FRAUD PREVENTION IN THE B2C E-COMMERCE MAIL ORDER BUSINESS: A FRAMEWORK FOR AN ECONOMIC PERSPECTIVE ON DATA MINING

TOBIAS KNUTH

Supervised by Doctor Richard Whitecross and
Professor Dennis Ahrholdt

Business School
Edinburgh Napier University

May 7, 2018

A thesis submitted in partial fulfilment of the requirements of Edinburgh Napier University, for the award of Doctor of Business Administration

Tobias Knuth: *Fraud Prevention in the B2C E-Commerce Mail Order Business,* A Framework for an Economic Perspective on Data Mining, May 7, 2018

DECLARATION

I declare that this work has not been submitted for any other degree or professional qualification. The thesis is the result of my own independent work.

Hamburg, May 7, 2018

Tobias Knuth

ABSTRACT

A remarkable gap exists between the financial impact of fraud in the B2C e-commerce mail order business and the amount of research conducted in this area — whether it be qualitative or quantitative research about fraud prevention. Projecting published fraud rates of only approx. one percent to e-commerce sales data, the affected sales volume amounts to $651 million in the German market, and in the North American market, the volume amounts to $5.22 billion; empirical data, however, indicate even higher fraud rates. Low profit margins amplify the financial damage caused by fraudulent activities. Hence, companies show increasing concern for raising numbers of internet fraud.

The problem motivates companies to invest into data analytics and, as a more sophisticated approach, into automated machine learning systems in order to inspect and evaluate the high volume of transactions in which potential fraud cases can be buried. In other areas that face fraud (e.g. automobile insurance), machine learning has been applied successfully. However, there is little evidence yet about which variables may act as fraud risk indicators and how to design such systems in the e-commerce mail order business.

In this research, mixed methods are applied in order to investigate the question how computer-aided systems can help detect and prevent fraudulent transactions. In the qualitative part, experts from fraud prevention companies are interviewed in order to understand how fraud prevention has been conventionally conducted in the e-commerce mail order business. The

quantitative part, for which a dataset containing transactions from one of the largest e-commerce firms in Europe has been analyzed, consists of three analytical components: First, feature importance is evaluated by computing information gain and training a decision tree in order to find out which features are relevant fraud indicators. Second, a prediction model is built using logistic regression and gradient boosted trees. The prediction model allows to estimate the fraud risk of future transactions. Third, because risk estimation alone does not equal profit maximization, utility theory is woven into prioritization of transactions such that the model optimizes the financial value of fraud prevention activities.

Results indicate that the interviewed companies want to use intelligent computer-aided systems that support manual inspection activities through the use of data mining techniques. Feature analysis reveals that some features, such as whether a shipment has been sent to a parcel shop, can help separate fraudulent from legitimate orders better than others. The predictive model yields promising results as it is able to correctly identify approximately 86% of the 2% most suspicious transactions as fraud. When the model is used to optimize the financial outcome instead of pure classification quality, results suggest that the company providing the dataset could achieve substantial additional savings of up to 87% through introduction of expected utility as a ranking measure when being constrained by limited inspection resources.

PUBLICATIONS

Knuth, T. and Ahrholdt, D. (2017). "How to Detect Fraud — Evaluation of B2C E-Commerce Transaction Data." In: *46th EMAC Annual Conference*. (May 23–26, 2017), p. 99.

It is of the highest importance in the art of detection to be able to recognize, out of a number of facts, which are incidental and which vital. Otherwise your energy and attention must be dissipated instead of being concentrated.

— Sherlock Holmes in "The Reigate Puzzle"
by Sir Arthur Conan Doyle

ACKNOWLEDGMENTS

I am grateful that I had the opportunity to spend more than three years on such an interesting research topic, and I appreciate the support I received from my supervisors, my family, and the company funding my research.

The thesis has been typeset with LaTeX; all analyses were computed with libraries from python and R. I would like to thank the developers of such open source projects for their commitment because their work provided the software this thesis relies on.

All figures are the result of the author's own work if not stated otherwise. When a reference is given, the figure has been reproduced to match the style of the thesis.

CONTENTS

I PART ONE 1

1 OVERVIEW OF THE THESIS 2

1.1 Introduction . . . 2

1.2 Thesis Structure . . . 9

2 THEORETICAL FOUNDATION 12

2.1 Introduction . . . 12

2.2 Literature Selection . . . 13

2.3 Fraud Prevention in Related Areas . . . 15

2.4 E-Commerce Fraud Prevention . . . 18

2.4.1 Overview of Literature . . . 18

2.4.2 Approaches to Fraud Prevention . . . 20

2.4.3 Implications for the Research Problem . . . 23

2.5 Decision-Theoretic Representation . . . 24

2.6 Data Mining Framework . . . 26

2.6.1 Introduction . . . 26

2.6.2 Application Classes . . . 27

2.6.3 Data Mining Methods . . . 31

2.6.4 Degree of Supervision . . . 34

2.7 Economic Perspective . . . 35

2.8 Conclusion . . . 38

3 METHODOLOGY 40

3.1 Introduction . . . 40

3.2 Philosophy . . . 41

3.2.1 Philosophical Stance . . . 41

3.2.2 Comparison of Philosophies . . . 43

3.3 Concept of Probability . . . 44

3.4 Research Implications . . . 45

3.5 Delimitations and Limitations 47
3.6 Ethical Considerations 48

II PART TWO 51

4 MODELING FRAUD 52
4.1 Introduction 52
4.2 Company Overview 53
4.3 Interview Setup 53
4.4 Findings 61
4.4.1 Introduction 61
4.4.2 Fraud Risk Factors 62
4.4.3 Investigation Process 63
4.4.4 Classification Errors 66
4.4.5 Impact Evaluation 69
4.5 Limitations 70
4.6 Conclusion 72

5 DATA PREPARATION 74
5.1 Introduction 74
5.2 Overview of the Dataset 75
5.3 Feature Overview 78
5.3.1 Introduction 78
5.3.2 Evidence-Based Features 79
5.3.3 Discovery Features 81
5.4 Transformation 83
5.4.1 Discretization 83
5.4.2 Missing Values 85
5.4.3 Feature Interdependence 86
5.5 Limitations 86
5.6 Conclusion 88

6 FEATURE ANALYSIS 89
6.1 Introduction 89
6.2 Feature Relevance 90

6.3 Information Gain . . . 93
6.3.1 Introduction . . . 93
6.3.2 Analysis . . . 97
6.3.3 Findings . . . 98
6.3.4 Limitations . . . 106
6.4 Decision Trees . . . 107
6.4.1 Introduction . . . 107
6.4.2 Analysis . . . 109
6.4.3 Findings . . . 110
6.4.4 Limitations . . . 112
6.5 Conclusion . . . 113
7 PREDICTION MODEL 115
7.1 Introduction . . . 115
7.2 Machine Learning Fundamentals . . . 117
7.2.1 Introduction . . . 117
7.2.2 Logistic Regression . . . 120
7.2.3 Gradient Boosted Trees . . . 125
7.2.4 Sampling with Cross Validation . . . 127
7.3 Findings . . . 130
7.4 Limitations . . . 135
7.5 Conclusion . . . 137
8 UTILITY APPROACH 139
8.1 Introduction . . . 139
8.2 Utility Concept . . . 140
8.2.1 Work of Torgo and Lopes . . . 140
8.2.2 Revised Concept . . . 142
8.3 Analysis . . . 147
8.3.1 Introduction . . . 147
8.3.2 Evaluation of Utility . . . 150
8.3.3 Interpretation of Results . . . 151
8.4 Limitations . . . 153
8.5 Conclusion . . . 156

9 CONCLUSIONS 157
9.1 Introduction 157
9.2 Contribution to Knowledge 159
9.3 Contribution to Professional Practice 164
9.4 Further Work 168
9.5 Conclusion 170

REFERENCES 173

III APPENDIX 187

A EVALUATION MEASURES 188
A.1 Introduction 188
A.2 Set-Based Evaluation Measures 189
A.3 Evaluation Measures for Ranked Lists 191

LIST OF FIGURES

Figure 1 Visualization of the thematic focus of the thesis. 14

Figure 2 Overview of the areas for which publications about data mining methods used for fraud prevention have been retrieved. 16

Figure 3 Overview of the most popular methods for fraud detection. 31

Figure 4 Fraud prevention process model. 64

Figure 5 Possible sources of classification errors. . 67

Figure 6 Visualization of how information gain can be used to estimate feature relevance. . . 95

Figure 7 Visualization of entropy H in a binary class case, such as *fraudulent* and *legitimate*. 96

Figure 8 Information gain per feature, plotted with entropy-based discretization, discretization with $\sqrt{n}$-sized bins, and χ^2-based discretization. 98

Figure 9 Fraud rate for features *address distance* and *total price* plotted in percentile bins of 10%, i.e. each bin contains one tenth of the data. The data have been ordered by value, i.e. with regard to the total price, the bin $[0; 10)$ contains the 10% cheapest transactions. 99

Figure 10 Fraud rates per attribute value (—) and frequency distributions (▭) in percent. The fraud rates do not sum to one, but the frequency distributions do. 102
Figure 11 Heatmap per weekday and hour, with fraud rate values shown in percent. The hours are left-inclusive, i.e. the value 7a includes all transactions from 6 a.m. to 7 a.m. . . . 105
Figure 12 Frequency distribution of transactions. This heatmap shows the relative volume of transactions per cell. Values are *not* percentages; they have been scaled such that 100 represents the highest hourly transaction volume seen in the data. 105
Figure 13 Decision tree with three layers with the fraud rate and the fraction of the data. . . 111
Figure 14 Visualization of 5-fold cross validation . . 129
Figure 15 Precision at k^{th} percentile. 134
Figure 16 ROC curves. 135
Figure 17 Visualization of the influence of risk and margin on the profit. 146
Figure 18 Utility analysis flowchart. 148
Figure 19 Precision at k^{th} percentile. 151
Figure 20 Utility at k^{th} percentile. 152
Figure 21 F values for combinations of precision and recall values. 192
Figure 22 Scores of discounted gains per relevance score and position. 193
Figure 23 Individual scores obtained from RMSE and MAE per relevance score and position. 194

LIST OF TABLES

Table 1 Decision matrix for fraud prevention. . . . 26

Table 2 Breakdown of one month's transactions into the count of transactions and the order value with regard to the fraud label. . 75

Table 3 Overview of features with basic descriptive metrics. Features used in the following analyses are marked with *Y* and features discarded are marked with *N*. 77

Table 4 Correlation table using Pearson correlation coefficient. 87

Table 5 List of key features according to information gain, decision tree, and logistic regression. 113

Table 6 Confusion matrix for binary fraud classification. 119

Table 7 Coefficients of logistic regression and associated p-values. Likelihood ratios are shown below. Since discrete features have been dummy-coded, their reference values are given in square brackets. 124

Table 8 Outcomes of fraud classification with states fraudulent (f) and legitimate (l) and acts accept (a) and reject (r). 144

Table 9 Average financial impact per transaction at different percentiles. 153

Table 10 Confusion matrix for binary fraud classification. 190

ACRONYMS

TPR	True Positive Rate
PPV	Positive Predictive Value
NPV	Negative Predictive Value
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAP	Mean Average Precision
ARHR	Average Reciprocal Hit Rank
NDCG	Normalized Discounted Cumulative Gain
CART	Classification and Regression Trees
ID_3	Iterative Dichotomiser 3
ISP	Internet Service Provider

PART ONE

This part contains three chapters: In the introduction, the research problem is discussed and the structure of this thesis is explained. In the following chapter, the literature review and the introduction of a data mining framework form the theoretical foundation of the thesis. In the third chapter, methodology is discussed, covering the research paradigm, limitations and delimitations of the research, and ethical aspects. These chapters serve as the basis for the subsequent qualitative and quantitative analyses.

1

OVERVIEW OF THE THESIS

1.1 INTRODUCTION

When criminals commit a theft by ordering products online without paying (called buyer fraud or just fraud in the following), the merchant is hurt financially. With the rise of online shopping in Germany, which has grown by 11.3% from 2016 ($58.52 billion) to 2017 ($65.13 billion) and is predicted to keep growing (eMarketer, 2017), criminals witness new opportunities to commit fraud. In 2016, the federal office of criminal investigation in Germany counted 183,529 cases of internet-related crimes (Bundeskriminalamt, 2016), and 52,000 cases of buyer fraud have been registered (Brendel, 2017). Projecting published fraud rates of only approx. 1% to e-commerce sales data (CyberSource, 2017; Quah and Sriganesh, 2008), the affected sales volume amounts to $651 million in the German market (eMarketer, 2017). The empirical transaction data used in this thesis, however, indicate even higher fraud rates (between 1% and 19%, cf. Chapter 4). In addition, inventions like parcel boxes are convenient but the accessibility and anonymity these services provide open up new ways of committing fraud (Brendel, 2017). Low profit margins amplify the financial damage caused by fraudulent activities. The problem is not new (Schmidt and Verbeet, 2004), and companies show increasing concern for rising numbers of internet fraud (Gassmann, 2015). These statistics are not complete, though, because they only cover cases that were officially reported and companies some-

times hesitate to file claims because they do not want to risk their customers' loyalty if they mistakenly regard a legitimate transaction as a fraud case (Anon, 2015a). Hence, internal data retrieved from companies show higher fraud rates than official statistics. Phua et al. (2010, p. 1) even state: "In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe."

Fraud is an important issue in many areas, and it can occur in various forms. In the Oxford Dictionary, it is defined as the "crime of deceiving someone to gain money or personal advantage" (Waite, 2012, p. 286). Consequently, e-commerce mail order fraud describes the practice of buying products online without the intention to pay for them, often communicating fake data to the online merchant. A detailed discussion of the definition will be presented in Chapter 2.

In order to protect themselves against fraud, companies can employ a variety of measures (CyberSource, 2017). In the literature, three terms, which summarize these activities, are mentioned: fraud detection (Torgo and Lopes, 2011), fraud prevention (Hinneburg, 2006), and fraud management (CyberSource, 2017). Due to a lack of semantic standardization, the following terminology is proposed: The sole purpose of fraud detection is to find out which e-commerce transactions are fraud cases. Fraud prevention aims at stopping delivery of such cases before financial damage occurs. Fraud management contains detection and prevention activities but also addresses the business problem. Decision makers must face a trade-off between maximizing the fraud detection rate and minimizing the amount of incorrectly blocked transactions. In particular, when fraud prevention is guided by economic principles, the damage caused

by fraud cases should be minimized and sales revenue should be maximized (cf. Chapter 8).

Some companies have established fraud prevention departments where employees identify and examine suspicious transactions (cf. Chapter 4). However, large companies process thousands of transactions per day and it would be economically infeasible to inspect every transaction manually. Therefore, computer-aided systems that analyze the high volume of transactions in which potential fraud cases are buried can support the fraud prevention work (Carneiro, Figueira, and Costa, 2017). According to the fraud report published by CyberSource (2017), which is part of the credit card company Visa, companies rely on validation services, analysis of customer data — sometimes they even share data with other retailers —, and device tracking techniques. The fraud report shows that simple techniques such as verification of the credit card number (used by 88% of the surveyed companies) or verification of addresses (82%) are most popular. Even though most companies employ fraud detection tools, automated scoring models that estimate the fraud risk are used in only approximately one out of two companies (CyberSource, 2017). Through the use of data mining techniques, these systems compute fraud risk scores of transactions automatically, searching for suspicious patterns (Phua et al., 2010). In general, data mining deals with examining large databases for generating new information (Stevenson, Pearsall, and Hanks, 2010) and will be discussed in detail in Chapter 2.

Surprisingly, with regard to academia, the financial impact of fraud in the B2C e-commerce mail order business is not appropriately reflected by the amount of research conducted in this area regarding both qualitative and quantitative research. Phua et al. (2010) mentioned a general lack of research with regard to fraud detection in all industries seven years ago, and only

two quantitative studies have focused on the e-commerce mail order business in particular until now: Brabazon et al. (2010) and Carneiro, Figueira, and Costa (2017) analyzed the use of machine learning techniques for fraud detection but solely considered credit card fraud. As such, there is no existing quantitative approach which is concerned with e-commerce transaction data in general, regardless of the specific payment method. Yet, this would be interesting because some companies heavily rely on payment per invoice (cf. Chapter 4). In the last years, data mining techniques have been used successfully in related areas, for example in insurance fraud (Bermúdez et al., 2008). The data mining methods used in such publications served as inspiration for this thesis. However, it is important to understand the conceptual differences between these related areas and the e-commerce mail order business (Carneiro, Figueira, and Costa, 2017). For example, which variables (also called features in the following) can act as effective fraud risk indicators depends on the context of application. Features used in automobile insurance fraud detection — such as whether a rental vehicle was involved in an accident (Viaene, Dedene, and Derrig, 2005) — are not helpful at all. In contrast, areas closer to e-commerce, such as credit card fraud, can be consulted; Chung and Suh (2009) included the amount of purchases and the duration of the customer relationship as variables in their analysis; these variables exist in the e-commerce mail order business as well. Nevertheless, it remains uncertain whether such features turn out to be valuable with regard to the e-commerce mail order business even if they do in other areas. In addition, it is unclear in how far variables which are only available in online retailing can be used for fraud detection, e.g. whether a parcel shop is used as the shipment address; such variables

have only partially been discussed by Hinneburg (2006) and Carneiro, Figueira, and Costa (2017).

In order to contribute to closing the research gap, the central research question is raised: How can automated data mining support profit-oriented B2C e-commerce mail order companies in dealing with fraud? The question is broken down into multiple parts: How is e-commerce mail order fraud prevention currently conducted? What are indicators of fraud? In how far can machine learning models be used to detect fraud cases? How can and should manual and automated inspection work together? How do economic interests shape fraud prevention? How should the performance of fraud prevention techniques be measured?

In order to answer the research questions, the thesis contains both a small qualitative part and a larger quantitative part divided into multiple steps. For the former, interviews with a few fraud prevention experts have been conducted. The latter is based on a dataset that contains e-commerce transactions from one of the interviewed companies, which is also one of Europe's largest online retailers.

Other publications have only touched aspects of the research problem: Hinneburg (2006) solely conducted interviews in order to investigate fraud prevention in the e-commerce mail order business twelve years ago; Carneiro, Figueira, and Costa (2017) developed a fraud detection model for online retailing but only considered credit card fraud. A holistic perspective that covers understanding the fraud prevention process, analyzing the relevance of potential fraud risk indicators, building a prediction model, and applying utility theory for profit maximization has not been taken yet. Therefore, after having fitted in the missing parts, this thesis aims at combining the individual pieces of the puzzle into a single consistent research work

in order to contribute to closing the research gap and to provide a link between the existing but conceptually different publications.

In the next section, the structure of the thesis is described. Extending the brief overview of the main components above, motivation for each chapter is provided and the methods used in the thesis are mentioned.

The thesis contains five components in order to address the research question, which are presented in the list below. For each component, a brief overview of results and contributions is provided.

- The literature review explores the current state of research regarding fraud detection in the e-commerce mail order business. Although comparative studies about data mining in fraud detection exist (Phua et al., 2010), the e-commerce mail order business has not yet received sufficient attention. The literature review provides a starting point for researchers and practitioners interested in fraud detection, prevention, and management (Chapter 2).

- The fraud prevention process in two of Europe's largest e-commerce mail order companies is investigated in order to find out how manual inspection is accomplished (Chapter 4). Results of the interviews show that such companies find it difficult to face the large number of transactions that by far exceeds their available manual inspection capacity. The interviewed companies also use other strategies for fraud prevention than mentioned in the literature, such as requiring identification of the customer on delivery. In general, they are interested in using data mining more strongly in order to support their fraud prevention work.

- Feature analysis (Chapter 6) aims at determining how useful individual features are for fraud detection. Analysis of information gain and a decision tree reveal that a handful of features may serve as promising fraud risk predictors, and combinations of feature values can represent specific patterns associated with extraordinarily high fraud rates. For example, amongst other insights, it is revealed that the fraud risk is increased if there exists a considerable distance between the invoice and the shipping address. Such insights can allow companies to focus their inspection activities on the most important features, regardless of whether they perform fraud prevention only manually or with a computer-aided system.

- Based on the previous insights, a predictive fraud detection model is developed for which two methods are used (Chapter 7): While the use of logistic regression is justified by its popularity regarding financial fraud detection (Ngai et al., 2011), tree-based models (such as gradient boosted trees) deal well with data of mixed type (Hastie, Tibshirani, and Friedman, 2013) such as found in the transaction dataset. Results show that the prediction model achieves satisfying performance and could be used to support existing fraud prevention work.

- Following the request of Carneiro, Figueira, and Costa (2017) to further research more appropriate evaluation techniques and to develop cost-based models, a utility-based model is proposed as an alternative to traditional methods purely based on the number of detected cases, which aim at maximization of detection rates (cf. Chapter 8). Torgo and Lopes (2011) demonstrated the economic superiority of utility-based fraud detection mod-

els with artificial and foreign trade data, and their approach is transferred to and modified for the e-commerce mail order business for the first time. When working with computer-aided systems and facing limited manual inspection resources, substantial savings can be achieved by using a utility-based model for identifying suspicious transactions.

1.2 THESIS STRUCTURE

The thesis consists of the following nine chapters: introduction, theoretical foundation, methodology, modeling fraud, data preparation, feature analysis, prediction model, utility approach, and conclusions. In the following paragraphs, the contents of these chapters are briefly described.

The theoretical foundation is presented in Chapter 2 and covers a range of aspects: First, approaches to fraud prevention in the e-commerce mail order business and — due to the scarcity of such publications — in related areas are discussed. Second, a conceptual framework taken from existing research on financial fraud detection is introduced and developed further in order to build a foundation for choosing the methods that are used in this thesis. Third, research about a utility-based approach to fraud detection is introduced and transferred to the e-commerce mail order business.

Chapter 3 focuses on methodology and contains four components: a justification of logical positivism as the research paradigm, a brief overview of methods, a discussion of the delimitations and limitations of the research design, and a discussion of ethical aspects. In order to keep the method explanations close to their actual application, methods are described in

detail in their respective chapters and not in the methodology chapter.

In Chapter 4, interviews with fraud prevention experts are presented in order to answer essential questions about fraud, including but not limited to the following questions: How are fraudulent transactions identified? What actions are taken if fraud is detected? How can fraud be distinguished from low creditworthiness?

A quantitative analysis can only be as good as the data fed into it. Thus, even though the thesis focuses on quantitative research, the qualitative part is an essential component of the research in order to understand how the transaction dataset was generated and which restrictions it is subdued to.

Data preparation is an important part of data mining, which is why the steps conducted to collect, select, and transform the data are described in Chapter 5. It may be considered an extension of the methodology chapter: The dataset is introduced, available features to identify fraud are described, and the data transformation steps — including handling of missing values and discretization of numeric features — are explained.

Feature analysis is conducted in Chapter 6 and aims at exploring to what extent the available features can help detect fraud cases. Two methods are applied: information gain and a decision tree. The former yields a value per variable, and the latter assigns a fraud risk estimate to combinations of feature values.

In Chapter 7, a prediction model is developed that can be used to forecast the fraud risk of future transactions. The quantitative dataset contains fraud reference labels (i.e. a flag whether a transaction is fraudulent or legitimate) from which the prediction model can learn. In this thesis, two techniques are used: logistic regression and gradient boosted trees.

For determining which transactions should be selected for inspection — more precisely: the order in which they should be inspected until inspection resources are exhausted —, expected utility is proposed as a ranking strategy in Chapter 8. This is an alternative to simply maximizing the detection rate. The estimated fraud probabilities, which are taken from the models built in the previous chapter, margins, and sales prices are used to compute expected financial utility, which transactions are sorted by. Then, the new, utility-based approach to ranking transactions is compared to the existing approach that considers the estimated risk only.

2

THEORETICAL FOUNDATION

2.1 INTRODUCTION

In this chapter, the literature review is presented, which aims at describing the current body of knowledge related to e-commerce mail order fraud prevention, and a conceptual framework is introduced that lays the theoretical foundation for the following empirical parts of the thesis. This chapter is organized as follows: First, it is explained how literature was selected and an overview of the literature relied upon is given. Second, publications that focus on fraud prevention in similar areas are presented and related to the research problem. Third, the act of fraud prevention is expressed as a decision-theoretic problem and data mining is proposed as an approach to the research problem. Both a conceptual data mining framework and methods most commonly used for fraud detection are presented. Fourth, utility theory is introduced in order to view the research problem from an economic perspective. Finally, the chapter closes with conclusions concerning the literature review itself and the contents of the presented literature.

To start with, however, it is discussed how the term fraud is defined and interpreted in this thesis. Ngai et al. (2011) point out that no universally accepted definition for fraud exists. According to Hinneburg (2006), fraud occurs when criminal intent and opportunity appear together. Most definitions contain the element of intentional deception and financial damage: "In the broadest sense, fraud can encompass any crime for gain that

uses deception as a principal modus operandi" (Spann, 2014, p. 1). In the Oxford Dictionary, fraud is defined as the "crime of deceiving someone to gain money or personal advantage" (Waite, 2012, p. 286). Coderre describes fraud as "a variety of acts characterized by the intent to deceive or to obtain an unearned benefit" (Coderre, 2009, p. 3), and the German law states that a fraudulent act must show certain characteristics including intention, deception, and financial harm (Hinneburg, 2006).

For a transaction to be considered fraudulent, financial damage is insufficient evidence because, if a customer faces an unforeseen financial bottleneck and is unable to clear his or her debts, the problem rather originates from a low creditworthiness than a criminal intention. These cases should not be targeted. Therefore, financial damage alone does not imply fraud.

Thus, fraud is defined as the practice of buying products online without the intention to pay for them, often communicating fake data to the online merchant. Using this definition, transactions can be analyzed for suspicious patterns that might indicate such an intention.

2.2 LITERATURE SELECTION

The context of the thesis can be described with the three pillars mail order business, fraud management, and data mining, shown in the Venn diagram in Figure 1. The reader is invited not to view the model as an unquestionable truth but rather as an illustration supporting an understanding for the position of the research problem in the literature.

In order to identify the existing body of knowledge, resources were searched online and offline. Regarding internet searches, LibrarySearch and Google Scholar were used. Both

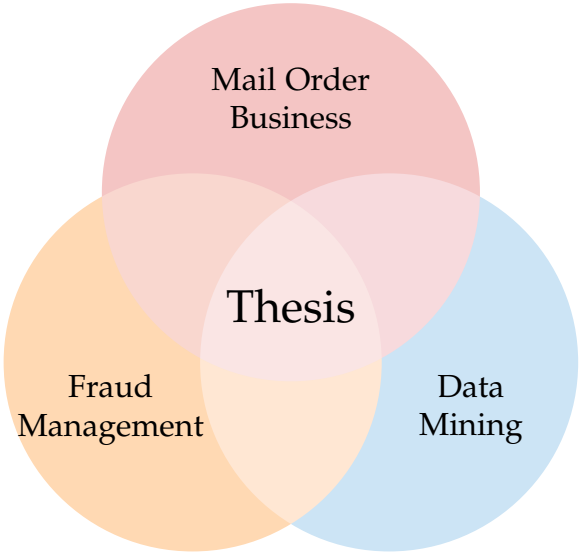

Figure 1: Visualization of the thematic focus of the thesis, which is located at the intersection of the three areas e-commerce mail order business, fraud management, and data mining.

services allow to search multiple digital and physical databases at once. Individual databases such as ScienceDirect, Emeraldinsight, and Springer Link were used. In addition, a set of journals related to e-commerce has been worked through directly, among them Decision Support Systems, Electronic Commerce Research, Expert Systems with Applications, the International Journal of Electronic Commerce, the International Journal of Electronic Business, the Journal of Electronic Commerce in Organizations, and the Journal of Electronic Commerce Research. Regarding offline search, local libraries in Hamburg and in Edinburgh were consulted.

Almost all searches contained at least one of the three terms *fraud detection*, *fraud prevention*, and *fraud management*. At the beginning, statistics about the e-commerce sector in Germany were looked through; then, the search was narrowed down to statistical data about fraud in the e-commerce mail order business. Regarding the theoretical foundation, literature about decision theory, utility theory, and risk management was searched

for applications in fraud detection. It was investigated what fraud risk indicators exist, both from a qualitative and a quantitative point of view. From the latter perspective, finding such fraud risk indicators is accomplished as part of feature analysis or feature selection in data mining. In terms of fraud classification and prediction models, unsupervised, supervised, and semi-supervised machine learning techniques were searched. The conceptual differences between these approaches are explained in the data mining part of this chapter. Supervised machine learning is the most obvious choice in this thesis because high quality fraud labels (i.e. fraudulent or legitimate) are available as part of the quantitative dataset, and supervised methods use such labels for learning which transactions are fraudulent. Approximately half of the resources considered in the thesis are journal articles and conference proceedings, and the other half mostly consists of specialized textbooks about data mining and economics.

Research about e-commerce mail order fraud prevention is still scarce. Therefore, the next sections move from the general to the specific in order to derive the research gap from the current body of knowledge.

2.3 FRAUD PREVENTION IN RELATED AREAS

Phua et al. (2010) state that fraud detection in general has received little attention. Ngai et al. (2011) explain that insurance fraud has been most popular, but other fraudulent activities deserve to be researched more deeply. However, the majority of the academic papers considered in this thesis has been published within the last ten years, indicating that the awareness of

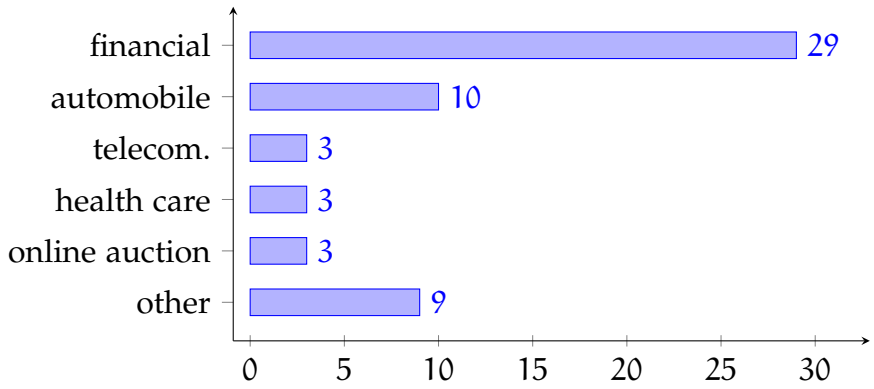

Figure 2: Overview of the areas for which publications about data mining methods used for fraud prevention have been retrieved (number of publications).

the importance of fraud detection has risen — at least in related areas.

Accompanying the rise of machine learning and big data in various industries (McAfee and Brynjolfsson, 2012), there are many publications that investigate the use of data mining for fraud prevention. Some of these areas are close to the research problem, such as dealing with credit card fraud, and others are farther apart, such as detection of management fraud. These related areas can serve as an orientation and inspiration for the design of the thesis, and in this section, an overview of them is presented. Later in this chapter, a data mining framework with fraud detection in mind will be introduced, providing a taxonomy for the approaches presented.

Figure 2 shows the distribution of publications which use data mining techniques for fraud prevention across different areas. The overwhelming majority of papers focused on financial fraud (i.e. credit card fraud, bank account fraud etc.), including but not limited to Halvaiee and Akbari (2014), Hartmann-Wendels, Mählmann, and Versen (2009), Jha, Guillén, and Westland (2012), Krivko (2010), Mahmoudi and Duman (2015), Sahin, Bulkan, and Duman (2013), Sánchez et al. (2009), Van

Vlasselaer et al. (2015), and Yeh and Lien (2009). Another well documented area is automobile insurance fraud (Bermúdez et al., 2008; Brockett, Xia, and Derrig, 1998; Brockett et al., 2002; Caudill, Ayuso, and Guillén, 2005; Viaene, Dedene, and Derrig, 2005; Viaene et al., 2002, 2007). Note that a couple of authors contributed to multiple of such publications. Telecommunication fraud (Hilas, 2009; Hilas and Mastorocostas, 2008), health care fraud (He et al., 1997; Li et al., 2008), and online auction fraud (Chang and Chang, 2014; Dong, Shatz, and Xu, 2009; Zhao et al., 2016) are each the focus of a couple of papers. Other papers include management fraud (Fanning, Cogger, and Srivastava, 1995), accounting fraud (Durtschi, Hillison, and Pacini, 2004), and ATM fraud (Li et al., 2012). Credit assessment was the focus of four papers and could be of interest as well (Chen, Ma, and Ma, 2009; Shi, Liu, and Ma, 2011), although not directly a problem related to fraud, as low creditworthiness alone does not suffice for a transaction to be considered a fraud case. In addition, some resources provide pragmatic advice: Mena (2003) discusses data mining in the context of crime detection, Spann (2014) introduces guidelines for fraud analytics, and Coderre (2009) focuses on online payment security and fraud prevention.

Applicability of such resources to the research problem is limited, however, because the contexts, i.e. features and processes, are often too dissimilar from the e-commerce mail order business. Carneiro, Figueira, and Costa (2017, p. 92) note that it is difficult to transfer fraud detection strategies from other areas, such as banking, to the e-commerce mail order business: "An important difference to the banking sector is that a single online retailer has very limited information about the customer it is doing business with. Moreover, few authors have dealt with the discussion of practical aspects such as which data to use." Such

resources have mostly been used to provide motivation for the choice of supervised machine learning methods presented in Chapter 7.

2.4 E-COMMERCE FRAUD PREVENTION

2.4.1 *Overview of Literature*

In the previous section, it has been concluded that fraud prevention approaches from other areas cannot be transferred to the e-commerce mail order business without restrictions. Therefore, it is important to consider fraud prevention literature that focuses on this particular area. Yet, it is found that such literature is much scarcer than fraud prevention literature in general.

Among the searched literature, only three studies focus on fraud prevention in the e-commerce mail order business. Carneiro, Figueira, and Costa (2017) investigated the use of data mining for partial automation of e-commerce mail order fraud prevention and refer to Brabazon et al. (2010) as the only existing source prior to their own research. Hinneburg (2006) undertook a qualitative study based on interviews to research e-commerce mail order fraud prevention in Germany.

Each of these studies has its limitations, though. Hinneburg (2006) researched fraud prevention from a rather descriptive point of view and did not explore computer-aided systems. In addition, her results were published more than ten years ago. Hinneburg (2006) mentions that the share of orders made online lies between 9 and 20 percent but the internet has become the dominant sales channel: With regard to the company that provided the dataset for the quantitative analyses, online sales represent more than 90% of total revenue (Anon, 2015a). Nev-

ertheless, Hinneburg (2006) provides insights about motivation and demographic patterns of e-commerce fraudsters, which are referred to in Chapter 6. Official statistics about e-commerce mail order fraud are generally available (Bundeskriminalamt, 2016), but statistics about fraud prevention efforts are rare. Carneiro, Figueira, and Costa (2017) research the use of multiple supervised machine learning techniques for fraud prevention in the e-commerce mail order business but limit their research to credit card fraud. More importantly, none of these publications addresses economic utility.

Therefore, from a business perspective, the aforementioned studies cover only parts of the problem. Companies usually strive for profit maximization — besides other possible aims such as sustainability. Understanding fraud (Hinneburg, 2006) is the first step to successful fraud prevention, and automating fraud detection (Brabazon et al., 2010; Carneiro, Figueira, and Costa, 2017) helps deal with the high volume of transactions large online retailers face every day. Yet, the total financial impact of fraud matters more than the pure number of fraud cases. Even fraud prevention departments usually measure their performance with revenue-based numbers (cf. Chapter 4). Therefore, fraud prevention should also be optimized with the financial outcome in mind. Torgo and Lopes (2011) have demonstrated the value of this perspective with artificial and foreign trade data. It seems obvious to transfer this approach to the e-commerce mail order business, but this has not been done so far.

