12-11-2016

Market Basket Analysis in the Financial Sector – A Customer Centric Approach

Gunjan Mansingh
The University of the West Indies, gunjan.mansingh@uwimona.edu.jm

Kweku-Muata Osei-Bryson
Virginia Commonwealth University, KMOsei@vcu.edu

Lila Rao
The University of the West Indies, lila.rao@uwimona.edu.jm

Maurice McNaughton
The University of the West Indies, maurice.mcnaughton@uwimona.edu.jm

Market Basket Analysis in the Financial Sector – A Customer Centric Approach

Completed Research Paper

Gunjan Mansingh
Department of Computing
The University of West Indies
Kingston, Jamaica
gunjan.mansingh@uwimona.edu.jm

Kweku-Muata Osei-Bryson
School of Business
Virginia Commonwealth University
Richmond, U.S.A.
kmosei@vcu.edu

Lila Rao
Mona School of Business and Management
The University of the West Indies
Kingston, Jamaica
lila.rao@uwimona.edu.jm

Maurice McNaughton
Mona School of Business and Management
The University of the West Indies
Kingston, Jamaica
maurice.mcnaughton@uwimona.edu.jm

Abstract

Organizations often struggle with their efforts to implement data mining projects successfully. This is often due to the fact that they are influenced by success stories of others that glamorize the outcome of successful initiatives, while understating the persistent rigour and diligence required. Although process models exist for the knowledge discovery process their focus is often on outlining the activities that must be done and not on describing how they should be done. While there is some research in addressing how to carry out the various tasks in the phases, the data preparation phase is thought to be the most challenging and is often described as an art rather than a science. In this study we apply a multi-phased integrated knowledge discovery and data mining process model (IKDDM) to a data set from the financial sector and a present a new approach to data preparation for Sequential Patterns (SP) that facilitated the identification of customer focused patterns rather than products focussed patterns in the modelling phase.

Keywords: Data Mining, KDDM, Financial Sector, Market Basket Analysis, Sequence Rules.

Introduction

Increasingly more organizations are recognizing the need for data-driven decision making to harness knowledge that is hidden in their large data assets. The strategic benefits that data mining can bring to organizations are numerous and its value increasingly apparent to business and technology leaders. However, these organizations have recognized that these benefits will only be realized if there is a clear understanding of the objectives of the data mining process and of the data itself. As data mining becomes an integral decision making tool it is important that Knowledge Discovery and Data Mining (KDDM) standards are established for this process (Kurgan and Musilek 2006; Okazaki 2006). Some standards for various aspects of the data mining process are emerging (e.g. models, settings and processes) and using these standards will help to ensure that the process of data mining is reliable and repeatable.
Data mining process models, such as CRISP-DM provide a comprehensive detailed description on what needs to be done in the various stages of the data mining process (Clifton and Thuraisingham 2001; Grossman et al. 2002) and provides a structure for organizing the data mining effort by describing the tasks involved in data mining. The phases include understanding the business problem, capturing and understanding data, applying data mining techniques, interpreting results, and deploying the knowledge gained in operations. The tasks and activities that have to be carried out are often presented in a checklist manner. The focus of these process models is on "what" needs to be done and not the "how". These process models have a fragmented approach and lack an integrated view (Sharma et al. 2012). To address these issues IKDDM (see figure 1) presented an integrated knowledge discovery process model which set out to identify the intra and inter task dependencies between the different phases in the process model. The phases identified by IKDDM which are similar to the ones in CRISP-DM are; Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment. Sharma et al. (2012) focused on explicating the task dependencies in the business understanding and the evaluation phases. IKDDM process model identifies the various dependencies between and across the different phases and provides a framework to organize the various tasks and activities that have to be carried out in the multiple phases. In this study we focus on the first five phases and present the derived results to the business analysts in the organization for further application / deployment considerations.

Financial institutions have been at the forefront of analytics adoption and numerous applications of data mining exist in a sector that accumulates an enormous amount of customer transactional, demographic and relational data. To understand existing and prospective customers, various analytical techniques have been applied to customer data in financial sector. The objective of these techniques is to find the "knowledge nuggets" in form of patterns, correlations and associations that are hidden in the data. This study provides an empirical case study of the implementation of the IKDDM process model in an organizational context. It emphasises the importance of each phase of IKDDM with a particular focus on data preparation which relies heavily on the knowledge of the domain and the ability of the data mining analyst to identify the appropriate data assets. We propose and demonstrate a method of preparing data for sequential patterns by including customer demographics with the products acquired by the customer across different offerings and business units (e.g. insurance type, credit cards, investment schemes). Analysis on such data ensures that the focus in more on the customers and their preferences rather than on products. We further contend that as more cases of this kind is published, that demonstrates a variety of data preparation techniques that enhance the data mining process, the more effective will be the science of data preparation.

