DEGREE PROJECT IN INDUSTRIAL ENGINEERING AND MANAGEMENT
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2023

Customer Acquisition Process Digitalization: A Case Study on the Use of Machine Learning in the Corporate Insurance Industry

Master Thesis

Klara Larsson and Freja Ling

KTH ROYAL INSTITUTE OF TECHNOLOGY

INDUSTRIAL ENGINEERING AND MANAGEMENT

10 JUNE 2023

Customer Acquisition Process Digitalization: A Case Study on the Use of Machine Learning in the Corporate Insurance Industry

by

Klara Larsson
Freja Ling

Master of Science Thesis TRITA-ITM-EX 2023:292
KTH Industrial Engineering and Management
Industrial Economics and Management
SE-100 44 STOCKHOLM

Digitalisering av kundanskaffningsprocessen: En fallstudie om användningen av maskininlärning inom företagsförsäkringsbranschen

Klara Larsson
Freja Ling

Examensarbete TRITA-ITM-EX 2023:292
KTH Industriell teknik och management
Industriell ekonomi och organisation
SE-100 44 STOCKHOLM

Authors

Freja Ling, Klara Larsson
frejal@kth.se, klalar@kth.se
Industrial Engineering and Management
KTH Royal Institute of Technology

Research Location

Stockholm, Sweden

Date

10 June 2023

Examiner

Frauke Urban
Stockholm
KTH Royal Institute of Technology

Supervisor

Emrah Karakaya

Stockholm

KTH Royal Institute of Technology

KTH Industrial Engineering and Management

Master of Science Thesis TRITA-ITM-EX 2023:292

Customer Acquisition Process Digitalization: A Case Study on the Use of Machine Learning in the Corporate Insurance Industry

Klara Larsson

Freja Ling

Approved	Examiner	Supervisor
2023-06-10	Frauke Urban	Emrah Karakaya
	Commissioner	Contact person
	Söderberg & Partners	Otto Åkerström

Abstract

This thesis explores the application of machine learning (ML) techniques in customer classification and their integration into customer relationship management (CRM) systems within the corporate insurance industry. The research aims to address the gap in the use of AI-CRM for the corporate insurance industry. It was conducted as a case study at a Swedish insurance broker company. The study leveraged external data sources to create a data set on customer information. The feature selection process included Variance Influence Factor (VIF) to remove collinearity and then Mutual Class Info and Random Forest, which are methods used to find which independent variables affect the dependent variable the most. Also, Recursive Feature Testing was applied to find the best feature combinations. Four different binary classification models were implemented and compared — Decision Tree, Random Forest, Support Vector Machine, and Artificial Neural Network. Note that Random Forest can be used both for feature selection and classification. The models were tested on four different feature combinations and evaluated using Accuracy, Recall, Precision, F1-score, and ROC-AUC. The study further conducted interviews at the partner company to evaluate their current CRM system. The findings show that ML-based customer classification can be leveraged to effectivize the customer acquisition process for corporate insurance. The Support Vector Machine model achieved the highest accuracy, at 80%. Depending on the available data and the use of metrics, different classifiers had the best performance. The study also found that when implementing classification into AI-CRM, the specific requirements at the company need to be examined. This study found it important to consider the data procurement process, the current customer acquisition process, the risks associated with misclassification, and present bias. The findings of this study have theoretical implications for the implementation of AI-CRM for customer acquisition. It demonstrates the practical benefits of integrating machine learning techniques into CRM systems, emphasizing the effectiveness of AI-CRM for customer classification. Further,

by comparing different classification models and evaluating their performance, the study enhances the theoretical understanding of model selection for customer classification tasks in this specific domain. Additionally, the research provides insights into effective feature selection methods, aiding researchers and practitioners in extracting relevant variables for customer classification.