2.4.2 *Approaches to Fraud Prevention*

Fraud prevention in the e-commerce mail order business is a difficult task for a number of reasons: Hinneburg (2006) states that detecting fraud poses a major challenge because only a fraction of the transactions that lead to payment default are actually fraud, and identifying the fraudulent transactions as such requires a substantial effort, leading to a conflict between effortless customer service and fraud prevention. The reason is that the two interests collide when transactions of honest customers are investigated and possibly stopped by mistake. Therefore, companies strive for pragmatic solutions, and the companies that have many years of experience in dealing with fraud have an advantage over those that do not (Hinneburg, 2006). In addition, "[f]raud detection is a challenging problem because fraudsters make their best efforts to make their behavior look legitimate" (Carneiro, Figueira, and Costa, 2017, p. 91).

E-commerce businesses can try to reduce the impact of fraud in a variety of ways — automated fraud detection systems are one of them. For example, they can insure themselves against fraud (Carneiro, Figueira, and Costa, 2017). Of course, this does not prevent fraud from happening and only works when there is little fraud because insurance fees would explode otherwise. Another approach is to require safe payment methods such as prepayment. Yet, since many honest customers appreciate or even demand payment via invoice, such actions could deter them from ordering as well (Anon, 2015a).

Among more sophisticated approaches are two-factor phone authentication, address verification services, and payer authentication (CyberSource, 2017). Yet, many of these methods heavily increase friction, i.e. they make the e-commerce experience less pleasurable for the customer. Companies therefore hesitate

to implement such methods (CyberSource, 2017). In contrast, fraud scoring models operate in the background, which makes them less intrusive and therefore more attractive. In order to determine whether the transaction is legitimate or fraudulent, they focus on gathering and evaluating data about the customer, the transaction, or technical information about the device from which the transaction has been issued, called device fingerprinting (Boda et al., 2012).

Often, suspicious transactions are inspected manually. However, due to the high throughput of orders large e-commerce companies face, it would be economically infeasible to inspect all transactions manually. Hinneburg (2006) refers to the use of data mining as a supporting element, but her research dates back to 2006. Today, data mining is a growing area due to the rapid advancement of technology (Witten, Frank, and Hall, 2011). Most transactions are processed without further action and only a small share is inspected by hand; CyberSource (2017) published an average review rate of 25% and even lower rates for larger companies — such as the one that provided the dataset used in this thesis. The smaller the fraction of manually inspected transactions, the more important automation through algorithms becomes. Algorithms can help detect fraud cases and, as experts state that active fraud prevention complicates the work of criminals, could also help deter criminals in general (Anon, 2015a).

According to Torgo and Lopes (2011, p. 1517), "[f]raud detection usually involves two main steps: [...] decide which cases to inspect, and [...] the inspection activity itself". The first step is important because inspection resources are limited and should be allocated in the best possible way. The strategy of automated scoring and manual inspection of suspicious cases is also presented by Carneiro, Figueira, and Costa (2017).

However, only few studies about fraud detection systems are actually implemented, and "none of them has considered the combination of data mining and manual revision in the whole system" (Carneiro, Figueira, and Costa, 2017, p. 92).

Of course, this is only one strategy how human experts and automated systems can interact with each other. For example, computer-aided analyses could as well support the manual inspection process itself instead of being prepended to it, or there could be applications in which no manual inspection exists at all and the process is fully automated. The fraud prevention process will be investigated as part of the empirical work in Chapter 4.

Also, fraud detection software can make mistakes which might harm the customer relation, and customers might be concerned about data privacy because algorithms can easily scan large datasets. Thus, such algorithms should be well-understood and monitored. In Germany, "[c]oncerns regarding confidentiality and security of online payment are slowly diminishing as more and more people use this channel on a regular basis" (Euromonitor International, 2014, p. 63). This development could lead to a broader acceptance of fraud prevention if consumers were less concerned about the use of their data for protection against crime. It could even be possible that fraud prevention measures which are communicated to the customers lead to an increase in customer trust and loyalty, as indicated with regard to credit card fraud (Carneiro, Figueira, and Costa, 2017) and in the banking sector (Hoffmann and Birnbrich, 2012). Although not concerned with fraud prevention in particular, Ahrholdt (2011) found that consumer trust is positively influenced by data privacy and security measures communicated by mail order companies. The negative consequences of fraudulent activities are more critical, though, and affect both companies

and customers, for example when accounts are stolen from honest customers in order to commit a fraud or legitimate sales are rejected as fraud by mistake. The trade-off faced in fraud prevention has been described regarding credit card fraud: "Fraud detection involves a fundamental trade-off. On the one hand, the company has to minimize the level of fraud, maximizing the detection of fraudulent transactions, thus avoiding chargebacks. On the other hand, it must provide high payment acceptance rates in order to convert as many sales as possible and minimize the number of customer insults (i.e. the number of legitimate transactions refused)" (Carneiro, Figueira, and Costa, 2017, p. 92).

In general, McAfee and Brynjolfsson (2012) emphasize that data-driven analyses allow to make better decisions and should override or at least support gut feelings and intuition. Fraud officers face the same conflict when their own knowledge contradicts predictions based on data mining (as pointed out in Chapter 4). Acknowledging this issue is one of the reasons why the empirical part of this thesis contains both qualitative and quantitative analyses.

2.4.3 *Implications for the Research Problem*

The research problem has been developed through the following line of argument so far: Having pointed out the financial impact of fraud in the e-commerce mail order business in Chapter 1, it has been shown in this chapter that numerous research articles have been published with regard to fraud prevention in related areas. However, it has been argued that such resources cannot be applied to the e-commerce mail order business without limitations and are therefore of limited use.

Only three studies were found which focus on fraud prevention in the e-commerce mail order business, and only two of them follow quantitative approaches in order to provide pragmatic solutions to the problem. None of these publications address that automated fraud prevention using data mining could and should be guided by economic principles. Alternative approaches to fraud prevention have been discussed, but using data mining for supporting manual fraud detection is considered the most appropriate one, which is not intrusive and still offers a pleasurable e-commerce experience to the customer.

Therefore, in the next three sections, important concepts are going to be presented that will be referred to throughout the thesis: Decision theory provides a conceptual model for expressing the fraud prevention problem in a formal way; a data mining framework is going to be introduced in order to provide an overview of methods and concepts that could be used for fraud prevention; lastly, it will be explained how economic goals can shape fraud prevention.

2.5 DECISION-THEORETIC REPRESENTATION

In the e-commerce mail order business, fraud prevention departments of large companies process thousands of transactions every day. For every transaction, it has to be decided whether the transaction should be accepted or rejected (other possible actions are going to be investigated in the qualitative part of this thesis) and, in order to make this decision, the transaction has to be analyzed either automatically, manually, or both. This thesis aims at both finding out how fraud prevention is accomplished at the companies which provided the transaction data and, assuming a rational decision maker, deriving how it

should be accomplished. Both aspects can be related to decision theory. The former aspect relates to descriptive decision theory, and the latter relates to normative decision theory. "Descriptive decision theories seek to explain and predict how people *actually* make decisions. [...] Normative theories seek to yield prescriptions about what decision makers are *rationally required* — or *ought* — to do" (Peterson, 2009, p. 3).

Decisions are "*rational* if and only if the decision maker chooses to do what she has most reason to do at the point in time at which the decision is made" (Peterson, 2009, p. 5). The assumption of rational behavior seems reasonable in the context of fraud management as a business problem, and it supports the objectivist approach to the research problem.

The presence of uncertainty is an essential part of the fraud prevention problem. Obviously, if there were no ambiguity about the truth, companies would stop all fraudulent transactions and limit their services to honest customers. Yet, criminals rely on deception, and thus companies have to make decisions with incomplete or erroneous data. Peterson (2009) defines two categories for decisions under uncertainty: decisions under risk and decisions under ignorance. For the former, the probabilities of the possible outcomes are known. For the latter, probabilities are not known.

The fraud prevention problem can be decomposed into two steps: estimating the probability that a transaction is fraud (or generally guessing its label, i.e. classifying the transaction) and deciding how to deal with the transaction based on the estimate, facing a problem under uncertainty. The first problem is a data mining problem, while the second is a decision problem.

Decision problems can be described using three components: states, acts, and outcomes (Peterson, 2009). In fraud prevention, there are two states: *fraudulent* and *legitimate*. This essential

	reject	accept
fraudulent	correctly reject	accept by mistake
legitimate	reject by mistake	correctly accept

Table 1: Decision matrix for fraud prevention.

information is unknown, and thus the decision maker has to guess the state. In a simple model, the two possible acts are *accept* or *reject*, based on the guess. Then, considering all possible combinations of states and acts leads to one out of four outcomes. This is shown in Table 1 (columns are acts and rows are states). In an optimal scenario, all fraud cases are rejected, all legitimate transactions are accepted, and no mistakes are made. Of course, this is difficult to achieve. The simple model is used by Torgo and Lopes (2011), while Carneiro, Figueira, and Costa (2017) introduce calling the customer as an additional action in order to verify the fraud status of transactions the status of which is unknown.

This simple decision-theoretic model shall serve as a foundation throughout the thesis. It will be discussed in Chapter 4 based on the interviews. The concepts of utility and expected value, which will be introduced in Chapter 8, build upon the model. First, however, a conceptual framework about data mining for fraud detection is presented.

2.6 DATA MINING FRAMEWORK

2.6.1 *Introduction*

In this section, a conceptual framework is presented according to which the methods used in this thesis are categorized. First, however, the term data mining requires a closer inspection. The

distinction between data mining, statistics, and machine learning will be discussed in Chapter 7.

> *In* data mining, *the data is stored electronically and the search is automated — or at least augmented — by computer. Even this is not particularly new. [...] What is new is the staggering increase in opportunities for finding patterns in data. The unbridled growth of databases in recent years, databases for such everyday activities as customer choices, brings data mining to the forefront of new business technologies (Witten, Frank, and Hall, 2011, p. 4).*

Ngai et al. (2011) present a classification framework for data mining methods in fraud detection, and they define two dimensions, application classes and data mining methods. Applications classes describe conceptual categories for solving data mining problems; data mining methods represent approaches to solving such problems. The following two sections present the two categories and corresponding literature. Other studies that provide an overview of fraud detection literature study are presented by Bhattacharyya et al. (2011) and Phua et al. (2010).

2.6.2 *Application Classes*

According to Ngai et al. (2011), application classes are classification, regression, prediction, clustering, outlier detection, and visualization. Although they are presented as alternatives, these classes are not generally interchangeable. For example, if a problem requires classification or prediction, choosing one of the two classes is almost always sufficient. In contrast, visualization can enhance an analysis and help understand results. In this section, the application classes are explained briefly, and they are referred to e-commerce mail order fraud literature.

The term classification refers to assigning a variable a nominal value (Ngai et al., 2011). In fraud detection, the two classes are usually *fraudulent* and *legitimate*. In the majority of the publications regarding both the literature of this thesis and the research of Ngai et al. (2011), classification is the primary goal and has been used in a number of studies (Brockett et al., 2002; Fanning, Cogger, and Srivastava, 1995; Humpherys et al., 2011; Phua, Alahakoon, and Lee, 2004; Sahin and Duman, 2011; Sahin, Bulkan, and Duman, 2013). However, instead of only guessing whether a transaction is fraudulent or not, sometimes the fraud risk, i.e. a probability estimate, is preferred in order to calculate risk-based measures (Yeh and Lien, 2009), such as shown in Chapter 8 when expected value is computed for ranking transactions according to an estimated utility value.

The aim of regression is to "determine the relationship between two (or more) variables so that we can gain information about one of them through knowing values of the other(s)" (Devore and Berk, 2012, p. 613). Regression, in contrast to classification, relies on continuous inputs and outputs. Logistic regression is one of the most popular data mining methods and will be explained and referenced in the next subsection. In Chapter 7, logistic regression and gradient boosted trees are trained to estimate fraud probabilities of transactions using historical data. Their performance is tested with historical data as well but their purpose is to predict the fraud status of future transactions. Since their output is continuous-valued, both methods are used in the sense of regression. However, tree-based models can be used for both regression (Torgo and Lopes, 2011) and classification (Sahin and Duman, 2011). Zhou and Kapoor (2011) review classification and regression approaches used for detecting financial statement fraud.

Regarding prediction, the values of continuous variables are estimated based on past experience. From a conceptual perspective, prediction resembles regression except that prediction is based on a temporal relation while regression is often based on a causal one. In this research and in contrast to Ngai et al. (2011), the term prediction shall not be limited to forecasting continuous-valued attributes. Thus, a classifier may be used to predict (categorical) future values. This is merely a semantic preference, which has no technical implications.

Clustering differs from the above terms as it "is used to partition objects into previously unknown conceptually meaningful groups" (Ngai et al., 2011, p. 562). In contrast to classification and regression, unlabeled data are used for the process. Groups (or classes) are defined within the clustering process, and each data point is assigned to one of the groups. Clustering does not allow to *detect* fraud directly; instead, it aims at exploring data and detecting new relationships. Even if the fraud status were already known, clustering could yield additional insights about patterns in the data. Combining such unsupervised techniques with supervised approaches leads to semi-supervised machine learning (Chapelle, Schölkopf, and Zien, 2010). Clustering is used in Chang and Chang (2014) and Hilas and Mastorocostas (2008).

Similar to clustering, outlier detection does not utilize data about group membership. Instead, it identifies values that are different from most of the data. In contrast to clustering, which separates data into groups, regarding outlier detection, it is rather assumed that there are two types of behavior, normal and anomalous. For example, Aggarwal (2013, p. 2) refers to the use of outlier detection in detecting credit card fraud, where "unauthorized use may show different patterns, such as a buying spree from geographically obscure locations. Such patterns

can be used to detect outliers in credit card data" (Aggarwal, 2013, p. 2). Ngai et al. (2011) state that outlier detection represents only 2.0% of the articles and conclude that this powerful method should be used more often. In e-commerce mail order fraud detection, most variables are categorical (see variable descriptions in Chapter 6). Outlier detection methods are natively designed for numeric attributes as they often rely on distance measures, but they can be adjusted to work with categorical data, even though a couple of challenges are faced; for example, dummy coding becomes infeasible when the attribute holds too many different unique values (Aggarwal, 2013). Outlier detection is performed by Torgo and Lopes (2011).

Finally, Ngai et al. (2011) introduce visualization as a term for presenting data in a comprehensible way. A slightly more demanding definition could include that visualization in data mining aims at generating value through presentation. Edward Tufte is an American statistician who pioneered the field of data visualization, and his work, *The Visual Display of Quantitative Information*, has influenced the design of this thesis in order to create meaningful data visualizations (Tufte, 2001).

Chapter 7 describes the use of two prediction methods for estimating the fraud risk as a probability value. Prediction is preferred to classification due to two reasons: First, the estimated fraud risk is used to derive a list for manual inspection that shows the most suspicious cases at the top of the list. Second, the estimated fraud probability will be used to compute expected value in Chapter 8. Probability estimates can not only be achieved through prediction methods, though; Torgo and Lopes (2011) computed similar scores through outlier rankings. However, prediction methods use the fraud labels of historical data in order to distinguish between fraudulent and legitimate transactions, which outlier detection methods do not.

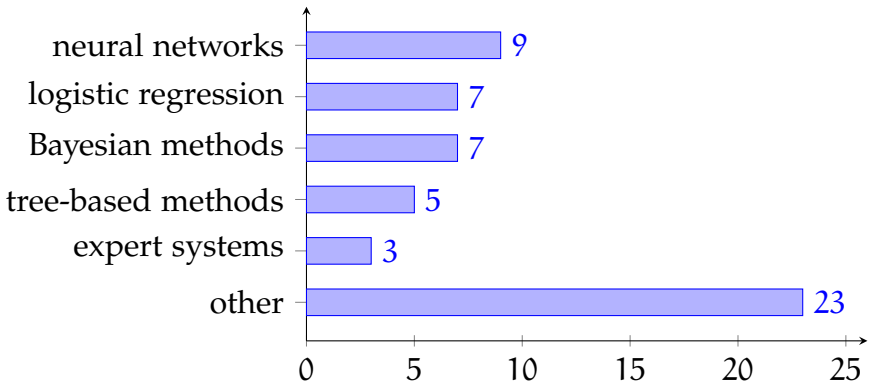

Figure 3: Overview of the most popular methods for fraud detection (number of publications).

2.6.3 *Data Mining Methods*

In this subsection, the most popular methods in financial fraud detection are described. Figure 3 shows how often these methods were used in the literature. In the next paragraphs, Bayesian methods, neural networks, regression, and tree-based models are introduced briefly.

"Bayesian reasoning provides a probabilistic approach to inference. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data" (Mitchell, 1997, p. 154). While Bayesian belief networks and naïve Bayes are two different methods (Sharma and Panigrahi, 2012), both share the concept mentioned above. However, as Mitchell (1997) points out, naïve Bayes requires a fundamental assumption of conditional independence of the input variables, which is loosened in Bayesian belief networks. Bayesian models are used for credit scoring (Antonakis and Sfakianakis, 2009), automobile insurance (Bermúdez et al., 2008; Viaene, Dedene, and Derrig, 2005), and ATM fraud (Li et al., 2012). With Bayesian networks, it is possible to model complex relationships between variables, and

it is possible to obtain estimates for the states of all the nodes in the network. In fraud prevention however, one is only interested in the fraud risk. Therefore, because Bayesian networks address more complex problems from the perspective of causal inference, they are not used in this thesis. Naïve Bayes, in contrast, seems to have no essential benefit over logistic regression, and thus it is also not employed in this thesis.

Artificial neural networks mimic the human brain. A network consists of interconnected units that each take a real-valued input and produce a real-valued output (Mitchell, 1997). Ngai et al. (2011, p. 563) state that "[neural networks] are adaptive; second, [they] can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are chiefly applied to credit card, automobile insurance[,] and corporate fraud." Neural networks come up in Antonakis and Sfakianakis (2009), Bose and Mahapatra (2001), Burge and Shawe-Taylor (2001), Fanning, Cogger, and Srivastava (1995), He et al. (1997), Lei and Ghorbani (2012), Ravisankar et al. (2011), and Serrano et al. (2012). Unfortunately, neural networks are difficult to interpret (Mitchell, 1997). However, when a transaction is considered fraudulent in particular, it is important to understand what led to that assessment. Because neural networks cannot provide such an explanation, they are not employed in this thesis.

Within the area of regression, various models exist, such as linear and logistic regression (Hastie, Tibshirani, and Friedman, 2013). Logistic regression relies on a linear model as well, but nonlinear models exist that are far more complex (Backhaus et al., 2011). Both regarding the literature collection of this thesis (Antonakis and Sfakianakis, 2009; Artís, Ayuso, and Guillén, 2002; Bell and Carcello, 2000; Bhattacharyya et al., 2011; Ravisankar et al., 2011) and the systematic review by Ngai

et al. (2011), logistic models are among the data mining techniques employed most often. Logistic regression is one of the techniques presented in Chapter 7 because it allows to reduce method-induced bias when comparing results with those of other studies.

A variety of tree-based models exists, the simplest of which are decision trees. "Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. [Learned trees] have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants" (Mitchell, 1997, p. 52). Trees are composed of if-then rules, increasing the method's comprehensibility. Decision trees rank fourth in the review of financial fraud detection methods (Ngai et al., 2011) and appear in various studies (Antonakis and Sfakianakis, 2009; Chang and Chang, 2012; Sahin and Duman, 2011; Sahin, Bulkan, and Duman, 2013; Viaene, Dedene, and Derrig, 2005). Gradient boosted trees, which depict the second technique described in Chapter 7, are ensembles of decision trees and can overcome some of the drawbacks that individual trees have. They can handle data of mixed types well, have high predictive power, and are robust to outliers (Pedregosa et al., 2011). Random forests, a similar method, have been used for fraud detection in the e-commerce mail order business (Carneiro, Figueira, and Costa, 2017) and for credit assessment (Shi, Liu, and Ma, 2011). Depending on the metrics used for evaluation, gradient boosted trees and random forests were shown to achieve results of similar quality (Caruana and Niculescu-Mizil, 2005).

The list of data mining methods is not exhaustive. Ngai et al. (2011) described only the most frequently used techniques in their framework; other techniques include support vector ma-

chines (Bhattacharyya et al., 2011; Chen, Ma, and Ma, 2009), expert systems (Hilas, 2009; Leonard, 1995), and association rule mining (Li et al., 2012; Tackett, 2013). In Chapter 7, a fraud detection model will be built using logistic regression and gradient boosted trees. At that point, these two methods will be explained in detail, including motivation for choosing them.

2.6.4 *Degree of Supervision*

The following aspects extend the classification framework introduced by Ngai et al. (2011), distinguishing between supervised, semi-supervised, and unsupervised learning.

First, supervised learning algorithms are fed historical data that contain a reference label (also called just labels in the following) such as whether a transaction is fraudulent or legitimate, i.e. a ground truth. Then, the algorithm learns to infer the fraud status of new cases from the patterns extracted from the historical data by awarding correct predictions and punishing mistakes. "Supervised methods have dominated the fraud detection literature" (Carneiro, Figueira, and Costa, 2017, p. 93). However, the authors add:

> *Often we do not know which class an observation belongs to. For example, take the case of an online order whose payment was rejected. One will never know whether this was a legitimate order or whether it had been correctly rejected. Such occurrences favor the use of unsupervised methods, which do not require data to be labeled [i.e. to contain a ground truth]. These methods look for extreme data occurrences or outliers. In order to get the best of two worlds, some solutions combine supervised and unsu-*

> *pervised techniques (Carneiro, Figueira, and Costa, 2017, p. 93).*

Such a combination of both approaches is semi-supervised learning, which can be appropriate when both techniques would not lead to the best results individually (Chapelle, Schölkopf, and Zien, 2010).

In this thesis, supervised machine learning is employed because reference labels of high quality are available. Including unsupervised methods through a semi-supervised approach could provide additional value to the analyses but is beyond the scope of this thesis since unsupervised methods form their own research area.

The above framework has been introduced in order to offer an overview of data mining and how it is used in fraud detection. As pointed out earlier, fraud *detection* is only one component of fraud management. Therefore, an economic perspective on fraud prevention is proposed in the next section.

2.7 ECONOMIC PERSPECTIVE

In almost all literature about data mining for fraud detection — including Carneiro, Figueira, and Costa (2017) —, performance of fraud detection models is measured with metrics based on the number of correct and incorrect classification (traditional evaluation metrics are presented in Chapter 7 and in the appendix). Antonakis and Sfakianakis (2009), who are concerned with credit risk estimation, refer to the *percentage of correctly classified* cases, which is the *accuracy* (Witten, Frank, and Hall, 2011). This is the amount of true negative (correctly labeled legitimate) and true positive (correctly labeled fraudulent) cases divided by the total number of transactions. In addition, Antonakis and

Sfakianakis (2009) introduce the *bad rate among accepts*, which is the ratio of undetected fraud cases in relation to accepted cases. Some sources use a different name for the same metric, *false negative rate* (Aral et al., 2012). Others focus on the detection rate, i.e. the share of fraud that is identified as such (Wheeler and Aitken, 2000; Wong et al., 2012). Bell and Carcello (2000) refer to a confusion matrix (similar to Table 1). In fact, all of these metrics can be constructed by calculating ratios of parts of the confusion matrix. Viaene, Derrig, and Dedene (2004) show that various other metrics exist for evaluation.

Yet, strictly speaking, companies run fraud prevention departments to reduce financial damage, which they achieve through detecting and stopping fraudulent transactions. Therefore, it might make sense to develop algorithms that try to maximize the financial impact of fraud prevention activities instead of the detection rate. An example shows the difference between the two approaches: It is financially more attractive to find one fraud case in which an expensive smartphone was ordered than to find five fraudulent transactions through which pairs of socks were stolen. In general, the higher the value of a transaction is, the more important it is to investigate properly because both an undetected fraud case and a missed sales opportunity (i.e. rejecting a legitimate transaction by mistake) lead to a reduction of profit.

In data mining, the utility-based perspective is not new (Chawla and Li, 2006; Friedman and Sandow, 2011; Yao, Hamilton, and Geng, 2006), and the idea of expected utility has been used for automobile fraud detection (Artís, Ayuso, and Guillén, 1999), for credit card assessment (Chung and Suh, 2009), and for fraud detection in general (Torgo and Lopes, 2011). However, according to the present knowledge, no literature exists regarding its application to fraud prevention in the e-commerce

mail order business. This is surprising because the cost of fraud and the benefit of legitimate sales are clearly imbalanced. The financial damage caused by a fraud case is much higher than the benefit of a successful sale because profit margins are usually below 50% (U.S. Census Bureau, 2014).

Generally, expected value is the sum of all possible outcomes of an experiment weighted by their probabilities (Fahrmeir et al., 2007). If a decision problem can be repeated arbitrarily, expected value indicates how favorable a decision is in the long term. When expected value refers to estimating utility values, it is called expected utility. Regarding fraud detection as a business problem, it makes sense to assume that utility resembles a financial measure.

In their research, Torgo and Lopes (2011) deal with artificial data and with foreign trade data. The authors propose that, when inspection resources for fraud detection are limited — an appropriate assumption regarding the e-commerce mail order business as well —, a utility-based perspective is financially more attractive than a traditional approach. In order to prove their point, Torgo and Lopes (2011) compare the achieved utility of two models with each other. The first is based only on the outlier ranking, and the second uses a combination of the outlier score as a probability and the expected cost and benefit of the fraud prediction activity. The authors show that the utility-based approach is superior to the traditional strategy. In particular, Torgo and Lopes (2011) demonstrate that, when inspection resources are limited, expected utility can lead to much better results than using traditional approaches. Chapter 8 focuses on building upon the ideas of Torgo and Lopes (2011) and transferring them to the e-commerce mail order business.

2.8 CONCLUSION

At the beginning of the chapter, fraud and related terms were defined in order to agree on basic vocabulary: Fraud has been defined as the practice of buying products online without the intention to pay for them, often communicating fake data to the online merchant.

Then, the strategy used for literature selection was explained. Literature has been collected about the e-commerce mail order business, about fraud detection (or detection and prevention, respectively), and about data mining techniques. Literature considered most important covers each of these aspects, i.e. data mining techniques for fraud management in the e-commerce mail order business.

In the following, publications that focus on fraud detection in similar areas were presented and related to the research problem. Even though the interest in fraud detection in general has increased, other areas like insurance fraud are conceptually different from online retailing and therefore insights cannot easily be transferred to the e-commerce mail order business. In addition, although the financial damage caused by fraud is substantial, few scientific publications deal with the e-commerce mail order business in particular, and they only cover parts of the problem. In particular, it has been shown that an economic perspective on e-commerce mail order fraud prevention has not been taken yet.

Furthermore, approaches to fraud prevention were introduced. Companies protect themselves by employing fraud prevention techniques such as manual inspection and computer-aided systems — because the high volume of e-commerce transactions often makes it impossible to inspect every transaction manually. The question was raised which techniques are ap-

propriate for the e-commerce mail order business, also given how manual inspection and automated processing typically interact. It was shown that decision theory was consulted in order to express the decision problem in a structured way. Then, a conceptual data mining framework was presented, consisting of data mining application classes and data mining methods. Both were used to guide the reader through the data mining concepts used to deal with fraud.

Finally, fraud prevention was described as a utility maximization problem. The value of such a perspective has been shown in related areas, encouraging to apply utility theory to fraud prevention in the e-commerce mail order business as well.

Altogether in this chapter, it has been shown that the research question focuses on a scientific gap, and the thesis aims at addressing that gap from multiple points of view. In the next chapter, methodology will be discussed, including research philosophy, limitations, and ethical considerations.

3

METHODOLOGY

3.1 INTRODUCTION

In this chapter, the methodology on which the thesis resides is presented. "A methodology is an approach to the process of the research, encompassing a body of methods." In contrast, "[a] method is a technique for collecting and/or analysing data" (Collis and Hussey, 2014, p. 59). In other words, the methodology sets the frame for which an appropriate set of methods has to be chosen. Note that, although the methods used in this thesis have already been mentioned in the previous chapters, they were derived from the methodology.

This chapter is composed of the following parts: First, the research philosophy is explained, followed by a brief excursion about a philosophical perspective on probability. Second, implications for the research approach are discussed. Third, delimitations and limitations of the research are introduced. Finally, ethical limitations are addressed.

The choice of methods is not discussed in this chapter, which may be unusual. Because mixed methods are applied in the thesis — including quantitative and qualitative research —, it seemed more appropriate to present the methods close to where they are used. Therefore, interviewing theory is found in Chapter 4, a discussion of data preparation is discussed in Chapter 5, theory about feature analysis is presented in Chapter 6, machine learning theory is approached in Chapter 7, and theory about utility-based decision making is discussed

in Chapter 8. This way, the empirical chapters can be read as rather self-contained units, and theory is presented in their respective chapters, which might offer a more pleasant reading experience.

3.2 PHILOSOPHY

3.2.1 *Philosophical Stance*

The point of this section is to show that the choice of methods is justified given the author's philosophical stance. Therefore, first, the chosen epistemology is presented, and second, the ontology is derived from it. Then, the chosen philosophy is compared with other, similar concepts.

In this thesis, logical positivism is chosen as an epistemology and objectivism is chosen as the ontology (Reichenbach, 2013). In the following evaluation it is assumed that the choice of a research philosophy is rooted in one's personal view of what shapes our conception of the world. Researchers might try to keep an objective distance to their research, but they are never truly objective because their personal philosophical assumptions, their perception of the world, and their values influence the way they approach it.

Two questions mark the starting point: Is there a real world that exists independently of one's perception of it (ontology) and how is knowledge gained (epistemology)? It may seem that the former question should precede the latter, but in this elaboration, epistemology is discussed first, and the chosen ontology will become the obvious conclusion of the logical assumptions associated with the epistemological point of view.

To answer any question, we must think. Dewey (2013, p. 6) defines reflective thought as the "[a]ctive, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends." But what supports a belief?

Dewey (2013) declares "past experience" and "prior knowledge" as the "sources of suggestion". In addition, "the scientific genius has never felt bound to the narrow steps and prescribed courses of logical reasoning" (Reichenbach, 2013, p. 5). Accordingly, Benton and Craib (2011) emphasize that mathematical and logical arguments do not provide new knowledge since they are true by definition. "They are statements in which we make explicit the implications of the way we define certain words, or mathematical operations" (Benton and Craib, 2011, p. 4). Logic cannot create new knowledge but only reshape it by deriving consequences from prior assumptions (Dewey, 2013, p. 12). Other sources of knowledge become necessary, such as experience.

But can past experience be trusted? From a philosophical perspective, it possibly cannot. Pyrrho, an ancient philosopher, was one of the first skeptics who kept challenging the consequences of even the most basic of his actions (Warburton, 2011). Although Pyrrho's level of skepticism may not have been sensible, and even though Kant developed arguments that suggest an innate understanding beyond empiricism (Benton and Craib, 2011), the question whether a world independent of one's conception exists still remains unresolved.

Reichenbach (2013) suggests to logically build our perspective on the world from the ground up by accepting certain postulates. As a logical positivist, he proposes that nothing is certain but that there are different degrees of probability regarding the truth. He concludes that, given an internally consistent set

of assumptions, a *real* world seems more probable than a subjective world.

Therefore, an objectivist ontology is derived from logical positivism. The line of argument and the associated assumptions are explained and supported with examples in Reichenbach (2013). A remarkable feature of logical positivism is that the latter is not a necessary consequence of the former, because the ontological implications depend on the logical assumptions made. Thus, logical positivism could — given a different set of assumptions — as well lead to a subjectivist ontology. However, Popper (1997, p. 210) strongly criticizes subjective approaches to science due to their *unconfined applicability* (translated by the author). The importance of an objectivist ontology will be explained later in this chapter.

3.2.2 *Comparison of Philosophies*

The philosophical stance described above shares important ideas with postpositivism and critical realism (Benton and Craib, 2011; Collier, 1994; Phillips and Burbules, 2000). Besides logical positivism, postpositivism as well addresses issues connected with positivism. While logical positivism uses pure logic to explain the limits of reasoning, postpositivism takes a more practical stance and lists the common pitfalls and problems with positivism, allowing the researcher to consider them in his or her research. Phillips and Burbules (2000) state that, regarding his reasoning, Hans Reichenbach — whose understanding of logical positivism forms the basis of this chapter — moves from logical positivism to non-foundationalist postpositivism.

Regarding the methodical consequences, besides postpositivism, critical realism resembles logical positivism as well. Col-

lier (1994) states that everybody is devoted to some kind of "realism". He or she may be "realistic" about sensations, a real world, or other things. According to Collier (1994), propositions of stronger critical realism are objectivity, fallibility, transphenomenality, and counter-phenomenality, which are all concepts that logical positivism can generally support. In essence, reality is not directly accessible, or even if it is, one cannot observe it directly. Therefore, conclusions drawn from observations might be wrong; they might be the consequences of invisible forces, or they might even contradict the rules that have been derived from other observations.

From a practical point of view — and regarding the consequences for this thesis — both critical realism and postpositivism could have been appropriate alternatives. Logical positivism is preferred to them because it provides an elegant approach to research philosophy. With logical positivism, any internally coherent set of assumptions can be chosen as a basis for deriving epistemological and ontological implications.

3.3 CONCEPT OF PROBABILITY

Probability is an essential component of the models developed in this thesis. Hence, a few thoughts should be given to what it really means. Neapolitan (2004, p. 17) presents two approaches to probability, the *relative frequency approach* and the *subjective probability approach*. The relative frequency approach states that if a random experiment is conducted a number of times, and that number approaches infinity, then the fraction of trials in which a certain outcome occurs will approach the probability of the outcome to occur. Thus, "if we tossed [a] thumbtack 10,000 times and it landed heads 3373 times, we would estimate the

probability of heads to be about [33.73%]" (Neapolitan, 2004, p. 18). The author points out that the probability of such an event is a physical property of the object.

In contrast to the relative frequency approach, subjective probabilities are based on beliefs. Neapolitan (2004, p. 18) mentions that, for example, a sequence of numbers can appear more or less random, such as *101110110* and *1010101010*. In this case, intuition can contradict pure combinatorial mathematics. In fraud detection, this aspect could be important when humans decide what looks suspicious. For a frequentist, it would not be possible to include such personal beliefs in a calculation. Neapolitan (2004) states that relative frequencies are an idealization, and even a coin's physical composition changes when it is tossed. Therefore, he points out, an interpretation of probabilities as physical properties can be difficult.

In conclusion, Neapolitan (2004) uses relative frequencies as if they were objective, but he acknowledges their subjective component. In this thesis, the same point of view on probabilities is taken. Chapter 7 shows how machine learning algorithms are trained to estimate fraud probabilities using historical data. It is important to acknowledge that even a quantitative analysis purely based on data — avoiding subjectivity through personal experience about the topic in question — produces probability estimates that are still based on the choices made by the persons who conducted the analysis.

3.4 RESEARCH IMPLICATIONS

In fraud management, it is assumed that an objective truth exists, i.e. that a transaction is either legitimate or fraudulent, but that this objective truth is unknown to the observer. Evidence

can be utilized to estimate the fraud risk of a transaction. For a thesis that aims at providing recommendations on how to conduct fraud prevention in the e-commerce mail order business — partially through automation —, the existence of an objective truth, and thus an objectivist ontology, is essential. This might favor the positivist paradigm, but it does not restrict the research to the use of quantitative methods as the work of Hinneburg (2006) and the qualitative parts of the work of Carneiro, Figueira, and Costa (2017) illustrate.