Background

Data mining has been defined as the process of extracting valid, non-trivial, previously unknown, interesting patterns or associations from large databases (Agrawal et al. 1993; Piatetsky-Shapiro and Matheus 1994). It is routinely being used in marketing, retail, banking, telecommunications, supply chain optimization, and fraud detection (Jourdan et al. 2008; Moin and Ahmed 2012; Rygielski et al. 2002). It is part of a larger multi-phase process, knowledge discovery and data mining (KDDM) that aims, at a minimum, to semi-automatically extract new knowledge from existing datasets. KDDM process has been described in various ways (Berry and Linoff 1997; Fayyad et al. 1996; Han and Kamber 2006; Sharma and Osei-Bryson 2010; Shearer 2000), essentially consists of the following steps: Business (or Application Domain) Understanding (which includes definition of business and data mining goals), Data Understanding, Data Preparation, Data Mining (or Modelling), Evaluation (e.g. evaluation of results based on Data Mining goals), and Deployment (Kurgan and Musilek 2006). CRISP_DM is the among the most popular process model that is used by data mining experts to solve real world data mining problems (Shearer 2000). In this process model "what" needs to be done is clearly articulated, however

they lack guidance on "how", therefore an Integrated Knowledge Discovery and Data Mining process model (IKDDM) was developed (Sharma and Osei-Bryson 2010). IKDDM has the same phases as CRISP-DM but was found to be more effective and efficient in performing the phases of KDDM. IKDDM improves the fragmented approach to the different tasks in the various phases and provides a framework which includes the output of the tasks, the tools that can be used to implement them and an indication as to whether a given task can be a candidate for semi-automation.

The data mining techniques that can be used in the modelling/data mining phase can be classified as either predictive or descriptive techniques. Predictive techniques are considered to be supervised learning methods which focus on either classifying data into a set of predefined classes or predicting future data states. Examples of predictive analytics are classification, regression and value prediction. Descriptive techniques are unsupervised learning methods which focus on describing behaviour, for example, sequential pattern and association rule mining.

Sequential pattern and association rule mining are techniques in data mining that are used to find recurring patterns in data of the form X => Y, where X and Y are concepts or sets of concepts which occur frequently together or in a sequence in the dataset (Agarwal and Srikant 1994; Chen et al. 1996). They have been used successfully when the focus is on market basket analysis, where a rule indicates relationships among items in a basket that is the occurrence of certain items in a transaction implies the occurrence of a certain other items in the same transaction. The sequential patterns also introduce a time sequence as a factor as the items in the basket may not have been purchased all at the same time but there is a time lag between the various purchases. The generation of these rules is a two-step process, first all the itemsets that satisfy a user-specified minimum support criterion are extracted from the dataset. Then the associations between the items that occur frequently together are identified using a user-specified minimum confidence criterion. Often the output is very large as there are no guidelines to determine the threshold values of these two criteria. High threshold may lead to missing useful rules, and low threshold may lead to large number of rules thus leading to users getting overwhelmed by the output (Au and Chan 2003; Wang et al. 2005). Also, since the number of rules generated is exponential to the number of itemsets, the task of finding the relevant rules from the generated rules is often a difficult task (Paranape-Voditel and Deshpande 2013). Several domain based techniques which can reduce the number of rules generated have been used to make the output more interesting and meaningful (Mansingh et al. 2010; Paranape-Voditel and Deshpande 2013).

There are several studies on applications of sequential pattern mining, and the typical data consists of the following fields, a unique identification field for a transaction, list of items and a corresponding date associated with each item. The extracted rules do not capture personal attributes of customers rather only their buying patterns. However, most purchase decisions made by customers are associated with their demographic attributes, therefore there is a need to include demographic variables in such an analysis. In this study by including demographic variables we demonstrate a novel method of preparing data for sequential pattern mining, which can facilitate the generation and identification of valuable customer centric patterns in the modelling phase. We focus on the data preparation phase of the IKDDM process model for association rule induction and sequential pattern mining and identify steps which assist in preparing the data in a way that maximizes the benefits of the data mining technique.

Proposed Method for Data Preparation for ARI/SPM

Of the six phases of IKDDM, the business understanding phase is considered to be the most important as it guides the subsequent phases in performing data mining (Mansingh et al. 2015; Sharma and Osei-Bryson 2010), and the Data Preparation (DP) phase is the most time consuming phase. The focus of the data preparation phase is on identifying quality data and formatting it appropriately, which can lead to generation of quality patterns by the chosen data mining algorithms (Zhang et al. 2003). Data preparation generates a dataset smaller than the original dataset but with better quality and relevant data which can significantly improve the efficiency of the modelling phase.

