference in the perceptions, cultures, language barrier and other globalized factors involved. Therefore, the need of a risk modelling framework such as proposed in this study, that is purely drive by the available data in globalized supply chain environment can be very efficient for aiding in decision making. The developed framework has been successfully applied on the procurement (supply-base) data of HVAC manufacturer that has global supply chain network. The supply risk scoring modelling approach provided a prospect to quantify the overall risk associated with supplier that can aid the decision maker to make final decision about procurement in globalized supply chain environment. The application of this novel risk modelling framework in supplier risk assessment will also help decision makers to visualize the holistic view of inter-relationship among risk factors and identify supply risk factors for making the decisions about proactive risk mitigation strategies.

References

1. Adanur, S. and Allen,B., (1995), First results on the effects of ISO 9000 in the US textile industry, Benchmarking for Quality Management & Technology, Vol. 2(3), pp.41-52.

2. Adhitya, A., Srinivasan, R. and Karimi, I.A., (2009), Supply chain risk identification using a HAZOP-based approach, AIChE Journal, Vol. 55(6), pp.1447-1463.

3. Alonso, E., Field, F., Gregory, J., and Kirchain, R., (2007), Materials Availability and the Supply Chain: Risks, Effects, and Responses, from http://hdl.handle.net/1721.1/35728 [Accessed 14th may 2014].

4. Anderson, R. (2007). The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.

5. Andrew, B.P., (1997), The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, Vol. 30(7), pp.1145-1159.

6. Arrow, K.J., (1965), The theory of risk aversion, Aspects of the theory of risk-bearing (Helsinki).

7. Atish P, S., and Huimin, Z., (2008), Incorporating domain knowledge into data mining classifiers: An application in indirect lending, Decision Support Systems Vol. 46(1), pp.287–299.

8. Atwater, C., Gopalan, R., Lancioni, R. and Hunt, J., (2014), Measuring supply chain risk: Predicting motor carriers' ability to withstand disruptive environmental change using conjoint analysis, Transportation Research Part C, Vol. 48, pp.360-378.

9. Badea, A., Prostean, G., Goncalves, G. and Allaoui, H. (2014), Assessing risk factors in collaborative supply chain with the analytic hierarchy process (AHP), Procedia-social and Behavioural Sciences, Vol. 124, pp.114-123.

10. Badurdeen, F., Shuaib, M., Wijekoon, K., Brown, A., Faulkner, W., Amundson, J., Jawahir, I.S. Goldsby, T.J., Iyengar, D. and Boden, B., (2014), Quantitative modelling and analysis of supply chain risks using Bayesian theory, Journal of Manufacturing Technology Management, Vol. 25(5), pp.631-654.

11. Banuls, V.A. and Turoff,M., (2011), Scenario construction via Delphi and cross-impact Analysis, Technological Forecasting and Social Change: An International Journal, Vol. 78, pp.1579-1602.

12. Barman, R.B (2005), Estimation of Default Probability for Basel II on Credit Risk.

13. Behdani, B., Adhitya, A., Lukszo, Z., and Srinivasan, R., (2012), How to Handle Disruptions in Supply Chains – An Integrated Framework and a Review of Literature, Available at SSRN 2012.

14. Berg, E., Knudsen, D. and Norrman, A., (2008), Assessing performance of supply chain risk management programmes: a tentative approach, International Journal of Risk Assessment and Management, Vol. 9(3), pp. 288 - 310.

15. Bharadwaj, B.K. and Pal, S., (2011), Data Mining: A prediction for performance improvement using classification, International Journal of Computer Science and Information Security (IJCSIS), Vol. 9(4), pp.136-140.

16. Bidault, F., Despres, C., and Butler, C., (1998), New product development and early supplier involvement (ESI): the drivers of ESI adoption, International Journal of Technology Management, Vol. 15(1/2), pp.49–69.

17. Blackhurst, J.V., Scheibe, K.P. and Johnson D.J., (2008), Supplier risk assessment and monitoring for the automotive industry, International Journal of Physical Distribution and Logistics Management, Vol.38(2), pp.143-165.

18. Blos, M. F., Quaddus, M., Wee, H. M. and Watanabe, K. (2009), Supply chain risk management (SCRM): a case study on the automotive and electronic industries in Brazil, Supply Chain Management: An International Journal, Vol. 14(4), pp.247-252.

19. Bogataj, D. and Bogataj, M. (2007), Measuring the supply chain risk and vulnerability in frequency space, International Journal of Production Economics, Vol. 108(1-2), pp.291-301.

20. Bolton, C. (2009), variable reduction and analysis in credit scoring, in Logistic regression and its application in credit scoring, Mater thesis, University of Pretoria.

21. Boyes, W.J., D.L. Hoffman and S.A. Low (1989), "An Econometric Analysis of the Bank Credit Scoring Problem", Journal of Econometrics, Vol. 40, pp. 3-14.

22. Brachman, R. J. and Anand, T. (1996), The process of' knowledge discovery in databases. In U.M. Fayyad, C. Piatetskv-Shapiro, P. Simyth, and B. Uthurusamy, editors, Advances in knowledge Discovery and Data Mining, chapter 2, pp. 37-57, AAAI Press/The MIT Press.

23. Braunscheidel, M.J. and Suresh, N.C. (2009), "The Organizational Antecedents of a Firm's Supply Chain Agility for Risk Mitigation and Response", Journal of Operations Management, Vol. 27(2), pp. 119-140.

24. Brazdil, P.B. and Soares, C. (2000), A comparison of ranking methods for classification algorithm selection, In Proceedings of 11th European Conference on Machine Learning, Springer Verlag.

25. Buckland, M. and Gey, F. (1994), The relationship between recall and precision, Journal of the American Society for Information Science, Vol. 45, pp.12-19.

26. Buhman, C., Kekre, S. and Singhal, J. (2005), Interdisciplinary and inter-organizational research: establishing the science of enterprise networks, Production and Operations Management, Vol. 14(4), pp. 493-513.

27. Canbolat, Y.B., Gupta, G., Matera, S. and Chelst, K. (2008), Analysing risk in sourcing design and manufacture of components and sub-systems to emerging markets, International Journal of Production Research, Vol. 46(18), pp.5145-5164.

28. Carter, P L. and Giunipero, L. C. (2010), Supplier Financial and Operational Risk Management, Tempe, AZ: CAPS Research, from http://www.mypurchasingcenter.com/files/8913/9524/5629/Caps-Research-Financial-Risk-paper-Dec-2010.pdf , [Accessed 12th August 2013].

29. Cavinato, L.J. (2004), Supply chain logistics risks: from the back room to the board room, International Journal of Physical Distribution and Logistics Management, Vol. 34(5), pp. 383-387.

30. Cessie, L. S. and Houwelingen, J. C. V. (1992), Ridge estimators in logistic regression, Applied Statistics, Vol. 41(1), pp.191-201.

31. Chapman, P., Bernon, M. and Haggett, P., (2011), Applying selected quality management techniques to diagnose delivery time variability, International Journal of Quality & Reliability Management, Vol. 28(9), pp. 1019-1040.

32. Chawla, N., Hall, L., K.W., B., and Kegelmeyer, W. (2002), SMOTE: Synthetic Minority Oversampling Technique, Journal of Articial Intelligence Research, Vol. 16, pp.321-357.

33. Chen, X., Sim, M., Simchi-Levi, D. and Sun, P., (2007), "Risk aversion in inventory management", Operations Research, Vol. 55 (5), pp. 828-842.

34. Chopra, S. and Meindl, P. (2006), Supply Chain Management, Strategy, Planning, and Operation, (3rd Edition), Prentice Hall, UK.

35. Chopra, S. and Sodhi, M.S. (2004). Managing Risk to Avoid Supply-Chain Breakdown, Sloan Management Review, Vol, 46(1), pp. 53-61.

36. Christopher, M. and Peck, H. (2004), Building the resilient supply chain, International Journal of Logistics Management, Vol. 15(2), pp.1-13.

37. Cohen, W. (1995), Fast effective rule induction, Proceedings of the 12th International Conference on Machine Learning, pp.115-123, Morgan Kaufmann.

38. Consumer Federation of America, (2002) Credit Scores Accuracy and Implications for Consumers, New York: Consumer Federation of America and National Credit Reporting Association.

39. Cooke, J.A., (2002), Brave new world, Logistics Management & Distribution Report, Vol. 41(1), pp.31-34.

40. Cossin, D. and Schellhorn, H., (2007), Credit risk in a network economy, Management Science, Vol. 53(10), pp.1604-1617.

41. Costantino, N. and Pellegrino, R., (2010), Choosing between single and multiple sourcing based on supplier default risk: A real options approach, Journal of Purchasing and Supply Management, Vol. 16(1), pp. 27-40.

42. Craighead, C.W., Blackhurst, J., Rungtusanatham, M.J. and Handfield, R.B., (2007), The severity of supply chain disruptions: Design characteristics and mitigation capabilities, Decision Sciences, Vol. 38(1), pp. 131–156.

43. Crook, J., Edelman D. and Thomas, L., (2007), Recent developments in consumer credit risk assessment, European Journal of Operational Research, Vol. 183(3), pp.1447-1465.

44. Cucchiella, F. and Gastaldi, M., (2006), Risk management in supply chain: a real option approach, Journal of Manufacturing Technology Management, Vol. 17(6), pp. 700-20.

45. Dani, S., (2009), Predicting and Managing Supply Chain Risks, In: Supply Chain Risk: A Handbook of Assessment, Management and Performance, Ed. 1, pp. 53-66, Springer.

46. Dani, S. and Deep, A. (2010). Fragile food supply chains- Reacting to risks, International Journal of Logistics Research and Applications, Vol. 13(5), pp. 395-410.

47. Díaz, V.G., Martinez, L.B., Fernandez, J.F.G. and Marquez, A.C., (2012), Contractual and quality aspects on warranty: Best practices for the warranty management and its maturity assessment, International Journal of Quality & Reliability Management, Vol. 29(3), pp.320-348.

48. Dickson, G.W. (1996), An analysis of vendor selection: systems and decisions, Journal of Purchasing, Vol. 2(1), pp. 5–17.

49. Doumpos, M., Kosmidou, K., Baourakis, G., and Zopounidis, C., (2002), Credit risk assessment using a multi-criteria hierarchical discrimination approach: A comparative analysis, European Journal of Operational Research, Vol. 138(2), pp.392-412.

50. Enyinda, C.I., Mbah, C.H.N. and Ogbuehi, A. (2010), An empirical analysis of risk mitigation in the pharmaceutical industry supply chain: A developing-

country perspective, Thunderbird International Business Review, Vol. 52(1), pp.45-54

51. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (1996a), From data mining to knowledge discovery: An Overview, In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds), 1996, Advances in Knowledge Discovery and Data Mining, AAAI Press, p.1-30.