Often, quantitative and qualitative research form two disjoint areas, and academics prefer one of the two branches of science to the other. However, "it must be emphasized that one cannot decide whether qualitative or quantitative studies are better or more useful. It is important to note that there are no pre-determinates for the appropriateness of either a qualitative or a quantitative study." (Blumberg, Cooper, and Schindler, 2014, p. 148) Therefore, "[m]any problems in business and management research can be researched both ways, qualitative and quantitative, although the answers obtained may have a different nature as the research questions asked depend often on the perspective." (Blumberg, Cooper, and Schindler, 2014, p. 148)

Yet, most of the literature that exists about fraud detection is of quantitative nature as shown in Section 2.3 in Chapter 2. This may be the case because often companies face large amounts of data, and businesses are not only interested in gaining insights but also in creating pragmatic solutions. With data mining, computer-aided systems can be developed that are able to analyze the high volume of transactions.

In order to research the problem in a holistic way, the thesis relies on mixed methods (Collis and Hussey, 2014) and includes both a small qualitative study — in order to understand the transaction data and the fraud management processes as

a preparatory step — and a more complex quantitative part. Knowledge can not only be gained through statistical evaluation of technical data but also through conversations with experts in the field; both are variants of the sources of suggestion Reichenbach (2013) refers to. This is not to be mistaken with triangulation: With the help of triangulation, validity of research results *of the same phenomenon* can be increased (Blumberg, Cooper, and Schindler, 2014). However, the qualitative and the quantitative part answer different questions each and the latter builds on top of the former.

3.5 DELIMITATIONS AND LIMITATIONS

The thesis focuses on e-commerce mail order fraud prevention, but fraud occurs in many other areas, such as insurance fraud or credit card fraud. Therefore, other areas have been consulted for inspiration due to the lack of literature in the field of interest, as pointed out in Chapter 2.

In order to explore the fraud prevention process, practitioners from the e-commerce mail order companies which provided the quantitative transaction data were interviewed (the interview setup will be discussed in detail in Chapter 4). The aim of the interviews is not to explore indicators of fraud risk; this is achieved through feature analysis in Chapter 6. It would not have been feasible to approach other companies' experts as well, because the quantitative core of the thesis relies on the transaction data that are available. Thus, the fraud prevention processes modeled in this thesis may differ from those at other companies, and results may only be generalizable with regard to e-commerce mail order companies of similar structure. In particular, a major assumption made in Chapter 7 and Chap-

ter 8 is that inspection resources are limited, which is why automation through machine learning models is explored in order to deal with large amounts of orders.

The thesis does not aim at providing a representative overview of fraud prevention in the e-commerce mail order business, and it does not aim at extending results beyond the questions faced in this particular area. Moreover, as mentioned by Carneiro, Figueira, and Costa (2017), data mining is only one strategy with which the problem of fraud in the e-commerce mail order business can be approached; examples for other strategies have been mentioned in Chapter 2. In addition to this brief overview, limitations of the research are discussed in each of the empirical parts of the thesis and in Chapter 9.

3.6 ETHICAL CONSIDERATIONS

"The use of data — particularly data about people — for data mining has serious ethical implications, and practitioners of data mining techniques must act responsibly by making themselves aware of the ethical issues that surround their particular application" (Witten, Frank, and Hall, 2011, p. 33).

For fraud detection and prevention, transaction data are regularly analyzed both manually and through automated systems. In this thesis, such data are used to find out which features (variables) may serve as fraud risk predictors and are used to train models that classify transactions as fraudulent or legitimate. Because the results of this thesis may influence the fraud prevention process — and therefore how customers are treated —, it is important to consider ethical problems that may arise as part of the analyses.

Mingers and Walsham (2010) provide an overview of the ethical issues connected with information systems and mention three perspectives on ethics and morality: consequentialism, deontology, and virtue ethics. In consequentialism, the results of an action are what define its value; the deontological approach focuses on whether the act itself is good or bad, i.e. what the intentions were; and virtue ethics emphasize on the attitude of the actor (Mingers and Walsham, 2010).

Regarding this thesis, ethical aspects should be evaluated from a consequentialist perspective because people who suffer from the *consequences* of an unethical act (e.g. accusing an honest customer of trying to commit fraud) are probably neither interested in the act itself nor in the intentions behind it. Thus, even though crime prevention is considered a virtuous aim that justifies the use of data in general, it is essential to deal with the data in an ethically acceptable way. In this section, possible ethical challenges are brought up and it is discussed how they are dealt with.

In the context of fraud management, there are three main issues: First, fraud analysis requires the use of possibly sensitive customer data. Although the German data privacy law is strict about when it is allowed to use such data (Recht, 2017), its ethical appropriateness is not necessarily implied. Therefore, a question to deal with is to which degree the socially accepted goal of crime prevention justifies the means of working with personal information. Second, discrimination could take place: Patterns in the data could possibly discriminate certain socio-demographic groups. Third, data from the interviews, e.g. about fraud officers' performance, could be held against them.

In order to deal with the first issue, the ethical use of sensitive customer data, all data are encrypted and anonymized to a degree that seems sensible; note that more anonymization also

reduces the information value of the data for fraud detection. Most importantly, no personal information is published.

The second issue relates to inductive reasoning through data mining; patterns found through data mining could discriminate certain groups of people. Two aspects address this possible issue: First, the dataset is composed of various kinds of features (see Chapter 5); thus, customer-related features represent only a small fraction of the data. Second, an approach to fraud detection based on data mining is usually regarded as more objective than judgments based on personal experience. According to the German data privacy law, scoring requires a scientifically acknowledged mathematical/statistical procedure (Recht, 2017). Thus, the issue of discrimination might even be reduced through data mining compared to the status quo.

Regarding the third issue, both the names of the interviewees and the companies are kept confidential, and guidelines from Edinburgh Napier University regarding informed consent have been followed. The interviews were transcribed in a summarized form in order to obscure the use of common terms which could be traced back to individual persons.

PART TWO

This part contains the empirical analyses: First, the results of interviews with fraud experts are presented. Second, the transaction dataset is prepared. Third, possible fraud risk factors are examined from a quantitative point of view in the form of feature analysis using information gain and a decision tree. Fourth, two supervised machine learning algorithms, logistic regression and gradient boosted trees, are trained. Fifth, in the last empirical chapter, the estimated fraud risks taken from the previous chapters are fed into a utility-based model in order to minimize the financial impact of fraud. Finally, the thesis closes with conclusions.

4

MODELING FRAUD

4.1 INTRODUCTION

Understanding the context which surrounds the quantitative transaction data is essential, because the results of any algorithm can only be as good as the quality of the input data, but not only that: Processes must be thoroughly understood in order to arrive at meaningful conclusions.

Two large e-commerce mail order companies and one company specialized in fraud prevention solutions have been interviewed in order to investigate their fraud prevention work. First, the interviews aim at determining the status quo of fraud prevention: Which kinds of transactions are considered suspicious? Which actions can be taken to identify fraud? Once a transaction is believed to be fraud, how can it be validated? Second, opportunities for improvement of the status quo are discussed: What is the role of machine learning in fraud prevention? Are there particular requirements regarding the choice of data mining algorithms? Do economic goals guide the fraud prevention work? Third, since one of the interviewed companies provided the quantitative dataset, limitations of the dataset are discussed. How the fraud labels of the transaction data have been created is an essential information because it determines what they can be used for and how analyses that are based on them can be interpreted.

This chapter is organized as follows: First, the company that provided the data for both the qualitative and the quantitative

part is introduced. Then, a theoretical overview of interviewing as a research technique is given and the interview setting is described. In the next step, results are presented. Following a discussion of limitations, the chapter closes with a summary.

4.2 COMPANY OVERVIEW

The company later referred to as company A of the two interviewed companies is of particular importance for this thesis because it was not only involved in the interviews but also provided the transaction dataset used in all quantitative analyses.

This company is one of Europe's largest e-commerce retailers, selling mostly clothing but also electronics, furniture, and *white goods* (e.g. refrigerators and washing machines). Offering millions of articles, the company faces up to 10 transactions per second and had a turnover of €3 billion in the fiscal year 2017/18, of which 90% were generated online. The company employs about 4,500 people. For these data, no references are given in order to maintain anonymity of the company.

The company's customers are mostly middle-aged (42 years on average), and the distribution is right-skewed, meaning that there are many customers slightly below the average age as well and there is a long tail of quite old customers. Recently, the company has started to target a younger customer base in order to maintain its share in a competitive market.

4.3 INTERVIEW SETUP

Although personal interviews produce rather subjective results, which cannot easily be generalized, they were considered most appropriate to explore the fraud prevention process:

> *The greatest value [of personal interviews] lies in the depth of information and detail that can be secured. It far exceeds the information secured from telephone and self-administered studies, such as mail surveys or web surveys. The interviewer can also do more things to improve the quality of the information received than with another method (Blumberg, Cooper, and Schindler, 2014, p. 213).*

Conducting personal interviews allowed to ask the company that provided the quantitative dataset detailed questions about it. Due to this opportunity, personal interviews were preferred to any other research method.

Helfferich (2011) presents a list of decisions which need to be made when planning to conduct an interview. In the following paragraphs, these decisions are introduced and commented on with regard to the research context.

The first task is to define the object of research (Helfferich, 2011). As the first of four empirical analyses, the interviews are supposed to lay a cornerstone for the subsequent parts. The quantitative dataset consists of transactions, which are labeled either fraudulent or legitimate. These fraud labels are set manually by the employees in the fraud prevention department, and they are essential to all quantitative analyses. Hence, it is important to thoroughly understand under which circumstances transactions have been labeled. The act of labeling is the consequence of a process chain, and that process chain is supposed to be investigated in the interviews. Only if it is understood how the fraud prevention process works, it will be possible to develop methods which improve it. Thus, the interviews aim at providing the foundation upon which the quantitative research can be built.

This aim leads to the next aspect mentioned by Helfferich (2011), which is *whom* to interview. In order to be able to relate results of the qualitative to the quantitative part, three interviewees were chosen from two e-commerce mail order companies, both of which work with similar datasets, and one of which provided the dataset for the quantitative part in this thesis. In this chapter, in order to ensure anonymity, these two companies are referred to as A and B. At the time of the interviews, it had not been clarified which of the two companies would allow the use of their data. Therefore, the heads of the fraud prevention departments from both companies who have extensive experience in the field were interviewed under the same premises. In the interview with company B, one of the fraud prevention employees who perform the actual inspection work attended as well in order to provide a more hands-on perspective on the questions asked. The interviewees know the practical problems related to fraud prevention but, at the same time, have a broad view on the departments and their roles within the companies. Collis and Hussey (2014) refer to this sampling method as *natural sampling*, because sometimes "only particular employees are involved in the phenomenon being investigated" (Collis and Hussey, 2014, p. 132). In addition, in a third interview, two account managers at Risk Ident were consulted. This company at which the author is employed sells software for device fingerprinting and fraud prevention. This interviewed aimed at providing a bird's eye view of the results from the two main interviews, as Risk Ident serves companies from different areas, and the account managers have a broad overview of the issues related to fraud prevention. In qualitative studies, it may be common to conduct a much higher number of interviews; Helfferich (2011) mentions that restricting factors are mostly time and money. However, the qualitative part of this thesis aims at

investigating the *specific* setting of the company that provided the dataset. From this perspective, a single interview would have been sufficient; in order to possibly put statements made by that particular company into perspective, two other parties were interviewed.

The third aspect addresses the type of interview (Helfferich, 2011). The interviews in this thesis aim at obtaining rather objective information. Of course, a notion of subjectivity always resonates with statements based on experience, but the research question does not deal with extremely personal views or challenging social constructs. Hence, problem-centered, semi-structured interviews seemed most appropriate, which are characterized by a dialogic form of communication. Helfferich (2011) states that, in problem-centered interviews, a set of prepared questions is used to guide the interviews, but spontaneous questions and discussions are appreciated. "Semi-structured interviews have two main objectives: on the one hand, the researcher wants to know the informant's perspective on the issue but, on the other, they also want to know whether the informant can confirm insights and information the researcher already holds" (Blumberg, Cooper, and Schindler, 2014, p. 307). This approach is deemed appropriate in order to explore the fraud prevention process.

The last main aspect Helfferich (2011) mentions focuses on how to evaluate the interviews: The fraud prevention process is outlined and populated with the retrieved insights. In the quantitative part of the thesis, this process documentation can be referred to where appropriate.

A remark should be given with regard to the relationship between the interviewer and the participants: Blumberg, Cooper, and Schindler (2014, p. 213) state that "[Interviewer and participant] are typically strangers [...]. The consequences of the

event are usually insignificant for the participant, who is asked to provide information but has little hope of receiving any immediate or direct benefit from this cooperation." However, Risk Ident is in a business relationship with both of the interviewed e-commerce mail order companies. The companies are interested in an improvement of the software Risk Ident sells to them, so it can be assumed that the heads of the fraud prevention departments were motivated to participate in the interviews. However, the participants had been informed before that the interviews were part of a research work independent of the business relationship.

The questions are aligned with the fraud prevention process and include: general ideas of fraud prevention, definitions of essential terms, inspection techniques, post-inspection actions, and evaluation measures. The questions are listed below, and each question is commented on in order to explain the motivation behind it.

	Question
1.	*What is fraud?* As the literature review in Chapter 2 has shown, various definitions of fraud exist. It is essential to share a common understanding of this central term. In addition, this thesis does not cover payment default which is related to creditworthiness. Thus, it is important to understand in how far — or whether at all — the companies distinguish between fraud and low creditworthiness. The dataset contains the labels *fraudulent*, *legitimate*, and *not creditworthy*.

	Question
2.	*How is the fraud label of transactions determined?* The fraud label of a transaction is the result of a manual inspection process. Thus, it is crucial to understand which actions lead to the decision whether a transaction is considered fraudulent or legitimate.
3.	*What kind of information is considered?* Extending the previous question, it should be discussed which parts of the quantitative data are used to investigate the fraud status.
4.	*How are transactions prioritized?* In e-commerce fraud prevention, the amount of transactions most likely exceeds available inspection resources of fraud prevention departments. If this is the case, then how are transactions prioritized?
5.	*Which actions follow the classification?* If a transaction is labeled fraudulent, which measures are taken? The question aims at exploring possible effects of the classification, including classification errors.
6.	*How much time does it take to inspect a transaction?* The feasibility of some of the actions may depend on the time invested into the inspection process. Hence, it may be helpful to know how time-consuming the inspection of transactions is.

	Question
7.	*Until when can a transaction be stopped?* In order to successfully stop a transaction that is considered fraudulent, the fraud prevention process has to be finished before delivery of the parcel at the latest. Therefore, it is to be found out until when the e-commerce companies can intervene in the delivery process. In addition, cancellation of a transaction becomes more difficult the further the delivery process has progressed.
8.	*How are the fraud department's decisions validated?* This question aims at evaluation of the labeling quality, i.e. how can reliability and validity of the labeling process be ensured?
9.	*Is there a difference between* legitimate *and* not inspected*?* Only a fraction of the total transaction data is labeled. In how far differ cases considered legitimate and cases that have not at all been inspected from each other? For example, does an inspected legitimate transaction increase credibility of the customer more than a transaction that has not been checked?
10.	*How is the performance of the fraud prevention department evaluated?* In Chapter 8, it will be argued that expected value is probably the most appropriate measure. This question aims at finding out the status quo.

	Question
11.	*Is A/B testing used in the fraud prevention department?* A/B testing might be a valuable tool in order to evaluate the fraud prevention process. If a transaction believed to be fraud is blocked, it becomes impossible to find out whether financial damage would actually have occurred. Therefore, it could make sense to let a fraction of transactions marked fraudulent pass nevertheless. Then, the value of fraud prevention could be measured more easily — at the cost of voluntarily allowing a defined amount of fraud cases, though. The applicability and acceptance of this concept shall be discussed in the interviews.

Since the interviews were semi-structured, the conversations were not tied to the list of questions. Instead, the questions were woven into a free-flowing dialogue.

Anonymity not only of the interviewees but also of the names of the companies was a prerequisite in order to be allowed to both conduct the interviews and to analyze the quantitative datasets. The interviews were conducted in accordance with the ethical guidelines of Edinburgh Napier University. The interviews were recorded, but the audio files were deleted upon request after transcription. The interviews lasted between one and two hours each.

In the findings section, the insights from the interviews are organized in four categories: fraud risk factors, investigation process, classification errors, and impact evaluation. In order to extract the information the transcribed interviews contain in a structured way, the interviews were coded. “A code in qualitative inquiry is most often a word or short phrase that symbolically assigns a summative, salient, essence-capturing and/or

evocative attribute for a portion of language-based or visual data" (Saldaña, 2009, p. 3). Such codes, for example *rule combinations*, *detection strategies*, and *fraud vs. creditworthiness* were assigned to the statements made in the interviews. The codes were then assigned to the four categories mentioned above.

4.4 FINDINGS

4.4.1 *Introduction*

Because the interviews were semi-structured and contain more information than the initial set of questions, findings are not presented in the order of the list of questions shown above. Instead, the process-related concepts discussed in this thesis are presented and supported with evidence from the interviews. References to the questions are given when they are answered.

First, however, a shared understanding of the central terms used in fraud prevention has to be established. The nature of fraud depends on the area in which it occurs. Thus, it is essential to define the terms used in the interview and to understand what meaning the interviewees attach to them.

Hinneburg (2006) states that fraud has to show certain characteristics: For example, the orderer has to have a criminal intention, and a financial damage has to occur. In contrast, the interviewed companies choose more practical definitions of fraud. According to company A, the intention itself already suffices for a transaction to be marked as fraud; fake data are considered an attempt to commit a fraud. Company B distinguishes between *fraud* and *fraud suspicion*. The former refers to just those cases that have been handed over to the police because financial damage has occurred, and the latter covers all cases

in which the company has only found evidence for a criminal intention, such as obviously faked data or customer data that are related to past fraud cases (question 1).

Although fraud and low creditworthiness are not the same, they are two different causes that share the same symptom (payment default), and thus they are hard to tell apart from each other. However, the distinction between fraud and low creditworthiness becomes particularly important when the company's business model relies on invoice and deferred payments (Anon, 2015a). Since this is the case for both interviewed e-commerce mail order companies, payment default should not be used for training algorithms to detect fraud.

On busy days, fraud prevention employees sometimes face up to 1,000 orders per person, but they inspect approximately 12 orders per hour, i.e. one inspection takes 5 minutes. Legitimate orders are usually identified quickly, but indistinct cases can consume hours or even days (Anon, 2015b). Therefore, it is crucial to work with a subset that consists of the most suspicious transactions (question 6).

4.4.2 *Fraud Risk Factors*

Based on the interviews, the author proposes to distinguish three types of risk factors: static data, connection data, and change-related data (question 3).

STATIC DATA In all interviews, it is mentioned that the distinction between new and old customer accounts plays a crucial role with regard to estimating the fraud risk (Anon, 2015a,b,c). In addition, the most promising questions to answer are: Is a parcel shop used? Does the shipment address deviate from the payment address? Is there a

Schufa entry available (Schufa is a private German credit bureau)? Moreover, company B mentions that the risk depends on the article group as well. In particular, in accordance with Hinneburg (2006), company A states that small, valuable items are associated with a higher fraud risk because they can be resold easily.

CONNECTION DATA Transaction connection data focus on connections between orders: Is the device connected to multiple customer accounts or are multiple devices used with the same customer account? Is the customer account connected to fraud cases already known (Anon, 2015a,c)?

CHANGE-RELATED DATA Have the data changed recently, e.g. has the phone number changed (Anon, 2015a)? Was there an unexpected increase of the order frequency (Anon, 2015c)?

4.4.3 *Investigation Process*

In the e-commerce mail order business, or at least in the companies that were interviewed, the fraud prevention process can be modeled as shown in figure 4. They face thousands of transactions per day but have only limited inspection resources, similar to the conditions described by Torgo and Lopes (2011).

Thus, the first step ① is to decide whether a transaction shall be manually inspected at all or whether it is processed automatically. Without any manual interference, all transactions are processed automatically. Fraud prevention employees select individual transactions using filters based on transaction properties, such as price ranges, address areas, etc. At the time of the interviews, no data mining techniques were used for guidance; the selection process was purely driven by experience and in-

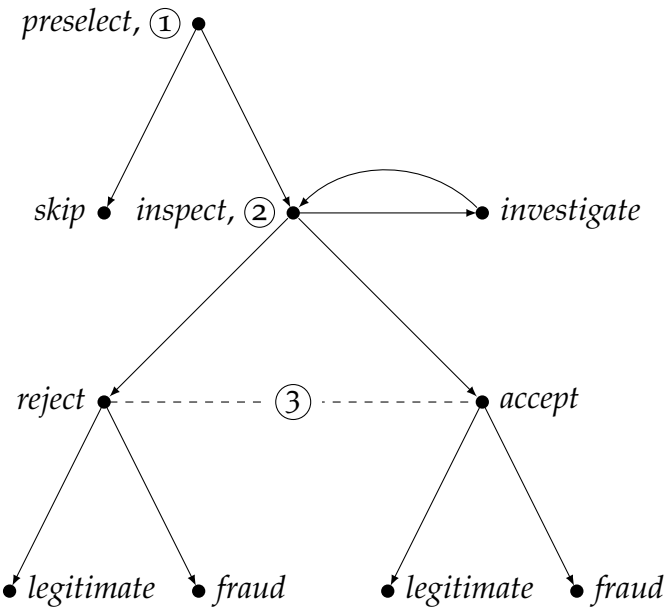

Figure 4: Fraud prevention process model.

tuition. Company A stated that a fraction of the transactions is also selected randomly (questions 2 and 4).

In the second step ②, inspected transactions may either be classified as fraudulent or legitimate. If the data available do not suffice in order to make a sound judgment, it may be investigated further in order to find stronger evidence for or against fraud (Anon, 2015a). In particular, the following scenarios have been mentioned (question 2):

1. When the legitimate customer is called and denies his or her order, the account may have been compromised. If the suspected fraudster answers the call and does not know the answer to a test question, such as what the date of birth registered with the account is, evidence strengthens the fraud hypothesis. The phone can be used for a couple of techniques to better assess the situation (Anon, 2015a). Calling the customer as an action to find out whether a transaction is fraud is mentioned by Carneiro, Figueira, and Costa (2017).

2. Further research on the customer can be conducted. For example, the fraud prevention officer can look up the name in social media (Anon, 2015a) or use Google to find out more (Anon, 2015b).

3. If available, a Schufa entry provides additional information regarding the creditworthiness of the customer. However, it is expensive and therefore only occasionally requested (Anon, 2015a). A negative entry leads to an immediate block of the transaction, even though low creditworthiness is not considered fraud as discussed above (Anon, 2015b).

If a transaction is considered fraudulent, multiple reactions are possible. The fraud model proposed here only distinguishes between *accept* and *reject* — as seen in step ③ — because, ultimately, fraud shall be blocked and legitimate orders shall pass. Yet, as the truth is sometimes nebulous, there are many shades of gray (question 5):

1. The order can be deleted either without further notice (Anon, 2015b) or the customer is informed that his or her order has been canceled (Anon, 2015a). Whether the customer is contacted depends on the reasons behind deletion (Anon, 2015b). At least, an incomprehensible risk estimate by a fraud detection software would not suffice for deletion (Anon, 2015b).

2. In extreme cases, i.e. when organized crime is faced, the police is contacted and charges are filed (Anon, 2015b).

3. If an order has been sent to the customer but not yet arrived, the logistics service can still try to stop delivery (Anon, 2015a), but the process becomes more expensive

(Anon, 2015b). Therefore, although a late cancellation is possible, detecting fraud before the delivery process has begun is clearly preferred (question 7).

4. The customer is offered only a safe payment method, such as prepayment, or a partial payment has to be made in advance (Anon, 2015a,b).

5. The delivery method can be changed to *delivery requires identification*. It is illegal to require an uncensored copy of personal documents (Anon, 2015b), but at least the postman can be asked to check the identification at delivery (Anon, 2015a,b). However, the process is time-consuming and expensive (Anon, 2015b).

6. If a transaction seems legitimate, it is marked as such and no other action is taken (Anon, 2015a). This also means that transactions labeled legitimate have been subjected to the same detailed inspection as transactions labeled fraudulent — in contrast to unlabeled transactions that have not been inspected at all (question 9). In machine learning, both types, fraudulent and legitimate transactions, will be needed to train the model.

4.4.4 *Classification Errors*

It can be difficult to find out whether the decision to accept or reject transactions was correct. This section deals with problems associated with the overall fraud prevention process.

Figure 5 shows four rows: Row ① states whether the transaction is actually fraud, i.e. the truth value; `F` stands for *fraudulent*, and `L` stands for *legitimate*. Row ② represents whether the transaction has been chosen for inspection (`I`), or has been

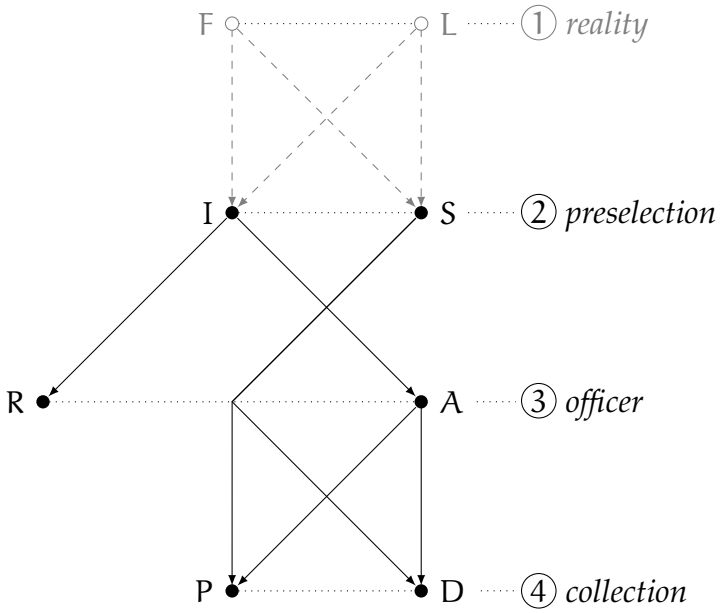

Figure 5: Possible sources of classification errors.

skipped (S). Row ③ depicts the officer's action, which is to reject (R) or accept (A) a transaction. ④ describes the financial outcome, i.e. whether a transaction was paid for (P) or whether it led to payment default (D). The different paths in the graph represent possible outcomes of the fraud prevention process. For example, the path F → I → R, or shorter notated as FIR, means that a fraudulent transaction has been preselected and the fraud officer regarded the order as fraud and blocked it. Of course, the collection step only exists for transactions that were not blocked, i.e. transactions that have either not been selected for inspection or have explicitly been accepted by the officer. The critical paths are described below.

Path	Description
`?IR`	A transaction should only be rejected if sufficient evidence exists. If transactions were falsely rejected, there would be no safe method to find out the truth afterwards. For example, an honest customer whose transaction was rejected by mistake could be disappointed that the promised service was not provided and just did not complain.
`F?P`	These paths should be impossible, because a transaction cannot be fraud if the products are fully paid for.
`FIAD`	If debt collection fails, reasons other than fraud may exist. For example, low creditworthiness could also lead to payment default. Thus, it is difficult to find out whether an accepted transaction should have been rejected as fraud.
`FSD`	Due to resource constraints, a transaction not selected for inspection in the first place is usually not brought up again even if it turned out to be fraud. This, however, would sometimes be worthwhile, because new patterns may emerge that would otherwise remain unrecognized (Anon, 2015a,b). In addition, transactions are rarely reinspected once a decision has been made (Anon, 2015a). In contrast to company A, company B checks whether the transactions inspected (and not blocked) appear in debt collection. In addition, as in the previous path, it is not possible to infer that a transaction is fraudulent only because of payment default.

The above paths show that, with regard to the fraud prevention process as the interviewed companies run it, multiple paths exist for which it is difficult to find out the truth retroactively. Therefore, the quality of the labels in the quantitative dataset heavily depends on the energy the fraud officers invest into inspection of cases (question 8). This is an important limitation that applies to how results of quantitative analyses can be interpreted. Due to the distinction between fraud and other forms of payment default, payment data alone cannot solve this issue.

Both companies state that A/B testing could alleviate the problem and help estimate the labeling error by letting a fraction of suspicious transactions pass. These methods are expensive, though, since letting fraud cases pass leads to immediate financial damage, and inspecting low-risk transactions is a waste of resources, at least in the short run. Therefore, the companies would expect a substantial improvement of their own fraud prevention activities as a compensation for the cost resulting from A/B testing (question 11).

4.4.5 *Impact Evaluation*

It is difficult to estimate the impact of the fraud prevention department. The most obvious reason are the classification errors mentioned above, but another problem requires attention as well. This section discusses how the impact of fraud prevention can be evaluated (question 10).

Sales and fraud prevention seem to create tension as their interests collide: Sales are supposed to increase, but fraud should be decreased; in particular, this is problematic when legitimate transactions are blocked by mistake that would normally

increase revenues (Anon, 2015a). The companies justify their work with revenue-based numbers, even though financial damage is not reflected by the sales prices but rather by the purchase prices plus operational costs (logistics etc.).

Instead, transactions could be preselected based on expected utility. This concept is described in detail in Chapter 8, and although it is not an evaluation measure, it shall be mentioned briefly. Both companies consider sorting transactions not based by perceived risk but by expected utility a valuable approach, even though the possibilities have not been explored yet (Anon, 2015a,b). Possibly, the concept of company-wide profit maximization contrasts with the department-focused view of fraud minimization.

Two years ago, an information bureau offered its services to company B, delivering data on-demand, i.e. the company would pay for additional information per transaction. The offer which included an implementation of expected utility was rejected because data protection issues arose and the benefits seemed intransparent (Anon, 2015b). Company B does not have product-related margins but holds approximately 1,000 group-related margins for more than 100,000 products. The company explicitly regards the concept as interesting if margins become more diverse, especially since some fraud risk scores already exist per article group. New customers are assigned a margin-like weight, because various costs associated with them, such as marketing costs, decrease the initial profits.

4.5 LIMITATIONS

In this section, possible limitations of the interviews are discussed and it is explained how they have been dealt with.

In general, measures commonly used in quantitative studies — such as reliability, validity, and generalizability (Collis and Hussey, 2014) — are not applicable when evaluating personal interviews because qualitative research is always subjective and depends on the context. Therefore, Helfferich (2011) argues, objectivity cannot be the goal of qualitative research; instead, one should aim at dealing with subjectivity in an appropriate way. This perspective guides the discussion of limitations.

First, in order to deal with subjectivity, a couple of strategies to support methodical control shall be mentioned (Helfferich, 2011): methodical control through openness, through reflexivity, and through traceability. Methodical control through openness aims at forcing as little structure as possible in order to let the interviewees unfold themselves without restrictions. In expert interviews, however, sometimes specific questions are of particular interest and it may make sense to ask them or to steer the interview into a certain direction if the interviewees do not bring them up by themselves. Considering expert interviews, more structure is possible when rather fact-based knowledge is to be obtained. The semi-structured interviews which have been conducted are considered a trade-off between the two positions. Methodical control through reflexivity focuses on actively thinking about how the assumptions the interviewer and the interviewee make and the values they hold influence the course of the interview. For example, the author of the thesis is employed in the company funding the research; thus, answers of the interviewees might have been confined to what seemed appropriate with regard to the business relationship. In order to deal with this possible issue, it was stated that the interviewees remain anonymous. Nevertheless, since in the interviews with company B and Risk Ident there were two interviewees in a single interview, they could have withheld information that might

have been divulged had they been alone. Methodical control through traceability focuses on using standardized procedures, e.g. how to act in the interview and how to deal with difficult situations. Before each interview, interviewing guidelines provided by Edinburgh Napier University were discussed.

Second, time constraints required to focus on the most important questions, although the explorative character might have allowed even more detailed insights into the process. However, being at all able to interview fraud prevention experts with the aim to support the quantitative analyses was a great opportunity.

Third, the retrieved statements may contain mistakes or might become outdated. In particular, the interviews were conducted in 2015, and the quantitative dataset is from 2016. Thus, if the meaning of the features in the dataset changed systematically within one year after the interviews had been conducted, there would be a mismatch between what has been said in the interviews and what is concluded from the quantitative analyses. Given that the structure of the data did not change in the past years, this is unlikely, though.

4.6 CONCLUSION

This chapter aims at laying a foundation for the following analyses by exploring the fraud prevention process and the quantitative dataset via interviews. The essential aspects which the following chapters build upon are summarized in this section.

First, a process model was developed. It was discovered that the fraud prevention process offers a variety of investigation mechanisms and that shades of gray between accepting and rejecting transactions exist. In Chapter 7, machine learning mod-

els will be trained that learn to detect fraud, and in Chapter 8, a utility-based perspective on fraud prevention will be taken. Both chapters build upon the simple process model developed through the interviews.

Second, it was discussed which feature values — and combination of feature values — fraud prevention employees consider particularly important. In Chapter 6, a mathematical approach to evaluating (static) features will be presented, building upon the statements about feature importance made by the fraud prevention experts.

Third, payment default was identified as an inappropriate representation of fraud. Instead, the manually set labels provide a promising basis for the quantitative part of the thesis. In all subsequent analyses, the fraud labels will be used as the ground truth.

Finally, the interviewed companies consider fraud an important problem and have developed countermeasures. However, e-commerce mail order fraud prevention seems to lack the application of scientific methods common in data mining. Thus, the next chapters focus on analyzing how such methods can support the fraud prevention process.

5

DATA PREPARATION

5.1 INTRODUCTION

In this chapter, the dataset which is used for the quantitative analyses presented in chapters 6 to 8 is prepared. It contains hundreds of thousands of transactions from one of the largest e-commerce mail order companies in Europe — referred to as company A in the interviews. Data preparation is an essential step of the data mining process: "Before the data can be analyzed, they must be organized into an appropriate form. Data preparation is the process of manipulating and organizing data prior to analysis" (Webb, 2010, p. 259). "Preparing input for a data mining investigation usually consumes the bulk of the effort invested in the entire data mining process" (Witten, Frank, and Hall, 2011, p. 51).

The chapter is organized according to the data preparation steps described by Webb (2010): First, a general overview of the dataset is provided, including the steps sourcing, selecting, and auditing of the data. Second, a complete list of e-commerce transaction features (order details, customer data, and technical data) is presented, and each feature is briefly discussed. Then, transforming data is explained in detail, covering discretization, and dealing with missing values through imputation. Finally, the chapter closes with a discussion of limitations and conclusions of the dataset itself and the data preparation process.

5.2 OVERVIEW OF THE DATASET

"It is necessary to review the data that are already available, assess their suitability to the task at hand, and investigate the feasibility of sourcing new data collected specifically for the desired task" (Webb, 2010, p. 259). This is reflected by the three aspects sourcing, selecting, and auditing.