As shown in fig. 1 the data preparation phase focuses on Cleaning the data, Construct the data (i.e. create derived variables, discretize where relevant, integrate if necessary), Convert data to the format that the selected tool requires to satisfy the requirements of the given DM tool. It is quite clear that these focus on "what" needs to be done but we recognise that there is not a great deal of literature on "how" it should be done. For example, there are many possibilities in constructing the data whether in terms of the derived variables that can be constructed, the ways in which the data can be discretised and the how

the data can be integrated. Converting the data to the format required for the selected tools also requires a great deal of expertise in understanding and identifying the various ways this can be done.

In this paper the scope is restricted to Association rule induction (ARI) and Sequential pattern mining (SPM). For ARI and SPM the features of a well-formulated data mining objective is "itemsets" (Sharma and Osei-Bryson 2010), and an association rule or sequential pattern shows relationships among items in an itemset (Chen et al. 1996). Therefore, while constructing and formatting data the focus is on determining what can be included as items. In many organizations, data that is relevant for decision making typically includes both time specific (i.e. behaviour of buying) and demographic variables. As organizations want to know the sequencing among the items in an "itemset" they also want to incorporate demographic variables is the rules/patterns. However, the demographic variables may be numeric and hence there is a need to discretize them so they can be added as an item in the basket.

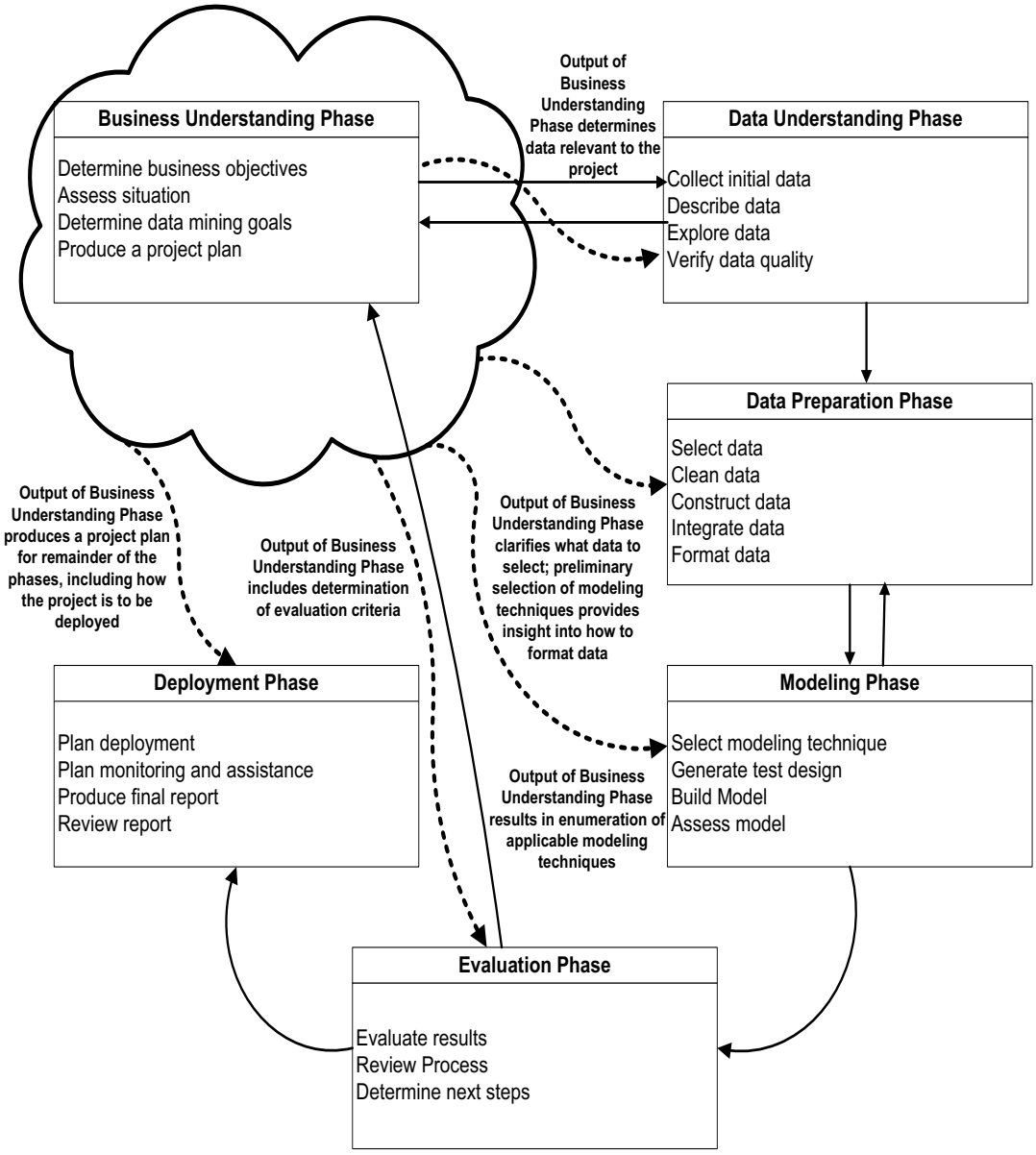

Figure 1 - Phases of the IKDDM Model ((Sharma and Osei-Bryson 2010)

A typical dataset for ARI/SPM contains:

- An id field.
- Sequence number (for sequential patterns).
- Target variable (i.e. items).