Keywords

Customer Relationship Management (CRM), Customer Classification, Customer Acquisition, Machine Learning, Insurance Industry, Corporate Insurance, B2B, AI-CRM

KTH Industriell teknik
och management

Examensarbete TRITA-ITM-EX 2023:292

Digitalisering av kundanskaffningsprocessen: En fallstudie om användningen av maskininlärning inom företagsförsäkringsbranschen

Klara Larsson
Freja Ling

Godkänt	Examinator	Handledare
2023-06-10	Frauke Urban	Emrah Karakaya
	Uppdragsgivare	Kontaktperson
	Söderberg & Partners	Otto Åkerström

Abstrakt

Denna studie utforskar tillämpningen av maskininlärning (ML) inom kundklassificering och dess integration i kundrelationssystem (CRM) inom företagsförsäkringsbranschen. Forskningen syftar till att fylla kunskapsluckan inom användningen av AI-CRM inom företagsförsäkringsbranschen. Studien genomfördes som en fallstudie på ett svenskt försäkringsmäklarföretag. Studien utnyttjade externa datakällor för att skapa en dataset av kundinformation. Processen för val av variabler inkluderade Variance Influence Factor (VIF) för att ta bort kollinearitet och sedan Mutual Class Info och Random Forest, som är metoder som används för att hitta vilka oberoende variabler som påverkar den beroende variabeln mest. Dessutom användes Recursive Feature Testing för att hitta de bästa kombinationerna av funktioner. Fyra olika binära klassificeringsmodeller implementerades och jämfördes - Decision Tree, Random Forest, Support Vector Machine och Artificial Neural Network. Observera att Random Forest kan användas både för val av funktioner och klassificering. Modellerna testades med fyra olika kombinationer av variabler och utvärderades med hjälp av Accuracy, Recall, Precision, F1-score och ROC-AUC. Studien genomförde även intervjuer på partnerföretaget för att utvärdera deras nuvarande CRM-system. Resultaten visar att ML-baserad kundklassificering kan användas för att effektivisera processen för kundanskaffning inom företagsförsäkring. Support Vector Machine-modellen uppnådde högst accuracy, 80%. Beroende på tillgängliga data och användning av evalueringsmått hade olika klassificerare bäst prestanda. Studien fann också att vid implementering av klassificering i AI-CRM måste de specifika kraven på företaget undersökas. Denna studie fann det viktigt att beakta processen för dataanskaffning, den nuvarande processen för kundanskaffning, riskerna med felklassificering och nuvarande partiskhet. Resultaten av denna studie har teoretiska implikationer för implementeringen av AI-CRM för kundanskaffning. Den visar på de praktiska fördelarna med att integrera maskininlärningstekniker i CRM-system och betonar effektiviteten hos AI-CRM för kundklassificering. Dessutom förbättrar studien

den teoretiska förståelsen för val av modeller för kundklassificeringsuppgifter i det specifika domänet genom att jämföra olika klassificeringsmodeller och utvärdera deras prestanda. Studien ger också insikter om effektiva metoder för val av variabler och hjälper forskare och utövare att extrahera relevanta variabler för kundklassificering.

Nyckelord

Kundrelationssystem (CRM), Kundklassificering, Nykundsbearbetning, Maskininlärning, Försäkringsbranschen, Företagsförsäkring, B2B, AI-CRM

Acknowledgements

We would like to express our sincere gratitude to our supervisor Emrah Karakaya, for his continuous support, encouragement, and guidance throughout the entire journey. His expertise, engagement, and dedication have been instrumental in shaping and refining this thesis. We appreciate the time you spent learning more about machine learning in order to provide us with feedback.

We are also deeply thankful to our examiner, Frauke Urban, for her valuable insights, guidance, and feedback during the seminars throughout the process of conducting this thesis. Her expertise and constructive comments have greatly contributed to the quality and rigor of this research.

Furthermore, we would like to extend our appreciation to Otto Åkerström and Nils Olsgärde from Söderberg & Partners for their cooperation, assistance, and willingness to share their knowledge and resources. Their insight into the technical parts of this study was of great importance for practical implementations. We would also like to thank the interviewees at the company. Their valuable inputs and collaboration have enriched this research and provided valuable industry insights.