Sourcing is the process of obtaining data, possibly from multiple sources (Webb, 2010). The raw dataset is stored in a non-relational database (Elasticsearch), and the documents are unstructured. This means that the data are not only composed of atomic key-value pairs (e.g. `total price: €199`) but contain nested elements as well, such as lists of order items or associated devices (e.g. `order items: [item 1, ..., item n]`), where each item may contain key-value pairs itself. Such a format is impractical for data analysis, and thus the first step is to reshape the data into a matrix, i.e. tabular form, where each column is a feature and each row is an observation (transaction) as described by Wickham (2014).

	inspected			other	
	fraud	legitimate	$\sum$		$\sum$
count in %	0.56	2.42	2.98	97.02	100
value in %	0.71	4.12	4.83	95.17	100

Table 2: Breakdown of one month's transactions into the count of transactions and the order value with regard to the fraud label.

The dataset consists of the e-commerce orders placed within a single month of early 2016, containing hundreds of thousands of transactions. 2.98% of such transactions (approximately 10,000 cases) have been manually inspected and labeled as fraudulent or legitimate, and this subset is selected

for the quantitative analyses in the following chapters. The remaining 97.02% of the transactions were either unlabeled or related to low creditworthiness. For each label (fraudulent, legitimate, other), Table 2 shows the relative amount and the relative value of transactions according to their total price (also called order value). Since the 2.98% of the transactions that have been inspected represent 4.83% of total revenue, the statement made in the interviews can be confirmed that fraud officers are more likely to inspect high-valued transactions. Interestingly, the difference between the relative share and the associated value of transactions is largest with regard to legitimate transactions (2.42% of transactions represent 4.12% of total value). This might indicate that fraud is not necessarily confined to expensive goods.

Table 3 shows the complete list of features, including basic descriptive analytics. Each feature used for analysis will be explained in detail in the next section. Three features — *order address*, *customer name*, and *phone number* — were not processed in the analysis due to concerns regarding data privacy and ethics.

Auditing aims at ensuring data quality. The data are composed of three sources: technical information, user-provided data, and the fraud labels. Technical data can be considered free of errors because their collection is automated. User-provided data does not always indicate the truth, since erroneous data might be provided accidentally or intentionally. However, the quantitative analyses still benefit even from fake data because not the truth is the determining factor but the pattern that the data show. For example, if criminals used fake phone numbers, that information could even provide additional value. Thus, fake user data are not considered a conceptual problem. Finally, the label quality is essential to all quantitative analyses. In order to evaluate it, the interviews presented in Chapter 4 were

Feature	Mean	Median	Used
Account Age	6.36 d	2.95 d	Y
Address Distance	63.33 km	2.04 km	Y
Customer Age	42 y	41 y	Y
Number of Articles	1.94	1.00	Y
Order Hour	12 pm	12 pm	Y
Total Price	759.85 €	589.00 €	Y

(a) Continuous Features

Feature	Distribution	Used
Browser	• 34% Google Chrome • 25% Mozilla Firefox • 14% Safari • 27% other	Y
Country	• 98% Germany • 2% other	Y
Customer Name	n/a	N
ISP	undisclosed due to data privacy reasons	Y
Operating System	• 61% Windows • 19% Android • 14% iOS • 6% other	Y
Order Address	n/a	N
Order Origin	• 85% internet • 15% telephone	Y
Known Address	94% known addresses	Y
New Device	40% via unknown device	Y
Parcel Shop	26% sent to a parcel shop	Y
Payment Type	• 98% invoice • 2% other	Y
Phone Number	n/a	N
Shipment Type	32% express orders	Y
Smartphone	33% via mobile/app	Y

(b) Discrete Features

Table 3: Overview of features with basic descriptive metrics. Features used in the following analyses are marked with *Y* and features discarded are marked with *N*.

conducted and it has been shown that the investigation process ensures high label quality: Once a transaction is considered suspicious, it will be inspected manually with care.

5.3 FEATURE OVERVIEW

5.3.1 *Introduction*

Information about features used as fraud predictors has been published with regard to credit card fraud (Bahnsen et al., 2016; Jha, Guillén, and Westland, 2012; Krivko, 2010), medical fraud (Aral et al., 2012), automobile fraud (Artís, Ayuso, and Guillén, 2002), and online auction fraud (Chang and Chang, 2012). Regarding the choice of features, it is difficult to draw inspiration from other areas of fraud because the raw data differ so strongly. For example, Carneiro, Figueira, and Costa (2017) mention that features used for fraud detection in the banking sector can barely be transferred to credit card fraud in the e-commerce mail order business because banks usually know much more about their customers than e-commerce businesses do, and thus the available data are quite different.

The features used in the quantitative analyses can be divided into two groups: The first group consists of features which fraud experts have declared as important fraud risk predictors, e.g. whether the order has been sent to a parcel shop, or which have been mentioned in the literature (called evidence-based features in the following). The second group contains features which simply were available as part of the dataset, e.g. the operating system from which the order has been placed (called discovery features in the following). Even though there is no empirical foundation that discovery features help detect fraud,

they might support the fraud detection process nevertheless. Philosophically, evidence-based features represent a deductive approach, i.e. theory comes first and is tested using real-world data (Benton and Craib, 2011), and discovery features represent an inductive approach because the data are used to identify patterns that might lead to new theory about fraud risk indicators (Cussens, 2010). Induction is often viewed from a critical perspective, referring to the problem of induction as described by David Hume (Phillips and Burbules, 2000).

5.3.2 *Evidence-Based Features*

ACCOUNT AGE The account age feature represents the time in days since the account's creation. Most accounts have been created within the last decade as frequent online shopping is a newer purchasing channel than the traditional mail order business. In order to face fraud with a long existing account, either the account had to be carefully set up by criminals years ago, or the account was stolen. Therefore, one can suspect that fraud occurs more often with recently created accounts. The interviewed companies confirm that the distinction between new and old customer accounts is crucial (Anon, 2015a,b,c), and the time since the first order — similar to the account age — is the most important feature according to Carneiro, Figueira, and Costa (2017).

COUNTRY The country feature indicates from which country an order has been issued. Some e-commerce shops do not allow products to be ordered from other countries in order to protect themselves against fraud (Hinneburg, 2006). Carneiro, Figueira, and Costa (2017) use the country fea-

ture as well but do not mention it as one of the most important features.

CUSTOMER AGE The customer age feature is a metric variable. According to Hinneburg (2006), criminals are mostly between 17 and 70 years old — which refers to most of the customers and, thus, the information seems to be of limited value.

ORDER ORIGIN The order origin is a categorical attribute and can be either internet, phone, or other (shop, physical mail, etc.). According to Hinneburg (2006), criminals switch between channels in order to disguise their intention.

ORDER HOUR The order hour describes when the transaction has been issued. Literature states that some criminals prefer to work from internetcafés at night (Hinneburg, 2006). It may be concluded that features which identify a device as a public computer, such as the IP address — which has not been disclosed, though —, are generally of interest as well. Carneiro, Figueira, and Costa (2017) list the order time as one of the most important features.

SHIPMENT TYPE Shipment types cover standard and express delivery. According to Hinneburg (2006), criminals prefer express delivery in order to quickly obtain the products, but Carneiro, Figueira, and Costa (2017) do not confirm this statement.

PARCEL SHOP This feature indicates whether a parcel shop has been used as a shipment address. A parcel shop is a usually small retail shop that receives and stores the orderer's parcel until it is picked up. In the interviews, it is

mentioned that orders which are sent to parcel shops are more likely to be fraudulent (Anon, 2015a).

PAYMENT TYPE Payment types include payment via invoice — which is the most popular option regarding the dataset —, cash transfer, and credit card. Invoice payments are regarded as the simplest way to commit a fraud as no payment information at all is required. According to Hinneburg (2006), criminals usually choose payment per invoice, but if this leads to significant restrictions, they will switch to direct debit. Deferred payments are associated with a higher risk because companies set higher security standards for such transactions. Hinneburg (2006) regards payments via credit card as unimportant, but its importance has increased during the last decade as multiple research articles show (Bahnsen et al., 2016; Carneiro, Figueira, and Costa, 2017; Dal Pozzolo et al., 2014; Duman and Ozcelik, 2011; Van Vlasselaer et al., 2015).

TOTAL PRICE The total price is the sum of the prices of all individual order items. Company A states that small, valuable items are associated with a higher risk because they can be resold easily, which accords with Hinneburg (2006). The total order price is the second most important feature according to Carneiro, Figueira, and Costa (2017).

5.3.3 *Discovery Features*

ADDRESS DISTANCE This feature is a metric value describing the distance between the shipping and the invoice address. If the two addresses are the same, it is set to zero. Large address distances could possibly indicate a stolen

account because an order is sent to an address far away from the initial account owner.

BROWSER The browser feature indicates which internet browser has been used to place the order. The initial attribute contained the exact browser version, but it was simplified in order to only state the browser type such as Chrome, Firefox, and Internet Explorer.

INTERNET SERVICE PROVIDER The ISP feature is a categorical variable that states from which internet service provider the transaction has been issued.

KNOWN ADDRESS An address is considered reliable if the address has been used before. In most cases, a new address is provided. Note that the condition does not imply successful delivery or even payment.

NEW DEVICE This feature states whether the orderer's computer has been registered before, and it is based on a digital fingerprint that considers a large set of technical variables (Boda et al., 2012).

NUMBER OF ARTICLES This feature is an integer counting the number of articles per transaction.

NUMBER OF DEVICES If the account from which the transaction at hand was issued has been used online, either in the current or a previous transaction, one or more devices are registered. The number of devices denotes the device count associated with the transaction's customer account.

OPERATING SYSTEM This feature is a categorical variable that states from which operating system (e.g. Windows, Linux, OS X) the order has been placed.

SMARTPHONE The categorical smartphone feature states whether the orderer used a smartphone. This can only be true if the order has been placed online.

The above-mentioned features have mixed types like categorical or metric ones. However, some analyses require discrete features. In addition, some features contain missing values. These issues are addressed in the next section, which focuses on transformation of features into an appropriate representation (Webb, 2010) through discretization and imputation.

5.4 TRANSFORMATION

5.4.1 *Discretization*

Discretization turns metric features into discrete ones. For example, the customer age can be converted to a list of intervals. Although metric data contain more information than discrete data, the information gain analysis following in Chapter 6 cannot deal with the former. Thus, feature values are (optionally) discretized such that metric data are used where possible and discrete data are used where needed.

Yang (2010) introduces dimensions with which discretization techniques can be described: In this thesis, all discretization techniques are global (vs. local), eager (vs. lazy), disjoint (vs. non-disjoint), and univariate (vs. multivariate). They are global because they use the full dataset in one single step. They are eager because discretization is performed before the application of data mining techniques (instead of being part of them). They are disjoint because the resulting value ranges do not overlap. Finally, they are univariate because each attribute is analyzed individually, which is important in order to conduct the analy-

sis in Chapter 6 on a per-feature basis. The following discretization techniques were used in order to provide results:

ENTROPY Cutpoints are set by minimizing entropy (Fayyad and Irani, 1993). Entropy is a measure for information uncertainty and will be explained in detail in Chapter 6. This technique is a supervised technique because the fraud labels are used to compute entropy values.

FIXED-WIDTH Fixed-width discretization splits the data into k equally wide intervals.

EQUAL-SIZED In equal-sized binning, the number of intervals is determined by the number of bins. As a special case, $\sqrt{n}$ determines both the number of intervals and the bin size (Yang and Webb, 2001).

CHI-SQUARED Chi-squared discretization uses the χ^2 statistic to find out whether adjacent intervals should be merged (Liu and Setiono, 1995).

In the following chapter, entropy-based, $\sqrt{n}$ (fixed-width and equal-sized), and Chi-squared discretization are applied alternatively prior to computing information gain values. Regarding the dataset, entropy-based discretization creates about two to five cutpoints, and Chi-squared discretization yields hundreds of cutpoints.

Of course, the thesis does not aim at exploring the differences between discretization techniques. However, they influence the results of the information gain analysis, and being aware of the conceptual differences between the discretization techniques helps better understand the results of the analysis.

5.4.2 *Missing Values*

Most datasets found in practice contain missing values, and one has to think carefully about why they occur and how they can be interpreted (Witten, Frank, and Hall, 2011). The dataset used in this thesis is first inspected with regard to missing values and then strategies to deal with these missing values are presented according to Bruha (2010).

Missing values are filled by performing *imputation*, which simply means that they are replaced by more appropriate ones. Consider this example in which rainfall is measured: "[A] missing value may mean there was no rain recorded on that day, and hence it is really a surrogate for 0mm of rain. Alternatively, perhaps the measuring equipment was not functioning that day and hence recorded no rain." (Williams, 2011, p. 161).

Three variables were found to contain missing values: address distance (65.10%), parcel shop (59.61%), and customer age (0.04%). Regarding the first two variables, the fraction of missing values is alerting. However, there are simple explanations, and the missing values can be replaced with meaningful ones. The address distance field is missing if there is no shipment address at all — the shipment address is an optional field, because by default, an order is shipped to the invoice address — or if there is a shipment address but it is identical to the invoice address. Therefore, missing values can safely be set to 0km.

With regard to the parcel shop, since customers do not live inside them, it can safely be assumed that an order is not sent to a parcel shop in case the value is missing, i.e. when no shipment address is provided. Therefore, the value can be set to *false* if a missing value is encountered.

The customer age variable is related to another kind of missing values. There is no truly correct replacement value such

as for missing values of the previous variables. The customer age variable contains missing values simply because the date of birth is an optional field (or was years ago). In this case, median-based imputation seems most appropriate. This means that, if a value is missing, it is replaced with the median as the best guess. The median is preferred to the mean because it produces stabler results when facing outliers (Fahrmeir et al., 2007), which is the case for many of the features (e.g. address distance and total price).

5.4.3 *Feature Interdependence*

Metric features were tested for correlations between them using the Pearson correlation coefficient (Fahrmeir et al., 2007). In Table 4, the correlation matrix is shown. The strongest correlations of $r = 0.38$ and $r = 0.35$ were found for the pair account age and customer age and for the pair total price and number of articles — which might be expected. Detailed scatter charts and histograms of variable pairs have been computed but are not included in the thesis because from them demographic data about customers can be inferred, endangering the anonymity of the company which provided the dataset.

5.5 LIMITATIONS

It is often difficult to obtain real-world e-commerce transaction data due to their confidentiality. Therefore, the value of being able to do so is not being questioned. Nevertheless, it is important to name possible limitations.

First, the data do not contain payment information. If fraud cases remained undetected, they would eventually be recog-

	PAR	AAG	ADD	NOD	TOP	SMP	OHO	NOA	NEW	KAD	CAG
PAR	100										
AAG	-22	100									
ADD	-9	2	100								
NOD	-4	-3	5	100							
TOP	-25	15	2	2	100						
SMP	-13	-6	-8	4	-2	100					
OHO	-8	-4	-11	-6	3	-4	100				
NOA	-9	9	-2	0	38	-1	5	100			
NEW	2	-15	0	-15	-3	15	0	-5	100		
KAD	-1	8	-9	20	2	4	-1	6	-8	100	
CAG	-5	35	5	-6	6	-21	-7	4	-9	1	100

Table 4: Correlation table using Pearson correlation coefficient (PAR: parcel shop, AAG: account age, ADD: address distance, NOD: number of devices, TOP: total order price, SMP: shipment type, OHO: order hour, NOA: number of articles, NEW: new customer, KAD: known address, CAG: customer age).

nized as payment default. Thus, despite the problems associated with payment information, which have been mentioned in the previous chapter, it could still be interesting to consider such information in future research.

Second, only a small fraction of the data is labeled. Therefore, although the labels were assigned carefully, scaling results to the size of the full dataset may be difficult. In particular, the experience-based inspection process may lead to selection bias. For example, fraud officers probably do not label transactions that are obviously legitimate. This might explain the surprisingly high average fraud ratio.

Third, data are provided by one e-commerce company only. Considering other datasets as well — for example from companies that sell different products or from companies of different size — could increase generalizability of results.

Fourth, the dataset contains one month's data. Thus, it is difficult to identify suspicious changes in behavior. In order to

achieve that, a longer time span would have had to be available.

5.6 CONCLUSION

In this chapter, data preparation was presented. First, an overview of the dataset was given, features were presented, and data transformation steps were explained.

The overview of the dataset covered the aspects sourcing, selecting, and auditing. The dataset was provided by one of Europe's largest e-commerce mail order companies and contains the transactions of one month. Only a fraction of the data is labeled fraudulent or legitimate, but the labels are of high quality. Then, features were divided into two groups, evidence-based features and discovery features, and each feature was presented briefly.

In the data transformation part, discretization was motivated and different techniques were introduced. In addition, it was explained how missing values were dealt with and which features were constructed from the raw data. Feature interdependence was tested, but no strong or unexpected correlations were found.

Finally, limitations of the dataset were discussed. Challenging aspects of the dataset could relate to the question of generalizability and to possible selection bias regarding the distribution of labels.

The data preparation lay the foundation for subsequent analyses. For every quantitative analysis, proper understanding of the input data is essential. In particular, the information gain analysis in the next chapter heavily depends on the discretization performed as a preparatory step.

6

FEATURE ANALYSIS

6.1 INTRODUCTION

The feature analysis chapter is the first of the three analyses which are based on the quantitative dataset. In this chapter, the question is examined to what extent the available features, i.e. properties of a transaction, can help detect fraud cases. Knowing what to look out for is not only helpful for a more focused manual inspection but can also improve the prediction quality of statistical models. Thus, the chapter has two purposes: On the one hand, it aims at being self-contained in the sense that its results can already be used to support the manual search for fraud. On the other hand, its findings are supposed to be used for fine-tuning machine learning algorithms such as those implemented in Chapter 7.

Besides the term *feature*, other names are commonly used as well: In the context of a statistical model where these properties are used as inputs to an algorithm, features are often referred to as *predictors*, "and more classically the *independent variables*. In the pattern recognition literature the term *features* is preferred" (Hastie, Tibshirani, and Friedman, 2013, p. 9).

This chapter is structured as follows: First, the question what makes a feature important is discussed, an overview of feature selection is given, and it is explained in how far it makes sense to rely upon such methods. Second, the main method used in this chapter, information gain, is presented in detail, the empirical analysis is conducted, and findings are discussed. Third,

the idea of feature combinations is approached by introducing decision trees. They serve as a thematic transition to Chapter 7, which focuses on supervised machine learning with gradient boosted trees and logistic regression.

6.2 FEATURE RELEVANCE

In order to approach the problem of feature relevance, a discussion of its theoretical concept is important. It is difficult to answer the question *what* makes a feature useful. From the perspective of e-commerce fraud prevention in the mail order business, an intuitive definition might be as follows: A feature is relevant if it helps better distinguish between fraudulent and legitimate transactions. However, the question remains how a metric could be defined that measures the ability to achieve this. A possibly counterintuitive remark shall be given upfront:

> *Practical [machine learning] algorithms [...] may benefit from the omission of features, including strongly relevant features. Relevance of a feature does not imply that it is in the optimal feature subset and, somewhat surprisingly, irrelevance does not imply that it should not be in the optimal feature subset (Kohavi and John, 1997, p. 279).*

It is even possible that features can be irrelevant when looked at individually but useful when combined (Guyon and Elisseff, 2003). Kohavi and John (1997) state that, eventually, the optimal feature set depends on the choice of the learning algorithm.

However, in general, algorithms usually perform better if provided with relevant features only (Kohavi and John, 1997; Witten, Frank, and Hall, 2011), and despite the aforementioned possibly confusing exceptions, there is a rule stating that it is generally favorable to reduce the amount of features. Liu (2010)

mentions the following thought experiment: Imagine four binary features and the fraud status as the dependent variable with values *fraudulent* and *legitimate*. Consequently, there are only $2^4 = 16$ possible instances, i.e. combinations of feature values. The set of all possible relations between the features and the dependent variable is called *hypothesis space*. In the example, the hypothesis space is $2^{(2^4)} = 65,536$ because each of the 16 possible instances is mapped to either *fraudulent* or *legitimate*, and thus, there are 2^{16} distinct mappings.

Feature selection techniques can be assigned to three categories: embedded models, wrapper models, and filter models. When feature selection is embedded, the machine learning algorithm *itself* is able to evaluate features. Wrapper models observe the influence of features on the prediction quality of a machine learning model, and filter models evaluate features by only looking at the data (Liu, 2010). In addition to such approaches that are related to feature selection and focus on classifier performance, of course, traditional descriptive statistics can be computed in order to explore the features and understand their influence on the fraud status.

One of the simplest approaches to determining feature relevance is to just measure the frequency-based fraud ratios: For example, an exact combination of values of the features *parcel shop*, *new customer*, and *express delivery* might apply to n cases of which f are fraud; then, of course, the frequency-based fraud ratio for $P(\text{fraud}|\text{parcel shop} = \text{true}, \text{new customer} = \text{true}, \text{express delivery} = \text{true})$ is f/n. However, the problem with this approach is that there are often insufficient or no data available for such specific combinations of feature values. Hence, other approaches are necessary.

With scheme-specific selection (wrapper models), "the performance of an attribute subset [...] is measured in terms of the

learning scheme's classification performance using just those attributes." (Witten, Frank, and Hall, 2011, p. 312). This means that an algorithm is trained again and again but each time a different feature subset is used. Even though "[g]ood results have been demonstrated on many datasets" (Witten, Frank, and Hall, 2011, p. 312), the approach is computationally expensive — it requires up to 2^m runs when an exhaustive search with m features is performed (Witten, Frank, and Hall, 2011).

Another aspect is considered even more important: Using wrapper or embedded models both requires knowledge about machine learning algorithms in order to meaningfully interpret results, but the thesis not only aims at automation of fraud prevention through machine learning but also at providing self-contained results about feature relevance.

With regard to filter methods, countless variants exist (He, Cai, and Niyogi, 2006; Mladenic and Grobelnik, 1999; Witten, Frank, and Hall, 2011). Which one to use seems to depend not only on the application context but on personal preference as well. The concept of confidence and support is used in association rule mining (Bayardo Jr. and Agrawal, 1999) and illustrates an important thought: With regard to fraud detection, confidence represents the fraud rate and support is the relative share of transactions to which the feature (rule) applies; note that confidence converges towards the average fraud ratio in the dataset the larger the support value becomes. On the one hand, it is important to consider both confidence and support: Even if the fraud rate of a specific pattern were close to 100%, the overall impact would be low if the pattern were rare. On the other hand, for classification of individual fraudulent transactions, it does not matter how frequent a pattern is as long as it is detected at all.

In this chapter, *information gain* is used, which can be considered a filter method of feature selection. It can not only be used to reduce the number of features prior to computationally more intense feature selection (Witten, Frank, and Hall, 2011), but it can easily be interpreted in an isolated way, i.e. the analysis is not tied to a machine learning algorithm. This means that companies interested in the information gain analysis can replicate it even if they do not want to deal with machine learning. In the next section, information gain is introduced, computed, and evaluated.

Despite arguing that information gain is an appropriate measure for determining feature relevance, it also faces limitations, which will be discussed later. One of them is that information gain is computed per individual feature and thus does not consider combinations of feature values. In order to overcome this particular issue, a decision tree is trained at the end of this chapter. This tree also provides a gentle transition to supervised machine learning, which is covered in Chapter 7.

6.3 INFORMATION GAIN

6.3.1 *Introduction*

Information gain has wide applications in areas such as machine learning and information retrieval. It measures the degree to which additional information reduces uncertainty. This uncertainty, or data impurity, is called entropy in information theory (Mitchell, 1997). From a mathematical point of view, information gain is the difference between two levels of uncertainty.

There are two concepts which must not be confused with each other: attribute levels and classes. A discrete feature, e.g. whether the order is sent to a parcel shop, has at least two attribute levels (parcel shop = true and parcel shop = false). Such a feature is then evaluated with regard to its ability to distinguish between the classes *fraudulent* and *legitimate*. This means that, in the case of fraud detection, there is only one bucket of data at the beginning, which includes fraudulent and legitimate transactions. The fraction of fraud within this bucket is used to compute base entropy. The entropy when considering the parcel shop feature is then calculated using two buckets (namely parcel shop = true and parcel shop = false), each of which contributes with an entropy value that is weighted according to the relative size of the bucket. This relation is illustrated in Figure 6. In the example, the original dataset has a (fictional) fraud rate of $7/16 = 43.75\%$. However, when the dataset is split using the parcel shop feature, i.e. conditioning the dataset on the attribute values parcel shop = true and parcel shop = false, the fraud rates of the two subsets are $6/8 = 75\%$ and $1/8 = 12.5\%$. This represents an information gain, because given the value of the parcel shop feature, the uncertainty with regard to the fraud status is reduced.

In general, entropy yields maximum uncertainty for 50%. Probabilities of 0% and 100% have no uncertainty because the outcome can be predicted reliably. The formulas for entropy (H) and information gain (ΔH) are depicted below (Mitchell, 1997):

$$H = -p \log_2(p) - (1-p) \log_2(1-p) \tag{1}$$

$$\Delta H = H_{\text{original}} - \sum_{i}^{n} w_i H_i \tag{2}$$

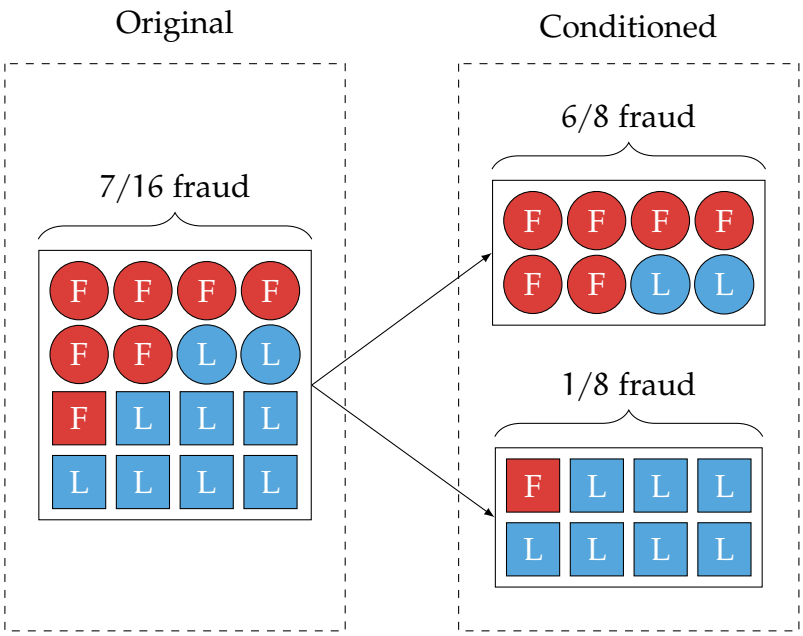

Figure 6: Visualization of how information gain can be used to estimate feature relevance. The dataset on the left side is divided into two groups, circles and rectangles, which represent two values of a binary feature, such as the parcel shop variable.

In the entropy formula, p denotes the fraction of fraud. For computation of information gain, the sum iterates over all n value groups of a feature, e.g. parcel shop $=$ true and parcel shop $=$ false. It weights the entropy per group according to the relative size of the group, thus measuring the difference between the old entropy before and the new value after the potential split. The entropy curve for a binary class case is visualized in Figure 7. The graph shows that jumping from p to $1-p$ would yield no information gain due to the symmetry of the function.

With regard to the example shown in Figure 6, entropy values and information gain are calculated as follows: First, entropy for the original dataset without any splits is computed.

$$H_{\text{orig.}} = -\frac{7}{16}\log_2\left(\frac{7}{16}\right) - \left(1-\frac{7}{16}\right)\log_2\left(1-\frac{7}{16}\right) = 0.99$$

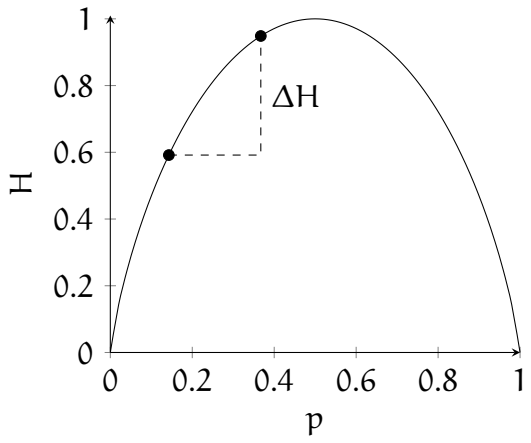

Figure 7: Visualization of entropy H in a binary class case, such as *fraudulent* and *legitimate*.

Then, entropy values for the individual buckets after the split (conditioned dataset) are computed and weighted according to their relative size. In this case, there are only two buckets (parcel shop = true and parcel shop = false).

$$\begin{aligned} H_{\text{cond.}} =& \frac{8}{16}\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right)-\left(1-\frac{6}{8}\right)\log_2\left(1-\frac{6}{8}\right)\right) \\ &+\frac{8}{16}\left(-\frac{1}{8}\log_2\left(\frac{1}{8}\right)-\left(1-\frac{1}{8}\right)\log_2\left(1-\frac{1}{8}\right)\right)=0.68 \end{aligned}$$

Finally, information gain is computed as the difference between the two components.

$$\Delta H = 0.99 - 0.68 = 0.31$$

In the example above, an information gain of 0.31 is achieved, because the original dataset has almost maximum entropy and the resulting split achieves a major reduction in uncertainty for both groups. The groups are equally important since they both have the same size. Note that the upper bound of the entropy value depends on the number of classes k. In this thesis, always only two classes are used (*fraudulent* and *legitimate*). Thus,

values are bounded by zero and one. The upper bound is calculated with $\log_2(k)$, which may be derived from the entropy formula.

6.3.2 *Analysis*

In the analysis, information gain is measured per feature, i.e. the entropy of the original dataset is compared with the entropy achieved when a given feature is used to split the data into subsets. The computation is performed with the Jupyter Notebook in Python using the library NumPy (Pérez and Granger, 2007; Walt, Colbert, and Varoquaux, 2011).

Figure 8 shows the information gain values. Some of the features had to be discretized, but since the choice of the discretization technique influences the results, three conceptually dissimilar techniques are applied. As stated earlier, entropy-based discretization tends to result in fewer bins than $\sqrt{n}$ discretization, and χ^2 tends to produce more bins. In general, the noisier a feature is, the higher the information gain values are that can be achieved when more bins are generated. For example, consider functions for which x is the feature's value (e.g. the order price) and y is the associated frequency-based fraud ratio of that value. Then, if a function has ups and downs, it has a non-zero information gain because the corresponding bins that are created through discretization have different fraud ratios. A horizontal line, in contrast, would yield no information gain, because discretization would lead to identical bins with the same fraud ratio in each bin. Filled circles (●) indicate that the feature has been discretized (e.g. account age), and empty circles (○) refer to natively discrete features (e.g. order origin). Note that, per natively discrete feature, the information gain

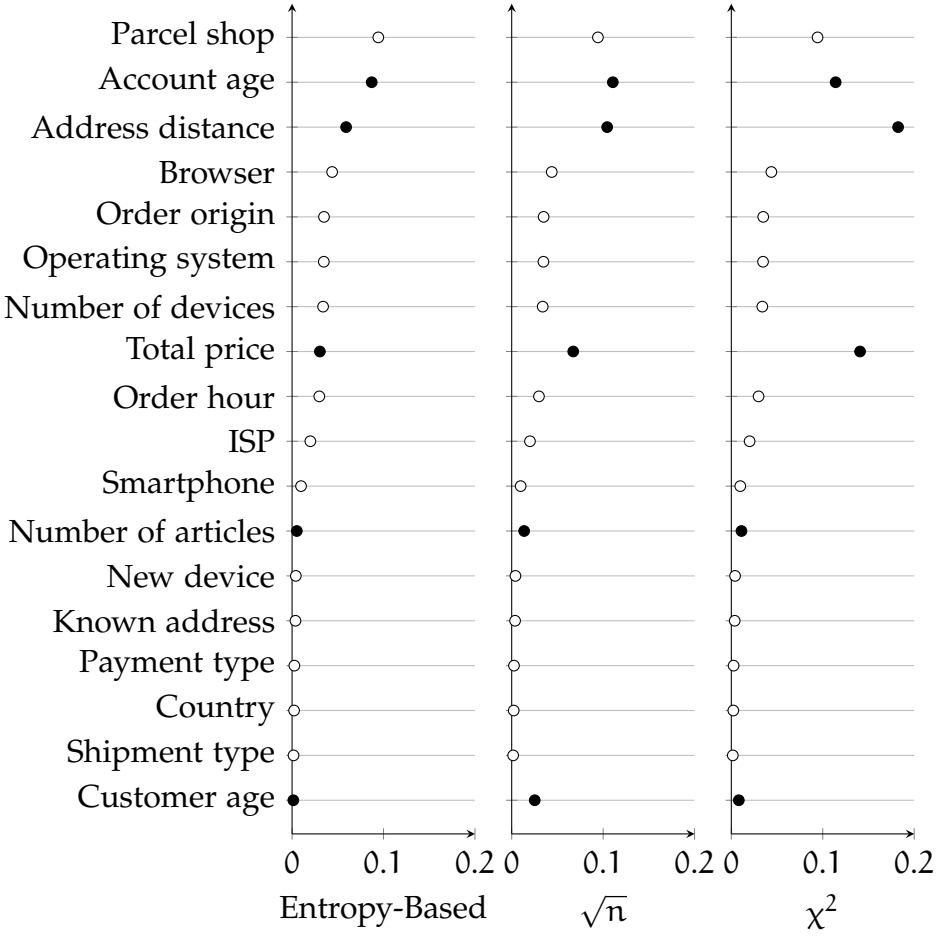

Figure 8: Information gain per feature, plotted with entropy-based discretization, discretization with $\sqrt{n}$-sized bins, and χ^2-based discretization.

values are the same across the columns in the figure because no discretization has been performed at all; thus, the information gain value is just shown as it is.

6.3.3 *Findings*

Through qualitative analysis, Hinneburg (2006) suggests that transaction properties exist which are associated with a higher fraud risk, such as certain preferred payment methods, delivery options, and demographic settings. Results show that information gain values vary strongly across the feature set and that

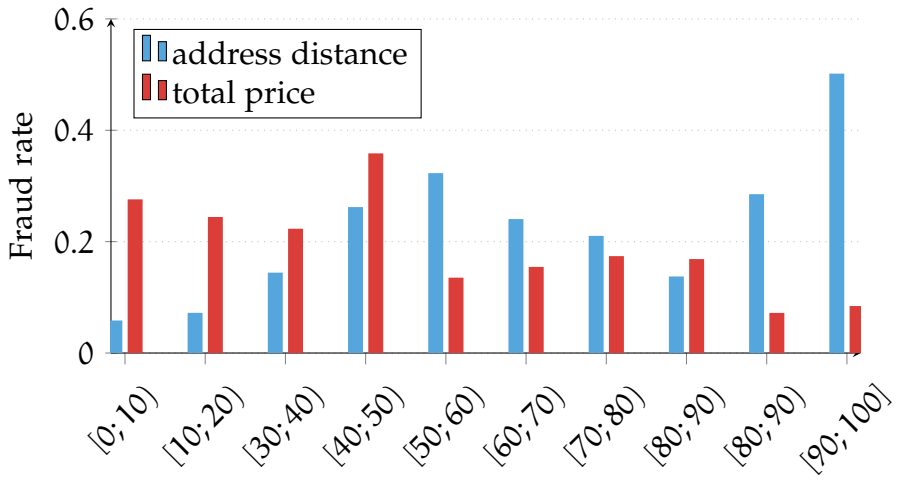

Figure 9: Fraud rate for features *address distance* and *total price* plotted in percentile bins of 10%, i.e. each bin contains one tenth of the data. The data have been ordered by value, i.e. with regard to the total price, the bin $[0;10)$ contains the 10% cheapest transactions.

there is slight variation among discretization techniques for at least some of the variables. The five most important features are *total price*, *account age* — which are both supported by Carneiro, Figueira, and Costa (2017) —, *address distance*, *browser*, and *parcel shop*.