For each id there can be multiple items which represent all the items that are associated with a single basket. A sequence number may or may not be present based on whether ARI or SPM is being performed.

In this paper we are proposing a method for constructing and formatting data for ARI and SPM as follows:

Step 1. The variable to be considered as an item (i.e. the original Target variable) in the dataset is categorical.
   a. Examine the frequency counts.
   b. Generate a concept hierarchy or if one exists examine to see its appropriateness.
   c. In collaboration with the business analyst determine if they categorical variables are being analyzed at the right level of granularity. Use the concept hierarchy to identify the level of analysis.
   d. By using the concept hierarchy a new categorical value will be generated instead of the original item e.g. XO-50" Smart TV and WHOSUNG – 52" Smart TV can be mapped to a concept "Big Size Smart TV".

For example, an appliance company contains data on several items that are bought at the shop. The objective is to see what items are more frequently bought together. However, the number of items is quite large (250) and that number needs to be reduced. For the subset of items shown in table 1, a concept hierarchy could be used to categorize the TV based on their sizes, brands, type or a combination. The business analyst will be needed to determine which concept should be used to reduce the number of items.

Items	**Count**
XO – 50" Smart TV	140
WHOSUNG – 52" Smart TV	45
XO – 52" TV	300
WHOSUNG – 32" TV	450
XO – 32" TV	200

Table 1. Example Frequency counts of items in a dataset

Step 2. The variables to be considered as an item (i.e. the original Target variable) in the dataset are numeric.
   a. Discretize the numeric variables.
   b. A Composite Variable is created that contains data from each original Target variable.
   c. There is thus multiple rows in the transformed dataset for each row in the original dataset (i.e. one row for each original Target variable).

   For example: Given the Zoo dataset, objective is to generate ARs that involve the Predator, Toothed, and Legs attributes, where the values of the Legs attribute is broken down into three (3) categories: 0, 1-2, and greater than 2 (see table 2).

Step 3. While preparing data some preliminary modelling has to be performed. Generate AR and determine whether co-occuring items are highly correlated, i.e. they occur together most of times only with each other (e.g. lift > 90% and confidence > 90%).
   a. If yes then and go back to step 1 and re-examine concept hierarchy with business analyst.
   For example, in an insurance company all persons who buy gold scheme get a free premium windshield coverage. It is likely that most occurrences of item "Gold scheme" will have item "premium windshield coverage" and vice versa hence it might be better to re-examine the concept hierarchy to determine how to remove the highly correlated items.

Attribute	**Discretization Function**					
	Value	**Discrete Value**	**Value**	**Discrete Value**	**Value**	**Discrete Value**
Predator	**0**	Non_Predator	**1**	Predator		
Toothed	**0**	Non_Toothed	**1**	Toothed		
Legs	**0**	No_Legs	**1 – 2**	1_2_Legs	**> 2**	More_than_2_Legs

Initial Data

Animal_ Name	hair	feathers	eggs	milk	aquatic	Predator	Toothed	Backbone	breathes	venomous	fins	Legs
Aardvark	1	0	0	1	0	1	1	1	1	0	0	4
Antelope	1	0	0	1	0	0	1	1	1	0	0	4

Data Preparation Steps

Step

2a *ID variable (e.g. Animal_Name) & relevant Target Variables (e.g. Predator, Toothed, Legs)*

Animal_Name	**Predator**	**Toothed**	**Legs**
aardvark	1	1	4
antelope	0	1	4

2b *Create & Add relevant Discretized Target Variables (e.g. D_, D_Toothed, D_Legs)*

Animal_Name	**Predator**	**Toothed**	**Legs**	**D_Predator**	**D_Toothed**	**D_Legs**
aardvark	1	1	4	Predator	Toothed	More_than_2_Legs
antelope	0	1	4	Non_Predator	Toothed	More_than_2_Legs

2c *Do Transformation resulting in new Transaction type dataset with ID variable & new Composite Target Variable (e.g. Animal_Category) and 1 row per original Target variable.*

Aardvark	Predator
Aardvark	Toothed
Aardvark	More_than_2_Legs
Antelope	Non_Predator
Antelope	Toothed
Antelope	More_than_2_Legs

Table 2. Activities in step 2

Step 4. Determine if demographic variables need to be added. If yes go back to either step 1 or 2 based on the type of variable. If categorical variable (e.g. Occupation) go to Step 1 and if they are numeric (e.g. age or income) go to step 2. Add these as items to the basket.
For sequential patterns a date field becomes a required field. Perform step 5 and 6 if data preparation is being done for SPM. If only doing ARI, skip to step 7.