We also express our gratitude to our peers for their constructive criticism and engaging discussions, which gave valuable insights and helped shape the research project.

We are grateful to all the individuals who have contributed in various ways to the successful completion of this thesis. Their support, encouragement, and contributions have been invaluable, and we are truly thankful for their assistance.

Acronyms

AI Artificial Intelligence

ANN Artificial Neural Network

BD Big Data

CRM Customer Relationship Management

DT Decision Tree

FP False Positive

FN False Negative

GA Genetic Algorithm

LCV Lifetime Customer Value

LR Logistic Regression

ML Machine Learning

MCI Mutual Class Info

NB Naive Bayes

NLP Natural Language Processing

RF Random Forest

SME Small-to-Medium Enterprise

SVM Support Vector Machine

VIF Variance Inflation Factor

Contents

Chapter 1

Introduction

This Chapter contains an introduction of the thesis. Beginning with a background of the field followed by the problem formulation, purpose, and research questions. Thereafter the delimitations are presented, and lastly, the outline of the report is described.

1.1 Background

The insurance industry is an important and complex sector that is critical in managing and mitigating risks in modern society (Grant 2012). Corporate insurance spreads the risk of unforeseen events across the insured business sector, protecting business owners from financial loss (Grant 2012). By transferring the risk of loss, corporate insurance promotes financial stability and economic growth by allowing corporations to operate without taking on the full burden of the financial risk (Grant 2012). Recently, the Covid-19 pandemic was an unforeseen event whose financial impact on the business sector was dampened by corporate insurance coverage (Nebolsina 2021; Przybytniowski et al. 2022).

The insurance industry encounters several challenges in today's fast-paced and competitive business landscape. The sector is currently undergoing a transformation towards increased digitalization (Eling and M. Lehmann 2018; Saxena and R. Kumar 2022), and the emergence of so-called "insurtechs" — digital insurance startups — has increased the competitive climate and sped up the digitalization of the sector (Stoeckli, Dremel, and Uebernickel 2018; Saxena and R. Kumar 2022).

Insurance brokers have traditionally relied on conventional customer acquisition strategies

such as direct mail, advertising, and referrals and typically have high costs related to acquisition (C.-C. Liu and Liao 2020). However, these strategies are growing increasingly ineffective due to new emerging technologies and digital insurance solutions. Customers can access a broad range of information and compare insurance products and services where companies offer digital acquisition channels such as recommendation systems and chatbots (Wong et al. 2020). Still, some consumers prefer personal advice regarding insurance products (Bryzgalov and Tsyganov 2022; C. Eckert, J. Eckert, and Zitzmann 2021), meaning that not all consumers want a fully digital experience when buying insurance. Business-to-business (B2B) relationships also differ from those of business-to-consumer (B2C), providing an additional challenge for corporate insurance companies. Limited research exists on the acquisition methods for corporate insurance brokers and on the preferences of acquisition and relationship management of corporate insurance customers regarding digitalization. However, transparency and trust are notable hygiene factors in B2B insurance relationships (Dexe, Franke, and Rad 2021). This is further supported by the fact that trust and personal relationships are cornerstones for B2B relationships in other sectors as well (I. Saura, Frasquet, and Cervera-Taulet 2009; Young, Wilkinson, and Smith 2015; Webster 1992; Zeeland-van der Holst and Henseler 2018).