52. Fayyad, U.M, Piatetsky-Shapiro, G. and Smyth, P., (1996b), from data mining to knowledge discovery in databases, AI Magazine, Vol. 17(3), pp.37-54.

53. Fayyad U.; Djorgovski, S.G.; and Weir, N. (1996), Automating the Analysis and Cataloging of Sky Surveys, in Advances in Knowledge Discovery and Data Mining. Fayyad U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. (Eds.), Cambridge, Mass.: MIT Press/AAAI Press.

54. Fayyad, U.M., Uthurusamy, R., (2002), Evolving data into mining solutions for insights, Communications of the ACM, Vol. 45(8), pp. 28-31.

55. Field, A.M., (2013), Preparing for the worst: developing a more structured, strategic approach to supply chain risk mitigation, from http://www.canadiansailings.ca/?p=6089 [Accessed 14th December 2014].

56. Finke, G.R., Schemitt, A.J., and Singh, M., (2010), Modelling and simulating supply chain schedule risk. In: Proceedings of the Winter Simulation Conference, Zurnich, Switzerland, pp. 3472–3481.

57. Frank, E., and Witten, I. (1998), Generating accurate rule sets without global optimization, Proceedings 15th International Conf. on Machine Learning, pp.144-151, Morgan Kaufmann, San Francisco, CA.

58. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H., (2004), Data mining in bioinformatics using Weka. Bioinformatics, Vol., 20, Pp. 2479–2481.

59. Frawley, W., Piatetsky-Shapiro, G. and Matheus, C., (1992), Knowledge Discovery in Databases - An Overview, in Knowledge Discovery in Databases edition , pp. 1-30, AI Magazine.

60. Freitas, A. A. (1999), on rule interestingness measures, Knowledge-Based Systems, Vol. 12, pp. 309–315.

61. Fried, A. and Linss, V., (2005), Towards an advanced impact analysis of intangible resources in organizations, Papers and Preprints of the Department of Innovation Research and Sustainable Resource Management, TU Chemnitz.

62. Furnkranz, J. and Flach, P. A., (2005), ROC 'n' rule learning: Towards a better understanding of covering algorithms, Machine Learning, Vol. 58 (1), pp. 39–77.

63. Furnkranz, J., and Widmer, G. (1994), Incremental reduced error pruning, Proceedings of the 11th International Conference on Machine Learning, pp.70-77, Morgan Kaufmann.

64. Galindo, J., and Tamayo, P., (2000), Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications, Basic Methodology and Risk Modelling Applications, Vol. 15(1-2), pp.107-143.

65. Ganguly, K.K., and Guin, K.K., (2013), A fuzzy AHP approach for inbound supply risk assessment, Benchmarking: An International Journal, Vol. 20(1), pp.129-146.

66. Gaonkar, R.S. and Viswanadham, N. (2007), Analytical Framework for the Management of Risk in Supply Chains, IEEE Transactions on Automation Science and Engineering, Vol. 4(2), pp.265-273.

67. Gartner Inc., 2011. Predicts 2012: supply chain predictions: talent, risk and analytics dominate. Accessed 2nd March, 2012. <www.gartner.com>.

68. Gaudenzi, B. and Borghesi, A. (2006), Managing risks in the supply chain using the AHP method, International Journal of Logistics Management, Vol. 17(1), pp.114-136.

69. Gereffi, G., Humphrey, J. and Sturgeon, T., (2005), The governance of global value chains, Review of International Political Economy, Vol. 12(1), pp.78-104.

70. Ghadge, A., Dani, S., Chester, M. and Kalawsky, R., (2013), A Systems approach for modelling Supply Chain Risks, Supply Chain Management: An International Journal, Vol.18 (5), pp. 523-538.

71. Goldschmidt, T.O., (2009). The rhetoric of hackers' neutralisations, In F. Schmalleger & M. Pittaro (Eds.), Crimes of the Internet, New Jersey: Pearson Education, Inc.

72. Godet, M., Pearse, J.D. and Lennon, H.K., (1979), the Crisis in Forecasting and the Emergence of the "Prospective" Approach: With Case Studies in Energy and Air Transport. Pergamon Press, New York.

73. Guertler, B. and Spinler, S., (2014), Supply risk interrelationships and the derivation of key supply risk indicators, Technological Forecasting & Social Change.

74. Hair J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L., (2006), Multivariate Data Analysis, 6th edition, Pearson Prentice Hall, New Jersey.

75. Hall, M. A. (1998), Correlation-based feature selection for machine learning, PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

76. Hallikas, J., Karvonen, I., Pulkkinen, U., Virolainen, V.M. and Tuominen, M., (2004), Risk management processes in supplier networks, International Journal of Production Economics, Vol. 90(1), pp.47-58.

77. Harland, C., Brenchley, R., and Walker, H., (2003), Risk in supply networks, Journal of Purchasing & Supply Management, Vol. 9 (2), pp. 51–62.

78. Harrison, G.W. and Rutström, E.E., (2008), Risk aversion in the laboratory, In Cox, J.C., Harrison, G.W. (eds.), Risk Aversion in Experiments, Emerald, Bingley.

79. Harrison, G. W., Moritz, S. and Pibernik, R., (2009), How Does the Risk Attitude of a Purchasing Manager Affect the Selection of Suppliers?, European Business School Research Paper No. 09-10.

80. Hayashi, Y., Smith, R. and Chozick, A., (2007), Quake bring safety issue to fore; plant standard in focus after radioactive leak; Japan's auto output hit, from http://www.wsj.com/articles/SB118466049034668761 , [Accessed 13th December 2013].

81. Hendricks, K. B. and Singhal, V.R., (2005)a, Association between supply chain glitches and operating performance, Management Science, Vol. 51, pp.695-711.

82. Hendricks, K. B. and Singhal, V.R., (2005) b, An empirical analysis of the effect of supply chain disruptions on long-run stock price performance and equity risk of the firm, Production and Operations Management, Vol.14, pp. 35-52.

83. Ho, C.-F., Chi, Y.-P., and Tai, Y.-M. (2005), A structural approach to measuring uncertainty in supply chains, International Journal of Electronic Commerce, Vol. 9, pp. 91-114.

84. Hoffmann, E., (2011), Natural hedging as a risk prophylaxis and supplier financing instrument in automotive supply chains, Supply Chain Management: An International Journal, Vol. 16(2), pp.128-141.

85. Hoffmann, P., Schiele, H. and Krabbendam, K., (2013), Uncertainty, supply risk management and their impact on performance, Journal of Purchasing & Supply Management, Vol.19, pp.199-211.

86. Horsemeat scandal: Tesco reveals 60% content in dish, from http://www.bbc.co.uk/news/uk-21418342 [Accessed 16th July 2014].

87. Hou, J., Zeng, A.Z., and Zhao, L., (2010), Coordination with a backup supplier through buy-back contract under supply disruption, Transportation Research Part E: Logistics and Transportation Review, Vol. 46 (6), pp. 881-895.

88. Huang, G.Q., Lau, J.S.K., Mak, K.L. and Liang, L. (2006), Distributed supply-chain project rescheduling: Part II - Distributed affected operations rescheduling algorithm, International Journal of Production Research, Vol. 44(1), pp. 1-25.

89. Huysmans, J., Baesens, B., Vanthienen, J., and Gestel, T. V., (2006), Failure prediction with self-organizing maps, Expert Systems with Applications, Vol. 30(3), pp.479-487.

90. Ince, H., and Aktan, B., (2009), A Comparison of Data mining Techniques for Credit Scoring in Banking: A managerial Perspective, Journal of Business Economics and Management, Vol. 10(3), pp.233-240.

91. ISO 9000 Introduction and Support Package: Guidance on the Documentation Requirements of ISO 9001:2008, Document: ISO/TC 176/SC 2/N 525R2, from http://www.iso.org/iso/02_guidance_on_the_documentation_requirements_of_iso_9001_2008..pdf [Accessed 6th June 2014].

92. Johnson-Speck, C. (2005), Abstracts of significant cases bearing on the regulation of insurance, Journal of Insurance Regulation, Vol. 23(4), pp.81-84.

93. Jung, K., Lim, Y. and Oh, J. (2011), A Model for Measuring Supplier Risk: Do Operational Capability Indicators Enhance the Prediction Accuracy of Supplier Risk, British Journal of Management, Vol. 22, pp. 609–627.

94. Jüttner, U. (2005), Supply chain risk management: understanding the business requirements from a practitioner perspective, The International Journal of Logistics Management, Vol. 16(1), pp. 120-141.

95. Jüttner, U., Peck, H., and Christopher, M., (2003), Supply chain risk management: outlining an agenda for future research, International Journal of Logistics, Vol. 6 (4), pp. 197–210.

96. Kantardzic, M. (2011), Data mining Applications, in Data Mining: Concepts, Models, Methods, and Algorithms, 2nd edition, pp.496-509.

97. Keramati, A. and Yousefi, N., (2011), A proposed classification of data mining techniques in credit scoring, Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia, January 22-24.

98. Kern, D., Moser, R., Hartmann, E., and Moder, M. (2012), Supply risk management: model development and empirical analysis, International Journal of Physical Distribution & Logistics Management, Vol. 42(1), pp. 60-82.

99. Khan, O. and Burnes, B., (2007), Risk and supply chain management: creating a research agenda, The International Journal of Logistics Management, Vol.18(2), pp.197-216.

100. Kim, D.Y., Kumar, V. and Kumar, U., (2012), Relationship between quality management practices and innovation, Journal of Operations Management, Vol. 30(4), pp.295-315.

101. King, W., Marks, P. and McCoy, S., (2002), the most important issues in knowledge management, Communications of the ACM, Vol. 45(9), pp. 93-97.

102. Kleindorfer, P. and Saad, G. H., (2005), Managing disruption risks in supply chains, Production and Operations Management, Vol. 14(1), pp.53-68.

103. Knemeyer, A.M., Zinn, W. and Eroglu, C., (2009), Proactive planning for catastrophic events in supply chains, Journal of Operations Management, Vol. 27(2), pp.141-153.

104. Kolari, P. and Joshi, A., (2004), Web mining: research and practice, IEEE Computing in Science and Engineering, Vol. 6(4), pp.49-53.

105. Kotsiantis, S. and Kanellopoulos, D., (2006), Discretization techniques: a recent survey, International Trans Computer Science Engineering, Vol. 32(1), pp.47-58.

106. Kouvelis, P., Chambers, C. and Wang, H., (2006), Supply chain management research and production and operations management: Review, trends, and opportunities, Production and Operations Management, Vol. 15(3), pp.449-469.