Step 5. Determine if demographic variables have to be added. If yes, add demographic variables with an arbitrary low sequence number where each demographic variable will have a unique sequence number. For example, education can have a sequence number 3 and income can have sequence number 4. All education values and all income values will have sequence number 3 and 4 respectively in the dataset.

Step 6. For transactional items in a dataset a date field exists, hence transform the date field associated with an item to a sequence number. Convert the date field in a way that the items with a later calendar date have a higher sequence number. Take the year, month and day to create this number. For example, 19 December 2004 will become 20041219.

Step 7. Format data based on data mining tool selected in BU phase. Some tools require data to be organized in column format while some want it in the row format.

- Tool: SAS Enterprise Miner – Column Format

Cust ID	**Items**
A10002	Income - VALUE
A10002	Car Insurance
A10005	Education- High School
A10005	House Insurance

- Tool: RapidMiner – Row Format

Cust ID	**Income-Emerging**	**Education - High School**	**Car Insurance**	**House Insurance**
A10002	Y	N	Y	N
A10005	N	Y	N	Y

In the next section we provide a description of the different phases while performing market basket analysis. The results presented in the case study are for illustrative purposes only and based on contrived data to protect the information privacy of the participating institution.

Case Study – Applying the IKDDM

The financial institution in the Caribbean is currently focussing on building a sales culture, as one of its strategic priorities is to improve sales. The fact that some of its products have moderate or low penetration suggests that both sales strategy and products need to be reviewed. Additionally, there is an awareness of repeat customer take-up of related products, the evidence of how this functions is merely anecdotal. However, there is no systematic means of identifying related products or directing sales personnel to offer product bundles. Based on various discussions with the business analyst we embarked on identifying what products can be sold together. This made us cognizant of the fact that the decision makers were not merely interested in which products could be sold together but they also needed the customer demographics to identify "who" will be interested in "what". Thus in this study we demonstrate how data mining techniques especially association rule and sequential pattern mining can assist in such

identification and help the financial institution maximize the value from their customers. We applied the different phases of IKDDM to multiple databases across various business units within the financial organization.

Business Understanding Phase

In this phase the business objectives were explicated by interviewing the institution's business leader and business analyst. Since the focus of this customer analytics was to improve marketing strategies, after briefly examining the datasets and the following business objectives were formulated.

1. Maximizing the customer value to the organization by cross selling / up-selling various products offered by the bank to existing customers.
2. Contribute to the bank's revenue targets by improving the efficiency of sales initiatives.

Both objectives focus on maximising customer transaction intensity and value by identifying patterns in their purchasing behaviour. In order to improve the success rate of current targeted marketing campaigns, market basket analysis was employed as a customer analytic technique. This technique uses association rule mining to identify the products which customers currently purchase together and can help to identify those products that go well together (in terms of bundles) and therefore should be marketed accordingly. Furthermore, since a customer does not buy a set of products at one time, rather the basket contains products bought over time, it is conducive to sequential rule mining which not only shows not only which products were bought together but also the sequence in which they were bought. Both association rules and sequential mining will help to increase the effectiveness of sales campaign management and targeting processes.

Data Understanding Phase

In this phase the datasets needed to build the baskets for customers were identified. The attributes and the data tables which contain the savings account details, loan, credit card, investment and insurance products were explored. Since both association rule and sequential rule mining were used, it was important to determine firstly if the baskets could be created for customers, and then, for sequence rule mining if the variable that could be used to represent a sequence was present. Frequency distributions for all products were examined to determine if the number of items in the baskets was too large. The measurement levels and distribution of demographic variables of the customers were analysed to determine whether they needed to be discretized to be included in the analysis.

Data Preparation Phase

This phase of the process required that the data for individual customers be prepared for market basket analysis. For each customer their demographics and the products they had bought and the date of purchase were extracted, this required accessing seven relational and transactional tables. In all 47,452 customer baskets were created. Since the number of products was quite large, there was a need to develop concept hierarchies in the domain and after a few iterations between this phase and modelling phase these hierarchies were used to reduce the possible number of items that could exist in a basket. For example, in figure 2 all occurrences of Prov-X, Prov-Y and Prov-Z will be replaced by Provision which is at a higher level in the concept hierarchy. The date the products were purchased was transformed and used as a sequence number in the format YYYYMMDD. The date field in this format was used to capture the sequence of the purchase of the products. The frequency distribution of demographics variables were examined and the numeric variables were discretised. The discretization was done in consultation with the financial institution's business analysts, who provided the linguistic terms used to describe each of the data ranges for the variables age and income. Each demographic variable was also given a sequence number. The inclusion of the demographic and product data together in a basket facilitates the discovery of multi-dimensional rules and frequent patterns in buying products (see table 3).