The highest information gain is obtained with regard to the address distance combined with χ^2 discretization. Figure 9 shows the distributions of the two features *address distance* and *total price*, which may help understand the corresponding information gain values. The graph depicts non-cumulative percentile ranges on the horizontal axis and the fraud rate on the vertical axis. For example, the lowest 10% of the feature *address distance*, i.e. the distances between the invoice and the shipping address, are associated with an average fraud rate of only 5.8%, as shown by the leftmost blue bar in the graph. On the contrary, the median address distance is close to 30% fraud, which means that half of the transactions have a fraud rate above and the other half have a fraud rate below 30%. Such a high fraud

rate may seem strange with regard to the e-commerce mail order business, and such values will occur occasionally during the rest of this chapter. Often, this is influenced by the fact that fraud is overrepresented in the dataset because clearly legitimate transactions are less likely to be selected for inspection, and consequently, they are less likely to be labeled, leading to an artificially high fraud ratio. However, even in relative terms, 30% fraud is still more than five times as much compared to 5.8%. In the third quartile of address distances, the fraud rate returns to a lower level, eventually rising to 50% with regard to the highest 10% of address distances that may cross country borders. This strong variation with multiple ups and downs explains the high information gain value; when the fraud rate differences between the bins that are created during discretization goes up, information gain also increases. One can suspect that transactions with extremely low address distances including a distance value of zero in which case the shipping address equals the invoice address are likely to be legitimate because criminals that steal accounts would probably send the products to addresses far away from the original account owner. This assumption, however, would need to be verified using additional data mining techniques.

Similar to the address distance, the distribution of the order price contains ups and downs, therefore emphasizing the effect of χ^2 discretization. At the lower end, the fraud rate ranges between 22% and 28%, and at the 30^{th} percentile, there is a peak of 36%. The fraud rate then declines until the 100^{th} percentile. This distribution may be counterintuitive: Why should higher-valued transactions be associated with lower fraud rates than lower-valued products? In contrast, Hinneburg (2006) explicitly states that criminals prefer small goods of high value and manageable size. The reason for these results could be that

new customers are required to provide at least a small prepayment if the order prices of their transactions surpass a certain threshold value. In other words, if new customers order expensive goods, they will have to pay a partial amount in advance (Anon, 2015a). Criminals would probably not do that. For criminals, it is thus easier to commit a fraud by staying below that critical value (the cut-off value is known to the author, but it was requested not to publish it).

Regarding the *parcel shop* feature, the 54% of the transactions which had not been sent to a parcel shop showed a fraud rate of only 8%, whereas the other 46% that had been sent to one had a fraud rate of 39%. It is assumed that parcel shops offer the anonymity criminals seek because they are usually used by numerous people. When a fraud case is detected, the associated shipment address is usually marked as related to fraud, but with regard to parcel shops, this would be impractical.

The account age starts with a high fraud rate of approximately 40% regarding the most recently created accounts and then steadily decreases to 6% for the longest existing accounts. Older accounts can seem more credible, because it is unlikely that a loyal customer starts committing fraud. Therefore, it is assumed that accounts of honest customers are involved in fraud when they are stolen. In contrast, new accounts can systematically be created by fraudsters, which might be the reason behind the importance of *account age*.

The *order origin* feature has a non-zero information gain because approx. 14% of the customers still order via phone and such transactions are mostly non-fraud. This is an intuitive result because ordering via phone involves a more personal level of communication and takes more time, thus possibly being less attractive to criminals.

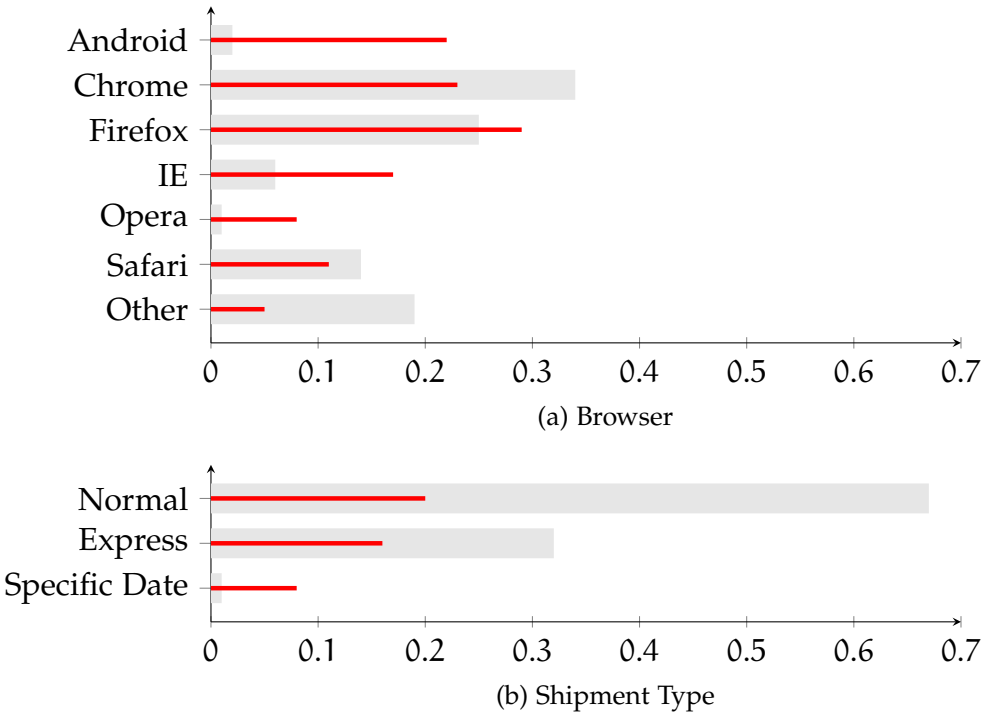

Figure 10: Fraud rates per attribute value (—) and frequency distributions (▭) in percent. The fraud rates do not sum to one, but the frequency distributions do.

The information gain value of the *number of devices* feature is only significant due to the low fraud risk of transactions with zero associated devices. This is the case if and only if an order is placed via phone as explained above. Thus, the feature is a bit deceptive because it makes almost no difference how many devices are connected with an account as long as there is at least one, which may indicate multicollinearity.

The *browser* feature is visualized in Figure 10(a): With regard to the discrete features, the browser has a relatively high information gain. It is shown that Chrome is used most often (23% fraud, 34% usage), but the highest fraud rate is observed with Firefox users (29% fraud, 25% usage). Significantly lower fraud rates are associated with Opera users, but the low popularity of Opera renders it insignificant with regard to the information gain value (8% fraud, 1% usage). A strong positive impact

on information gain is observed regarding Safari users (11% fraud, 14% usage) and users of other, less common browsers (5% fraud, 19% usage). The browser feature is difficult to interpret, if a sound explanation exists at all. The Safari browser is tied to Apple products, and owners of Apple products might have higher incomes and thus be less likely to commit fraud. Yet, a socio-demographic explanation like this would have to be researched properly.

The *operating system* feature has one of the most interesting distributions: Again, OS X users (i.e. Apple users) show a low fraud rate of 8%, but only 5% of all transactions were issued from OS X. Linux is used in less than 1% of the cases and has an above-average fraud rate of 35%. 57% of the transactions are associated with Windows users (except Windows XP) and have an average fraud rate of 24%. Only 4% use Windows XP, but this small fraction is fraud in 59% of the cases, and it is one of the highest fraud rates found in the whole dataset.

Hinneburg (2006, p. 54) states that criminals prefer express deliveries. However, the data at hand cannot confirm this claim. In fact, the fraud rate associated with express shipments (16%) is even slightly lower than the fraud rate corresponding to normal deliveries (20%), and the feature has one of the lowest information gain values. This can be confirmed by looking at Figure 10(b). Requesting to have the parcel delivered at a certain day reduces the fraud rate more strongly, leading to a higher entropy difference, but this affects so few cases that the corresponding weight (w_i) is low, hence scaling down the contribution to information gain. Hinneburg (2006) does not mention such type of delivery at all (possibly specific-date delivery was not available ten years ago). Regarding the difference between normal and express delivery, there are many possible explanations: For example, Hinneburg (2006) could be gener-

ally right and the sample used in this analysis is coincidentally not able to confirm such results. Another explanation could be that criminals might have learned that express deliveries are regarded as more suspicious than regular orders and have therefore adapted their behavior. No matter what the actual reason is, the information gain value of the shipment type feature is almost zero, and thus, the feature seems less relevant.

Figure 11 shows a heatmap for the feature *order hour* per weekday. While the order hour already seems to be useful by itself, it is even more meaningful when viewed as a matrix as depicted in the figure. The figure shows that the highest fraud rates are observed late at night during weekdays (particularly Monday to Wednesday after midnight). Figure 12 complements the fraud heatmap by depicting the relative transaction volume per weekday and hour. Note that the values in Figure 12 have been scaled such that the highest hourly transaction volume observed in the data represents a value of 100, ensuring that both heatmaps are consistent in terms of value range and coloring. It can be seen that most transactions have been issued during the day. The distribution provides evidence for the assumption that the low (relative) fraud rates observed at the weekend do not necessarily imply less fraudulent activity but could also be the result of increased legitimate transactions.

Just like the shipment type feature, the features *internet service provider (ISP)*, *smartphone*, *number of articles*, *new device*, *known address*, *payment type*, *country*, and *customer age* have information gain values of almost zero. In general, this has at least one of the two possible reasons: First, if fraud rates differ insignificantly between feature values — even with sufficient support (i.e. relative frequency) —, information gain will be low. Second, if fraud rates differ more strongly but one of the

	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12p
Mon	0	33	100	0	0	0	7	16	10	17	21	23
Tue	43	56	78	17	27	8	10	8	23	20	18	31
Wed	67	20	50	0	0	5	6	15	26	23	20	25
Thu	11	40	0	33	7	14	10	16	24	22	15	35
Fri	25	0	33	50	17	16	6	10	22	6	18	21
Sat	22	25	0	0	22	0	0	3	8	7	13	18
Sun	0	50	0	0	0	0	20	13	3	12	5	11
	1p	2p	3p	4p	5p	6p	7p	8p	9p	10p	11p	12a
Mon	18	30	30	19	22	19	7	15	19	39	50	19
Tue	19	17	25	15	22	8	5	18	11	27	65	33
Wed	30	43	11	14	15	10	6	21	13	26	36	57
Thu	23	18	16	23	22	0	18	15	50	18	33	25
Fri	32	17	17	12	9	0	0	6	42	19	13	62
Sat	14	18	20	8	16	10	7	15	11	6	33	12
Sun	14	16	6	14	10	9	3	7	24	13	5	25

Figure 11: Heatmap per weekday and hour, with fraud rate values shown in percent. The hours are left-inclusive, i.e. the value 7a includes all transactions from 6 a.m. to 7 a.m.

	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12p
Mon	5	3	1	0	10	13	27	48	60	75	50	69
Tue	7	9	9	6	11	25	38	51	63	100	59	86
Wed	12	5	6	0	9	18	32	47	64	87	59	75
Thu	9	5	3	6	14	14	29	53	74	92	40	61
Fri	8	6	3	2	6	18	17	40	61	46	39	38
Sat	9	4	3	2	9	3	17	32	35	42	30	38
Sun	3	2	1	1	6	5	10	30	38	58	59	53
	1p	2p	3p	4p	5p	6p	7p	8p	9p	10p	11p	12a
Mon	59	54	54	75	48	25	29	26	26	48	33	20
Tue	68	61	66	83	44	13	19	21	17	54	39	17
Wed	86	88	62	87	58	19	16	23	22	37	27	20
Thu	55	43	36	46	31	3	11	19	8	21	29	12
Fri	46	46	47	41	21	16	14	17	12	30	15	25
Sat	36	32	39	48	44	38	41	52	35	16	15	8
Sun	43	55	65	75	65	34	33	29	57	37	21	8

Figure 12: Frequency distribution of transactions. This heatmap shows the relative volume of transactions per cell. Values are *not* percentages; they have been scaled such that 100 represents the highest hourly transaction volume seen in the data.

feature values dominates all the others in terms of support, information gain will be low as well.

6.3.4 *Limitations*

Even though information gain offers a good overview of feature relevance, the results of the analysis should be viewed with regard to a couple of limitations. The first limitation is a general restriction of information gain, and the second refers to how the analysis was conducted.

The first aspect refers to the following idea: Think of a fictional boolean feature stating whether the sun shone at the time the order was placed. Further assume that fraud is committed if and only if the day is cloudy. Then, the conditional probability of fraud given a sunny day would be zero ($P(\text{fraud}|\text{sunny}) = 0$) and the conditional probability for fraud given a cloudy day would be one ($P(\text{fraud}|\text{cloudy}) = 1$). This setting *could* lead to a high information gain, but it does not have to. Consider the case in which the sun shines almost always: Although fraud could perfectly be predicted, the weighting factor (w_i) would scale down the impact of cloudy days so that there would be almost no information gain. This circumstance is not necessarily a disadvantage of information gain, but certainly one would have to ask whether this behavior is desired. On the one hand, the given feature is a perfect predictor, but on the other hand, it is only valuable with regard to a tiny amount of transactions. In general, it is argued that information gain is not supposed to find such corner cases but instead to rate the feature's overall ability to separate fraud from legitimate orders.

The other aspect to consider refers to the combination of feature values. With regard to information gain, all features are

evaluated individually. However, reality is more complex. For example, the fraud risk associated with new customers that send expensive orders to a parcel shop might be disproportionately higher than the corresponding individual risk values. As mentioned in the introduction, feature combinations could turn out to provide additional value like the multivariate visualization in Figure 11 indicates. This concept will be addressed in the next section.

6.4 DECISION TREES

6.4.1 *Introduction*

As has been stated in the preceding section about limitations, the information gain analysis examines features individually and does not consider combinations of feature values. In this section, the idea of feature combinations is implemented via a decision tree; alternative techniques such as association rule mining could lead to similar results (Witten, Frank, and Hall, 2011). In Chapter 7, two methods of supervised machine learning, logistic regression and gradient boosted trees, will be introduced, which serve as classifiers for automated fraud detection and form the foundation for the utility-based perspective in Chapter 8. Conceptually, gradient boosted trees build upon the theory of decision trees. In this section, the theoretical background of decision trees is introduced, an instance is trained, and the tree is evaluated and interpreted in order to estimate the value of feature combinations compared to analyzing features in an isolated way, such as done in the information gain analysis.

Decision trees have been applied to health care fraud (Li et al., 2008), credit card fraud (Sahin, Bulkan, and Duman, 2013), credit assessment (Chung and Suh, 2009), and automobile insurance fraud (Viaene et al., 2002). Therefore, it seems reasonable that the method could work well in the e-commerce mail order business, too.

From a theoretical point of view, Mitchell (1997, p. 52) states: "Decision tree learning is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of disjunctive expressions." This is another way of stating that decision trees can learn rules based on discrete-valued data that may contain errors. Such rules can represent an arbitrary combination of feature values.

Several algorithms for decision trees exist (ID3, C4.5, and CART), but as decision trees are a means to an end, the algorithms are not described in detail. In general, decision trees follow a *divide-and-conquer* approach (Witten, Frank, and Hall, 2011). "Nodes in a decision tree involve testing a particular attribute" (Witten, Frank, and Hall, 2011, p. 64). For example, the first node could — and, as the results show, will — split the dataset into two subsets, one containing transactions issued from freshly created customer accounts and one containing transactions issued from more mature customer accounts.

In order to determine which attribute is most appropriate for splitting, a performance metric such as information gain or Gini impurity is calculated at each node until a stopping criterion is met. The simplest stopping criterion is that there are no more data to divide; others include a maximum tree depth or ratios between frequencies of the attributes' levels. After construction of the tree, it can be pruned in order to reduce the risk of overfitting (Mitchell, 1997). "If the number of coefficients is

large relative to the number of training instances, the resulting model will be 'too nonlinear' — it will overfit the training data" (Witten, Frank, and Hall, 2011, p. 224). This means that the model is too specific, i.e. it does not generalize well. Therefore, overfitting should be avoided.

Decision trees offer an attractive knowledge representation, which can be visualized in a convenient manner. In addition, the tree-based layout can easily be expressed as a set of rules.

6.4.2 *Analysis*

For the analysis, the python machine learning library *scikit-learn* is used. The tree is built as follows: First, the algorithm looks for the feature that creates the highest information gain, which is the *account age* feature. With Figure 8 in mind, this might be surprising because, although *account age* was among the most relevant features, it did not clearly stand out. However, although the decision tree algorithm relies on information gain in order to select which attribute is most valuable, features are discretized in a slightly different way than shown above. The algorithm tries *every possible* cut-off value, which leads to a marginally different result than in the information gain analysis and puts *account age* at the top of the list. Second, once the root has been set, the data are split into two groups. Then, for each group, the process is repeated until a tree of satisfying length has been retrieved. In this case, the tree has only a depth of three levels in order to generate well arranged rules. Figure 13 not only shows results but might also help understand the process. In this analysis, the full dataset has been used for training the tree, because this chapter aims at exploring and evaluating the features. It may seem unusual not to reserve a fraction of

the data for model testing, but the decision tree would be penalized compared to the information gain analysis if it were trained using only a subset of the data. In contrast, in the following chapter, which aims at training prediction models, k-fold cross-validation will be used in order to avoid overfitting of the models (Segaran, 2007; Witten, Frank, and Hall, 2011).

6.4.3 *Findings*

The beginning of the tree is shown in Figure 13, plotted from left to right. The first bifurcation splits the data according to the account age. The upper branch contains the data associated with accounts older than 4.5 days, and the lower branch contains the complementary set. There are two metrics built into the visualization: an arc above and an arc below each node (circle). The red arc represents the fraud rate, and the black arc represents the relative share of the data. 100% corresponds to a value of 180 degree. This is why the black arc is a semicircle — the first node represents the whole dataset. The node is labeled *account age* because the following split is performed with regard to that feature. The average fraud rate with regard to the full dataset is 18.9%, hence the red arc spans $0.189 \cdot 180° = 34.02°$.

The figure shows just the first three levels of the tree — the other features are hidden in deeper levels. However, one can already find a combination of feature values (rule) with one of the highest fraud rates of 75.6% discussed so far: It is the subset of transactions that corresponds to the node labeled ⑥ in the tree visualization and to which the following setting applies: The account from which the order has been issued is not older than 4.5 days; the order has been sent to a parcel shop; and it

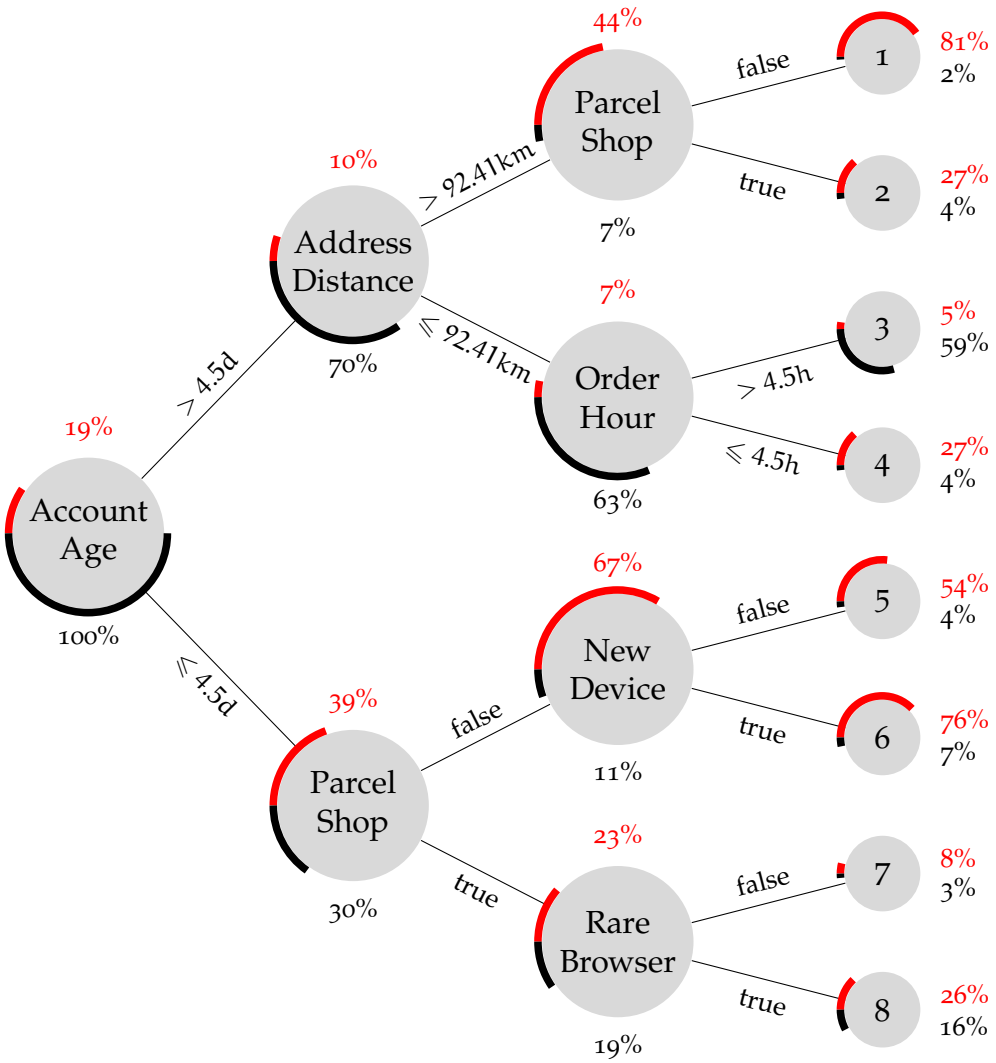

Figure 13: Decision tree with three layers with the fraud rate (red) and the fraction of the data (black). 100% corresponds to a value of 180 degree. Fraud rates (red) and relative frequencies (black) are rounded.

has been issued from an unknown device, i.e. a computer or mobile phone not yet registered with a fingerprint.

In contrast, one of the lowest fraud rates of 5.25% can be found at the subset of transactions labeled (3) in the figure. The corresponding rule is: The account associated with the transaction is older than 4.5 days; the shipping address lies within 92.41km of the invoice address; and the transaction has not been issued between midnight and 4:30 a.m.

6.4.4 *Limitations*

In the previous sections, sharp decision borders, such as distinguishing between transactions with a distance greater than 92.41km between shipping and invoice address, or less, respectively, may seem arbitrary. Yet, they are not — the decision tree decided to choose such cut-off values with optimization of information gain in mind. Yet, a binary decision border may not be the best fit for a natively metric feature such as the address distance. Some types of decision trees support multiway splits, i.e. splitting into more than two subsets per junction. However, "[w]hile this can sometimes be useful, it is not a good general strategy. The problem is that multiway splits fragment the data too quickly, leaving insufficient data at the next level down. Thus, one would want to use such splits only when needed. Since multiway splits can be achieved by a series of binary splits, the latter are preferred" (Hastie, Tibshirani, and Friedman, 2013, p. 311).

Decision trees impose a hierarchical structure on the data. Thus, at least one feature — the account age in this case — is always set. In contrast, in association rule mining, rules are derived that do not necessarily share at least one common feature (Witten, Frank, and Hall, 2011). The fact that decision trees rely on a hierarchy does not have to be a disadvantage, though, as long as the features assigned to the higher-level nodes are actually most important. The gradient boosted trees implemented in Chapter 7 alleviate this problem by building more than just one tree and by randomizing the design of the trees.

An important remark is that the fraud labels are overrepresented in the dataset, which leads to possibly unrealistically high fraud rates. This circumstance is problematic if the values

Features with highest information gain:
• Total price • Account age • Address distance • Browser • Parcel shop
Key combinations according to decision tree:
• Old accounts, high address distances, not parcel shop • Fresh accounts, not parcel shop, new device • Fresh accounts, not parcel shop, old device
Strongest coefficients of logistic regression (cf. Chapter 7)
• Parcel shop • Operating system: Windows XP • Payment type: deferred payments • Browser: other (lowers risk) • Order origin: phone (lowers risk)

Table 5: List of key features according to information gain, decision tree, and logistic regression.

are interpreted as estimates of actual probabilities. The issue will be picked up again in Chapter 8.

6.5 CONCLUSION

At the beginning of the feature analysis chapter, the concept of feature relevance was introduced and motivated. A theoretical overview was given, and information gain was presented first. Results showed that a subset of the available features could serve as promising fraud risk indicators. Moreover, results showed that some of the features could possibly be discarded, although feature relevance depends on the context, and with regard to other algorithms, other sets of features could be

relevant. Descriptive statistics were presented both numerically and visually in order to interpret results. Following a discussion of limitations, the idea of feature combinations was discussed, for which decision trees were introduced briefly. A single tree was trained on the dataset and visualized in order to point out the exploratory power of the tree and the set of rules that can be derived from it.

Table 5 summarizes the key features with regard to information gain, feature combinations of the decision tree, and coefficients of logistic regression. The latter results will be discussed in detail in the next chapter, which deals with machine learning. The decision tree was the first touchpoint for supervised machine learning, but as part of the information gain analysis, it was purely used for descriptive purposes. In the next chapter, two machine learning algorithms will be trained in order to actually predict the fraud status of new transactions.

7

PREDICTION MODEL

7.1 INTRODUCTION

In Chapter 4, the fraud prevention process was examined from a qualitative point of view, and in the chapters thereafter, the quantitative dataset was introduced and the features' ability to help detect fraud was analyzed. Both the qualitative and the quantitative research that has been conducted so far have a descriptive character.

In contrast, this chapter deals with building a prediction model that can estimate the fraud risk of future transactions on the basis of the same dataset that has been used in Chapter 6. In order to achieve this, machine learning is utilized, which "is concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell, 1997, p. XV).

> *Machine learning is a subfield of artificial intelligence (AI) concerned with algorithms that allow computers to* learn. *What this means, in most cases, is that an algorithm is given a set of data and infers information about the properties of the data — and that information allows it to make predictions about other data that it might see in the future. This is possible because almost all nonrandom data contains patterns, and these patterns allow the machine to generalize. In order to generalize, it trains a*

> model *with what it determines are the important aspects of the data (Segaran, 2007, p. 3).*

In the context of fraud detection, machine learning algorithms can learn patterns frequently associated with fraud and thus help identify suspicious cases. In the stricter sense, *learning* refers to supervised methods (the distinction between supervised, unsupervised, and semi-supervised methods has been addressed in the literature review in Chapter 2).

> *There are many different machine-learning algorithms, all with different strengths and suited to different types of problems. Some, such as decision trees, are transparent, so that an observer can totally understand the reasoning process undertaken by the machine. Others, such as neural networks, are* black box*, meaning that they produce an answer, but it's often very difficult to reproduce the reasoning behind it (Segaran, 2007, p. 4).*

Data mining, statistics, and machine learning are often not clearly separated from each other. According to Witten, Frank, and Hall (2011, p. 8), "[d]ata mining is a topic that involves [machine] learning in a practical, non-theoretical sense." Regarding the distinction between statistics and machine learning, Witten, Frank, and Hall (2011, pp. 29 sq.) state:

> *In truth, you should not look for a dividing line between machine learning and statistics because there is a continuum — and a multidimensional one at that — of data analysis techniques. Some derive from the skills taught in standard statistics courses, and others are more closely associated with the kind of machine learning that has arisen out of computer science. If forced to point to a single difference of emphasis, it might be that statistics has been*

> *more concerned with testing hypotheses, whereas machine learning has been more concerned with formulating the process of generalization as a search through possible hypotheses. But this is a gross oversimplification: Statistics is far more than just hypothesis testing, and many machine learning techniques do not involve any searching at all.*

The chapter is organized as follows: First, machine learning is presented from a conceptual point of view. Second, algorithms used in this thesis are discussed in greater detail. Third, the machine learning procedure is conducted, and findings are presented. Fourth, limitations of the chosen approach are discussed. Finally, the chapter closes with conclusions drawn from the preceding findings.

7.2 MACHINE LEARNING FUNDAMENTALS

7.2.1 *Introduction*

In machine learning, supervised, unsupervised, and semi-supervised machine learning methods exist, as described in the literature review in Chapter 2. In the following paragraphs, the three strategies are revisited.

Supervised methods use reference labels, i.e. a ground truth (fraud labels), to learn patterns from historical data. In contrast, unsupervised techniques do not require such labels; they only look at the transactions' properties (even if labels were available). Clustering (Hastie, Tibshirani, and Friedman, 2013) and outlier detection (Aggarwal, 2013) are examples for such unsupervised learning techniques. In fraud detection, outlier detection could identify transactions with suspiciously high or-

der values. Also, sudden changes in buying behavior, such as much more frequent orders, could constitute suspicious behavior (Aggarwal, 2013). Semi-supervised machine learning combines both strategies and is appropriate in a setting where both strategies would not lead to the best results used by themselves (Chapelle, Schölkopf, and Zien, 2010), for example when only a subset of the data is labeled.

In this thesis, supervised techniques are used for two reasons: First, fraud labels are available. Second, the aim of the analysis is to predict the fraud status of future transactions. Finding anomalies (outliers) *could* also indicate fraud, but learning from fraud labels is a more direct inference than assuming that anomalous behavior equals fraud. Of course, it could be valuable to extend the analysis with unsupervised methods in future research. However, this extension would exceed the scope of this thesis. The choice of specific supervised learning techniques will be discussed later.

According to Mitchell (1997), most of these supervised learning problems can be characterized through four generic modules: a type of training experience, a target function, a representation of the learned function, and a learning algorithm. Mitchell (1997) explains these components by referring to a chess game.

Regarding the training experience, a machine learning algorithm can, for example, gather experience by playing against itself or against experts (Mitchell, 1997). In fraud prevention, the former could be difficult to achieve since, in contrast to a game like chess, there is no defined set of rules according to which criminals create new fraudulent transactions. Thus, the algorithm must learn through externally induced data such as the fraud reference labels contained within the transaction dataset.

	predicted fraudulent	predicted legitimate
fraudulent	true positive (TP)	false negative (FN)
legitimate	false positive (FP)	true negative (TN)

Table 6: Confusion matrix for binary fraud classification.

In chess, the target function can either evaluate the value of a *move* to a new board state (a state is a description of every chessman's position on the board) or the new board state *itself* (Mitchell, 1997). The difference is that the former approach considers the change and the latter considers the status quo of the game regardless of the move, i.e. other moves could have led to the same state. Unlike in chess, however, where different moves can lead to the same state and different states can precede the same move, the problem is less complex in fraud prevention because each move (either considering a transaction fraudulent or legitimate) leads to one out of only four outcomes, depending on the actual fraud status of the transaction. This relation is depicted in Table 6. Note that the confusion matrix is structurally identical to the decision matrix (Table 1) in Chapter 2 — the cells are just phrased differently. Therefore, the target function only measures whether the move was correct or not or, if a fraud score is computed, how close it is to the real value (such as 0 for legitimate and 1 for fraud).

The representation of the learned function is the entity that is capable of making decisions after it has been trained. For example, this could be an artificial neural network, a linear function or a polynomial (Mitchell, 1997). With regard to fraud, it should be possible to understand predictions in order to justify decisions. While neural networks are black-box models and cannot be understood on a per-case basis, predictions based on decision trees can be explained easily (Mitchell, 1997).

Finally, the learning algorithm must be chosen. It determines *how* the target function is estimated. Mitchell (1997) mentions gradient descent and linear programming as examples. Therefore, the *learning* in machine learning can also be viewed as an optimization of a target function.

Two different supervised machine learning methods are trained and tested in this chapter: logistic regression and gradient boosted trees. The methods are conceptually dissimilar in order to increase generalizability of results. As argued in the literature review, supervised machine learning has proven its effectiveness in various related applications. In addition, logistic regression and random forests (a method similar to gradient boosted trees) have been used by Carneiro, Figueira, and Costa (2017) in order to build a model for credit card fraud detection in the e-commerce mail order business. In the following, logistic regression and gradient boosted trees will be introduced. Note that the notation of the formulas introduced in the following paragraphs has been modified in order to maintain a consistent presentation and may therefore differ slightly from the original sources with regard to what symbols are used and how indices are placed.

7.2.2 *Logistic Regression*

Logistic regression is well known in both statistics and machine learning, and according to Ngai et al. (2011), it is the most frequently used technique in financial fraud detection. "The logistic regression model arises from the desire to model the posterior probabilities of the [...] classes via linear functions [...], while at the same time ensuring that they sum to one and remain in [0,1]" (Hastie, Tibshirani, and Friedman,

2013, p. 119). In particular, since there are only two classes (i.e. $K = 2$), fraudulent and legitimate, the model is simplified because "there is only a single linear function" (Hastie, Tibshirani, and Friedman, 2013, p. 119). Logistic regression accepts a vector of numeric feature values as input and yields a value between zero and one in order to estimate the fraud risk of a transaction. Logistic regression is implemented using the Python package *scikit-learn* (Pedregosa et al., 2011), which is actively developed and has 982 contributors at the time of writing (`https://github.com/scikit-learn/scikit-learn`). The logistic model is formulated as follows (Hastie, Tibshirani, and Friedman, 2013):

$$\Pr(G = k|X = x_i) = \frac{1}{1 + exp(\beta_0 + \beta^T x_i)} \tag{3}$$

- k class (i.e. fraudulent or legitimate)
- x_i feature vector of instance i
- β_0 intercept value
- β variable coefficients

For logistic regression, Equation 3 depicts the representation of the learned function (which is one of the four generic modules of learning problems mentioned above). Given a transaction i with the feature vector x_i, the probability that a transaction is fraudulent is represented by $\Pr = (G = 1|X = x)$. In order to actually compute such values, the parameters of the logistic function β are needed. Obtaining such values is an optimization problem, which *scikit-learn* implements with the *LIBLINEAR* solver (Fan et al., 2008). The target function (another component of the four generic modules) is defined as follows (Fan et al., 2008; Hsieh et al., 2008):

$$\min_{\omega} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l} \zeta(\omega; x_i, y_i) \tag{4}$$

- ω model parameters (called β above)
- C penalty term
- x_i feature vector of instance i
- y_i class of instance i, $y_i \in \{-1, +1\}$
- l number of instances
- ζ loss function

The minimization consists of two parts: $\frac{1}{2}\omega^T\omega$, which is the regularization term (L_2 regularization), and the term that contains the loss function $C\sum_{i=1}^{l} \zeta(\omega; x_i, y_i)$. The former, i.e. the regularization term, punishes very large parameters in order to avoid overfitting of the function. Note that $\omega^T\omega$ is the sum of the squares of the parameters. The latter term simply contains the loss function, which is the following (Hsieh et al., 2008):

$$\zeta(\omega; x_i, y_i) = \max(1 - y_i\omega^T x_i, 0)^2 \tag{5}$$

As Equation 5 shows, correct classifications yield zero loss, and incorrect classifications yield a loss value that corresponds to the difference between the actual (1 or -1) and the estimated value ($\omega^T x_i$). In order to reach the minimum of the target function, *LIBLINEAR* employs dual coordinate descent (Hsieh et al., 2008), which is the learning algorithm — the final component of the four generic modules of machine learning.

Logistic regression is chosen as a method due to its popularity in machine learning. In addition, it will be possible to compare results with those of Carneiro, Figueira, and Costa (2017), who developed a model to detect credit card fraud in the e-

commerce mailorder business and used logistic regression for the same reason.