107. Kraljic, P. (1983), Purchasing must become supply management, Harvard Business Review, Vol. 5, pp. 109-117.

108. Kull, T. and Closs, D. (2008), The risk of second-tier supplier failures in serial supply chains: Implications for order policies and distributor autonomy, European Journal of Operational Research, Vol. 186(3), pp.1158-1174.

109. Kull, T. J. and Talluri, S. (2008), A supply risk reduction model using integrated multi-criteria decision making, IEEE Transactions on Engineering Management, Vol. 55(3), pp.409-419.

110. Kumar, S.K., Tiwari, M.K. and Babiceanu, R. F. (2010), Minimisation of supply chain cost with embedded risk using computational intelligence approaches, International Journal of Production Research, Vol. 48(13), pp. 3717–3739.

111. Kumar, S., Himes, K.J. and Kritzer, C.P., (2014), Risk assessment and operational approaches to managing risk in global supply chains, Journal of Manufacturing Technology Management, Vol. 25(6), pp.873-890.

112. Lambert, D. M. and Cooper, M. C., (2000), Issues in Supply Chain Management, Industrial Marketing Management, Vol. 29, pp. 65–83.

113. Lee, A.S., Gutierrez-Arcelus, M., Perry, G.H., Vallender, E.J., Johnson, W.E., Miller, G.M., Korbel, J.O. and Lee, C., (2008), Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies, Human Molecular Genetics: Oxford journal, Vol. 17, pp.1127-1136.

114. Levary, R. R. (2007), Ranking foreign suppliers based on supply risk, Supply Chain Management: An International Journal, Vol. 12(6), pp.392-394.

115. Levary, R.R. (2008), using the analytic hierarchy process to rank foreign suppliers based on supply risks, Computers and Industrial Engineering, Vol. 55(2), pp.535-542.

116. Li, F. C., (2009), Comparison of the Primitive Classifiers without Features Selection in Credit Scoring, International Conference on Management and Service Science, City: IEEE.

117. Lindner, G. and Studer, R., (1999), AST: Support for algorithm selection with a CBR approach, In Proceeding of the 3rd Practice of Knowledge Discovery in Databases (PKDD) Conference, pp.418-423.

118. Linss, V. and Fried, A., (2009), Advanced Impact Analysis: the ADVIAN method –an enhanced approach for the analysis of impact strengths with the consideration of indirect relations, Papers and Preprints of the Department of Innovation Research and Sustainable Resource Management (BWL IX), Chemnitz University of Technology.

119. Linss, V. and Fried, A., (2010), The ADVIAN classification- a new classification approach for the rating of impact factors, Technological Forecasting and Social Change,Vol.77 (1), Pp.110–119.

120. Lockamy III, A., (2014), Assessing disaster risks in supply chains, Industrial Management & Data Systems, Vol. 114(5), pp.755-777.

121. Lockamy III, A. and McCormack, K., (2012), Modelling supplier risks using Bayesian networks, Industrial Management & Data Systems, Vol. 112(2), pp. 13-33.

122. Lunney, S., (2011), How to Negotiate with Sole Sources, at my purchasing centre, from http://www.mypurchasingcenter.com/purchasing/blogs/single-and-sole-source-suppliers/ [Accessed 19 July 2014]

123. MacGillivray, A., Begley, P. and Zadek, S., (2007), The State of Responsible Competitiveness, Accountability, London.

124. Maimon, O. and Rokach, L., (2005), Introduction to Knowledge Discovery and Data Mining in Handbook of Data Mining and Knowledge Discovery in Databases, ed. (2), pp. 1-18, Springer.

125. Malone, D.W. (1975), An introduction to the application of interpretive structural modelling, Proceedings of the IEEE, Vol. 63(3), pp. 397-404.

126. Mangla, S.K., Kumar, P. and Baru, M.K., (2014), Monte Carlo Simulation Based Approach to Manage Risks in Operational Networks in Green Supply Chain, Procedia Engineering, Vol. 97, pp.2186-2194.

127. Manuj, I. and Mentzer, J.T. (2008), Global supply chain risk management strategies, International Journal of Physical Distribution and Logistics Management, Vol. 38(3), pp.192-223.

128. March, J.G. and Shapira, Z. (1987), Managerial perspectives on risk and risk taking, Management Science, Vol. 33(11), pp. 1404–1418.

129. Markmann, C., Darkow, I. and Gracht, H., (2013), A Delphi-based risk analysis Identifying and assessing future challenges for supply chain security in a multi-stakeholder environment, Technological Forecasting & Social Change, Vol. 80, pp.1815-1833.

130. Martha, J., and Subbakrishna, S. (2002), Targeting a Just-In-Case Supply Chain for the Inevitable Next Disaster, Supply Chain Management Review, Vol. 6(5), pp.18-23.

131. Matheus, C., Chan, P. and Piatetsky-Shapiro, G., (1993.), Systems for Knowledge Discovery in Databases IEEE Transactions on Data and Knowledge Engineering, Vol. 5(6), pp.914-925.

132. Matook, S., Lasch, R. and Tamaschke, R. (2009), Supplier development with benchmarking as part of a comprehensive supplier risk management framework, International Journal of Operations and Production Management, Vol. 29 (3), pp. 241-67.

133. McKechnie J.L. (1983), Webster's New Twentieth Century Dictionary of the English Language, Unabridged, Prentice Hall Press, New York.

134. Mester, L., (1997), What's the point of credit scoring?, Federal Reserve Bank of Philadelphia Business Review, September/October, pp.3-16.

135. Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994), Machine Learning, Neural and Statistical Classification, Ellis Harwood Limited.

136. Miller, H.J. and Han, J., (2001), Geographic data mining and knowledge discovery-An Overview, In: Miller, H.J., Han, Geographic Data Mining and Knowledge Discovery, pp.3-32, Taylor and Francis.

137. Miller, M. (2003), Research confirms value of credit scoring (Another Perspectives), in National Underwriter Property & Casualty-Risk & Benefits Management.

138. Mitchell V.W. (1995), Organisational risk perception and reduction: A literature review, British Journal of Management, Vol. 6(2), pp. 115–133.

139. Mohanty M. K. and Gahan P., (2012), Buyer Supplier Relationship in Manufacturing Industry - Findings from Indian Manufacturing Sector, Business Intelligence Journal, Vol. 5(2), pp.319-333.

140. Mohtadi, H. and Murshid, A. P. (2009), Risk analysis of chemical, biological, or radionuclear threats: Implications for food security, Risk Analysis, Vol. 29(9), pp.1317-1335.

141. Munoz, A. and Clements, M.D. (2008), Disruptions in information flow: A revenue costing supply chain dilemma, Journal of Theoretical and Applied Electronic Commerce Research, Vol. 3(1), pp.30-40.

142. Narasimhan, R. and Talluri, S., (2009), Perspectives on risk management in supply chains, Journal of Operations Management, Vol. 27 (2), pp.114-118.

143. Neiger, D., Rotaru, K. and Churilov, L., (2009), Supply chain risk identification with value-focused process engineering, Journal of Operations Management, Vol. 27, pp.154-168.

144. Neter, J., (1966), Financial Ratios as Predictors of Failure: Discussant, Journal of Accounting Research Supplement, Vol. 4, pp.112-118.

145. Norrman, A. and Jansson, U., (2004), Ericsson's proactive supply chain risk management approach after a serious sub-supplier accident, International Journal of Physical Distribution & Logistics Management, vol. 34 (5), pp. 434–456.

146. Norrman A. and Lindroth R. (2002), Supply chain risk management: purchasers' vs. planners view on sharing capacity investment risks in the telecom industry Proceedings of the IPSERA 11th International Conference Enschede Holland 25–27 March, pp. 577–595.

147. Odette, P., (2013), Solving the Sole-Source Quandary – Mitigate the Risks, at Global Supply Chain Solutions (GSCS), from http://costflexrisk.com/2013/02/26/solving-the-sole-source-quandry-mitigate-the-risks/ [Accessed 16 July 2014].

148. Oehmen, J. Ziegenbein, Alard, A. and R. Schonsleben, P. (2009), System-oriented supply chain risks management, Production Planning and Control, Vol. 20(4), pp. 343-361.

149. Oke, A. and Gopalakrishnan, M., (2009), Managing disruptions in supply chains: a case study of a retail supply chain, International Journal of Production Economics, Vol. 118 (1), pp. 168–174

150. Olson, D.L. and Wu, D.D. (2010), A review of enterprise risk management in supply chain, Kybernetes, Vol. 39(5), pp.694-706.

151. O'Marah, K. (2009), Supply Chain Risk, 2008–2009: As Bad as It Gets, available at www.amrresearch.com.

152. Page, S. L., and Hawley, R. S., (2004), the genetics and molecular biology of the synaptonemal complex, Annual Review of Cell and Developmental Biology, Vol. 20, pp.525-558.

153. Paleologo, G., Elisseeff, A., and Antonini, G., (2010), Subagging for credit scoring models, Journal of Operational Research, Vol. 201(2), pp.490-499.

154. Pechenizkiy, M., Tsymbal, A. and Puuronen, S., (2005), Knowledge Management Challenges in Knowledge Discovery Systems, Proceedings of the 16th International Workshop on Database and Expert Systems Applications.

155. Peck, H. (2005), Drivers of supply chain vulnerability: An integrated framework, International Journal of Physical Distribution and Logistics Management, Vol. 35(4), pp. 210-232.

156. Pereira, N.S., (2005), Why Sole-Supplier Vaccine Markets May Be Here to Stay, Health Affairs, Vol. 24(3), pp.694-696.

157. Perrine, D., (2007), What is a Scoring Model?. [Online] Available at: http://www.scoringmodels.com/ [Accessed on 17th June 2011].

158. Pfohl, H.C., Gallus, P. and Thomas, D., (2011), Interpretive structural modeling of supply chain risks, International Journal of Physical Distribution & Logistics Management, Vol. 41(9), pp.839-859.

159. Piatetsky-Shapiro, G., (2007), Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics", Data Mining and Knowledge Discovery, Vol. 15, pp.99-105.

160. Piatetsky-Shapiro, G., (1991), Discovery, analysis and presentation of strong rules, in: G. Piatetsky-Shapiro, W.J. Frawley (Eds.), Knowledge Discovery in Databases, AAAI, pp. 229.

161. Piatetsky-Shapiro, G., (1991), Knowledge discovery in real databases: A report on the UCAI-89 Workshop, AI Magazine, Vol. 11(5), pp.68-70.

162. Piramuthu, S., (2004), Evaluating feature selection methods for learning in data mining applications, European Journal of Operation Research, Vol.156, pp.483-494.

163. Prakash, S. (1995) Mortgage lenders see credit scoring as key to hacking through red tape, American Banker (August), pp.1.

164. Primo, M.A.M. and Amundson, S.D., (2002), An exploratory study of the effects of supplier relationships on new product development outcomes, Journal of Operations Management, Vol. 20(1), pp.33-52.