Cust Id	Sequence	Target
A10002	1	College / University
A10002	2	Clerical/Administrative
A10002	3	VALUE
A10002	4	Male
A10002	5	Married
A10002	6	YOUNG
A10002	20120320	XYZ Card
A10002	20150314	Provision
A10002	20150324	Car Insurance
A10005	1	College / University
A10005	2	Teacher/Lecturer
A10005	3	EMERGING
A10005	4	Female
A10005	5	Married
A10005	6	MIDDLE
A10005	20150122	Provision
A10005	20100429	ABC Card

Table 3. Sample of prepared data set

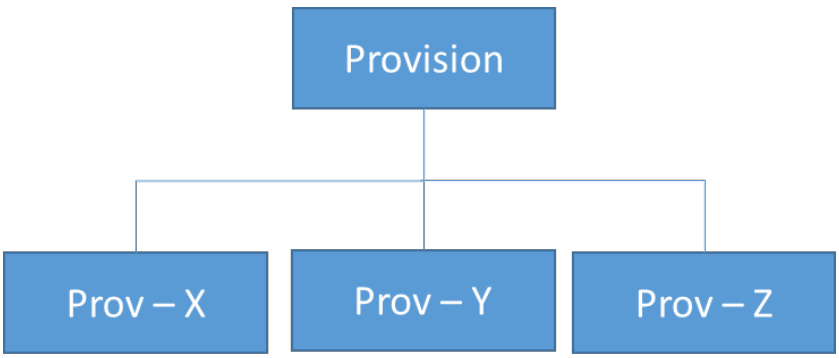

Figure 2: An example of a Concept Hierarchy

Modelling Phase

There were several iterations between the data preparation phase and the modelling phase (see figure 1). The items in the basket are of two types, the demographics and the products. First association rule mining was performed on customer baskets data containing just the products (i.e. cards, insurance, investment and loans). The generated association rules provided strong evidence that several products were co-occurring (see table 4). Due to the relationship between the items it became imperative to explore the distribution of all the product groups (i.e. card, loan, insurance and investment). It was found that there was a large number of products under each of these

product groups (51, 9, 53, and 15 respectively). Since having such a large number of products and also products that are highly correlated can affect the ability to identify interesting patterns in the data, a possible taxonomy (concept hierarchy) was sought for these product groups (see figure 2). These hierarchies were provided by the business analyst at the institution and were used to classify each product into a product sub group which was then used in the development of the models. The generated association rules were presented to the business analyst, however the association rules were focussing more on the products rather than customers. Therefore the demographics together with all the products were included as basket items and sequential rule mining was performed.

CONF	SUPPORT	LIFT	COUNT	RULE
100	1.11	90.16	255	M FUND ==> B FUND
100	1.11	90.16	255	E FUND ==> B FUND
100	1.11	90.16	255	M FUND & E FUND ==> B FUND

Table 4. Association Rules – Similar RHS

A number of sequential pattern rules were generated, each of which had a measure of strength and interestingness in terms of Support, Confidence and Lift. One of the issues with association rules/sequential pattern mining is that the number of rules generated are exponential based on the number on frequent itemsets. Therefore, the output of these rules is often very large. Presenting the large number of rules to business analysts posed a challenge as there was no context to present the output. Hence, we decided to prune the ruleset by dividing them into subsets based on the recommendation that subsets can be created based on either prior knowledge of domain rules (Mansingh et al. 2010) or just numerical value partitions. They also suggest the following partitions: novel rules with high strengths; known rules with high strengths; known rules with low strength; missing rules; and contradictory rules. In the present study, no prior rules were known therefore we created four partitions following the given criteria:

1. Rules with high Lift (> 1) and low Support (Support <10).
2. Rules with high Lift (> 1) and high Support (Support >= 10).
3. Rules with low Lift (< = 1) and high Support (Support >= 10).
4. Rules with low Lift (< = 1) and low Support (Support <10).

The partitioned ruleset was presented to the business analyst for evaluation. The business leaders saw value in the output that was given to them as they could easily identify the "who" and the corresponding "what" for increasing their sales.

Evaluation Phase

The rules in the four partitions provided context for business interpretation. Rules in partition 1 are considered to be interesting because high Lift value and low Support values signify that the association between the left and right hand side of the rule is higher than expected but the number of occurrences of all the items in the rule is low. Strong associations with low occurrences suggest that there are opportunities for the institution to increase the number of occurrences by marketing to persons with similar profiles. Such prospecting of customers will reduce the number of failures in sales efforts to sales people and ultimately drive sales performance. This partition can be interpreted as "novel rules" as the rules in the partition may not have been obvious to the business analyst. See table 3 rules 1 to 4, the rules have high Lift and low Support.