Insurance brokers are turning to technical solutions such as customer classification and recommendation systems to improve their acquisition and retention strategies (Saxena and R. Kumar 2022; Chiu 2002). These systems employ sophisticated data analytics and machine learning (ML) algorithms to scrutinize customer behavior and preferences, thus delivering customized suggestions for insurance products and services (Pisoni and Díaz-Rodríguez 2023; Saxena and R. Kumar 2022). These technologies typically run on large amounts of data to be beneficial. As technologies such as data mining and big data (BD) are increasingly being used by companies in order to improve analysis, strategic decisions, and gain customer insight (Khade 2016; Satish and Yusof 2017; Ngai, Xiu, and Chau 2009), it provides an opportunity to combine classification systems and BD in order to improve customer relationships. These data are commonly collected and stored in Customer Relationship Management (CRM) Systems, which integrate customer data across several functions (Tohidi and Jabbari 2012; Anshari et al. 2019; Zerbino et al. 2018; Ngai, Xiu, and Chau 2009). Among other things, CRM systems are used to improve customer relationships, retention, and acquisition rates (I. J. Chen and Popovich 2003; Xu et al. 2002). As artificial intelligence (AI) and ML algorithms utilize large

amounts of data, researchers have explored the possibility of using CRM together with AI (Chatterjee, Rana, et al. 2021; Libai et al. 2020). There have also been previous research into the use of AI-CRM in the B2B context, where companies implement AI-CRM to stay competitive in the dynamic business landscape enabled by the digitalization of business and emerging digital technologies (Chatterjee, Chaudhuri, and Vrontis 2022; Chatterjee, Chaudhuri, Vrontis, and Jabeen 2022; Rahman et al. 2023).

There is thus an opportunity for a solution to digitalize and automate the customer acquisition process for B2B insurance brokers, which also respects the need for personal contact and trust in B2B relationships. This study explores the use of AI-CRM in the B2B insurance sector through a case study and implementation of a digital B2B acquisition solution. This contributes to the literature studying the use of CRM in the B2B context and the digitalization of the insurance industry by examining the use of AI-CRM solutions for corporate insurance brokers. By leveraging customer data and offering personalized recommendations, the study demonstrates how insurance brokers can improve customer engagement and satisfaction, increasing customer retention and acquisition rates, and revenue in the business insurance industry.

1.1.1 Case Study Background

The authors partner with a Swedish insurance broker company to provide a more practical approach to this research. This partnership enables the conduction of a case study on how customer classification can be applied in the business insurance industry to acquire new customers. The case study is focused on recommending the combined business insurance product to potential customers by classifying them into relevant segments based on their business needs and risks. There is a close collaboration with the insurance broker company to collect data on potential customers and use ML algorithms to classify them into different segments based on their characteristics.

The process

The considered process is a customer acquisition process in which insurance brokers seek out and cold call prospective businesses, suggesting that they switch their current insurance to the product featured in this study, a type of combination insurance product described in detail further down. To comprehend the process, interviews were conducted with three departments at the partner company. As described by the partner company,

the resulting process can be viewed in Figure 1.1.1.

The acquisition process involves three stakeholders: the insurance broker, the insurance company, and the customer. The insurance broker is the partner company and the process leader, and initiates the process by manually searching for prospective customers. Individual insurance brokers use their personal knowledge and intuition to find customers who they believe would be suitable for the product by filtering on different categories, such as specific industries. The brokers' goal is to find companies that will be accepted by the insurance company as they hand them over in the next stage of the process.

The brokers cold-call potential customers and suggest that they switch their business insurance for a more competitive price. Interested customers share their current insurance information, including their current premium and coverage, which is then used to filter further. A list of potential customers is then sent to the insurance company for another customer selection process since they take on all the risk in the final stage of becoming the insurer. The insurance company either gives a product offer or declines to offer the customer the product. Finally, the customer either accepts the offer or does not.