165. Pujawan, I. N. and Geraldin, L. H. (2009), House of risk: a model for proactive supply chain risk management, Business Process Management Journal, Vol. 15(6), pp.953-967.

166. Punniyamoorthy, M., Thamaraiselvan, N. and Manikandan, L., (2013), Assessment of supply chain risk: scale development and validation, Benchmarking: An International Journal, Vol. 20(1), pp.79-105.

167. Quayle, M., (2002), E-Commerce: the challenge for the UK SME's in the twenty-first century, International Journal of Operation and Production Management, Vol. 22, pp.1148–1161.

168. Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos.

169. Radjou, N. (2002), Adapting to Supply Network Change, Forrester Research Tech Strategy Report, Cambridge, MA.

170. Rao, S. and Goldsby, T. (2009), Supply chain risks: a review and typology, The International Journal of Logistics Management, Vol. 20(1), pp. 97-123.

171. Ravindran, A.R., Bilsel, R.U., Wadhwa, V. and Yang, T. (2010), Risk adjusted multicriteria supplier selection models with applications, International Journal of Production Research, Vol. 48(2), pp.405-424.

172. Riddalls, C., Bennett, S. and Tipi, N., (2000), Modelling the dynamics of supply chains, International Journal of Systems Science, Vol. 31 (8), pp.969-976.

173. Ritchie, B. and Brindley, C., (2007), Supply chain risk management and performance: a guiding framework for future development, International Journal of Operations and Production Management, Vol. 27 (3), pp. 303–322.

174. Rokach, L. and Maimon, O. (2002), Top-Down Induction of Decision Trees Classifiers- A Survey, IEEE transactions on systems, man and cybernetics: part c, Vol. 1(11), pp.1-12.

175. Roth, A.V., Tsay, A.A., Pullman, M.E. and Gray, J.V. (2008), Unraveling the food supply chain: Strategic insights from China and the recalls, Journal of Supply Chain Management, Vol. 44(1), pp.22-39.

176. Rumbaugh, J., Jacobson, I. and Booch, G., (2005), The Unified Modelling Language Reference Manual, Second Edition, Pearson Education.

177. Samson, S., Reneke, J.A. and Wiecek, M.M., (2009), A review of different perspectives on uncertainty and risk and an alternative modelling paradigm, Reliability Engineering & System Safety, Vol 94(2), pp. 558–567.

178. Sanchez-Rodrigues, V., Potter, A. and Naim, M.M., (2010), The impact of logistics uncertainty on sustainable transport operations, International Journal of Physical Distribution and Logistics Management, Vol. 40(1), pp.61-83.

179. Sandoval, J., (2009), RESTful Java Web Services, Packet Publishing.

180. Satchidanand, S. S., and Jay, B. S., (2007), Empirical evaluation of sampling and algorithm selection for predictive modelling for default risk, ISIS.

181. Sawik, T., (2011), Selection of supply portfolio under disruption risks, Omega, Vol. 39, pp.194-208.

182. Schoenherr, T., Rao Tummala, V.M. and Harrison, T.P. (2008), Assessing supply chain risks with the analytic hierarchy process: Providing decision support for the offshoring decision by a US manufacturing company, Journal of Purchasing and Supply Management, Vol. 14(2), pp.100-111.

183. Schreiner, M., (2004), Benefits and Pitfalls of Statistical Credit Scoring for Microfinance, Centre for Social Development, Washington University in St. Louis, USA.

184. Sheffi, Y., (2001), Supply Chain Management under the Threat of International Terrorism, The International Journal of Logistics Management, Vol. 12 (2), pp.1-11

185. Sheffi, Y., (2005), The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage. Cambridge, MA, MIT Press.

186. Sheffi, Y., and Rice, J. (2005), A supply chain view of the resilient enterprise, MIT Sloan Management Review, Vol. 47(1), pp. 41–48.

187. Sinha, P. R., Whitman, L. E. and Malzahn, D. (2004), Methodology to mitigate supplier risk in an aerospace supply chain, Supply Chain Management: An International Journal, Vol. 9(2), pp.154-168.

188. Singhal, A. and Jajodia, S., (2005), Data Mining for Intrusion Detection, in in Handbook of Data Mining and Knowledge Discovery in Databases, ed. (2), pp. 1171-1180, Springer.

189. Siddiqi, N., (2006), Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring, New Jersey: John Wiley & Sons.

190. Smillie L, and Blissett A., (2010), A model for developing risk communication strategy, Journal of Risk Research, Vol. 13(1), pp.115-134.

191. Sitkin, S.B. and Pablo, A.L., (1992), Re-conceptualising the determinants of risk behaviour, Academy of Management Review, Vol. 17 (1), pp. 9–38.

192. Sodhi, M.S. and Lee, S. (2007), An analysis of sources of risk in the consumer electronics industry, Journal of the Operational Research Society, Vol. 58(11), pp.1430-1439.

193. Stecke, K.E. and Kumar, S. (2009), Sources of supply chain disruptions, factors that breed vulnerability, and mitigating strategies, Journal of Marketing Channels, Vol. 16(3), pp. 193-226.

194. Supply-Chain-Digest (2006), the 11 Greatest Supply Chain Disasters, from http://www2.isye.gatech.edu/~jjb/wh/tidbits/top-sc-disasters.pdf dated (25-11-2012).

195. Svensson, G. (2000), A conceptual framework for the analysis of vulnerability in supply chains, International Journal of Physical Distribution and Logistics Management, Vol. 30(9), pp.731–749.

196. Tang, C. (2006), Perspectives in supply chain risk management, International Journal of Production Economics, Vol. 132 (2), pp. 451–488.

197. Tang, C. and Musa, S. N., (2011), Identifying risk issues and research advancements in supply chain risk management, International Journal of Production Economics, Vol. 33, pp.25-34.

198. Tang, C. and Tomlin, B. (2008), The power of flexibility for mitigating supply chain risks, International Journal of Production Economics, Vol. 116(1), pp. 12-27.

199. Tazelaar, F. and Snijders, C. (2013), Operational risk assessments by supply chain professionals: Process and performance, Journal of Operations Management, Vol. 31, pp. 37–51.

200. Thun, J., Drüke, M. and Hoenig, D., (2011), Managing Uncertainty: an empirical analysis of supply chain risk management in small and medium-sized enterprises, International Journal of Production Research Vol. 49(18), pp. 5511-5525.

201. Thun, J. and Hoenig, D. (2009), An Empirical Analysis of Supply Chain Risk Management in the German Automotive Industry, International Journal of Production Economics, Vol. 131(1), pp. 242-249.

202. Thun, J., and Hoenig, D., (2011), An empirical analysis of supply chain risk management in the German automotive industry, International Journal of Production Economics, Vol.131, pp.242–249.

203. Todorovski, L. and Dzeroski, S., (1999), Experiments in Meta-level Learning with ILP. In Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases.

204. Tomlin, B. (2006), On the value of mitigation and contingency strategies for managing supply-chain disruption risks, Management Science, Vol. 52(5), pp.639-657.

205. Trkman, P. and McCormack, K. (2009), Supply chain risk in turbulent environments a conceptual model for managing supply chain network risk, International Journal of production Economics, Vol. 119(2), pp. 247-258.

206. Tse, Y.K. and Tan, K.H., (2012), Managing product quality risk and visibility in multi-layer supply chain, International Journal of Production Economics, Vol. 139, pp.49-57.

207. Tuncel, G. and Alpan, G. (2010), Risk assessment and management for supply chain networks- A case study, Computers in Industry, Vol. 61(3), pp.250-259.

208. Twala, B., (2010), Multiple classifier application to credit risk assessment, Expert Systems with Applications, Vol. 37(4), pp.3326-3336.

209. Verbeke, W., Martens, D. , Mues, C., and Baesens, B., (2011), Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications, Vol. 38, pp. 2354–2364.

210. Vlajic, J. V., Vorst, Jack G. A. J. V. and Haijema, R., (2012), A frame work for designing robust food supply chains, International Journal of Production Economics, Vol. 137, pp. 176-189.

211. Wacker, J.G., (2004), A theory of formal conceptual definitions: developing theory building measurement instruments, Journal of Operation Management, Vol. 22 (6), pp.6290-650.

212. Wagner, S.M. and Bode, C. (2006), An empirical investigation into supply chain vulnerability, Journal of Purchasing and Supply Management, Vol. 12(6), pp.301-312.

213. Wagner, S.M. and Bode, C. (2008), An empirical examination of supply chain performance along several dimensions of risk. Journal of Business Logistics, Vol. 29 (1), pp.307–32.

214. Wagner, S. and Bode, C. (2009), Dominant Risks and Risk Management Practices in Supply, International Series in Operations Research & Management Science, Vol. 124, pp.271-290.

215. Wang, K., Zhou, S. and Han, J. (2002), Profit mining: From patterns to actions, In Proceedings of the 8th Conference on Extending Database Technology (EDBT), Prague, Czech Republic, pp.70–87.

216. Waters, D., (2007), Supply chain risk management, Vulnerability and resilience in logistics, Kogan Page, London.

217. Watts, C. A., Kim, K. Y., and Hahn, C. K., (1992), "Linking purchasing to corporate competitive strategy," international journal purchasing material management, Vol. 28, pp. 2–8.

218. Weele, A. J.V. and Rozemeijer, F. A., (1996), Revolution in purchasing, Building competitive power through proactive purchasing, European Journal of Purchasing & Supply Management, Vol. 2 (4), pp. 153-160.

219. Wei, H., Dong, M. and Sun, S. (2010), Inoperability input-output modeling (IIM) of disruptions to supply chain networks, Systems Engineering, Vol. 13(4), pp.324-339.

220. Wiendahl, H.-P., Selaouti, A. and Nickel, R. (2008), Proactive supply chain management in the forging industry, Production Engineering, Vol. 2(4), pp.425-430.

221. Wilson, M.C. (2007), The impact of transportation disruptions on supply chain performance, Transportation Research Part E: Logistics and Transportation Review, Vol. 43(4), pp.295-320.

222. Witten, I. H. and Frank, E., (2005), Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, San Francisco: Morgan Kaufmann.

223. Witten I. Frank E. and Hall M., (2011), Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann Publishers.

224. Witten, I. H., Frank, E. and Hall, M.A., (2011), Implementations: Real Machine Learning Schemes, in Data Mining Practical Machine Learning Tools and Techniques, ed. (3), pp. 191-303, Morgan Kaufmann Publishers.

225. Wu, T., Blackhurst, J. and Chidambaram, V. (2006). A model for inbound supply risk analysis, Computers in Industry, vol. 57(4), pp. 350-365.

226. Wu, T., Blackhurst, J. and O'Grady, P. (2007), Methodology for supply chain disruption analysis, International Journal of Production Research, Vol. 45(7), pp.1665-1682.