Partition 2 consists of rules with high Lift and high Support values. This again indicates a high association between the left and right hand side of the rule. However, the support is also high so there are already quite a few customers that have already bought the products. This means that as a pool for targeting this partition identifies a less likely set of customers as compared to those with low Support. However, the rules in this partition should still be examined by business analyst as they probably confirm what is already known. This can be seen as a partition of "known rules" (see table 5 – rule 5).

Rules in partition 3 with low Lift indicate that the items occur together less frequently than expected. These rules also have a relatively high Support, which could indicate that there is a random high incidence of the items in the rule within the dataset and that it may not be a good rule to use for targeting customers. These are seen as "contradictory rules" because although the high incidences may create a perception that these items are bought together, they occur together only because of their large occurrences in the data. Finally, the rules in partition 4

can be seen as "outliers" as there are only a few customers who follow that buying pattern and the items in the baskets occur less frequently than expected.

In this phase the rules in the partitions were presented to the business leader and business analysts, who saw tremendous business value in applying these sequential patterns in their sales prospecting efforts. The analysts concurred that these rules provided more actionable knowledge from traditional market basket analysis which typically focus on what products are bought together.

Initially the number of rules generated by both association and sequence rule mining made the interpretation very difficult, but the context provided by partitions made it easier to interpret the results. The inclusion of demographics in the rules shifts the focus from what products are bought together to what products and by who thereby making the focus more customer centric. The decision makers were not just getting "What products customers typically buy in a sequence" but rather "What products a particular type of customer buys and in what sequence". The knowledge embedded in these rules can be used by the organization to cross sell/upsell and prospect new customers. Also applying these rules would drive their sales focus from being product oriented to customer oriented.

No.	Rule	Pseudolift	Support
1.	High School ==> VALUE ==> XGENCARD	3.34	2.39
2.	Female ==> XYZCARD ==> PROVISION	2.62	2.16
3.	EMERGING ==> XYZCARD ==> PROVISON	2.37	2.34
4	College / University ==> MIDDLE ==> XGOLD	1.86	2.01
5	College / University ==> EMERGING => XCLASSIC	1.3	41.73

Table 5: Sample Sequential Rules

Conclusion

This paper provides an empirical case study of an implementation of the IKDDM process model in a practical organizational context. By applying IKDDM in the real world one can examine the benefits and weaknesses of the process model and propose further developments (Alter 2013). IKDDM provides a structure to ensure that accurate, valid, relevant and quality data is pushed through the various stages of the process model. In this study we applied market basket analysis to financial data in a manner that allows business analysts to easily identify the "who" and the "what" during cross-selling and up-selling. This allows the business professionals to develop profiles of customer preferences for specific products, thereby offering only those products and services to a specific set of customers the organization can make substantial savings on promotions and offerings that would otherwise be unprofitable.

Including both the demographics and items/products in the sequential pattern mining allows not only determining the sequence of buying but also the characteristics of the customers exhibiting that specific behaviour. The method proposed in this paper provides a mechanism to incorporate the demographic variables with the behaviour variables (i.e. buying patterns) which will facilitate the organization to take decisions which are more customer oriented rather than product oriented. The method also provides a mechanism to connect static variables such as demographics with attributes that change over time with the buying behaviour of the customers. This method was applied to financial sector but it can be applied to other decision making domains, such as healthcare and human resources.

In association rules induction and sequential pattern mining the number of rules that are produced are large as in these techniques several strong rules are generated. Whereas in other techniques such as decision trees induction due to the recursive nature of these algorithms only the strongest rules are generated for predicting/classifying the target variable. The strong rules from ARI and SPM allows business analyst to identify several items of interest (i.e. by examining the right most item in the rule), and method proposed in the paper allows the inclusion of demographic and behavioural variables in the left hand side of rules. This provides the organisation the opportunity

to explore multiple avenues and then choose the one they find most appropriate to them in terms of sales/promotions of products.

Based on the insights from this study we plan on refining the IKDDM process and propose development of a new artifact. In the data preparation phase we posit that there is a need for an additional task that checks to see if the particular case being considered matches a previous case. This phase is highly dependent on the expertise of the data mining analyst and on the type of data. To enable this an organization will need to develop a knowledgebase of past projects. This will improve one of the most tedious and critical phases of IKDDM and further ensure that the experiential (tacit) organizational knowledge is stored and made accessible for future use, thereby ensuring variability from one project to another is minimized within an organization.

References

Agarwal, R., and Srikant, R. 1994. "Fast Algorithm for Mining Association Rules in Large Databases," *International Conference on Very Large Databases,* J.B. Bocca, M. Jarke and C. Zaniol (eds.), Santiago de Chile, Chile, pp. 487-499.

Agrawal , R., Imielinski, T., and Swami, A. 1993. "Database Mining: A Performance Perspective," *IEEE Transactions on Knowledge and Data Engineering* (5:6), pp. 914-925.