Using evaluation metrics independent of the algorithm, performance of logistic regression will be measured and compared to that of gradient boosted trees in the findings section of this chapter. However, the coefficients that are specific to logistic regression — and corresponding significance values — are presented in Table 7.

Most of the features are significant at a significance level of $\alpha = 0.05$. It could be assumed that a strong coefficient is correlated with the information gain value as both measures represent the influence of fraud. However, this does not have to be the case: Analysis shows one of the largest coefficents for orders that were made via phone call (-1.4308), but regarding the information gain analysis, the order origin feature shows only average values. A possible reason could be that most orders are made online, and information gain considers this fact when estimating feature importance, leading to a decrease of relevance. In contrast, the coefficent for orders via phone measures the effect on the fraud rate only for phone orders, and it is strong. This example shows that coefficients should be interpreted differently and provide additional insights besides information gain analysis. In general, the estimated coefficients are consistent with the fraud rates measured per feature and results of the information gain analysis.

Feature	coefficient	p
Account age (years)	-0.0904	0.00
Address distance (km)	0.0041	0.00
Browser [Chrome]		
– *Android*	0.5815	0.48
– *Firefox*	0.1395	0.00
– *IE*	-0.5293	0.41
– *Opera*	-0.9984	0.05
– *Safari*	-0.6370	0.00
– *other*	-1.1354	0.00
Country is Germany	-0.4115	0.00
Customer age	-0.0051	0.00
Express shipment	-0.1138	0.01
Known address	-0.1448	0.86
New device	-0.4701	0.22
Number of articles	-0.0731	0.00
Order hour	-0.0288	0.00
Operating system [Win., not XP]		
– *Android*	-0.0860	0.04
– *iOS*	0.2246	0.00
– *Linux*	-0.0916	0.03
– *Mac OS*	-0.3909	0.00
– *Windows XP*	1.5633	0.00
– *other*	0.3253	0.00
Order origin is phone call	-1.4308	0.00
Parcel shop	1.2108	0.00
Payment type [invoice]		
– *credit card*	-0.4877	0.01
– *rates*	1.2843	0.01
Smartphone used	-0.0726	0.00
Total price	-0.0003	0.00

Likelihood Ratios	LR+
at 5^{th} percentile	12.32
at 15^{th} percentile	12.01
at 25^{th} percentile	8.01

Table 7: Coefficients of logistic regression and associated p-values. Likelihood ratios are shown below. Since discrete features have been dummy-coded, their reference values are given in square brackets.

For the logistic regression model, likelihood ratios are calculated. Positive likelihood ratio (LR+) is defined as sensitivity divided by 1 − specificity (Florkowski, 2008). In other words, it is the ratio of two ratios: (1) The ratio of transactions considered fraudulent that are actually fraud compared to all fraud cases divided by (2) the ratio of legitimate cases considered fraudulent compared to all legitimate cases. "The further likelihood ratios are from 1 the stronger the evidence for the [target feature]. Likelihood ratios above 10 [...] are considered to provide strong evidence [...]." (Deeks and Altman, 2004, p. 168).

Since fraud prevention departments process lists of transactions ordered by the estimated fraud risk, likelihood ratios are calculated for multiple positions in the list at different percentiles, which is shown at the bottom of Table 7. Regarding the 5 percent most suspicious transactions, LR+ is 12.32. Therefore, regarding the transactions up to the 5^{th} percentile, it is 12.32 times more likely to classify a fraudulent transaction as fraud than to classify a legitimate transaction as fraud.

7.2.3 *Gradient Boosted Trees*

Tree-based models are quite different to common statistical procedures such as logistic regression. A single decision tree has been presented in Chapter 6 in order to estimate the importance of feature combinations. However, decision trees as implemented in Chapter 6 "can lead to difficulties when there is noise in the data or when the number of training examples is too small to produce a representative sample of the true target function. In either of these cases, this simple algorithm can produce trees that *overfit* the training examples" (Mitchell, 1997, pp. 66 sq.).

In this section, gradient boosted trees are introduced, which better deal with these possible issues and therefore provide more robust solutions. They are implemented using the Python package *xgboost* (DMLC, 2016). Like *scikit-learn*, it is actively developed and has 259 contributors at the time of writing (`https://github.com/dmlc/xgboost`).

Generally speaking, gradient boosted trees are ensembles of trees. These ensembles are iteratively extended such that each iteration improves the results of the current set of trees (Chen and Guestrin, 2016). Each tree is built to complement the existing ensemble and may thus specialize in detecting certain sub-patterns. Equation 6 is the representation of the learned function, and it shows that the estimated score for a transaction i is the sum of the Q additive functions, i.e. trees (Chen and Guestrin, 2016):

$$\hat{y}_i = \sum_{q=1}^{Q} f_q(x_i) \tag{6}$$

- Q number of functions (trees)
- f_q instance q of the space of possible trees
- x_i feature vector of instance i
- $\hat{y}_i$ estimated score for instance i

The resulting values are not automatically probability estimates. However, such probabilities can be obtained by applying the logistic transformation, which is explained in the *xgboost* documentation (`http://xgboost.readthedocs.io/en/latest/model.html`).

Each function f_q represents a single decision tree with weights ω and the total number of leaves L. In order to find the best ensemble, the following target function is minimized (Chen and Guestrin, 2016):

$$\min_{\omega} \sum_{q=1}^{Q} \gamma L + \frac{1}{2}\lambda\omega^{T}\omega + \sum_{i=1}^{n} l(\hat{y}_i, y_i) \tag{7}$$

ω branch weights

γ, λ constants

l differentiable, convex loss function

$\hat{y}_i$ estimated score for instance i

Minimization of the target function is performed with gradient descent of a second order derivative of the target function (Chen and Guestrin, 2016). This procedure represents the learning algorithm and is the reason for *gradient* in gradient boosted trees. The full procedure is documented by Chen and Guestrin (2016).

7.2.4 *Sampling with Cross Validation*

The dataset as it has been prepared in Chapter 5 is sampled with cross validation. In this section, it is explained what cross validation is and why it is important. The prediction model developed in this chapter aims at best forecasting the fraud status of *future* transactions, i.e. it aims at learning a concept (depending on the type of algorithm, this could be a set of rules) that best detects fraud not only regarding the currently available but also future, yet unseen data. In order to achieve this transfer from historical to future data, performance of machine learning models should not be measured using the training dataset. The reason is that it can contain errors (i.e. incorrectly assigned fraud labels), and learning such errors decreases performance in a real application but not regarding the training dataset. Consider the following analogy: A student prepares for an exam

and reads a textbook that contains a couple of errors. The student completes the exercises in the textbook without any problems and probably feels quite confident about the upcoming exam. However, he or she then receives a much lower grade than expected because part of what the student had learned was actually wrong. In machine learning, it is important to test algorithms using another dataset than the one the model has been trained with in order to reduce the error introduced by generalization from a limited number of cases (i.e. the size of the sample).

Therefore, using the training dataset for measuring performance of machine learning models is insufficient (Russell and Norvig, 2012). In order to estimate prediction quality, such models have to be tested with transactions that have not been used for training. One possible method to achieve this is to simply split the dataset into training and test data. However, doing so drastically reduces the amount of data available for both training and testing, which can lower generalizability of results. The sampling method k-fold cross validation alleviates this problem by using each transaction both for training and testing.

> *The standard way of predicting the error rate [i.e. the fraction of false positives and false negatives] of a learning technique given a single, fixed sample of data is to use stratified tenfold cross-validation. The data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus, the learning procedure is executed a total of 10 times on different training*

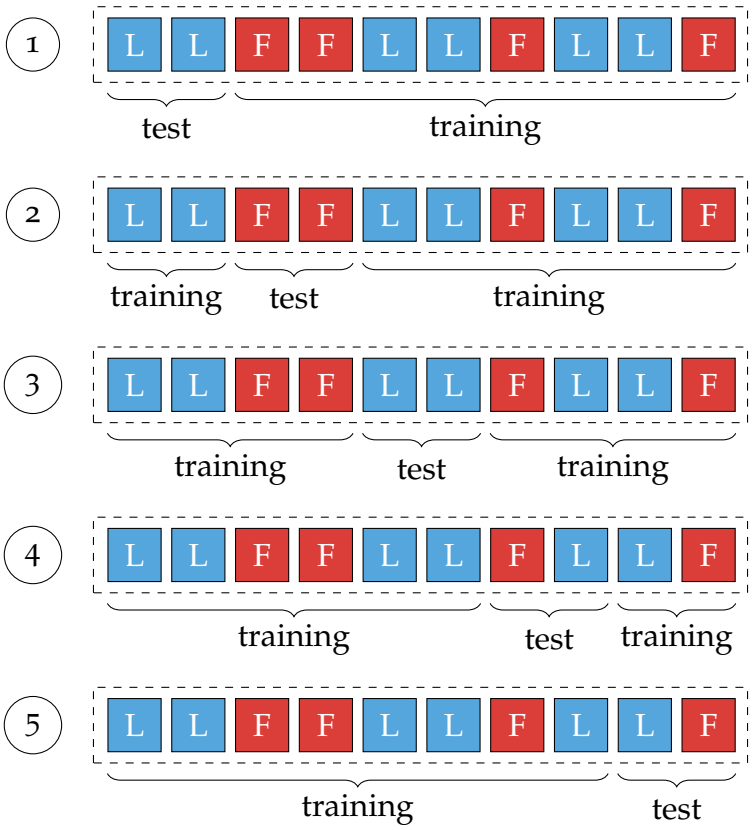

Figure 14: Visualization of 5-fold cross validation; the 10 transactions (L for legitimate and F for fraudulent transactions) are iterated over 5 times, and each time, another fraction is used for training and the remainder is used for testing. In each iteration, the full dataset is used.

> *sets (each set has a lot in common with the others). Finally, the 10 error estimates are averaged to yield an overall error estimate (Witten, Frank, and Hall, 2011, p. 153).*

In the analysis, 10-fold cross validation is performed. The process is also visualized in Figure 14 for $k = 5$ (which is easier to show). Other numbers of folds would have been possible, of course, but Witten, Frank, and Hall (2011) state that there is both theoretical and practical evidence that 10-fold cross validation yields the best results.

7.3 FINDINGS

In order to assess performance of logistic regression and gradient boosted trees for fraud detection, evaluation measures have to be defined. Such a measure (also called metric in the following) should be able to provide immediate insights about an algorithm's relevance with regard to fraud detection. However, not all measures are able to provide this. For example, the coefficient of determination, R^2, is computed for evaluation of regression models, and it answers the question what fraction of the observed variance is explained by the regression model (Fahrmeir et al., 2007). Therefore, models with higher R^2 values are preferred. Although the metric can be used to choose from competing models, it does not enable decision makers to understand the immediate impact of the chosen model on the success of their fraud detection activities. Thus, other metrics will be introduced that allow decision makers to better interpret how the use of an automated fraud detection model influences their work. A detailed overview of various evaluation measures can be found in Chapter A in the appendix. In this section, a brief conceptual overview is provided, and the measures used in the analysis are presented in detail.

Manning, Raghavan, and Schütze (2009) present two categories for evaluation measures, *set-based measures* and *measures for ranked retrieval*. Set-based measures require classification of transactions into distinct groups (sets). Then, relations between these groups can be computed, such as the fraction of correctly classified (labeled) transactions. In contrast, measures for ranked lists consider the order of such lists.

Logistic regression and gradient boosted trees both produce ranked lists with estimated fraud probabilities. This means that a transaction at position k in the list is considered more suspi-

cious (i.e. more likely to be fraudulent) than a transaction at position $k + 1$. Ranked lists can be converted to sets: Carneiro, Figueira, and Costa (2017) search a threshold-value to cut the list into fraudulent and legitimate transactions, and these sets can then be evaluated with set-based measures. For Carneiro, Figueira, and Costa (2017), this is an appropriate strategy, because they develop an automated system which deals with the most suspicious cases and the remainder is inspected manually. Thus, it is inevitable to form such distinct groups in order to assign transactions to either automated processing or manual inspection.

However, as presented in Chapter 4, the fraud detection process explored in this thesis *always* contains manual revision. Fraud officers inspect transactions until their capacity is exhausted. Consequently, it matters in which order transactions considered fraudulent are inspected, and strict assignment into groups (sets) seems inappropriate. The model developed in this chapter aims at supporting that process by ordering transactions according to the estimated fraud risk, indicating which (potentially) fraudulent transactions should be inspected first.

Confusion matrices, the basis for many metrics (Shani and Gunawardana, 2011), are introduced first, from which the measures used in this thesis will be developed. An example for a confusion matrix is shown in Table 6. It depicts the four possible outcomes of classification given the true fraud status. For example, if a fraudulent transaction is correctly predicted as fraudulent, the outcome is a true positive. A perfect prediction model would have only true positives and true negatives, because all cases would be classified correctly. However, it is almost impossible not to make any prediction errors in the given context. In general, "[g]ood results correspond to large numbers down the main diagonal and small, ideally zero, off-

diagonal elements [of the matrix]" (Witten, Frank, and Hall, 2011, p. 164). From the four outcomes shown in the table, additional metrics depicting relations between them can be computed, such as *precision at k*, a measure for ranked lists extending regular precision, which is based on the confusion matrix. Precision at k is regarded as the most important metric for algorithm performance — arguments for this claim will be provided below.

Regular precision is defined as the number of correctly detected fraud cases (i.e. true positive cases) divided by the number of cases that are classified as fraudulent (Shani and Gunawardana, 2011):

$$\text{precision} = \frac{\text{true positives}}{\#(\text{labeled fraud})} \tag{8}$$

Then, precision at k is defined as the precision value for the set of transactions up to the k^{th} percentile of a ranked list:

$$\text{precision at } k = \frac{\text{true positives within } k^{\text{th}} \text{ percentile}}{\#(\text{labeled fraud within } k^{\text{th}} \text{ percentile})} \tag{9}$$

Assume that an e-commerce mail order company is able to inspect approximately k percent of its transactions, and using a prediction model such as developed in this chapter, the company creates a ranked list ordered by the fraud risk. Ideally, all transactions up to the inspection capacity (corresponding to the k^{th} percentile) should be fraud cases, or alternatively all fraud cases contained in the dataset should be among the top k percent of the cases presented. Hence, in this particular application, precision at k is just the fraud ratio of all cases until the k^{th} percentile. The measure does not only describe the status quo

but can also indicate an optimal degree of utilization because one can plot a precision curve and understand how many fraud cases different levels of inspection capacity could yield.

The precision curves for logistic regression and gradient boosted trees are shown in Figure 15. Gradient boosted trees slightly outperform logistic regression, but the more k increases — moving down the lines to the right —, the less the difference between the two is. Eventually, when all transactions are considered, it does not matter which algorithm is used for sorting transactions and precision at k is the average fraud ratio of the whole dataset. An advantage of precision as an evaluation metric is that the relative number can be converted to an absolute one in order to estimate the impact on a business (the coefficient of determination was criticized for lacking this property): For example, if the 2% most suspicious transactions were reviewed, gradient boosted trees would yield a precision rate of 85% (75% for logistic regression). Imagine that a company processes 5,000 transactions per day; then, the top 2% correspond to 100 transactions, of which approximately 85 would be fraudulent (75 cases using logistic regression). In contrast, if 100 transactions were randomly sampled, only approximately 19 would be fraudulent because the fraud rate in the dataset is 19%.

Although precision at k is considered the most appropriate metric regarding the fraud prevention process, two alternative evaluation techniques will be discussed below. The first is the Receiver Operating Characteristics (ROC) curve, which is shown in Figure 16. "ROC curves depict the performance of a classifier without regard to class distribution or error costs. They plot the true positive rate on the vertical axis against the [false positive] rate on the horizontal axis" (Witten, Frank, and Hall, 2011, p. 172). The former is the fraction of true positives compared to

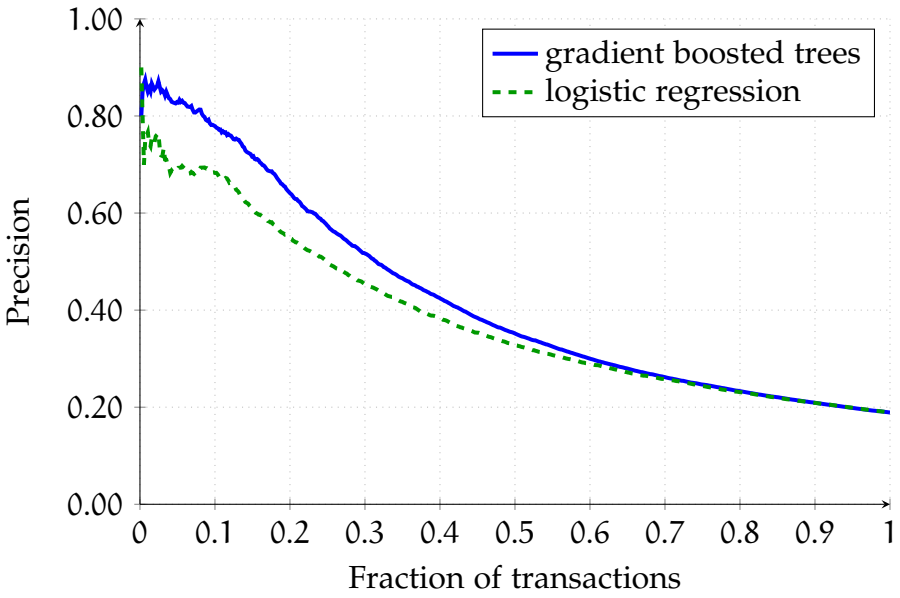

Figure 15: Precision at k^{th} percentile.

all fraud cases; the latter is the number of false positives compared to all legitimate cases (Witten, Frank, and Hall, 2011). The best possible curve is closest to the upper left corner (Witten, Frank, and Hall, 2011). The worst possible curve is a straight line from the lower left to the upper right.

"To summarize ROC curves in a single quantity, people sometimes use the area under the curve (AUC) because, roughly speaking, the larger the area the better the model. The area also has a nice interpretation as the probability that the classifier ranks as a randomly chosen positive instance above a randomly chosen negative one" (Witten, Frank, and Hall, 2011, p. 177). The AUC values for logistic regression and gradient boosted trees are 0.85 and 0.9, respectively. The AUC metric *summarizes* the whole ROC curve, and with regard to fraud detection, i.e. in the case of limited inspection resources, only as many transactions are of interest as can actually be inspected.

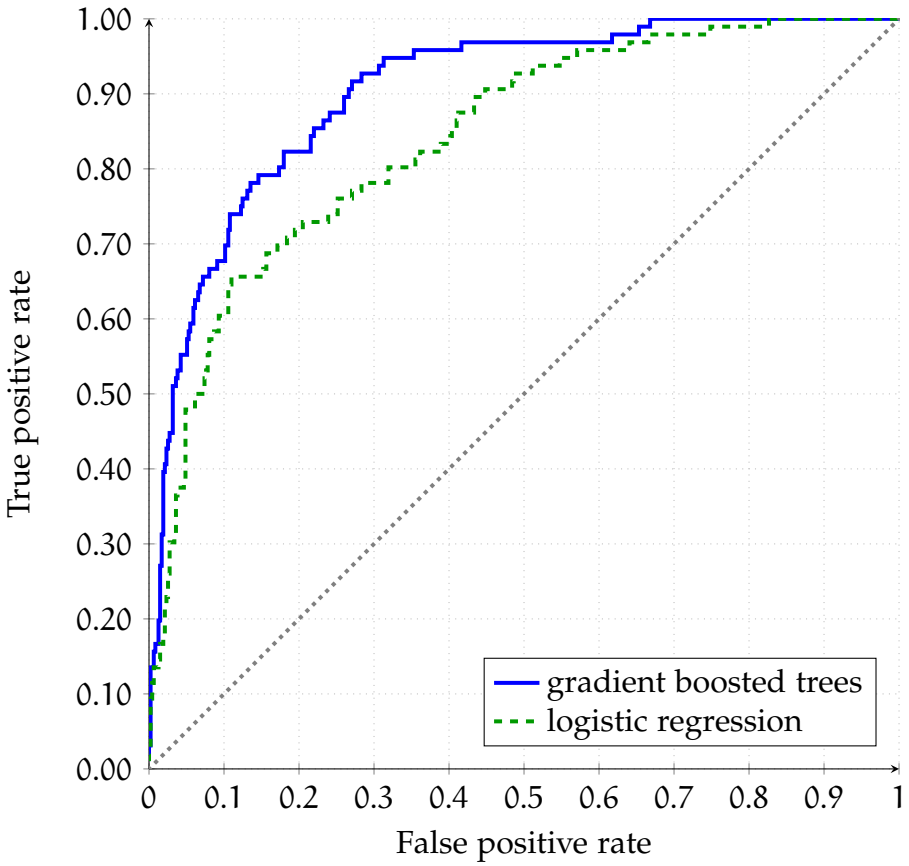

Figure 16: ROC curves.

7.4 LIMITATIONS

Before drawing conclusions from the results of the analysis, a couple of aspects have to be considered. The mentioned limitations refer to two areas: the choice of methods and their evaluation.

First, as pointed out in Chapter 4, labels in the dataset are created through manual inspection, and the process of deciding which transactions are selected for inspection is based on experience of fraud officers (in addition, a fraction of the transactions is selected randomly). This explains the remarkable fraud ratio of 18.9% among the labeled transactions, although the fraud ratio in e-commerce businesses is usually lower; for example, CyberSource (2017) states a fraud rate of approximately

1%, and Quah and Sriganesh (2008) refer to statistics by Garner that mention fraud rates between 0.8% and 0.9%. Thus, the dataset is biased toward fraudulent transactions, and the prediction model might therefore face an increased amount of false positive results when evaluating new, unlabeled transactions because the model would expect more fraud than actually exists in the dataset. In addition, some patterns might not be detected simply because fraud prevention employees never select the transactions that would help prediction models learn them.

Second, the analysis relies on supervised learning methods. Both methods, logistic regression and gradient boosted trees, are well established in machine learning (Chen and Guestrin, 2016; Hastie, Tibshirani, and Friedman, 2013; Witten, Frank, and Hall, 2011). Yet, unsupervised or semi-supervised machine learning could allow to approach the problem from another perspective and lower the impact of the aforementioned selection bias. For example, in contrast to automating the existing inspection process with supervised machine learning, unsupervised methods such as outlier detection could help find new kinds of anomalous patterns in the data and thus create new insights (Aggarwal, 2013).

Third, in this thesis, precision at k heavily depends on class imbalance, i.e. the high fraud ratio. This means that a high fraud rate may correspond to generally higher precision values. Therefore, if multiple datasets with different fraud ratios were to be compared with one another, precision would not be an appropriate measure anymore. This is a hypothetical limitation, though, because only one dataset is used in this analysis.

7.5 CONCLUSION

In this chapter, a prediction model based on supervised machine learning was developed. Supervised methods were preferred because the dataset contains fraud labels which are considered reliable, and they should be utilized for model training. However, other techniques like unsupervised machine learning methods could be used in future work in order to approach the problem from another perspective. Various different supervised machine learning algorithms exist, but logistic regression and tree-based models are among the most popular methods used in other studies in related areas (Ngai et al., 2011).

With regard to evaluation of the prediction models, many performance measures seem inappropriate because they do not reflect the way fraud departments work. Two categories of evaluation measures, set-based measures and measures for ranked lists, were introduced. It was explained that, facing limited inspection resources, a good model produces a ranked list that positions the most suspicious cases at the top of the list. Results were presented through the following measures: precision at k, ROC curves, and the AUC value. The first metric, precision at k, best fits the problem, although it makes comparison of multiple datasets with different fraud ratios difficult. It was found that both models perform well, with gradient boosted trees being slightly ahead of logistic regression. Thus, the fraud probability estimates needed in the following chapter are exclusively retrieved from gradient boosted trees.

Results show that the variables presented in Chapter 6 are useful and allow to create prediction models that are capable of supporting fraud departments in preselecting suspicious transactions. Until now, the prediction model simply aimed at detecting fraud cases. However, e-commerce companies might be

more interested in a reduction of the the actual financial damage. In the next chapter, the fraud probability estimates provided by the prediction model will be used in conjunction with sales price data and margin data in order to introduce an economic perspective on fraud prevention.

8

UTILITY APPROACH

8.1 INTRODUCTION

In the previous chapter, it was shown how two machine learning models (logistic regression and gradient boosted trees) have been trained using historical transaction data, and these models are able to estimate the fraud risk of new transactions. Their performance was evaluated with metrics often used in data mining in general and for fraud detection in the e-commerce mail order business by Carneiro, Figueira, and Costa (2017). However, from an economic point of view, such measures are regarded problematic: Even though less fraud is always better, sometimes priorities must be set owing to a lack of inspection resources. Consequently, it might sometimes be worthwhile to tolerate a certain amount of minor fraud cases in order to focus on a few cases that have a high financial impact. For example, as mentioned in the literature review, a few stolen pairs of socks should be accepted when a fraudulent transaction containing an expensive smartphone can be detected and stopped instead. Therefore, it is proposed to use performance measures that reflect the financial impact of fraud cases instead of purely count-based measures both for scoring and sorting of transactions and for evaluation of the models. Torgo and Lopes (2011) discussed a utility-based perspective on fraud detection using artificial and foreign trade data. In this chapter, their ideas are introduced, further developed, and applied to the context of the e-commerce mail order business.

The chapter is structured as follows: First, the work of Torgo and Lopes (2011) is discussed, and their approach is reviewed from a critical stance with the e-commerce mail order business in mind. Based on the results of the review, a formalization of the utility concept is tailored to the context of the thesis. Third, the empirical analysis is presented. Finally, results are discussed.

8.2 UTILITY CONCEPT

8.2.1 *Work of Torgo and Lopes*

Torgo and Lopes (2011, p. 1517) state that "[i]nspection activities associated with [fraud detection] are usually constrained by limited available resources. Data analysis methods can provide help in the task of deciding where to allocate these limited resources in order to optimize the outcome of the inspection activities." The authors argue that, when an algorithm assesses transactions, besides the estimated fraud probabilities, "the inspection costs and expected payoff if the case is confirmed as a fraud" should as well be considered (Torgo and Lopes, 2011, p. 1517). Their idea borrows concepts from utility theory and emphasizes that the risk alone is an insufficient measure from the perspective of profit-oriented companies.

It is important to note that Torgo and Lopes (2011) do *not* focus on the e-commerce mail order business. Thus, possible remarks are not criticisms *per se* but view their research from a more specialized perspective.

Fraud detection involves two parts: An algorithm ranks transactions according to a criterion, and fraud officers review the list, inspecting the highest-ranked transactions first (Torgo and

Lopes, 2011), regardless of whether transactions are ranked by the estimated fraud risk or a utility metric. This scenario fits the fraud prevention process described in Chapter 4 well.

In general, utility can — but it does not have to — represent a financial value (Friedman and Sandow, 2011; Peterson, 2009; von Neumann and Morgenstern, 1953). In the thesis, this is always the case for the sake of simplicity. Torgo and Lopes (2011, p. 1518) propose the formula below:

$$E[U_i] = \hat{P}_i \cdot u(\hat{B}_i - \hat{C}_i) + (1 - \hat{P}_i) \cdot u(-\hat{C}_i) \tag{10}$$

$E[U_i]$	expected utility of transaction i
$\hat{P}_i$	estimated probability of i being a fraud
$\hat{B}_i$	estimated benefit of case i if confirmed fraudulent
$\hat{C}_i$	estimated inspection cost of case i
$u(.)$	utility function

The formula covers two cases, which are weighted with the estimated fraud risk: Either the transaction at hand is fraudulent, in which case cost and benefit are considered (first part of the equation), or it is legitimate, meaning that only the cost are considered (second part of the equation). Please notice that the formula assumes that the fraud status of a transaction is known for sure once it has been inspected. This assumption aligns well with how the fraud prevention process was described in the interviews in Chapter 4.

All components of the equation are estimates: In this thesis, the estimates for the fraud risk $\hat{P}_i$ come from the classifiers in Chapter 7. The components $\hat{B}_i$ and $\hat{C}_i$ "are clearly application dependent" (Torgo and Lopes, 2011, p. 1518) and require further discussion. Thus, the utility functions $u(.)$ in the equation will be replaced with refined terms that are tailored to the problem faced in the e-commerce mail order business.

With regard to the research problem, the estimated cost of inspection can be considered constant for all transactions; this assumption is inevitable for the empirical part because no data about variable inspection cost are available, as it has been stated in the interviews in Chapter 4. Even if the assumption were false, it would be a simple task to add the inspection cost to the formula.

In the formula, estimated benefits refer to the positive effect that occurs when a fraud case is detected. In the e-commerce mail order business, the benefit covers at least the sum of the purchase prices of all items contained in the order. In addition, the benefit could include parts of the value chain like cost of logistics, operations, and service; to sum up, activities required to deliver the order increase the benefit of a fraud case which is detected before it is too late. It has been discussed in Chapter 4 that fraud prevention is often performed while an order is being processed, and the later a transaction is tried to be canceled, the more expensive the cancellation becomes. There are specific events like dispatching the order in the warehouse that increase the cost of cancellation such that the cost function takes the shape of a step function. It is difficult to estimate such costs, and it is simpler to focus on the purchase price. Note that, with regard to the benefit of fraud prevention, the sales price by itself is irrelevant because it is a hypothetical value, which does not reflect financial damage because it includes the merchant's profit margin as well.

8.2.2 *Revised Concept*

In this section, the remarks made in the previous section are referred to in order to tailor the utility approach to the e-

commerce mail order business. First, an essential conceptual question is discussed. Carneiro, Figueira, and Costa (2017, p. 100) conclude in their research: "We recommend the exploration of other ways to choose the score threshold". The authors continue: "[A] cost-based performance measure could be used in training the algorithm in order to further steer the learning process towards the intended business outcome". The authors recommend to use such a cost-based performance measure for *training* of the machine learning model itself. Then, the model such as logistic regression or gradient boosted trees would not only learn to differentiate between fraudulent and legitimate transactions but also to consider the value of the transaction. In this thesis, it is argued that such an approach introduces unnecessary complexity to the problem. As long as the prediction model yields estimated fraud probabilities, utility theory can be used to implement an economic perspective. This strategy allows to extend any existing model that produces estimated probabilities. Therefore, the problem can be formulated as two clearly separated subproblems, the prediction model that uses machine learning (Chapter 7) and the utility approach (this chapter). This thought also accords with the process view that Torgo and Lopes (2011) hold.

As explained earlier, the inspection process is a classic decision problem (Peterson, 2009): An actor, the fraud officer, has two choices — he or she can accept (a) or reject (r) an order —, and the transaction has one of two states — it is fraudulent (f) or legitimate (l). This decision problem leads to four possible outcomes, which are visualized in Table 8. Note that such a matrix has first been presented in Chapter 2, and it has been revisited for the purpose of algorithm evaluation in Chapter 7. In Table 8, the cells are modified in order to consider the utility terms associated with each outcome.

	fraudulent	legitimate
accept	A fraud case is labeled legitimate by mistake, resulting in $u_i(a \mid f)$.	A legitimate order is processed, resulting in $u_i(a \mid l)$.
reject	A fraud case is successfully prevented, resulting in $u_i(r \mid f) = 0$.	A legitimate order is rejected by mistake, resulting in $u_i(r \mid l) = 0$.

Table 8: Outcomes of fraud classification with states fraudulent (f) and legitimate (l) and acts accept (a) and reject (r).

For the four outcomes, the notation $u_i(\text{choice} \mid \text{state})$ is established. The problem resembles the formulation of the classification task in Chapter 7, and it could be tempting to refer to the outcomes as true positives, false negatives etc.; however, this is avoided in order not to confuse the outcomes of the decision problem with the outcomes of the classifiers.

If a transaction is rejected, its utility is assumed to be always zero because the items are not shipped at all or shipping is canceled. This assumption might not be true, though: Even if the cost of inspection are ignored — because they are considered constant —, cancellation may be expensive if the order is already being delivered. However, data about such measures were not available; therefore, as a simplification, utility of rejected transactions is assumed to be zero. The outcomes that matter financially and vary per transaction are $u_i(\text{accept} \mid \text{fraudulent})$ and $u_i(\text{accept} \mid \text{legitimate})$. In the former case, the value of the purchase price is lost, and in the latter case, a successful sale is made and approximately the difference between the selling and the purchase price is earned. This perspective focuses on the actual cash flow. It would also be possible to set the utility values of the aforementioned two outcomes to zero and instead assign $u_i(\text{reject} \mid \text{fraudulent})$ and $u_i(\text{reject} \mid \text{legitimate})$ non-zero values. In this thesis, the first ap-

proach is preferred because it allows to make statements about the actual cash flow.

The formula below, Formula 11, is similar to the one proposed by Torgo and Lopes (2011), but it takes a different perspective. It allows to answer the question: Given the fraud status of a transaction, what would the financial impact be if it were accepted? The two cases (accept, reject) are weighted with the estimated fraud risk. Therefore, it is estimated that $\hat{E}[U_i|\text{accept}]$ would be earned (or lost if negative) per accepted transaction.

With the e-commerce mail order business in mind, the utility functions $u(.)$ are implemented as follows: For a transaction composed of n items (i.e. products), the product of the sales price S_{ki} and the margin M_{ki} is iterated over in order to compute $u_i(\text{accept} \mid \text{fraudulent})$ and $u_i(\text{accept} \mid \text{legitimate})$. i and k are indices, which refer to the k^{th} item of the i^{th} transaction. The relation between risk and margin is visualized in Figure 17: Given a constant sales price, the smaller the margin, the larger the potential loss if a fraud case remains undetected. In contrast, high margins alleviate the impact of fraud. The sales price acts as a regular factor and scales the expected value.

$$\hat{E}[U_i|\text{accept}] = (1 - \hat{P}_i) \cdot u_i(a \mid l) - \hat{P}_i \cdot u_i(a \mid f) \tag{11}$$

$$= (1 - \hat{P}_i) \sum_{k=1}^{n} M_{ki} S_{ki} - \hat{P}_i \sum_{k=1}^{n} (1 - M_{ki}) S_{ki} \tag{12}$$

$$\hat{E}[U_i|\text{reject}] = 0 \tag{13}$$

An interesting side effect of the notation is the following insight: If a transaction contains exactly one item, the equation can be simplified, and it is found that the margin has to exceed the risk in order to observe positive expected utility:

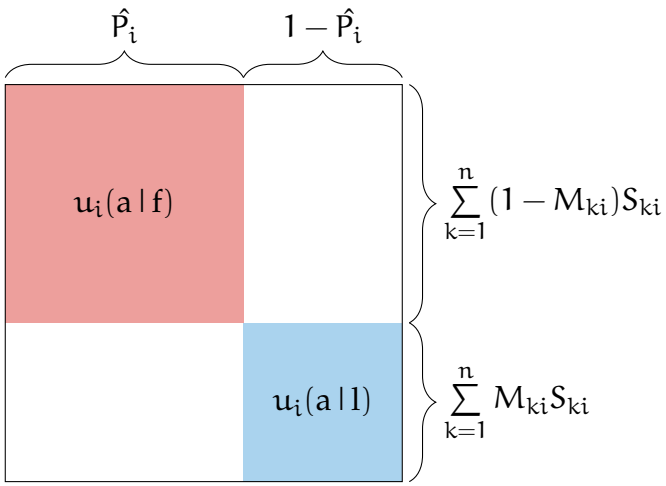

Figure 17: Visualization of the influence of risk and margin on the profit.

$$\begin{aligned} & \hat{E}[U_i|\text{accept}] > \hat{E}[U_i|\text{reject}] = 0 \\ \Leftrightarrow \quad & (1 - \hat{P}_i) M_i S_i - \hat{P}_i (1 - M_i) S_i > 0 \\ \Leftrightarrow \quad & M_i S_i - \hat{P}_i S_i > 0 \\ \Leftrightarrow \quad & M_i > \hat{P}_i \end{aligned} \tag{14}$$

However, it could be overhasty to accept a transaction only if a positive expected value is observed, although this strategy seems to be optimal from a decision-theoretic point of view. The consequences of misclassification are not covered by the formulas because it is assumed that manual inspection always precedes a rejection, and it is further assumed that manual inspection does not lead to rejection of legitimate cases (i.e. that no false positives exist). Automated processing, though, could severely damage customer satisfaction and hurt companies' reputation if honest customers were denied service.