227. Wu, D., and Olson, D.L., (2008), Supply chain risk, simulation and vendor selection, International Journal of Production Economics, Vol. 114 (2), pp. 646–655.

228. Xiao, W., Zhao, Q., and Fei, Q., (2006), a comparative study of data mining methods in consumer loans credit scoring management, journal of systems science and systems engineering, Vol. 15(4), pp.419-435.

229. Yang, B. and Yang, Y. (2010), Postponement in supply chain risk management: A complexity perspective, International Journal of Production Research, Vol. 48(7), pp.1901-1912.

230. Yang, J., Tiyyagura, A., Chen, F. and Honavar, V., (1999), Feature subset selection for rule induction using RIPPER, Proceedings of the Genetic and Evolutionary Computation Conference, volume 2, p. 1800, Orlando, Florida, USA: Morgan Kaufmann

231. Zimmermann, H. J., (2000), An application-oriented view of modelling uncertainty, European Journal of Operational Research, Vol. 122(2), Pp. 190–198.

232. Zsidisin, G. (2003)a, A grounded definition of supply risk, Journal of Purchasing and Supply Management, Vol. 9(5-6), pp. 217–224.

233. Zsidisin, G. A. (2003)b, Managerial Perceptions of Supply Risk, Journal of Supply Chain Management, Vol. 39(1), pp. 14-25.

234. Zsidisin, G.A. and Ellram, L.M., (2003), An Agency Theory Investigation of Supply Risk Management, Journal of Supply Chain Management, Vol.39 (2), Pp. 15-27.

235. Zsidisin, G., Ellram, L., Carter, J. and Cavinato, J., (2004), An analysis of supply risk assessment techniques, International Journal of Physical Distribution and Logistics Management, Vol. 34 (5), pp.397–413.

236. Zsidisin, G. A., Panelli, A. and Upton, R., (2000), "Purchasing organization involvement in risk assessments", Supply Chain Management: An International Journal, 5(4), pp. 187-197.

237. Zsidisin, G. A. and Smith, M. E. (2005), Managing supply risk with early supplier involvement: a case study and research propositions, The Journal of Supply Chain Management, Vol. 41(4), pp.44-57.

Appendix I

Key risk factors

The proposed methodology for supplier risk assessment system requires a list of risk factors (supply chain characteristics), which can be possible source of risk. The knowledge from literature review, personal experience, case-study companies past experience and standard industrial risk factors were used to facilitate the risk factor list development process. The initial list of risk factors was very vast. Brain storming sessions were conduct with the mangers and purchasing persons at case-study companies to reduce the number of factors at suitable size. Only those factors were included for which the data is available. However it is taken into consideration that during the factors attenuation process, it does not loss any crucial information source. The proposed methodology is data driven, so the data availability is key factors for the success implementation of proposed methodology. All the indentified risk factors are evaluated on their effect on multiple supply risk outcome (quality, delivery, price). These risk factors were divided into four main categories: financial risk factors, operational risk factors, network risk factors and environmental risk factors.

Financial Risk Factors:

The goal of the financial risk factors is not only to capture the financial standing of supplier but also other financial condition necessary to draw the conclusion about supplier ability to fulfil the customer requirement and contracted performance. Supplier financial reports as well as other financial factors like exchange rate will be considered. In addition the supplier may be aggressive in the providing the proposed price, so it is also necessary to include the price competitiveness of supplier with respect to other suppliers in bidding process. There can be different financial measurement which can be used to analyse the supplier are given table.

FACTORS	Description	Unit
T1=WORKING CAPITAL/TOTAL ASSETS	Measure the liquidity available to supplier	Ratio
T2=RETAINED EARNINGS/ TOTAL ASSETS	Measure the development level of supplier	ratio
T3= EARNINGS BEFORE TAX	Measure the operational efficiency of supplier	ratio

AND INTEREST/ TOTAL ASSETS		
T4= TOTAL EQUITY/TOTAL LIABILITIES	Measure the solvency position of supplier	ratio
T5=SALES/ TOTAL ASSETS	Measure the assets turnover	ratio
PRICE COMPARISON	The supplier price offer standing with compare to other supplier's offer	
COMMODITY PRICE	The upward variation in commodity price index during purchase cycle time	percentage
EXCHANGE RATE INDEX	Measure the impact of exchange rate	
GROSS PROFIT MARGIN	This parameter will used to measure the company (supplier) ability to produce the profit to overcome its cost expenses. This parameter can be calculated by **(Gross Profit/Revenue)*100**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
NET PROFIT MARGINE	This parameter measures the ability of the company (supplier) to generate the profit after all the expenditures. This parameter can be calculated by the **(Net Profit/Revenue)*100.** The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
RETURN ON ASSETS (ROA)	This parameter measure the ability of company to utilize its assets to generate the profit. This can be calculated **(EBIT/Total Assets)*100**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
RETURN ON EQUITY(ROE)	This parameter measure the ability of company to return the investors money. This can be calculated by **(EBIT /Owner's Equity employed)*100.** The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)

RETURN ON CAPITAL EMPOLYED (ROCE)	This parameter will assess the companies ability to generate the profit efficiency on total investment employed (Equity +Debt). This can be calculated **(EBIT/total Investment)*100**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
DEBT RATIO	This parameter gives the overlook on financial structure of supplier or debt position. This can be calculated by **(Total liabilities (Debt)/Total Assets)*100**.The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
DEBT TO EQUITY RATIO	This parameter gives the companies financial structure by showing the debt contribution in total investment. This can be calculated by **(Total Debt/Total Equity)*100.**The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
CURRENT RATIO	This parameter will measure the risk or ability of company to continue its operation financially. This can be calculated by **(Current-Assets/Current-Liabilities)*100.**The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
ASSETS TURNOVER	This parameter gives the insight into assets utilization by company. This can be calculated as **(Sales-Revenue/Fixed Assets)*100**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
INVENTORY TURNOVER	This parameters measure the effectiveness of company inventory utilization. This can be calculated as **(cost of good sold/Total-Inventory)*100.**The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	% ($ or £)
DEBTOR AGE	This parameter measure the numbers of days for payment receive from customer. This can be calculated by **(receivable/credit sales)*365**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	Days

CRIDETOR AGE	This parameters measure the number of days to payment made by company. This can be calculated by **(Payable/credit purchase)*365**. The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	Days
STOCK HOLDING PERIOD	This gives the insight stock holding period of company. This can be calculated by **(365/inventory turnover).**The weight-age of parameter can be assigned according to the average value of parameter in current economy and with the help of financial expert.	Days

Operation risk factors:

The supplier Operational capabilities can also contribute toward supply risk and risk profile of supplier. The main operations capabilities characteristics identified as risk factors deal with the supplier's upstream supply chain, quality standards at the supplier and supplier's manufacturing capabilities. The list of operation risk factors measurement and their description with units are given table.

FACTORS	**Description**	Unit
ISO CERTIFICATION	The information about the supplier has ISO certification.	Binary (yes /No)
QUALITY AWARD	The information about supplier has national or international quality award or certification	Binary (yes /No)
WARRANTY	Does supplier offer the warranty cost or services for the defective parts/products	Binary (yes /No)
QUALITY PLANNING DOCUMENT AND RECORD	Does supplier have the complete quality planning documentation and record of quality control	Binary (yes /No)
QUALITY IMPROVEMENT	Does supplier has implemented any quality improvement philosophy like TQM, Six Sigma, Lean or JIT etc.	Binary (yes /No)
QUALITY INSPECTION	The method of quality inspection is implemented by supplier for production quality control	1:Statistical process control SPC

		2: batch inspection 3:in-work
TECHNICAL CAPABILITIES	The ability of supplier to provide the design and technical support	score from 1to 5 (high to low)
MANUFACTURING YIELD	The capability of supplier to produce the defect free product to total production (percentage)	percent
PRODUCTION FACILITY	The condition and technology of supplier production facility	score from 1to 5 (high to low)
CYCLE TIME	time between order placed and received	week
CAPACITY UTILIZATION	The percentage of supplier total production capacity is being utilized during the period.	percent
STRIKE HOUR RATIO	This parameters measure the hours lost due to staff strike. This can be calculated **(Hours lost due to strike/Total working-hours)*100.** The parameter weight-age can be assigned by considering the average value in HR index for industry type and location with help of HR Expert/specialist.	%(HOURS)
ACCIDENT HOURS RATIO	This parameter measure the hours of work lost due to accidents or staff injuries, which reflect company working environment. This can be calculated by **(Hours lost due to accident/injuries/Total working-hours)*100**. The parameter weight-age can be assigned by considering the average value in HR index for industry type and location with help of HR Expert/specialist.	%(Hours)
EMPLOYEE TURNOVER	This parameter measure the employee output and provide the work for utilization information. This can be calculated as **(Net-sales/Total Employee)*100**.The parameter weight-age can be assigned by considering the average value in HR index for industry type and location with help of HR Expert/specialist.	%($,£)
LABOUR COST	This parameter measure labour cost of company. This can be calculated as **(Labour Cost/Total cost of goods)*100**. The parameter weight-age can be assigned by considering the average value in HR index for industry type and location with help of HR Expert/specialist.	%($,£)
JOB SATISFACTION	This parameter reflects the employee's interest in their job. This calculated by company surveys. The parameter weight-age can be assigned by considering the average value in HR	%

INDEX	index for industry type and location with help of HR Expert/specialist.	
TRAINING HOURS RATIO	This parameter measure training given per employee. This parameter can be calculated as **(Total training Hours/Total staff)*100**. The parameter weight-age can be assigned by considering the average value in HR index for industry type and location with help of HR Expert/specialist.	% (Hr)
ADVANCE SHIPMENT NOTIFICATION (ASN) RATIO	This parameter measure the information sharing rate for the delivery. This can be calculated as **(no. of ASN/total deliveries)*100**. The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	% (Number of deliveries)
SHIPMENT RATIO	This parameter measure the on-time shipment ability of the company for a order. This can be calculated by formula **(No. of shipments made with-in lead time /Total no. of deliveries)*100**. The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	%(Number of deliveries)
DELIVERY ADHERENCE	This parameter measure the on-time deliver ratio in full order as committed/requested. This can be calculated as (**on-time (full) deliveries/Total no. of deliveries)*100**. The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	%(Number of deliveries)
FILL- RATE	This parameter will observe the no. of deliveries sent by seller in full order quantity of items. This can be calculated as **(no. of full-order deliveries/total no. of deliveries)*100.** The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	%(Number of deliveries)
BACKORDER RATIO	This parameter measure the performance level of back-order per delivery. This can be calculated as **(total items received/total items ordered)*100.** The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	%(no. of items/products)
DAMAGE/LOST RATIO	These parameter measures the no. of damage/LOST parts per delivery in logistic process. This can be calculated as **(no. of products damaged/total no. of deliveries)*100**. The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	%(Number of deliveries) %(no. of items/products)