Alter, S. 2013. "Work Systems Theory: Overview of Core Concepts, Extensions, and Challenges for the Future," *Journal of the Association for Information Systems* (17:2), pp. 72-121.

Au, W. H., and Chan, K. C. C. 2003. "Mining Fuzzy Association Rules in a Bank-Account Database," *IEEE Transaction on Fuzzy Systems* (11), pp. 238-248.

Berry, M., and Linoff, G. 1997. *Data Mining Techniques for Marketing, Sales and Customer Support*, (3rd ed.). U.S.A.: John Wiley & Sons, Inc.

Chen, M., Han, J., and Yu, P. S. 1996. "Data Mining: An Overview from a Database Perspective," *IEEE Transaction on Knowledge and Data Engineering* (8:6), pp. 866-883.

Cios, K., Teresinska, A., Konieczna, S., Potocka, J., and Sharma, S. 2000. "Diagnosing Myocardial Perfusion from Pect Bull's-Eye Maps - a Knowledge Discovery Approach," *IEEE Engineering in Medicine and Biology Magazine* (19:4), pp. 17-25.

Clifton, C., and Thuraisingham, B. 2001. "Emerging Standards for Data Mining," *Computer Standards & Interfaces* (23:3), pp. 187-193.

Delen, D. 2010. "A Comparative Analysis of Machine Learning Techniques for Student Retention Management," *Decision Support Systems* (49:4), pp. 498-506.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* (17:3), pp. 37-54.

Grossman, R. L., Hornick, M. F., and Meyer, G. 2002. "Data Mining Standards Initiatives " *Communications of the ACM* (45:8), pp. 59-61.

Han, J., and Kamber, M. 2006. *Data Mining: Concepts and Techniques*, (second ed.). Morgan Kaufmann.

Jourdan, Z., Rainer, R. K., and Marshall, T. E. 2008. "Business Intelligence: An Analysis of the Literature," *Information Systems Management* (25:2), pp. 121-131.

Kurgan, L. A., and Musilek, P. 2006. "The Survey of Knowledge Discovery and Data Mining Process Models," *The Knowledge Engineering Review* (21:1), pp. 1-24.

Li, S.-T., and Shue, L.-Y. 2004. "Data Mining to Aid Policy Making in Air Pollution Management," *Expert Systems with Applications* (27:3), pp. 331-340.

Mansingh, G., Osei-Bryson, K.-M., and Asnani, M. 2015. "Exploring the Antecedents of the Quality of Life of Pateints with Scile Cell Disease: Using a Knowledge Discovery and Data Mining Proess Model-Based Framework," *Health systems*), pp. 1-14.

Mansingh, G., Osei-Bryson, K.-M., and Reichgelt, H. 2010. "Using Ontologies to Facilitate Post-Processing of Association Rules by Domain Experts," *Information Sciences* (181:3), pp. 419-434.

Mariscal, G., Marban, O., and Fernandez, C. 2010. "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies," *The Knowledge Engineering Review* (25:2), pp. 137-166.

Moin, K. I., and Ahmed, Q. B. 2012. "Use of Data Mining in Banking," *International Journal of Engineering Research and Applications* (2:2), pp. 738-742.

Okazaki, S. 2006. "What Do We Know About Mobile Internet Adopters? A Cluster Analysis.," *Information and Management* (43:2), pp. 127-141.

Paranape-Voditel, P., and Deshpande, U. 2013. "A Stock Market Portfolio Recommender System Based on Association Rule Mining," *Applied Soft Computing* (13), pp. 1055-1063.

Piatetsky-Shapiro, G., and Matheus, C. J. 1994. "The Interestingness of Deviations," *AAAI'94 Workshop on Knowledge Discovery in Databases*, Seattle, WA, U.S.A., pp. 25-36.

Rygielski, C., Jyun-Cheng, W., and Yen, D. C. 2002. "Data Mining Techniques for Customer Relationship Management," *Technology in Society* (24), pp. 483-502.

Sharma, S., and Osei-Bryson, K.-M. 2010. "Toward an Integrated Knowledge Discovery and Data Mining Process Model," *The Knowledge Engineering Review* (25:1), pp. 49-67.

Sharma, S., Osei-Bryson, K.-M., and Kasper, G. M. 2012. "Evaluation of an Integrated Knowledge Discovery and Data Mining Process Model," *Expert Systems with Applications* (39:13), pp. 11335-11348.

Shearer, C. 2000. "The Crisp-Dm Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* (5:4), pp. 13-22.

Wang, J., Han, J., Lu, Y., and Tzvetkov, P. 2005. "Tfp: An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets," *IEEE Transaction on Knowledge and Data Engineering* (17:5), pp. 652-664.

Zhang, S., Zhang, C., and Yang, Q. 2003. "Data Preparation for Data Mining," *Applied Artificial Intelligence* (17), pp. 375-381.