8.3 ANALYSIS

8.3.1 *Introduction*

The aim of the analysis is to compare the novel, utility-based approach to fraud prevention with a traditional, risk-only approach. However, the utility-based approach should not simply be evaluated with a traditional performance metric: If it is argued that profit-oriented companies should adopt a utility-based perspective on fraud prevention, and if it is concluded that, consequently, transactions should be ranked by *expected* utility, then the quality of such a ranking should be measured with a performance metric based on *actual* utility. In other words: For ranking transactions, the traditional, risk-only approach and a utility-based approach are used. Then, the two techniques are evaluated with both traditional measures and with the financial impact of the fraud cases that have been prevented through inspection based on the ranked lists. It is hypothesized that, with regard to the utility-based evaluation, the utility-based ranking will provide better results than the traditional, risk-only ranking. Hence, in order to conduct an unbiased analysis, both approaches are evaluated with both evaluation techniques, leading to four combinations of ranking technique and evaluation measure.

The analysis section is structured as follows: First, the process is described in detail, and second, results are presented and discussed. k-fold cross-validation is applied as in Chapter 7, but in order to focus on the value contributed by the utility approach, cross-validation will not be mentioned any further. The full procedure is shown in Figure 18 and consists of two main steps, the score computation step and the evaluation step. The estimated fraud probabilities $\hat{P}_i$ form the starting point for

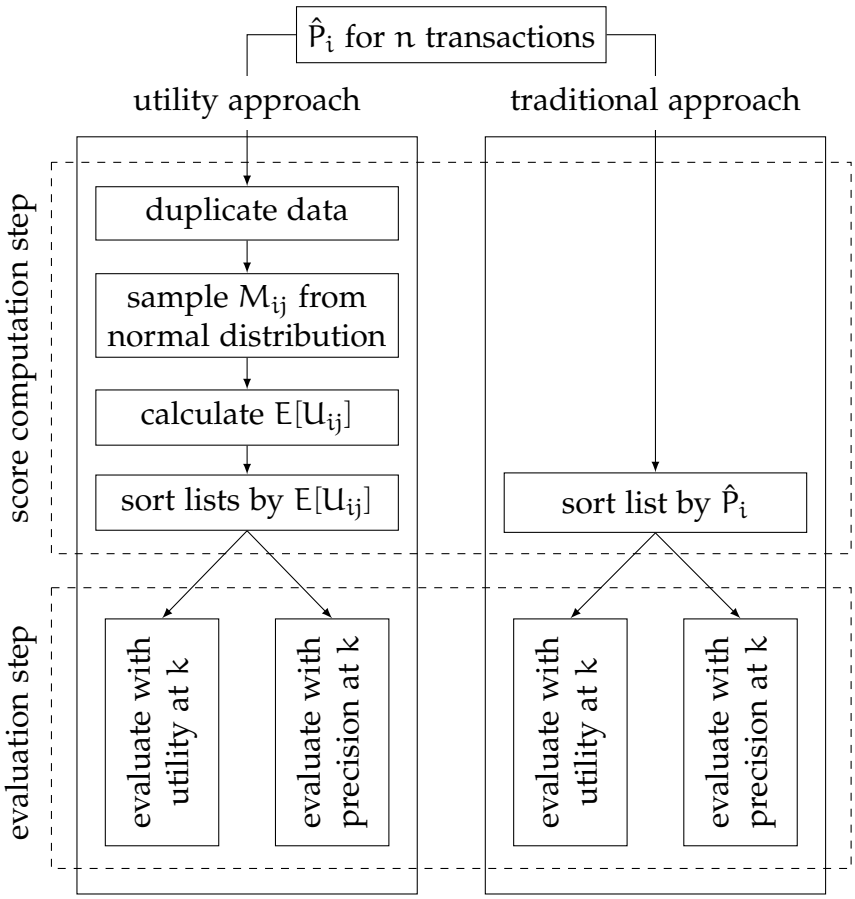

Figure 18: Utility analysis flowchart.

both approaches. The full analysis has been conducted with estimated fraud probabilities from both logistic regression and gradient boosted trees, but results were quite similar, and since the latter method performs slightly better than the former — as shown in Chapter 7 —, results based on gradient boosted trees are presented in this chapter.

For the traditional approach, the ranked list is just sorted by $\hat{P}_i$ in decreasing order, i.e. the most suspicious transactions come first. The traditional approach corresponds with the procedure presented in Chapter 7 and will be picked up again when discussing the evaluation technique.

For the novel ranking approach relying on expected utility, sales prices and product margins are needed besides the fraud

probability. The former are available as part of the dataset, but the latter have to be estimated. Although product margins are not available, category margins are estimated with the help of census data (U.S. Census Bureau, 2014). Four categories with the following margins are established: 45% for clothing, 27% for electronics, 48% for luxury goods, and 36% for other product types. Of course, margins may differ across companies. In order to account for this, the margins are sampled from a normal distribution that introduces an element of variation. Therefore, the score computation step is designed as follows: First, the set of transactions is duplicated q times. Second, for each group of duplicated transactions, different margins are sampled from a normal distribution $\mathcal{N}(\mu = \text{estimated margin}, \sigma = 0.1)$. Setting $\sigma = 0.1$ means a standard variation of 10 percentage points. For example, with an estimate of 45%, one standard deviation ranges from 35% to 55%. Note that the choice of σ is arbitrary and depends on the desired maximum variation for a given confidence level. Third, expected utility values are calculated. Fourth, the enlarged dataset is partitioned into q groups of size n/q such that each group contains all transactions once and no duplicates. At the end, in contrast to the simpler traditional approach, not only one but q ranked lists are retrieved. This allows to inspect sensitivity of the procedure by calculating a confidence interval (95%), showing the robustness of the model. The confidence interval is computed by retrieving the lower bound (2.5^{th} percentile) and the upper bound (97.5^{th} percentile).

In Figure 18, the two steps corresponding to this procedure, duplication of data and sampling margin values from a normal distribution, are part of the score computation step because they have to be conducted before expected utility values can be calculated. However, the procedure is only performed in order

to increase the explanatory power of the evaluation and not a necessary component of the expected utility approach.

8.3.2 *Evaluation of Utility*

In the previous chapter, precision at k was introduced among other evaluation measures. Precision at k ignores the economic perspective, though. However, it can be enhanced to reflect utility with minor changes. In order to achieve this, actual utility of a transaction (in contrast to expected utility) needs to be defined in the first place. Two cases are possible: Either a transaction is detected as fraud, which means that it has correctly been placed at the top of the ranked list and, consequently, the order has been rejected, or it is a false positive and nothing happens. In the former case, utility u_i is defined as the vendor's buying price of the products ordered, which equals $u_i(a \mid \text{fraud})$. The value of the latter case is just zero.

$$u_i = \begin{cases} u_i(a|f) & \text{if transaction is a fraud case} \\ 0 & \text{if it is legitimate} \end{cases} \tag{15}$$

Utility at k is then defined as the sum of the utility values u_i until the k^{th} position, which allows to interpret utility at k as the average value saved by the fraud prevention activity, given the inspection capacity k.

$$\text{utility at } k = \frac{1}{k} \sum_{i}^{k} u_i \tag{16}$$

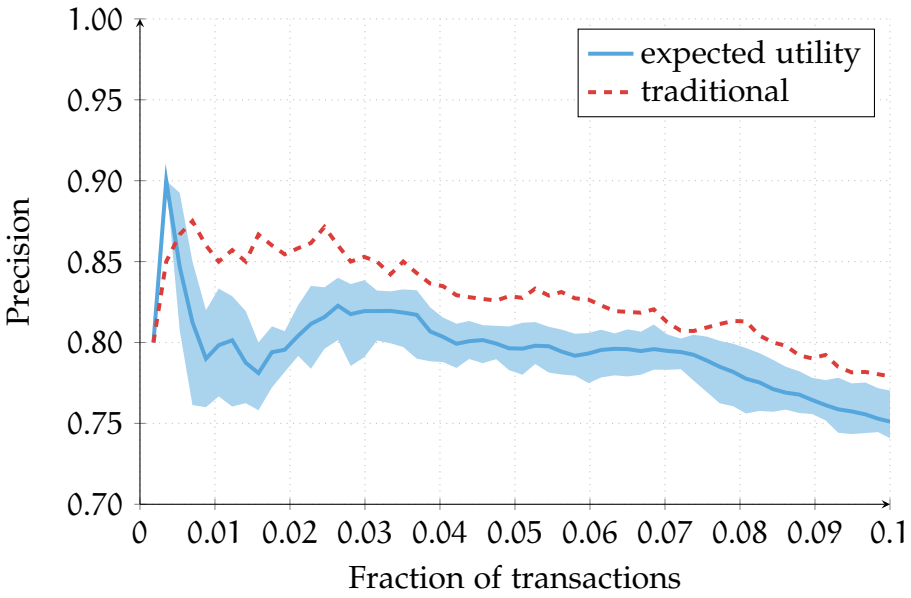

Figure 19: Precision at k^{th} percentile.

8.3.3 *Interpretation of Results*

The two evaluation measures precision at k and utility at k have been introduced, and therefore the empirical analysis is going to be evaluated in the upcoming paragraphs. First, the two approaches, the novel and the traditional ranking, are evaluated using traditional precision at k. This is shown in Figure 19. For example, when inspecting the 2% highest-scored transactions, the traditional ranking approach achieves a precision value of approximately 85%: If there were 5,000 transactions, 2% reflect 100 cases, and 85 of such cases would be fraudulent. A couple of aspects should be emphasized: There seem to be only slight differences between the traditional and the expected utility approach, and the former seems to even outperform the latter. However, this is not surprising as indicated at the beginning of the evaluation section: Measuring expected utility with regular precision is unfair, because expected utility picks up the elements margins and sales prices that the precision measure does not consider. Therefore, the utility-based ranking should

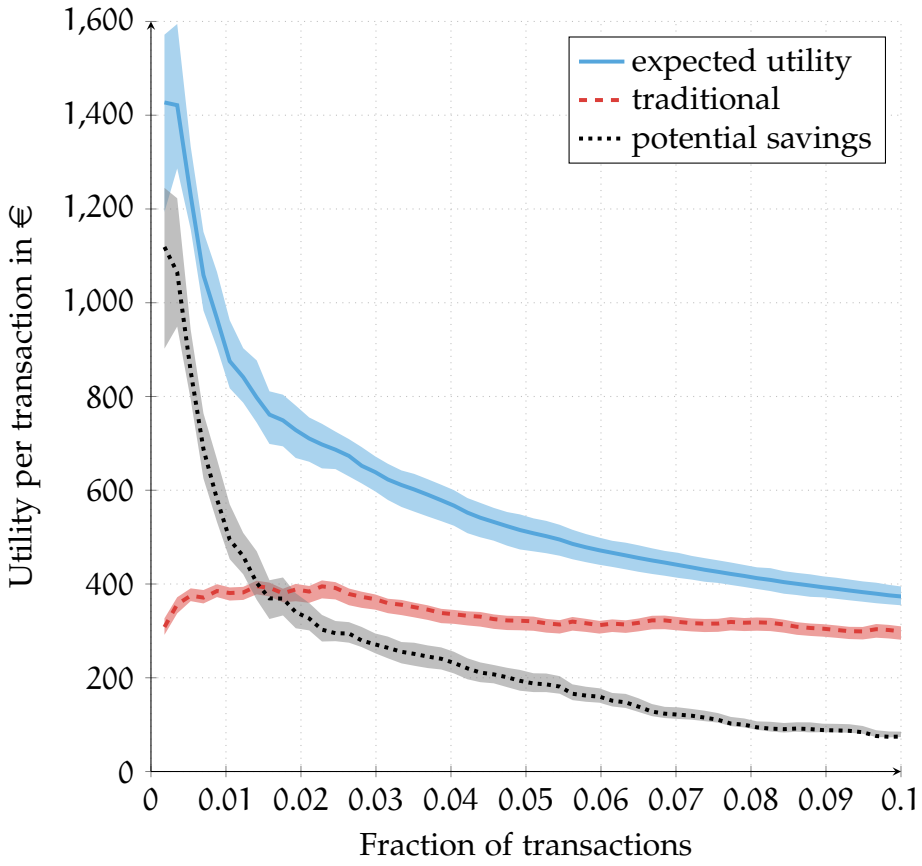

Figure 20: Utility at k^{th} percentile.

be inferior to the traditional approach when relying on a traditional evaluation measure. It is interesting, though, that results are worse only around approximately 5-10 percentage points. Note that, eventually, the two curves converge (not shown in the figure), because the more data are considered, the less the order of the ranked list matters. Thus, both ranking algorithms yield the same precision at 100%, which is simply the average fraud ratio.

Figure 20 shows utility at k per transaction, again for both approaches. In this case, expected utility yields much better results. Exemplary values are presented in Table 9. The traditional approach remains fairly stable at €380. In contrast, the expected utility approach starts at extraordinary high levels (approximately €1,400) and decreases, eventually converging to-

k^{th} percentile	Traditional	Utility-based	Savings
2%	380 €	710 €	330 €
5%	316 €	502 €	186 €
10%	299 €	373 €	74 €

Table 9: Average financial impact per transaction at different percentiles.

wards the traditional approach. The results show that, under the assumption of limited inspection resources, significant savings could be achieved when a utility-based approach to ranking transactions is implemented. For example, the graph states that, if 5,000 orders were placed per day, and resources sufficed to inspect the highest-ranked 2% (i.e. 100 cases), € 38,000 could be saved via the conventional ranking, but € 71,000 could be saved using the expected utility ranking. Thus, potential additional savings amount to € 33,000. This is also indicated by the dotted black line which shows the difference between the values of the two approaches. Of course, the values are based on the underlying data. With regard to other datasets and branches of trade, margins and sales prices could be different. In relative numbers, at the beginning, the expected utility approach yields a benefit of 300% compared to the traditional approach. At the 10^{th} percentile, it is still 25%.

8.4 LIMITATIONS

The results presented in this chapter seem promising. Therefore, it is important not only to interpret the results but also to think about possible criticisms or drawbacks of the approach. These are discussed in this section.

First, in the dataset, the transactions with the highest risk were not those with the highest utility, which is why results

of the two ranking techniques differ. This incongruity does not have to hold for every dataset. Possibly, in other datasets, correlation between risk and value might be higher, which would decrease the superiority of the expected value approach over the traditional technique.

Second, using cross-validation, a random subset of the dataset was used to perform the analysis. However, the fraud prevention departments inspect transactions on a daily basis. If the temporal distribution of high-value and high-risk cases differs, the surplus value of the expected value approach may vary in both directions, meaning it could increase or decrease.

Third, margins could differ. Possible errors regarding the estimation of margins have been accounted for, and sensitivity analysis shows little volatility. Nevertheless, one should keep in mind that the margin would have to be adjusted if the analysis were repeated with a dataset from another branch of trade.

Fourth, Elovici and Braha (2003) discuss a mathematical notation that represents utility-based decisions under uncertainty using matrix algebra, which can be projected onto the fraud prevention problem. In general, matrix algebra could help bridge the gap between the conceptual idea, mathematical notation, and its software implementation. In this thesis, the code written for the analysis has not been presented because it relies on functional programming and cannot be interpreted easily.

Fifth, other costs could be associated with rejecting legitimate transactions: When honest customers are denied service, customers could become dissatisfied, and therefore their corresponding customer lifetime value could decrease due to a loss of trust. This represents another kind of costs which is connected to classification errors.

Sixth, fraud is overrepresented in the dataset. A base rate of 18.9% is rather unrealistic. Hence, since the estimated fraud

probabilities rely on the fraud rates in the data, expected value may be lower than the results presented in the analysis, and thus, the difference between the utility approach and the traditional approach could decrease.

Seventh, some fraud prevention departments might face a conflict of interest with regard to implementing the utility-based approach. As mentioned in the interviews, sometimes the sum of the sales prices of successfully prevented fraud cases is used as a performance measure. On the one hand, this increases the perceived impact of the fraud prevention work because sales prices usually exceed utility values, but on the other hand, the utility-based approach better represents the department's contribution to profit. Therefore, such a conflict of interest would have to be resolved before switching to another performance measure like the utility-based approach.

Eighth, return ratios have not been considered in the analysis. Returns reduce the profitability of legitimate sales, but they do not reduce the damage from fraud because criminals usually do not return their loot. Therefore, it may be assumed that a high return ratio amplifies the importance of the utility approach presented in this chapter.

Ninth, decision makers are sometimes risk-averse (Kahneman, 2012). In this case, maximization of utility is no longer the primary goal. Given that each transaction resembles a play of a repeatable game, companies are likely to prefer utility maximization, but it is important to note that, due to risk aversion, even a rational decision maker could choose a different behavior compared to what has been proposed in this chapter.

Finally, limitations that relate to the estimation of the fraud probabilities as described in Chapter 7 apply to the analysis in this chapter as well, because the estimated fraud probabilities are used for the calculation of expected value.

Although such aspects should be considered in future work, the algorithm produces a ranked list that is manually examined by fraud officers. Hence, since the algorithm aims at supporting the manual fraud detection process, the negative impact of the aforementioned limitations is smaller than if the system were allowed to block transactions considered fraudulent without manual supervision.

8.5 CONCLUSION

In this chapter, an economic perspective on the previous work has been adopted. The estimated fraud probabilities — which had been retrieved using gradient boosted trees —, margins, and sales prices were used to compute results ranked by expected utility. The traditional and the novel approach were evaluated both with conventional precision at k and the proposed measure utility at k. Results show that, under the assumption of limited inspection resources, a substantial amount of money could be saved if expected utility were used for ranking of transactions instead of conventional fraud probability estimates.

9

CONCLUSIONS

9.1 INTRODUCTION

Buyer fraud has become a prevalent problem in the e-commerce mail order business: The financial importance of fraud is described in Chapter 1 by referring to statistics about the e-commerce market, reported crimes, and fraud rates. The importance of the problem is underlined through the interviews (Chapter 4) and the number of documented fraud cases in the transaction dataset (Chapter 5).

In order to protect themselves against it, some companies run fraud prevention departments where suspicious transactions are inspected. However, almost no literature exists concerning the questions how manual fraud prevention is performed, which features can be used as effective fraud risk indicators, how automated data mining can be realized in order to support the fraud prevention process, and how economic principles can be woven into decision making. The publications that focus on the e-commerce mail order business touch individual aspects of these questions at best. Therefore, this thesis addresses the central research question how automated data mining can support profit-oriented B2C e-commerce mail order companies in dealing with fraud.

In order to answer it, the thesis contains five parts that build on top of one another. Chapters not mentioned in this overview provide a theoretical background or serve to prepare the data basis for the research. First, the current state of research in

e-commerce mail order fraud detection and prevention is explored in the theoretical foundation presented in Chapter 2. A conceptual data mining framework is introduced in order to provide an overview of methods used for fraud detection and in order to structure existing literature. Second, a small qualitative study presented in Chapter 4 investigates the fraud prevention process by conducting interviews with experts from e-commerce companies. Third, Chapter 6 shows how information gain and a decision tree were used to estimate the importance of features both individually and combined. Fourth, Chapter 7 documents how a predictive fraud detection model was developed in order to help fraud prevention departments focus on the most suspicious cases by ranking transactions according to estimated risk. Finally, a utility-based model is compared with such traditional, risk-based orderings, and it is shown that substantial savings could be achieved through introduction of utility presented in Chapter 8. While conclusions have been provided in each of these parts, this chapter focuses on highlighting the contribution of the thesis to theoretical and practical aspects of fraud prevention in the e-commerce mail order business, critically reflecting upon the whole picture, and providing an outlook for future work.

The chapter is structured accordingly. It starts with a description of the thesis' contribution to theoretical knowledge followed by a discussion of its contribution to professional practice. Then, topics of potential future work are discussed.

9.2 CONTRIBUTION TO KNOWLEDGE

This thesis contributes to knowledge through the following key aspects, which will be explained in detail below:

- The review of literature provides an overview of the current state of knowledge regarding fraud prevention in the e-commerce mail order business and may serve as a starting point for researchers interested in this area.
- The interviews with fraud prevention experts from two large mail order companies offer insights about the fraud prevention process, which has rarely been covered in existing literature.
- The analysis of fraud risk indicators challenges and extends existing knowledge about feature relevance in the e-commerce mail order business.
- The fraud detection model strengthens the evidence that that logistic regression and gradient boosted trees are appropriate techniques for predicting e-commerce mail order fraud.
- The development of a utility-based approach adds a novel perspective to fraud prevention in the e-commerce mail order business.

The review of literature as the first key aspect showed that Hinneburg (2006) first investigated fraud prevention in the e-commerce mail order business in Germany from a qualitative point of view and discussed when and how such fraud may occur. Since then, research in related areas of fraud detection such as insurance fraud, medical fraud, or credit card fraud

in general has advanced through publication of various studies, but the e-commerce mail order business has mostly been left out. In order to provide a contextual frame and to offer a starting point for researchers interested in fraud prevention in the e-commerce mail order business, this thesis provides an overview of relevant publications.

Mentioned as the second key aspect, the interviews shed light on the question what e-commerce mail order companies regard as fraud and how they approach to detect and prevent it. It has been demonstrated that fraud prevention can be performed following and possibly combining various strategies; data mining is only one of them. Some strategies focus on true prevention, i.e. they prevent fraud from occurring at all — instead of detecting and then canceling transactions that are found to be fraudulent. For example, if only safe payment methods such as prepayment can be chosen, fraud becomes much more difficult to commit if not even impossible. Other strategies include consistency checks such as address verification services or two-factor authentication (CyberSource, 2017). The former strategy requires customers to use real addresses for invoicing and delivery, reducing the amount of fake data. The latter strategy requires customers to connect a second device such as a mobile phone with their accounts, making authentication more secure and account theft more difficult.

However, many of these strategies force *all* customers to pay the price for extra security. Although measures such as prepayment and two-factor authentication may make it more difficult to commit fraud, they require additional effort from all customers. Since only a fraction of transactions are fraud cases, most of the time such methods only increase the necessary effort to place a legitimate order. Thus, companies face a trade-off between smooth customer service and fraud prevention. As

discussed in the interview with company B, additional services that help reduce fraud are generally of interest given that they are transparent, convincing, and that they outweigh associated costs (Anon, 2015b).

Due to the problems associated with the strategies stated above, in the e-commerce mail order business, it is common practice to run fraud prevention departments (Anon, 2015a,b; Carneiro, Figueira, and Costa, 2017; Hinneburg, 2006). Over the years, computer-aided systems have been integrated into the fraud prevention process more strongly. While Hinneburg (2006) merely mentions their existence, today companies either already use them or are interested in implementing such systems (Anon, 2015a,b; Carneiro, Figueira, and Costa, 2017). Nevertheless, manual inspection continues to play an essential role because companies hesitate to let computers alone decide on the fate of e-commerce orders (Anon, 2015a).

Therefore, the fraud prevention process consists of two steps: the selection of transactions to be inspected and the actual inspection of such transactions itself. In the interviewed companies, the fraud prevention process is more complex than stated in some of the literature: Torgo and Lopes (2011) only refer to manual inspection without modeling the process, and Carneiro, Figueira, and Costa (2017) mention that the customer can be called in order to find out the fraud status of a transaction. However, the interviews showed that — at least in the companies A and B — fraud prevention employees can choose from a variety of investigation strategies and multiple actions, including but not limited to simply accepting or rejecting the transaction.

Regarding the third key aspect, the performance of available features was analyzed through computing information gain and building a decision tree. Results show that certain features

and also feature combinations stand out with regard to their ability to separate fraudulent from legitimate orders. Hinneburg (2006) states that fraud is more frequent among particular socio-demographic groups, e.g. people with low income, unemployed people, and asylum seekers. However, relying on such patterns may raise ethical issues. In the interviews, rather pragmatic features were emphasized such as whether a parcel shop is used as the shipment address and whether the shipment address differs from the invoice address. The companies A and B both confirmed Hinneburg's statement that expensive, small goods which can easily be moved and resold are more likely to be subjected to fraud than other article groups (Anon, 2015a,b; Hinneburg, 2006). For feature analysis in this thesis, technical features (e.g. operating system), transaction details (e.g. number of articles), and customer data (e.g. customer age) have been used and yielded promising results. Through information gain analysis, the five most useful features for fraud detection have been identified as *total price*, *account age*, *address distance*, *browser*, and *parcel shop*. By constructing a decision tree, it is possible to analyze combinations of feature values, which by far achieved the best distinction between fraudulent and legitimate transactions. The features considered important for fraud detection in the e-commerce mail order business overlap with those highlighted by Carneiro, Figueira, and Costa (2017). However, the authors focus on credit card fraud, which is why the feature sets differ from each other. On the one hand, this makes comparison of research results more difficult; on the other hand, more diverse insights about feature importance can be gained when multiple publications focus on different sets of features each, offering a broader perspective to the reader who considers all of such publications.

Various publications researched how data mining can support fraud detection in related areas, but during the last years, only a single study has been published that deals with the e-commerce mail order business: Carneiro, Figueira, and Costa (2017) used credit card data for training logistic regression and random forests to detect fraud. However, the question remained open whether transaction-centered features could be used for prediction as well, which is particularly important when customers pay via invoice because in that case credit card data is not available.

Referring to the fourth key aspect, logistic regression and gradient boosted trees are used in this thesis because they have proven their predictive power in many other applications. Results suggest that relying on features usually available to e-commerce mail order companies achieves promising performance, similar to that achieved by Carneiro, Figueira, and Costa (2017) with credit card data.

With regard to the fifth key aspect, this thesis conceptually enhances the value of machine learning for fraud detection in the e-commerce mail order business by adopting an economic perspective on the problem through the application of utility theory. Such an approach had been proposed in general but Torgo and Lopes (2011) but not yet brought to the e-commerce mail order business.

The next section focuses the thesis' contribution to professional practice. It will be explained how the empirical parts of the thesis may influence practitioners to reflect upon and improve their fraud prevention systems.

9.3 CONTRIBUTION TO PROFESSIONAL PRACTICE

Besides the insights discussed regarding the contribution to knowledge, which is recommended to be considered by practitioners, this thesis takes a holistic perspective on fraud prevention in the e-commerce mail order business that may guide practitioners towards a better understanding of fraud prevention processes and towards a more business-driven, economic perspective on data mining techniques. The central aspects of contribution to professional practice are as follows:

- Investigation of the fraud prevention process at the interviewed companies may help consider strengths and weaknesses of manual inspection, and it may help better understand how the quality of fraud labels can be estimated.
- Feature analysis shows how the relevance of fraud risk indicators can be determined and offers evaluation criteria in order to decide which of them companies might want to include into their evaluation.
- The fraud prediction model shows how established machine learning models can be applied to e-commerce mail order fraud detection and how the work of fraud prevention departments can be supported by partial automation.
- Utility analysis suggests that substantial savings can be achieved when expected cost and benefit are used for ranking of transactions instead of a purely risk-based approach.

The interviews as the first key aspect revealed that the fraud prevention process can be more complex than it is often modeled: While most quantitative fraud detection studies distin-

guish between fraudulent and legitimate, and consequently recommend either rejecting or accepting a transaction, fraud prevention strategies are much more diverse. Future data mining models could become much more powerful by being able to consider the whole spectrum of prevention strategies. For example, customers for which a medium fraud risk is estimated could automatically be required to use two-factor authentication or other techniques to identify themselves as the proper owners of their accounts instead of being denied service directly. With such differentiation, appropriate anti-fraud measures could be assigned to the corresponding levels of risk dynamically and on a per-case basis, which would allow fraud prevention departments to spend more time on investigating the most suspicious cases. The interviews helped consider the whole picture instead of only focusing on optimization of fraud detection rates. To sum up, practitioners should not simply implement a standalone fraud detection algorithm; instead, they should evaluate existing as well as potential future anti-fraud strategies in order to benefit most from the introduction of an automated fraud detection system.

A major challenge is obtaining high-quality reference data, because it is difficult to identify fraud cases clearly: The definition of the term *fraud* is not standardized across publications. For example, Carneiro, Figueira, and Costa (2017) regard credit card chargebacks as fraud, i.e. the occurrence of financial damage is sufficient for the fraud verdict. In contrast, Hinneburg (2006) states that a criminal intention is required for an activity to be labeled fraudulent. The interviewed companies hesitated to regard payment default as sufficient evidence for fraud because there can be various reasons behind it. Most importantly, they want to avoid flagging customers with low creditworthiness as criminals by mistake (Anon, 2015a,b). Hence, in order to

find evidence for criminal intention, such as fake data or stolen accounts, the fraudulent transactions considered in this thesis have been subjected to a detailed investigation. The investigation of the fraud prevention process may help practitioners set up guidelines that lead to high-quality reference data, which are needed in every subsequent analysis.

Mentioned as the second key aspect, feature analysis not only shows which features might be of interest in the e-commerce mail order business, but it also supplies practitioners with techniques about how to determine relevant features in their own datasets. The latter aspect is even more important, because fraud patterns most likely differ across companies of different size, structure, and branch of trade. Knowing which features best help separate fraudulent from legitimate orders allows companies to compare costs and potential benefits of adding certain features to their dataset and maintaining them — or removing them from it.

For smaller e-commerce companies that have to deal with fraud but do not want or are not able to implement a fraud prevention system, knowledge about patterns indicating fraud might suffice in order to reduce financial damage. Larger businesses, however, should rely on computer-aided systems in order to deal with the large volume of fraud, partially automating the prevention process. As has been shown for two of the largest e-commerce companies in Europe, the amount of transactions to be inspected clearly exceeds available inspection capacity even though fraud prevention is performed in departments solely dedicated to that purpose.

Referring to the third key aspect, results of the machine learning part of the thesis suggest that established methods, such as logistic regression and gradient boosted trees, can be used for prediction of fraud cases. It is concluded that the choice of a

machine learning algorithm is not the most essential decision. Different algorithms have shown similar performance across different publications and datasets, of course given that the algorithm can deal with the data at hand. Therefore, practitioners are encouraged to invest more time into obtaining an essential understanding of processes and gathering the most helpful features instead of horse-racing machine learning algorithms.

In this thesis, it is argued that choosing precision at k for evaluation of model performance is most appropriate — considering only typical metrics for classifier evaluation. From an economic point of view, however, such measures are regarded problematic: For a company it might be sensible to accept a certain number of fraud cases if the stolen products are relatively cheap and instead concentrate on a few cases with a high financial impact.

Therefore, representing the fourth key aspect, a more business-driven approach is introduced by taking an economic perspective. The idea is reflected through the introduction of expected utility at k for ranking transactions and actual utility at k for evaluation of results. It is shown that substantial financial savings are possible if a utility-based perspective on fraud prevention is taken. Practitioners may find the barrier to include the utility-based approach to be relatively low because such a computation can be added to any existing fraud prevention systems that produces estimated fraud risk values.

To sum up, this thesis provides a framework that covers an investigation of the fraud prevention process, multiple methods of feature analysis, development of a predictive model, and the introduction of an economic perspective on fraud prevention. Practitioners may use any subset of the presented techniques to strengthen the fraud prevention work within their companies.

9.4 FURTHER WORK

This thesis contains qualitative and quantitative work, and while it answers questions, it also raises new ones. This section contains a discussion about which topics could benefit from further research.

The qualitative part consisted of interviews with two large B2C e-commerce mail order companies and served as a preparation for the quantitative part. The interviews show how insightful qualitative analysis can be with regard to a topic that is mostly approached with quantitative methods, and they show how important it is to understand the fraud prevention process when interpreting results of data mining. Therefore, two recommendations are made: First, more qualitative research should be conducted in order to increase generalizability of insights about fraud prevention in the e-commerce mail order business. For example, it could be explored to what extent fraud occurs in smaller companies, companies with a focus on a specific range of products (e.g. luxury goods or food), or companies in other countries. Second, when conducting quantitative research, the importance of preparatory qualitative inquiry should not be underestimated.

Regarding the quantitative analyses, general recommendations for further work are presented first. Then, the individual analyses are addressed. The original transaction dataset contains hundreds of thousands of transactions, but only a few percent contain the essential fraud labels. Transactions were mostly chosen for inspection by experience and partially by random chance. Therefore, further research could benefit from a transaction dataset constructed in a truly random way, which would probably be reflected by a lower fraud rate closer to what has been mentioned in the literature.

For feature analysis, information gain and a decision tree were used. Information gain provides a good overview of relevance on a per-feature basis. The decision tree suggests that feature combinations may be more powerful predictors, which was also confirmed in the interviews. In the future, researchers could further investigate alternative ways to evaluate such combinations.

The predictive model relies on supervised machine learning methods. However, unsupervised or semi-supervised machine methods could add value to fraud prevention: Unsupervised methods, such as outlier detection, may detect anomalous patterns. This would alleviate the dependency on labeled data. Combining unsupervised and supervised techniques could result in utilizing the best of both worlds and is therefore a promising topic of future work.

The utility analysis introduces an economic perspective on fraud prevention in the e-commerce mail order business and achieves promising results, which should be challenged in future research. For example, product category margins were extracted from the literature and varied in order to account for possible estimation errors, but product-based margins might yield more precise results. In addition, other influences could be considered by the utility function, such as the return rate, which decreases the profitability of legitimate transactions. Another example is that honest customers might become disappointed and buy elsewhere if they are mistakenly regarded as fraudsters, which could also be considered in a utility model.

To sum up, potential areas for future research refer to either better dealing with some of the limitations of this thesis or extending promising concepts. Fraud prevention in the e-commerce mail order business still offers many opportunities

for insightful research that has direct implications for a financially relevant real-world problem.

9.5 CONCLUSION

Multiple authors have underlined that a remarkable research gap exists concerning fraud prevention in the e-commerce mail order business and that the area deserves more attention. This thesis contributes to closing the gap in several ways, which will be described in the following.

In the literature review, the current state of research is explored, providing an overview of the fraud prevention problem and existing approaches to solving it. Researchers interested in the topic might consider the literature review a useful starting point.

Results of the interviews of fraud prevention practitioners help understand the fraud prevention process, not only laying the foundation for the subsequent analyses but also discussing the structural problems faced in fraud prevention. In addition, the limitations of the dataset have been explored through the interviews as well. Researchers investigating fraud prevention in the e-commerce mail order business are encouraged to thoroughly understand the characteristics of their specific problem setting prior to any quantitative analysis.

Information gain analysis and a decision tree demonstrate that a careful selection of features is sensible because some are much more useful than others. This may help companies make better decisions when designing their fraud prevention systems.

A prediction model has been developed that is able to identify fraud cases with satisfying performance. It strengthens ex-

isting research, which advocates the use of machine learning for fraud detection in the e-commerce mail order business. It is also shown that, even though the application of the latest machine learning techniques in business applications is often advertised, the performance across methods is similar. This leads to the conclusion that the choice of a specific learning algorithm is not the greatest leverage for improving fraud prevention. In contrast, choosing the right features and evaluating them with the most appropriate metrics seems far more essential.