AVERAGE LEAD TIME	This parameter will be measure the variation of company average-lead time from the market for specific product. This can be calculated by formula (1+ ((average time-lead time)/Average-Time))*100. The weight-age can be assigned according to required average value in specific industry with help of purchasing expert.	No. of Hours
QUALITY CHECK RATIO	This parameter measure the company commitment to quality assurance. This parameter can be calculated as **(No. of quality checks/total target value)*100**. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%(percentage of number)
PRODUCT SAFETY	This parameter measure the ability of company providing safe product to market. This can be calculated by two mean i.e. no of un-safe product incidents or (loss due to un-safe product/revenue generation). The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	Number of incidents %($^£)
COST OF QUALITY	This parameter measure the supplier commitment to quality improvement. This can be calculated as **(cost of quality/revenue)*100**. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%($^£)
DOCUMENTATION	This parameter reflects the company's documentation evidence of performance. This value can be obtained from the manager. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%
DOCUMENTATION UP-GRADATION	This parameter reflects the documentation up-gradation of supplier to keep the record clear and on-time. The value can be obtained from the manager. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	Days
DEFECT RATIO(MATERIAL)	This parameter will measure material compliance of supplier-I or Supplier-II. This can be calculated as (defect material (unit/value)/total material)*100. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%(unit^$^£)
DEFECT PER MILLION OPPORTUNITY (DPMO)	This parameter will measure the production quality of supplier. This is calculation normally use in SIX Sigma can be calculated as (no. of defects /total no. of opportunities)*1Million. The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%
	This will measure the product percentage with defect to total production. This can be calculated as (no. of defect units/total no. of defects)*100. The weight-age of this parameter	%(units)

DEFECT RATIO (PRODUCT)	can be assigned by advice of purchasing/ quality expert.	
REJECTION RATIO	This will measure the no. of rejected parts inspected by buyer-company and also probability of defected quality. This can be calculated as **(no. of rejected parts/total no. of parts)*100. AND (no. of rejected parts/total no. of deliveries)*100.** The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%
RE-WORKED RATIO	This parameter will measure the amount of re-worked done by buyer company or re-worked done by supplier to make order correct. This can be calculated as **(no. of re-work parts/total no. of parts)*100. AND (no. of re-work parts/total no. of deliveries)*100.** The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%
EQUIPMENT QUALITY	This will measure the equipment ability. This can be calculated as **(Good Units/total units)*100.** The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%
SCRAP COST	This parameter measures the supplier warranty cost expenditure due to unreliability of service/product. This can be calculated as **(Scrap cost/COGS)*100.** The weight-age of this parameter can be assigned by advice of purchasing/ quality expert.	%(£^$^units)
SCHEDULE ADHERENCE	This parameter will measure the variance in planned production schedule and ability of supplier facility to meet the delivery time. This can be calculated as **(actual production/schedule production)*100**.	%(unit, hours)
EQUIPMENT PERFORMANCE	This parameter will be used to assess the ability of equipment to produce good quantity and quality of equipments at supplier's facility. This can be calculated as (**ideal cycle time*good no. of units/operating cycle time)*100**. The weight-age can be assigned by average value of tolerant in specific industry after the production expert advice.	%(units per time)
WIP RATIO	This parameter will measure the WIP inventory. This can be calculated as **(average no. of units in WIP/total Production)*100.**The weight-age can be assigned by average value of tolerant in specific industry after the production/inventory expert advice.	%
	This parameter will give information about the available inventory risk at supplier end according to demand. This can be calculated as **(no. of units in stock/total production)*100**.	%

STOCK RATIO	The weight-age can be assigned by average value of tolerant in specific industry after the production/inventory expert advice.	
BUFFER RATIO	This parameter will measure the capacity to meet the demand fluctuation. This can be calculated as (**buffer stock/total stock)*100**. The weight-age can be assigned by average value of tolerant in specific industry after the procurement/inventory expert advice.	%(units)
INVENTORY ACCURACY	This parameter will measure misrepresentation of data at specific period of time. This can be calculated as (**stock in book/actual stock)*100**. The weight-age can be assigned by average value of tolerant in specific industry after the procurement/inventory expert advice.	% (units^$,£)
BUFFER-USAGE RATIO	This parameter will measure the usage of buffer stock during specific high demand period. This can be calculated as **(buffer stock used/total buffer stock)*100**. The weight-age can be assigned by average value of tolerant in specific industry after the production/inventory expert advice.	%
STOCK-OUT	This parameter will give insight in risk of out-stock due to un-availability by supplier. This can be calculated for specific period of time as **(1+(demand-stock)/stock)*100**.The weight-age can be assigned by average value of tolerant in specific industry after the procurement/inventory expert advice.	%
DEMAND RATIO	This parameter will measure the ability to meet the demand at specific period of time. This can be calculated as **(demand/(inventory + production))*100**. The weight-age can be assigned by average value of tolerant in specific industry after the procurement/inventory expert advice.	%(units)
EQUIPMENT AVAILABILITY	This parameter will measure the flexibility of facility to produce or start new product production by supplier. This can be calculated as **(operating time/planned operating time)*100.** The weight-age can be assigned by average value of tolerant in specific industry after the production expert advice.	%(hours)
INNOVATION INDEX	This parametr will measure the ability of supplier to meet the innovation requirement. This can be measured as **(new/ changed products introduced/total no. of products)*100.**The weight-age can be assigned by average value of tolerant in specific industry after the production/design expert advice.	%
AVAILABLE	This parameter will measure the flexibility of facility. This can be calculated as **(un-utilized capacity/total capacity)*100.**The weight-age can be assigned by average value of tolerant in	%

CAPACITY RATIO	specific industry after the production expert advice.	
PRODUCTION FLEXIBILITY	This parameter measure the ability of supplier to meet the change in demand. This can be calculated by No. of Days require to meet the demand.	No. of days
ECO CYCLE TIME	This measure the required time to make the changes in the design or blueprint released by engineering	No. of days

Network risk factors:

Network risk factors represent buyer's purchasing policy, purchasing market and purchasing network characteristics. For example does the company have sole supplier purchasing policy or dual, as it can increase the dependency on the supplier and can provide the opportunistic behaviour for supplier.

FACTORS	Description	Unit
AVAILABILITY	The availability of part or material in the global market	1:High 2:Med 3:Low
RELATIONSHIP	The number of year Buyer is relationship with the supplier	years
SUPPLIER LOCK	The dependency of buyer on the supplier	1:Sole supplier 2:Dual supplier 3:multiple
INFORMATION SHARING	The level of information sharing between the supplier and Buyer for purchased order	1:Good 2:Average 3:Low
RESPONSE	The time taken to response the request of purchase or quotation	days
ORDER TYPE	Order placed to new supplier or old supplier	1:new order 2:repetive purchase
SHIPMENT ROUTE	The type of shipment route from supplier to buyer	1: Air 2:Surface 3:Other

Environmental Risk factors

The previous risks factors are related to supplier and purchasing company purchasing network policy. Environmental risk factors are directly address concerns outside the scope of the companies involved. These factors include geo-political and social risk factors, which hope to address whether political and social issues in the supplier's country could affect the contract between the two companies.

FACTORS	Description	Unit
NATURAL DISASTERS	The warning (probability) of natural disaster at supplier location during each purchase order time	Low to high (white green yellow red)
MANMADE DISASTER	The impact of manmade disasters during purchase order time like terrorist attack ,war and crime situation	Low to high
POLITICAL ENVIRONMENT	The situation of political stability in country or location of supplier. Political effectiveness are measured by calculating the Regime/Governance Stability score	Low to high 4 point scale Green , yellow, orange and red
INFRASTRUCTURE	The assessment of general infrastructure (e.g., transport, telephony, and energy) in suppliers country	1 = extremely underdeveloped; 7 = extensive
ECONOMIC FREEDOM	The economic freedom level of supplier's which reflect the effectiveness of economic regulation	score from 0 to 100 (Low to high)
CUSTOM REGULATION	The level of custom regulation and other rules for international import and export	score from 1to 5 (Low to high)
LOGISTICS PERFORMANCE INDEX	The International LPI provides qualitative evaluations of supplier country in six areas (Customs, Infrastructure, International shipments, Logistics competence, Tracking & tracing, Timeliness)	score from 1to 5 (Low to high)

Appendix II:

The proposed methodology in current research thesis is a data driven approach that depends upon the availability of the data, data can be obtained from both company specific data bases and publically available data-sources. The rational for inclusion the publically available data is: the global supply chain operates in global environment that include different countries and their demographic and other factors. These demographic and other factors can disrupt the normal flow of material, information and money within a supply chain. The following table identified the possible data-sources for given factors.

Company specific data-sources

Data Source	Risk Factors
Supplier evaluation reports • This data source can integrated with procurement database • The factors selection depends upon the choice of buyer company supplier selection criteria	T1=working capital/total assets, T2=retained earnings/ total assets, T3= earnings before tax and interest/ total assets, T4= total equity/total liabilities, T5=sales/ total assets, **Z-score**, Gross profit margin, Net profit margin, Return on assets, Return on equity, Return on capital employed, Debt ratio, Debt to equity ratio, Current ratio, Assets turnover, Inventory turnover, Debtor age, Creditor age, Stock holding period, **ISO certification, Quality award, Warranty, Quality planning documents and records, Quality improvement, Quality inspection, Technical capabilities, Manufacturing yield, Production facility, Capacity utilization,** Strike hour ratio, Accident hours ratio, Employee turnover, Labour cost, Job satisfaction index, Training hours ratio, Quality check ratio, Documentation up-gradation, Defect per million opportunity, Scrap cost, schedule adherence, Equipment performance, WIP ratio, Stock ratio, Buffer ratio, Inventory accuracy, Buffer-usage ratio, demand ratio, equipment availability, innovation index, Available capacity ratio, Production flexibility, ECO cycle time
Supplier's Annual-Report (Financial and Non-financial)	T1=working capital/total assets, T2=retained earnings/ total assets, T3= earnings before tax and interest/ total assets, T4= total equity/total liabilities, T5=sales/ total assets, Gross profit margin, Net profit margin, Return on assets, Return on equity, Return on capital employed, Debt ratio, Debt to equity ratio, Current ratio, Assets turnover, Inventory turnover, Debtor age, Creditor age, Stock holding period, Labour cost, Strike hour ratio, Accident hours ratio, Employee turnover, Labour cost, ISO certification, Quality award, Accident hours ratio, Training hours ratio, Cost of quality, Scrap cost, innovation index, Available capacity ratio, Production flexibility, ECO cycle time
Supplier's Production and Quality control Reports	ISO certification, Quality award, Warranty, Quality planning documents and records, Quality improvement, Quality inspection, Technical capabilities, Manufacturing yield, Production facility, Capacity utilization, Quality check ratio, Documentation up-gradation, Defect ratio (material and product) , Defect per million opportunity,