Utility theory is applied to the problem in order to reflect that companies aim for profit maximization rather than for pure fraud minimization. The evaluation of this approach shows that substantial savings could be achieved if fraud prevention activities were guided by economic principles.

To sum up, companies facing fraud in the e-commerce mail order business should focus on the features that were found to be relevant in this thesis, and they should conduct their own feature analysis in order to evaluate the relevance of features specific to their dataset. Results should be used both for manual inspection and development of an automated scoring system. Scoring should be achieved through expected economic utility rather than pure estimated risk.

Besides the questions this thesis answers, it raises many new as well. As the experience of criminals grows, so must fraud detection systems. The ongoing advance of strategies to commit fraud necessitates continuous development of such systems. Only when fraud detection systems are able to outplay even the most innovative criminals fraud can successfully be prevented. In order to achieve this, it is recommended to conduct further research in three main areas arising from: First, the variation of the contextual setting, investigating the differences of fraud prevention across various branches of trade, companies of different

size, and other countries. Second, the transition from computer-aided systems with manual inspection to fully automated fraud prevention. Third, further developing the utility-based perspective on fraud prevention by considering more data, such as cost of order cancellation at different stages of delivery. This thesis shall support other scientists to find a starting point for their individual research subjects on fraud prevention, and it shall provide practitioners access to a scientific approach in order to deal with fraud.

REFERENCES

Aggarwal, C. C. (2013). *Outlier Analysis*. New York, New York, USA: Springer.

Ahrholdt, D. C. (2011). "Empirical Identification of Success-Enhancing Web Site Signals in E-Tailing: An Analysis Based on Known E-Tailers and the Theory of Reasoned Action." In: *Journal of Marketing Theory and Practice* 19.4, pp. 441–458.

Anon (2015a). *Expert Interview with Company A*. German.

— (2015b). *Expert Interview with Company B*. German.

— (2015c). *Expert Interview with Risk Ident*. German.

Antonakis, A. C. and Sfakianakis, M. E. (2009). "Assessing Naive Bayes as a Method for Screening Credit Applicants." In: *Journal of Applied Statistics* 36.5, pp. 537–545.

Aral, K. D. et al. (2012). "A Prescription Fraud Detection Model." In: *Computer Methods and Programs in Biomedicine* 106.1, pp. 37–46.

Artís, M., Ayuso, M., and Guillén, M. (1999). "Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market." In: *Insurance: Mathematics and Economics* 24.1–2, pp. 67–81.

— (2002). "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims." In: *Journal of Risk and Insurance* 69.3, pp. 325–340.

Backhaus, K. et al. (2011). *Multivariate Analysemethoden*. German. 13th edition. Berlin, Germany: Springer.

Bahnsen, A. C. et al. (2016). "Feature Engineering Strategies for Credit Card Fraud Detection." In: *Expert Systems with Applications* 51.Supplement C, pp. 134 –142.

Bayardo Jr., R. J. and Agrawal, R. (1999). "Mining the Most Interesting Rules." In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, California, USA, pp. 145–154.

Bell, T. B. and Carcello, J. V. (2000). "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting." In: *Auditing* 19.1, pp. 169–184.

Benton, T. and Craib, I. (2011). *Philosophy of Social Science: The Philosophical Foundations of Social Thought (Traditions in Social Theory)*. 2nd edition. Basingstoke, Hampshire, England: Palgrave Macmillan.

Bermúdez, L. et al. (2008). "A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance." In: *Insurance: Mathematics and Economics* 42.2, pp. 779–786.

Bhattacharyya, S. et al. (2011). "Data Mining for Credit Card Fraud: A Comparative Study." In: *Decision Support Systems* 50.3, pp. 602–613.

Blumberg, B. F., Cooper, D. R., and Schindler, P. S. (2014). *Business Research Methods*. 4th edition. Berkshire, England: McGraw-Hill Education.

Boda, K. et al. (2012). "User Tracking on the Web via Cross-Browser Fingerprinting." In: *Information Security Technology for Applications: 16th Nordic Conference on Secure IT Systems, NordSec 2011, Tallinn, Estonia, October 26-28, 2011, Revised Selected Papers*. Ed. by P. Laud. Berlin, Germany: Springer, pp. 31–46.

Bose, I. and Mahapatra, R. K. (2001). "Business Data Mining — A Machine Learning Perspective." In: *Information & Management* 39.3, pp. 211 –225.

Brabazon, A. et al. (2010). "Identifying Online Credit Card Fraud Using Artificial Immune Systems." In: *IEEE Congress on Evolutionary Computation*, pp. 1–7.

Brendel, M. (2017). "Betrug auf Rechnung." German. In: *Der Spiegel* 41, p. 73.

Brockett, P. L., Xia, X., and Derrig, R. A. (1998). "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud." In: *The Journal of Risk and Insurance* 65.2, pp. 245–274.

Brockett, P. L. et al. (2002). "Fraud Classification Using Principal Component Analysis of RIDITs." In: *Journal of Risk and Insurance* 69.3, pp. 341–371.

Bruha, I. (2010). In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. New York, New York, USA: Springer. Chap. Missing Attribute Values.

Bundeskriminalamt (2016). *Lagebild Cybercrime 2016*. German. URL: `https://goo.gl/773Dwv` (visited on 10/23/2017).

Burge, P. and Shawe-Taylor, J. (2001). "An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection." In: *Journal of Parallel and Distributed Computing* 61.7, pp. 915–925.

Carneiro, N., Figueira, G., and Costa, M. (2017). "A Data Mining Based System for Credit-Card Fraud Detection in E-Tail." In: *Decision Support Systems* 95.Supplement C, pp. 91–101.

Caruana, R. and Niculescu-Mizil, A. (2005). "An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics." In: *Proceedings of the 23rd International Conference on Machine learning (ICML'06)*, pp. 161–168.

Caudill, S. B., Ayuso, M., and Guillén, M. (2005). "Fraud Detection Using a Multinomial Logit Model with Missing Information." In: *Journal of Risk and Insurance* 72.4, pp. 539–550.

Chang, J.-S. and Chang, W.-H. (2014). "Analysis of Fraudulent Behavior Strategies in Online Auctions for Detecting La-

tent Fraudsters." In: *Electronic Commerce Research and Applications* 13.2, pp. 79 –97.

Chang, W.-H. and Chang, J.-S. (2012). "An Effective Early Fraud Detection Method for Online Auctions." In: *Electronic Commerce Research and Applications* 11.4, pp. 346–360.

Chapelle, O., B. Schölkopf, and A. Zien, eds. (2010). *Semi-Supervised Learning*. Cambridge, England: The MIT Press.

Chawla, N. and Li, X. (2006). "Pricing Based Framework for Benefit Scoring." In: *Second International Workshop on Utility-Based Data Mining*, pp. 65–69.

Chen, T. and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." In: *CoRR* abs/1603.02754.

Chen, W., Ma, C., and Ma, L. (2009). "Mining the Customer Credit Using Hybrid Support Vector Machine Technique." In: *Expert Systems with Applications* 36.4, pp. 7611–7616.

Chung, S.-H. and Suh, Y. (2009). "Estimating the Utility Value of Individual Credit Card Delinquents." In: *Expert Systems with Applications* 36.2, Part 2, pp. 3975–3981.

Coderre, D. G. (2009). *Computer-Aided Fraud Prevention and Detection: A Step-By-Step Guide*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Collier, A. (1994). *An Introduction to Roy Bhaskar's Philosophy*. London, England: Verso.

Collis, J. and Hussey, R. (2014). *Business Research: A Practical Guide for Undergraduate & Postgraduate Students*. 4th edition. Basingstoke, England: Palgrave MacMillan Higher Education.

Conan Doyle, Sir A. (2009). "The Reigate Puzzle." In: *The Penguin Complete Sherlock Holmes*. London, England: Penguin.

Cussens, J. (2010). In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. New York, New York, USA: Springer. Chap. Induction.

CyberSource (2017). *Online Fraud Benchmark Report*. URL: http://www.cybersource.com (visited on 11/30/2017).

DMLC (2016). *XGBoost Python Package*. URL: https://goo.gl/fYTshF (visited on 01/19/2017).

Dal Pozzolo, A. et al. (2014). "Learned Lessons in Credit Card Fraud Detection From a Practitioner Perspective." In: *Expert Systems with Applications* 41.10, pp. 4915 –4928.

Deeks, J. J. and Altman, D. G. (2004). "Diagnostic Tests 4: Likelihood Ratios." In: *British Medical Journal* 329 (7458), pp. 168–269.

Devore, J. L. and Berk, K. N. (2012). *Modern Mathematical Statistics with Applications*. Ed. by G. Casella and O. I. Fienberg S. and. Springer Texts in Statistics. Springer New York.

Dewey, J. (2013). *How We Think*. Original work published 1910. London, England: Forgotten Books.

Dong, F., Shatz, S. M., and Xu, H. (2009). "Combating Online In-Auction Fraud: Clues, Techniques and Challenges." In: *Computer Science Review* 3.4, pp. 245–258.

Duman, E. and Ozcelik, M. H. (2011). "Detecting Credit Card Fraud by Genetic Algorithm and Scatter Search." In: *Expert Systems with Applications* 38.10, pp. 13057–13063.

Durtschi, C., Hillison, W., and Pacini, C. (2004). "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data." In: *Journal of Forensic Accounting* 5, pp. 17–34.

Elovici, Y. and Braha, D. (2003). "A Decision-Theoretic Approach to Data Mining." In: *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans* 33.1, pp. 42–51.

Euromonitor International (2014). *Internet vs Store-Based Shopping: The Global Move Towards Omnichannel Retailing*. URL: http://www.euromonitor.com (visited on 06/23/2015).

Fahrmeir, L. et al. (2007). *Statistik: Der Weg zur Datenanalyse*. German. 6th edition. Berlin, Germany: Springer.

Fan, R.-E. et al. (2008). "LIBLINEAR: A Library for Large Linear Classification." In: *Journal of Machine Learning Research* 9, pp. 1871–1874.

Fanning, K., Cogger, K. O., and Srivastava, R. (1995). "Detection of Management Fraud: A Neural Network Approach." In: *Intelligent Systems in Accounting, Finance and Management* 4.2, pp. 123–126.

Fayyad, U. M. and Irani, K. B. (1993). "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry France, August 28 – September 3, 1993*. Ed. by R. Bajcsy. Morgan Kaufmann, pp. 1022–1029.

Florkowski, C. M. (2008). "Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests." In: *The Clinical Biochemist Reviews* 29 (Supplement 1), pp. 83–87.

Friedman, C. and Sandow, S. (2011). *Utility-Based Learning from Data*. Boca Raton, Florida, USA: CRC Press.

Gassmann, M. (2015). *Moderne Diebe lassen sich Beute frei Haus liefern*. German. URL: `https://goo.gl/VRzq8C` (visited on 11/30/2017).

Guyon, I. and Elisseff, A. (2003). "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3, pp. 1157–1182.

Halvaiee, N. S. and Akbari, M. K. (2014). "A Novel Model for Credit Card Fraud Detection Using Artificial Immune Systems." In: *Applied Soft Computing* 24, pp. 40–49.

Hartmann-Wendels, T., Mählmann, T., and Versen, T. (2009). "Determinants of Banks' Risk Exposure to New Account Fraud – Evidence from Germany." In: *Journal of Banking & Finance* 33.2, pp. 347–357.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning*. Springer.

He, H. et al. (1997). "Application of Neural Networks to Detection of Medical Fraud." In: *Expert Systems with Applications* 13.4, pp. 329–336.

He, X., Cai, D., and Niyogi, P. (2006). "Laplacian Score for Feature Selection." In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, P. B. Schölkopf, and J. C. Platt. MIT Press, pp. 507–514.

Helfferich, C. (2011). *Die Qualität qualitativer Daten*. German. 4th edition. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften, Springer.

Hilas, C. S. (2009). "Designing an Expert System for Fraud Detection in Private Telecommunications Networks." In: *Expert Systems with Applications* 36.9, pp. 11559–11569.

Hilas, C. S. and Mastorocostas, P. A. (2008). "An Application of Supervised and Unsupervised Learning Approaches to Telecommunications Fraud Detection." In: *Knowledge-Based Systems* 21.7, pp. 721–726.

Hinneburg, H. (2006). *Prävention von Kriminalität im E-Commerce*. German. Frankfurt am Main, Germany: Peter Lang.

Hoffmann, A. O. and Birnbrich, C. (2012). "The Impact of Fraud Prevention on Bank-Customer Relationships: An Empirical Investigation in Retail Banking." In: *International Journal of Bank Marketing* 30.5, pp. 390–407.

Hsieh, C.-J. et al. (2008). "A Dual Coordinate Descent Method for Large-Scale Linear SVM." In: *Proceedings of the 25th In-*

ternational Conference on Machine Learning. ICML '08. New York, New York, USA: ACM, pp. 408–415.

Humpherys, S. L. et al. (2011). "Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis." In: *Decision Support Systems* 50.3, pp. 585–594.

Jha, S., Guillén, M., and Westland, J. C. (2012). "Employing Transaction Aggregation Strategy to Detect Credit Card Fraud." In: *Expert Systems with Applications* 39.16, pp. 12650–12657.

Kahneman, D. (2012). *Thinking, Fast and Slow*. London, England: Penguin.

Kohavi, R. and John, G. H. (1997). "Wrappers for Feature Subset Selection." In: *Artificial Intelligence* 97.1-2, pp. 273–324.

Krivko, M. (2010). "A Hybrid Model for Plastic Card Fraud Detection Systems." In: *Expert Systems with Applications* 37.8, pp. 6070–6076.

Lei, J. Z. and Ghorbani, A. A. (2012). "Improved Competitive Learning Neural Networks for Network Intrusion and Fraud Detection." In: *Neurocomputing* 75.1, pp. 135–145.

Leonard, K. J. (1995). "The Development of a Rule Based Expert System Model for Fraud Alert in Consumer Credit." In: *European Journal of Operational Research* 80.2, pp. 350–356.

Li, J. et al. (2008). "A Survey on Statistical Methods for Health Care Fraud Detection." In: *Health Care Management Science* 11.3, pp. 275–287.

Li, S.-H. et al. (2012). "Identifying the Signs of Fraudulent Accounts Using Data Mining Techniques." In: *Computers in Human Behavior* 28.3, pp. 1002–1013.

Liu, H. (2010). In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Springer. Chap. Feature Selection.

Liu, H. and Setiono, R. (1995). "Chi2: Feature Selection and Discretization of Numeric Attributes." In: *Proceedings of the*

Seventh International Conference on Tools with Artificial Intelligence, pp. 388–391.

Mahmoudi, N. and Duman, E. (2015). "Detecting Credit Card Fraud by Modified Fisher Discriminant Analysis." In: *Expert Systems with Applications* 42.5, pp. 2510–2516.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

McAfee, A. and Brynjolfsson, E. (2012). *Big Data: The Management Revolution*. Boston, MA.

Mena, J. (2003). *Investigative Data Mining for Security and Criminal Detection*. Butterworth-Heinemann.

Mingers, J. and Walsham, G. (2010). "Toward Ethical Information Systems: The Contribution of Discourse Ethics." In: *MIS Quarterly* 34.4, pp. 833–854.

Mitchell, T. (1997). *Machine Learning*. Ed. by C. L. Liu and A. B. Tucker. McGraw-Hill Series in Computer Science. Singapore: McGraw-Hill.

Mladenic, D. and Grobelnik, M. (1999). "Feature Selection for Unbalanced Class Distribution and Naive Bayes." In: *Proceedings of the Sixteenth International Conference on Machine Learning*. ICML '99. San Francisco, California, USA: Morgan Kaufmann Publishers Inc., pp. 258–267.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Upper Saddle River, New Jersey: Pearson Education, Inc.

Ngai, E. W. T. et al. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature." In: *Decision Support Systems* 50.3, pp. 559–569.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pérez, F. and Granger, B. E. (2007). "IPython: a System for Interactive Scientific Computing." In: *Computing in Science and Engineering* 9.3, pp. 21–29.

Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge, England: Cambridge University Press.

Phillips, D. C. and Burbules, N. C. (2000). *Postpositivism and Educational Research*. Lanham, Maryland, USA: Rowman & Littlefield Publishers.

Phua, C., Alahakoon, D., and Lee, V. (2004). "Minority Report in Fraud Detection: Classification of Skewed Data." In: *SIGKDD Explorations Newsletter* 6.1, pp. 50–59.

Phua, C. et al. (2010). "A Comprehensive Survey of Data Mining-based Fraud Detection Research." In: *CoRR* abs/1009.6119.

Popper, K. (1997). *Lesebuch: Ausgewählte Texte zu Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. 2nd edition. Tübingen, Germany: Mohr Siebeck.

Quah, J. T. S. and Sriganesh, M. (2008). "Real-Time Credit Card Fraud Detection Using Computational Intelligence." In: *Expert Systems with Applications* 35.4, pp. 1721–1732.

Ravisankar, P. et al. (2011). "Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques." In: *Decision Support Systems* 50.2, pp. 491–500.

Recht, G., ed. (2017). *Bundesdatenschutzgesetz (BDSG)*. German. 3rd edition. Merseburg, Germany: CreateSpace Independent Publishing Platform.

Reichenbach, H. (2013). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Original work published pre-1945, year unknown. London, England: Forgotten Books.

Russell, S. and Norvig, P. (2012). *Künstliche Intelligenz: Ein moderner Ansatz*. German. 3rd edition. München, Germany: Pearson Deutschland GmbH.

Sahin, Y. and Duman, E. (2011). "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines." In: *Proceedings of the International MultiConference of Engineers and Computer Scientists 2011* 1.

Sahin, Y., Bulkan, S., and Duman, E. (2013). "A Cost-Sensitive Decision Tree Approach for Fraud Detection." In: *Expert Systems with Applications* 40.15, pp. 5916–5923.

Saldaña, J. (2009). *The Coding Manual for Qualitative Researchers*. London: SAGE Publications Ltd.

Sánchez, D. et al. (2009). "Association Rules Applied to Credit Card Fraud Detection." In: *Expert Systems with Applications* 36.2, Part 2, pp. 3630–3640.

Schmidt, C. and Verbeet, M. (2004). "Betrügen leicht gemacht." German. In: *Der Spiegel* 31, p. 41.

Segaran, T. (2007). *Programming Collective Intelligence*. Ed. by M. T. O'Brien. Sebastopol, California, USA: O'Reilly Media, Inc.

Serrano, A. et al. (2012). "Neural Network Predictor for Fraud Detection: A Study Case for the Federal Patrimony Department." In: *The Seventh International Conference on Forensic Computer Science — ICoFCS 2012*, pp. 61–66.

Shani, G. and Gunawardana, A. (2011). "Recommender Systems Handbook." In: ed. by F. Ricci et al. Springer US. Chap. Evaluating Recommender Systems, pp. 257–297.

Sharma, A. and Panigrahi, P. K. (2012). "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques." In: *International Journal of Computer Applications* 39.1.

Shi, L., Liu, Y., and Ma, X. (2011). In: *Emerging Research in Artificial Intelligence and Computational Intelligence*. Ed. by H. Deng et al. Vol. 237. Communications in Computer and Information Science. Springer Berlin Heidelberg. Chap. Credit Assessment with Random Forests, pp. 24–28.

Spann, D. D. (2014). *Fraud Analytics: Strategies and Methods for Detection and Prevention*. Hoboken, New Jersey, USA: John Wiley & Sons.

Stevenson, A., J. Pearsall, and P. Hanks, eds. (2010). *Oxford Dictionary of English*. 3rd edition. Oxford University Press.

Tackett, J. A. (2013). "Association Rules for Fraud Detection." In: *Journal of Corporate Accounting & Finance* 24.4, pp. 15–22.

Torgo, L. and Lopes, E. (2011). "Utility-Based Fraud Detection." In: *Proceedings of 22th International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 1517–1522.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. 2nd edition. Cheshire, Connecticut, USA: Graphics Press.

U.S. Census Bureau (2014). *Gross Margin as a Percentage of Sales (1993-2014)*. URL: `https://goo.gl/Se2nF9` (visited on 02/16/2017).

Van Vlasselaer, V. et al. (2015). "APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection Using Network-Based Extensions." In: *Decision Support Systems* 75.Supplement C, pp. 38–48.

Viaene, S., Dedene, G., and Derrig, R. (2005). "Auto Claim Fraud Detection Using Bayesian Learning Neural Networks." In: *Expert Systems with Applications* 29.3, pp. 653–666.

Viaene, S., Derrig, R. A., and Dedene, G. (2004). "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis." In: *IEEE Transactions on Knowledge and Data Engineering* (5), pp. 612–620.

Viaene, S. et al. (2002). "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection." In: *Journal of Risk & Insurance* 69.3, pp. 373–421.

Viaene, S. et al. (2007). "Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry." In: *European Journal of Operational Research* 176.1, pp. 565–583.

Waite, M., ed. (2012). *Paperback Oxford English Dictionary*. 7th edition. Oxford University Press.

Walt, S. van der, Colbert, S. C., and Varoquaux, G. (2011). "The NumPy Array: A Structure for Efficient Numerical Computation." In: *Computing in Science & Engineering* 13.2, pp. 22–30.

Warburton, N. (2011). *A Little History of Philosophy*. Padstow, Cornwall, England: Yale University Press.

Webb, G. I. (2010). In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Springer. Chap. Data Preparation.

Wheeler, R. and Aitken, S. (2000). "Multiple Algorithms for Fraud Detection." In: *Knowledge-Based Systems* 13.2—3, pp. 93–99.

Wickham, H. (2014). "Tidy Data." In: *Journal of Statistical Software* 59 (10).

Williams, G. (2011). *Data Mining with Rattle and R*. New York: Springer New York.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. The Morgan Kaufmann Series in Data Management Systems. Burlington, Massachusetts: Morgan Kaufmann Publishers.

Wong, N. et al. (2012). "Artificial Immune Systems for the Detection of Credit Card Fraud: An Architecture, Prototype

and Preliminary Results." In: *Information Systems Journal* 22.1, pp. 53–76.

Yang, Y. (2010). In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Springer. Chap. Discretization.

Yang, Y. and Webb, G. I. (2001). "Proportional k-Interval Discretization for Naive-Bayes Classifiers." In: *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings*. Ed. by L. De Raedt and P. Flach. Berlin: Springer, pp. 564–575.

Yao, H., Hamilton, H. J., and Geng, L. (2006). "A Unified Framework for Utility Based Measures for Mining Itemsets." In: *Proc. of ACM SIGKDD 2nd Workshop on Utility-Based Data Mining*, pp. 28–37.

Yeh, I.-C. and Lien, C.-H. (2009). "The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients." In: *Expert Systems with Applications* 36.2, Part 1, pp. 2473–2480.

Zhao, J. et al. (2016). "Extracting and Reasoning about Implicit Behavioral Evidences for Detecting Fraudulent Online Transactions in E-Commerce." In: *Decision Support Systems* 86.Supplement C, pp. 109–121.

Zhou, W. and Kapoor, G. (2011). "Detecting Evolutionary Financial Statement Fraud." In: *Decision Support Systems* 50.3, pp. 570–575.

eMarketer (2017). *Retail Ecommerce in Germany to Top $65 Billion*. URL: `https://goo.gl/v8XzAy` (visited on 01/02/2018).

von Neumann, J. and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. 3rd edition. Princeton, New Jersey, USA: Princeton University Press.

APPENDIX

A

EVALUATION MEASURES

A.1 INTRODUCTION

In the light of the research problem, some classifier evaluation techniques seem more appropriate than others. Evaluation measures can be divided into two groups: set-based measures and measures for ranked lists (Manning, Raghavan, and Schütze, 2009). Ranked lists can be converted to discrete data by using a threshold value to divide the list into groups.

Because manual inspection is desired and therefore total automation is not required, and because such resources are limited, evaluation measures which are able to focus on the top of the ranked list are preferred. The following metrics and concepts are discussed below:

- confusion matrix (precision, recall, ...)
- precision at k
- utility at k
- ROC curves
- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Average Precision (MAP)
- Average Reciprocal Hit Rank (ARHR)
- Normalized Discounted Cumulative Gain (NDCG)

In order to discuss evaluation measures in a structured way, criteria for their evaluation need to be defined. The following

questions may be considered for differentiation: How does the measure's formula influence relevance values? Is the measure appropriate for the research problem? Is the measure capable of evaluating ranked lists? Does the measure support a notion of cost-sensitivity or utility? Is it easily understandable?

Measures are drawn from multiple areas: statistics, machine learning, information retrieval, and recommender systems. The former two have been discussed throughout the thesis. In addition, information retrieval is considered because fraud prevention can be seen as trying to retrieve relevant documents, i.e. fraudulent transactions, from data. Recommender systems are considered since fraud cases may be viewed as the relevant transactions that shall be offered to the user, in this case the fraud officer. Because there is no consistent formal style across the different sources, if necessary, equations are adapted to the fraud prevention problem using the notation below:

- k describes the k^{th} position in a ranked list of transactions
- $\delta(k)$ is the estimated fraud class at position k, either 1 for fraud or 0 for legitimate
- $\rho(k)$ describes a relevance score between 0 and 1, e.g. an estimated fraud probability at position k, i.e it is a continuous version of $\delta(k)$
- $\tau(k)$ is the actual fraud class at position k, either 1 for fraud or 0 for legitimate

A.2 SET-BASED EVALUATION MEASURES

"In multi-class prediction, the result on a test set is often displayed as a two-dimensional *confusion matrix* with a row and

	$\delta(k) = 1$	$\delta(k) = 0$
$\tau(k) = 1$	true positive (TP)	false negative (FN)
$\tau(k) = 0$	false positive (FP)	true negative (TN)

Table 10: Confusion matrix for binary fraud classification.

column for each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements" (Witten, Frank, and Hall, 2011, p. 164). The four outcomes shown in Table 10 lead to additional metrics that represent relations. Description of the metrics has been collected from Manning, Raghavan, and Schütze (2009), Shani and Gunawardana (2011), and Witten, Frank, and Hall (2011).

Precision, also called Positive Predictive Value (PPV), describes the fraction of correctly labeled fraud cases with regard to all cases labeled fraud. Recall (sensitivity, True Positive Rate (TPR)) describes the fraction of correctly labeled fraud cases with regard to all fraud cases. Whether precision or recall is preferred depends on the problem context (Manning, Raghavan, and Schütze, 2009).

$$\begin{aligned} \text{precision} \equiv \text{PPV} &= \frac{\text{TP}}{\#(\text{labeled fraud})} \\ \text{recall} \equiv \text{TPR} &= \frac{\text{TP}}{\#(\text{fraud})} \end{aligned} \tag{17}$$

An alternative to precision and recall is accuracy. Accuracy considers both true positives and true negatives equally important. Thus, it is the ratio of correct classifications regardless of the label. Yet, in the case of fraud prevention, this is not appropriate: The data are skewed, because the fraction of fraud cases is small compared to the total number of transactions. Hence, if accuracy were maximized when data are skewed, no

cases would be labeled fraud anymore (Manning, Raghavan, and Schütze, 2009).

$$\begin{aligned} \text{accuracy} &= \frac{TP + TN}{\#(\text{all transactions})} \\ \text{specificity} &= \frac{TN}{\#(\text{legitimate})} \\ NPV &= \frac{TN}{\#(\text{labeled legit.})} \\ FPR &= \frac{FP}{\#(\text{legitimate})} \\ F &= \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \end{aligned} \tag{18}$$

Likewise, specificity and Negative Predictive Value (NPV) are not appropriate for fraud detection. Another single measure is the F measure. Since the formula is less straightforward than the formulas introduced above, the F measure is visualized in Figure 21. It trades off precision versus recall and allows to emphasize one of them using the α value. The figure shows two graphs; on the left side, recall is favored ($\alpha = 0.1$), and on the right side, precision and recall are balanced ($\alpha = 0.5$). However, it sometimes makes sense to maintain multiple measures instead of aiming at a single value (Manning, Raghavan, and Schütze, 2009).

A.3 EVALUATION MEASURES FOR RANKED LISTS

The set-based measures can be converted into measures for ranked lists. For example, *precision at* k can be considered an extension of regular precision. Precision is defined as the number of correctly detected fraud cases (i.e. TP cases) divided by the number of cases that are thought to be fraud (Shani and Gunawardana, 2011), and precision at k is defined as the precision

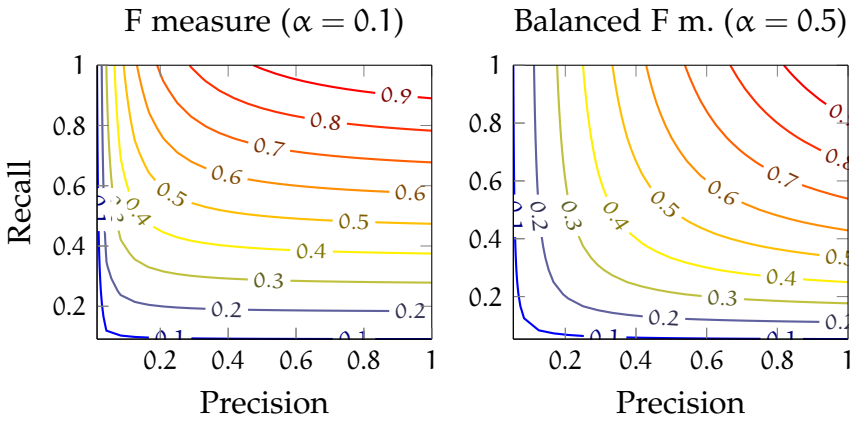

Figure 21: F values for combinations of precision and recall values.

value for the set of transactions up to the k^{th} percentile (or the k^{th} transaction) of a ranked list:

$$\text{precision at } k = \frac{\text{TP within } k^{th} \text{ percentile}}{\#(\text{labeled fraud within } k^{th} \text{ p.})} \tag{19}$$

With regard to the problem context, the application of precision at k is even simpler: All transactions up to the inspection capability k should be fraud cases, or at least all fraud cases should be among the top k cases presented. Hence, in this particular application precision at k is just the fraud ratio of all cases up to k. Manning, Raghavan, and Schütze (2009) state that precision can also be converted to R-precision, where a perfect score of one is possible if all transactions in the subset are labeled correctly. The concept of "at k" also allows to plot precision-recall curves and ROC curves. Precision at k ignores the economic perspective, though. This issue is addressed in Chapter 8, where *utility at* k is introduced.

MAP is a single measure that deals with different *information needs* (Manning, Raghavan, and Schütze, 2009). In information retrieval, different search queries are usually meant to produce different results. However, in fraud prevention, the query is always to identify fraud cases. Although the measure can still be

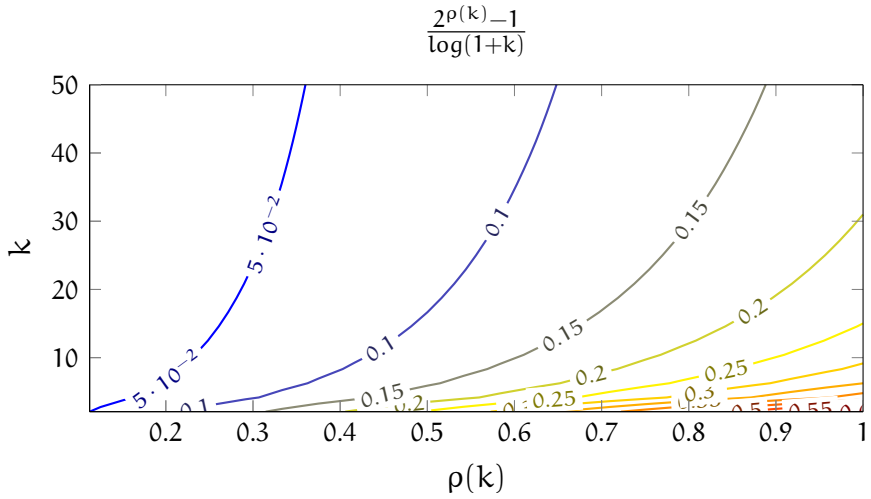

Figure 22: Scores of discounted gains per relevance score and position.

used, it solves a problem absent in fraud prevention. Thus, MAP is not explained any further.

NDCG, short for Normalized Discounted Cumulative Gain, is another single measure (Manning, Raghavan, and Schütze, 2009). It uses a relevance score (utility) that is discounted according to k, i.e. transactions at the top of the list are weighted stronger than those farther away. In fraud prevention, relevance could differ across fraud cases, for example when considering potential financial damage. The formula has been simplified, omitting the part dealing with different information needs.

$$\mathrm{NDCG}(k) = \sum_{k=1}^{n} \frac{2^{\rho(k)} - 1}{\log_2(1+k)} \tag{20}$$

Observing Figure 22, it becomes obvious that the farther away from the top a transaction is ranked, the weaker the influence of its relevance score becomes. For example, approximately, a relevance score of 0.5 at position 5 influences the DCG as strongly as a score of 0.8 at position 10 while a score of 0.8 at position 2 is twice as much as the latter. The method

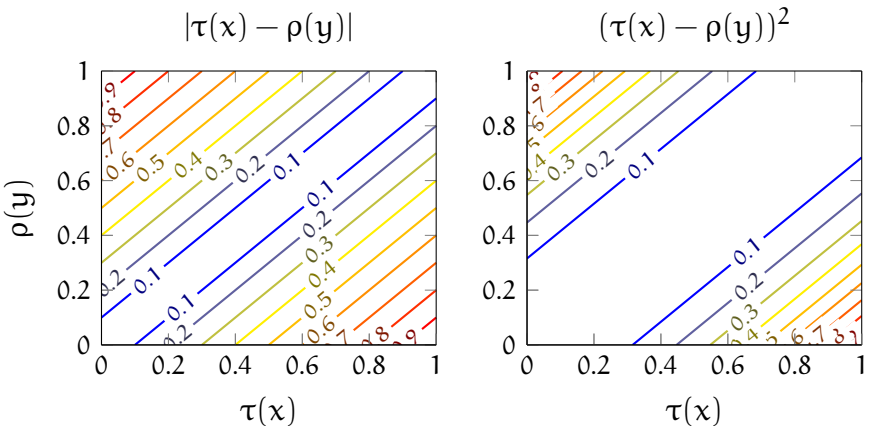

Figure 23: Individual scores obtained from RMSE and MAE per relevance score and position.

of discounting a gain value resembles discounted cash flow in economics. However, NDCG does not produce meaningful financial numbers, and therefore, other techniques are preferred to consider the financial impact of fraud cases. An alternative is ARHR, where each transaction's utility is calculated by $1/k$ if and only if the transaction is relevant, with k being its position (Shani and Gunawardana, 2011).

Two of the most common measures of errors in statistics are RMSE and MAE as introduced by Shani and Gunawardana (2011, pp. 273 sqq.).

$$\begin{aligned} \mathrm{RMSE}(k) &= \sqrt{\frac{1}{k}\sum_{n=1}^{k}(\tau(n) - \rho(n))^2} \\ \mathrm{MAE}(k) &= \sqrt{\frac{1}{k}\sum_{n=1}^{k}|\tau(n) - \rho(n)|} \end{aligned} \tag{21}$$

Of course, instead of the relevance score $\rho(k)$, the binary classification $\delta(k)$ could be used, which is a special case of the former. The square function of RMSE emphasizes the importance of large errors. The difference between the two measures is shown in Figure 23.