	Rejection ratio, WIP ratio, Buffer ratio, Inventory accuracy, Buffer-usage ratio, Stock-out, demand ratio, equipment availability, Available capacity ratio, Production flexibility, ECO cycle time
Supplier's HR document/reports	Strike hour ratio, Accident hours ratio, Employee turnover, Job satisfaction index, Training hours ratio,
Buyer's procurement database Including (*Quality control reports, Order delivery reports and Procurement audits)*	**Price comparison, Cycle time,** Advance shipment notification ratio, Shipment ratio, **Delivery adherence, Fill-rate,** Backorder ratio, Damage/lost ratio, Defect ratio (material and product), **Rejection ratio**, re-worked ratio, **Availability, Relationship, Supplier lock, Information sharing**, Response, Order type, Shipment route

Highlighted risk factors in above table are used in current case study

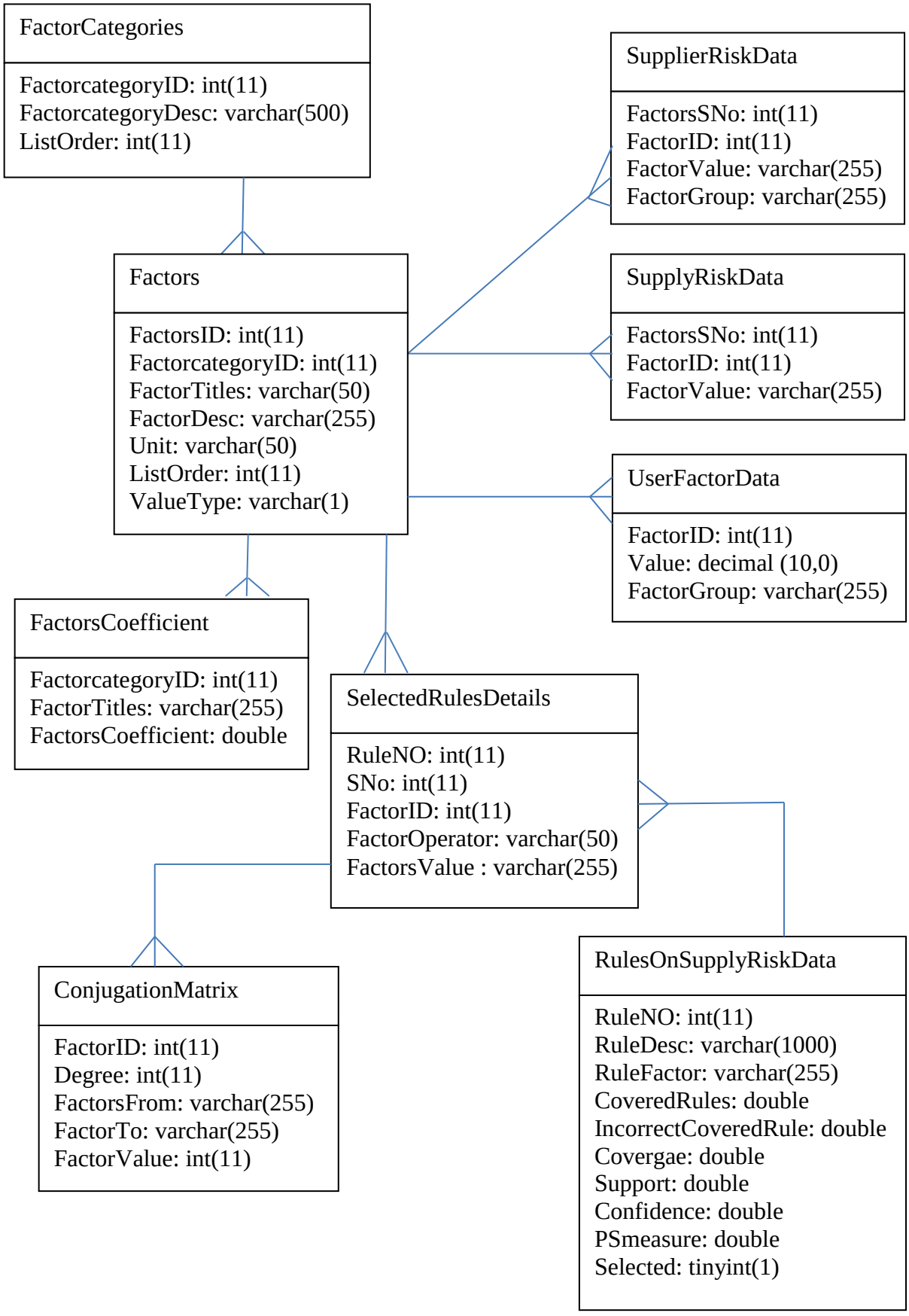
Publically available data sources

Data Source	Risk Factors
• Global Disaster Alert and Coordination System http://www.gdacs.org • Centre for Research on the Epidemiology of Disasters http://www.emdat.be • Dartmouth Flood Observatory http://floodobservatory.colorado.edu/Archives/index.html • World Meteorological Organization (WMO) http://www.wmo.int/pages/index_en.htmlN	Natural Disaster
• Global Terrorism Database (GTD) http://www.start.umd.edu/gtd/ • Centre for Research on the Epidemiology of Disasters http://www.emdat.be • LABORSTA (national series on strikes and lockouts) http://laborsta.ilo.org/applv8/data/c9e.html • GRIPWEB's Data & Informational Portal	Man-made Disaster

http://www.gripweb.org/gripweb/?q=data-information	
• World bank http://lpisurvey.worldbank.org/	Infrastructure, Custom regulation, logistics performance index
• World Economic forum http://www.weforum.org/issues/competitiveness-0/gci2012-data-platform/	Infrastructure, Custom regulation
• Heritage Foundation http://www.heritage.org/index/ranking	Economic freedom Index
• IMF Primary Commodity Monthly Reports data-base http://www.imf.org/external/ns/cs.aspx?id=313 • World Bank commodity monitoring database http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDECPROSPECTS/0,,contentMDK:21574907~menuPK:7859231~pagePK:64165401~piPK:64165026~theSitePK:476883,00.html	Commodity price
• Real effective exchange rate index from World Bank database http://databank.worldbank.org/data/views/reports/tableview.aspx • Exchange rate index from state bank of supplier's country such as for UK 's supplier Bank of England http://www.bankofengland.co.uk/boeapps/iadb/Index.asp?first=yes&SectionRequired=I&HideNums=-1&ExtraInfo=true&Travel=NIx	Exchange rate index
• Global Observatory http://theglobalobservatory.org/2012/09/indices/	Conflict, Fragility, and political instability, Environment, Gender, Freedoms and Rights, Governance, Socio-Economics
• United Nations Commodity Trade Statistics Database • http://comtrade.un.org/db/default.aspx • Third party Product specific database such semiconductor Fab-database	Availability

Appendix III:

The logical design of database for supplier risk assessment system



Appendix IV:

Results of model developed without using knowledge discovery approach

```
Algorithm:    Logistic Regression with ridged setting at 1.0E-8
Dataset:     supplier data training dataset
Total observations:    684
Total variables:   26
              Z-Score
              Commodity price
               Price comparison
              Exchange rate
              ISO certification
              Quality Award
              Warranty
              Quality record
              Quality improvement
              Quality inspection
              Technical capabilities
              Manufacturing Yield
              Production Facility
              Cycle Time
              Capacity utilization
              Availability
              Relationship
              Supplier lock
              Information Sharing
              Natural Disasters
              Manmade Disaster
              Political Stability
              Infrastructure
              Economic Freedom
              Logistics Performance Index
              CLASS
Test method:    10-fold cross-validation

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         509               74.4152
%
Incorrectly Classified Instances       175               25.5848
%
Kappa statistic                          0.4655
Mean absolute error                      0.3551
Root mean squared error                  0.4334
Relative absolute error                 72.768 %
Root relative squared error             87.7389 %
Coverage of cases (0.95 level)          98.9766 %
Mean rel. region size (0.95 level)      97.3684 %
Total Number of Instances              684
```

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
0.626     0.170     0.730       0.626    0.674       0.770     0
0.830     0.374     0.752       0.830    0.789       0.770     1
0.744     0.287     0.743       0.744    0.741       0.770  Weighted
Avg.

=== Confusion Matrix ===

   a   b   <-- classified as
 181 108 |   a = 0
  67 328 |   b = 1
```

=== Re-evaluation on test set ===

```
User supplied test set
Relation:     supplier data without KD testing
Instances:     unknown (yet). Reading incrementally
Attributes:   26

=== Summary ===

Correctly Classified Instances         101               74.2647
%
Incorrectly Classified Instances        35               25.7353
%
Kappa statistic                          0.4584
Mean absolute error                      0.3422
Root mean squared error                  0.4366
Coverage of cases (0.95 level)          99.2647 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
0.586     0.141     0.756       0.586    0.660       0.757     0
0.859     0.414     0.736       0.859    0.793       0.757     1
0.743     0.297     0.744       0.743    0.736       0.757
Weighted Avg.

=== Confusion Matrix ===

  a  b   <-- classified as
 34 24 |  a = 0
 11 67 |  b = 1
```

Appendix V:

Results of model developed without knowledge discovery and using discretiztion approach

```
Algorithm:    Logistic Regression with ridged setting at 1.0E-8
Dataset:     supplier data training dataset
Total observations:    684
Total variables:   26
              Z-Score
              Commodity price
               Price comparison
              Exchange rate
              ISO certification
              Quality Award
              Warranty
              Quality record
              Quality improvement
              Quality inspection
              Technical capabilities
              Manufacturing Yield
              Production Facility
              Cycle Time
              Capacity utilization
              Availability
              Relationship
              Supplier lock
              Information Sharing
              Natural Disasters
              Manmade Disaster
              Political Stability
              Infrastructure
              Economic Freedom
              Logistics Performance Index
              CLASS
Test method:    10-fold cross-validation

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         500               73.0994
%
Incorrectly Classified Instances       184               26.9006
%
Kappa statistic                          0.4388
Mean absolute error                      0.3522
Root mean squared error                  0.4395
Relative absolute error                 72.1621 %
Root relative squared error             88.9809 %
Coverage of cases (0.95 level)          99.1228 %
Mean rel. region size (0.95 level)      96.345  %
Total Number of Instances              684
```

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
0.616     0.185     0.709       0.616    0.659       0.762      0
0.815     0.384     0.744       0.815    0.778       0.762      1
0.731     0.300     0.729       0.731    0.728       0.762
Weighted Avg.

=== Confusion Matrix ===

   a   b   <-- classified as
 178 111 |   a = 0
  73 322 |   b = 1
```

=== Re-evaluation on test set ===

```
User supplied test set
Relation:     supplier data without KD testing
Instances:     unknown (yet). Reading incrementally
Attributes:   26

=== Summary ===

Correctly Classified Instances          98               72.0588
%
Incorrectly Classified Instances        38               27.9412
%
Kappa statistic                          0.4263
Mean absolute error                      0.37
Root mean squared error                  0.4459
Coverage of cases (0.95 level)          98.5294 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
0.655     0.231     0.679       0.655    0.667       0.746      0
0.769     0.345     0.750       0.769    0.759       0.746      1
0.721     0.296     0.720       0.721    0.720       0.746
Weighted Avg.

=== Confusion Matrix ===

  a  b   <-- classified as
 38 20 |  a = 0
 18 60 |  b = 1
```

Appendix VI:

Table shows the result of equal width binning

Variable	Group	Value (Range)	Variable	Group	Value (Range)
Z-Score	1	$RF_1 < 2.7$	Cycle time	1	$RF_{14} < 4$
	2	$4.6 \leq RF_1 \geq 2.7$		2	$8 \leq RF_{14} \geq 4$
	3	$RF_1 > 4.6$		3	$RF_{14} > 8$
Commodity Price	1	$RF_2 < -14.9$	Capacity utilization	1	$RF_{15} < 65.9$
	2	$5.1 \leq RF_2 \geq -14.9$		2	$76.8 \leq RF_{15} \geq 65.9$
	3	$RF_2 > 5.1$		3	$RF_{15} > 76.8$
Exchange rate	1	$RF_4 < -2.6$	Relationship	1	$RF_{17} < 2.7$
	2	$2.6 \leq RF_4 \geq -2.6$		2	$5.3 \leq RF_{17} \geq 2.7$
	3	$RF_4 > 2.6$		3	$RF_{17} > 5.3$
Technical capabilities	1	$RF_{11} < 2.3$	Infrastructure	1	$RF_{23} < 4.2$
	2	$3.7 \leq RF_{11} \geq 2.3$		2	$5.6 \leq RF_{23} \geq 4.2$
	3	$RF_{11} > 3.7$		3	$RF_{23} > 5.6$
Manufacturing Yield	1	$RF_{12} < 62.31$	Economic Freedom	1	$RF_{24} < 61.7$
	2	$47.6 \leq RF_{12} \geq 62.3$		2	$73.3 \leq RF_{24} \geq 61.7$
	3	$RF_{12} > 74.6$		3	$RF_{24} > 73.3$
Production Facility	1	$RF_{13} < 2.3$	Logistics Performance Index	1	$RF_{25} < 3.3$
	2	$3.7 \leq RF_{13} \geq 2.3$		2	$4.2 \leq RF_{25} \geq 3.3$
	3	$RF_{13} > 3.7$		3	$RF_{25} > 4.2$

RF_1= Z-Score, RF_2= Commodity price, RF_4= Exchange rate, RF_{11}= Technical capabilities, RF_{12}= Manufacturing Yield, RF_{13}= Production Facility, RF_{14}= Cycle Time, RF_{15}= Capacity utilization, RF_{17}= Relationship, RF_{23}= Infrastructure, RF_{24}= Economic Freedom Index, RF_{25}= Logistics Performance Index

Appendix VII:

Results of model developed using variable selection approach

```
Algorithm:   Logistic Regression with ridged setting at 1.0E-8
Dataset:     supplier data training dataset
Total observations:    684
Total variables:       11
              Z-Score
              ISO certification
              Quality Award
              Quality record
              Quality improvement
              Availability
              Relationship
              Natural Disasters
              Political Stability
              Logistics Performance Index
              CLASS
Test mode:    10-fold cross-validation

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         509               74.4152
%
Incorrectly Classified Instances       175               25.5848
%
Kappa statistic                          0.4615
Mean absolute error                      0.3527
Root mean squared error                  0.4238
Relative absolute error                 72.2607 %
Root relative squared error             85.7959 %
Coverage of cases (0.95 level)          99.5614 %
Mean rel. region size (0.95 level)      98.1725 %
Total Number of Instances              684

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  ROC Area  Class
0.599    0.149    0.746      0.599    0.664      0.787     0
0.851    0.401    0.743      0.851    0.793      0.787     1
0.744    0.295    0.744      0.744    0.739      0.787
Weighted Avg.

=== Confusion Matrix ===

   a   b   <-- classified as
 173 116 |   a = 0
  59 336 |   b = 1
```

=== Re-evaluation on test set ===

```
User supplied test set
Relation:     supplier data without KD testing
Instances:     unknown (yet). Reading incrementally
Attributes:   11

=== Summary ===

Correctly Classified Instances         105               77.2059
%
Incorrectly Classified Instances        31               22.7941
%
Kappa statistic                          0.5203
Mean absolute error                      0.3567
Root mean squared error                  0.436
Coverage of cases (0.95 level)          99.2647 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
0.621     0.115     0.800       0.621    0.699       0.754      0
0.885     0.379     0.758       0.885    0.817       0.754      1
0.772     0.267     0.776       0.772    0.766       0.754
Weighted Avg.

=== Confusion Matrix ===

  a  b   <-- classified as
 36 22 |  a = 0
  9 69 |  b = 1
```

Appendix VIII:

Results of model developed using variable selection and discretization approach

```
Algorithm:   Logistic Regression with ridged setting at 1.0E-8
Dataset:     supplier data training dataset
Total observations:    684
Total variables:       11
              ISO certification
              Quality Award
              Quality record
              Quality improvement
              Manufacturing Yield
              Availability
              Relationship
              Natural Disasters
              Political Stability
              Logistics Performance Index
              CLASS
Test mode:    10-fold cross-validation

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         511               74.7076
%
Incorrectly Classified Instances       173               25.2924
%
Kappa statistic                          0.4686
Mean absolute error                      0.3444
Root mean squared error                  0.4217
Relative absolute error                 70.569  %
Root relative squared error             85.3685 %
Coverage of cases (0.95 level)          99.7076 %
Mean rel. region size (0.95 level)      97.6608 %
Total Number of Instances              684

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  ROC Area  Class
0.609    0.152    0.746      0.609    0.670      0.792     0
0.848    0.391    0.748      0.848    0.795      0.792     1
0.747    0.290    0.747      0.747    0.742      0.792
Weighted Avg.

=== Confusion Matrix ===

   a   b   <-- classified as
 176 113 |   a = 0
  60 335 |   b = 1
```

=== Re-evaluation on test set ===

```
User supplied test set
Relation:     supplier data without KD testing
Instances:     unknown (yet). Reading incrementally
Attributes:   11

=== Summary ===

Correctly Classified Instances         101               74.2647
%
Incorrectly Classified Instances        35               25.7353
%
Kappa statistic                          0.4632
Mean absolute error                      0.3697
Root mean squared error                  0.4399
Coverage of cases (0.95 level)          99.2647 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
0.621     0.167     0.735       0.621    0.673       0.744      0
0.833     0.379     0.747       0.833    0.788       0.744      1
0.743     0.289     0.742       0.743    0.739       0.744
Weighted Avg.

=== Confusion Matrix ===

  a  b   <-- classified as
 36 22 |  a = 0
 13 65 |  b = 1
```

Appendix IX:

Statistics of the available numerical type training data

Variable	Minimum	Maximum	Mean	Std. Dev.
Z-Score	0.77	6.46	2.59	1.30
Commodity price	-34.83	24.99	-1.59	12.79
Exchange rate	-7.9	7.92	0.51	3.87
Technical Capabilities	1.01	4.99	3.01	1.04
Manufacturing Yield	50	86.92	70.54	8.73
Production Facility	1	4.99	3.04	1.00
Cycle Time	0.01	11.96	3.89	2.98
Capacity utilization	55.01	87.74	72.82	6.98
Relationship	0	8	3.12	2.03
Infrastructure	2.8	7	4.63	1.18
Economic Freedom	50	85	66.19	8.76
Logistics Performance Index	2.5	5	3.52	0.57

Statistics of the available categorical type training data

Variable	Number of categories	Value of each category and count			
		1	2	3	4
ISO certification	Binary	No (316)	Yes (368)		
Quality Award	Binary	No (279)	Yes (405)		
Warranty	Binary	No (311)	Yes (373)		
Quality record	Binary	No (190)	Yes (494)		
Quality improvement	Binary	No (264)	Yes (420)		
Quality inspection	3	SPC (188)	BI (324)	Jud. (178)	
Availability	3	High(161)	Med(428)	Low (95)	
Supplier lock	3	Sole (168)	Dual (197)	Multi.(319)	
Information Sharing	3	High (226)	Average (326)	Low (132)	
Price comparison	3	Lower (167)	Average (330)	High (187)	
Natural Disasters	4	Green (127)	Yellow (216)	Orange (244)	Red(97)
Manmade Disaster	4	Green (127)	Yellow (237)	Orange (215)	Red (105)
Political Stability	4	Green (110)	Yellow (242)	Orange (215)	Red (117)

Appendix X:

Statistics of the available numerical type testing data

Variable	Minimum	Maximum	Mean	Std. Dev.
Z-Score	0.77	6.46	2.59	1.30
Commodity price	-34.83	24.99	-1.59	12.79
Exchange rate	-7.9	7.92	0.51	3.87
Technical Capabilities	1.01	4.99	3.01	1.04
Manufacturing Yield	50	86.92	70.54	8.73
Production Facility	1	4.99	3.04	1.00
Cycle Time	0.01	11.96	3.89	2.98
Capacity utilization	55.01	87.74	72.82	6.98
Relationship	0	8	3.12	2.03
Infrastructure	2.8	7	4.63	1.18
Economic Freedom	50	85	66.19	8.76
Logistics Performance Index	2.5	5	3.52	0.57

Statistics of the available categorical type testing data

Variable	Number of categories	Value of each category and count			
		1	2	3	4
ISO certification	Binary	No (66)	Yes (70)		
Quality Award	Binary	No (58)	Yes (78)		
Warranty	Binary	No (72)	Yes (64)		
Quality record	Binary	No (46)	Yes (90)		
Quality improvement	Binary	No (46)	Yes (90)		
Quality inspection	3	SPC (50)	BI (41)	Jud. (45)	
Availability	3	High(32)	Med(81)	Low (23)	
Supplier lock	3	Sole (30)	Dual (48)	Multi.(58)	
Information Sharing	3	High (68)	Average (42)	Low (26)	
Price comparison	3	Lower (38)	Average (66)	High (32)	
Natural Disasters	4	Green (33)	Yellow (42)	Orange (43)	Red(18)
Manmade Disaster	4	Green (28)	Yellow (49)	Orange (34)	Red (25)
Political Stability	4	Green (18)	Yellow (73)	Orange (31)	Red (14)