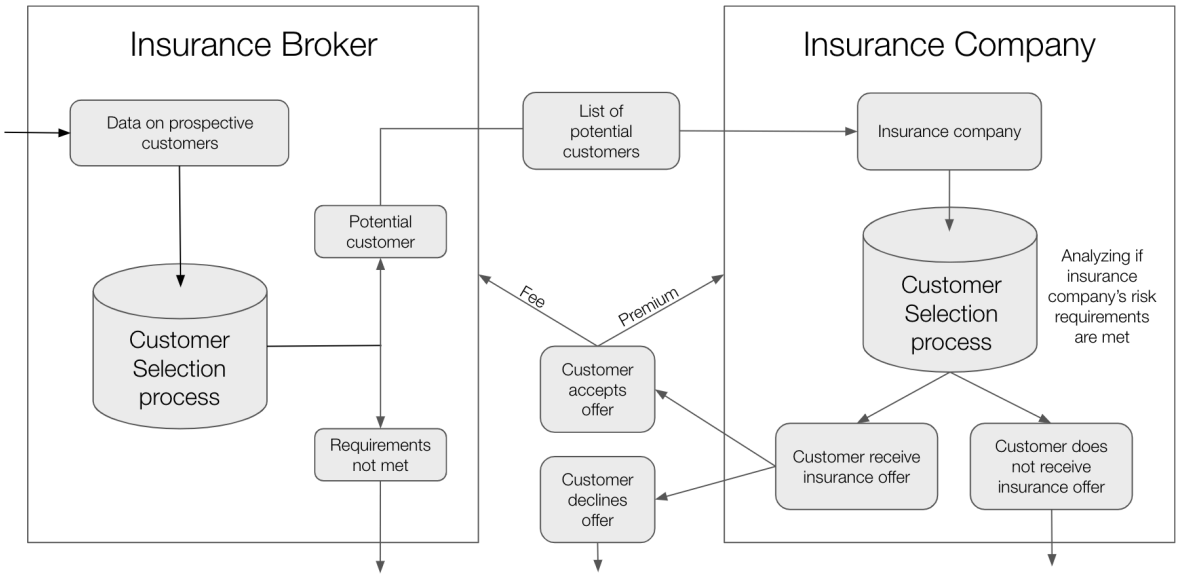


Figure 1.1.1: The customer acquisition process of SME customers at the Swedish insurance company

The Product

The product is a combined company insurance tailored for small to medium-sized enterprises (SMEs) that can include several types of insurance: title insurance, interruption insurance, liability insurance, and others.

The Case

The customer selection process on the broker's side is currently conducted manually and is not standardized, leading to time-consuming research and the potential omission of specific corporate customers from certain industries due to individual bias. Further, finding the customers that are most likely to be accepted by the insurance company's selection process is directly linked to the brokers' profits. Thus, they believe there is room for improvement in terms of efficiency as well as higher acceptance rates from the insurance company due to more thoroughly selected customers. To improve process efficiency, the partner company aims to adopt emerging technologies within the field, which can reduce the time required for the process and increase the acceptance ratio of proposed customers by the insurance company.

1.2 Problem Formulation

The business insurance sector encounters a number of difficulties in attracting new clients. Insurance brokers frequently rely on conventional sales techniques and lack the resources to segment potential clients based on their particular traits and preferences. The main challenge for the brokers lies in not missing potential customers for the product and avoiding recommending non-suitable customers for the product. The result is either missed potential revenue for the former or wasted time and resources for the latter.

There is a potential to utilize data collected through the insurance brokers' CRM systems with emerging classification systems technologies to improve the customer acquisition process. The goal is to improve customer acquisition rates internally through faster processing and externally through an increased acceptance rate by the insurance company.

Additionally, there is a dearth of prior research on using customer classification and recommendation systems for new clients in the corporate insurance sector, particularly for small to medium-sized enterprises. Despite the fact that these technologies have gained widespread adoption in other sectors like e-commerce and entertainment, their potential in the insurance industry has not been explored to the same degree.

Integrating customer classification systems into CRM systems can pose a challenge for insurance brokers, requiring significant technological and organizational changes. Furthermore, there is a lack of guidance on how to effectively implement these systems

in the acquisition strategy of insurance brokers.

1.3 Purpose

The research aims to explore the potential of customer classification in the corporate insurance industry, and their integration into CRM systems. It will investigate the current state of these systems, explore their potential benefits for collecting and utilizing customer data in customer acquisition, and develop an ML-based system for the study partner company. The research also aims to provide practical insights and recommendations for insurance brokers looking to implement these systems in their acquisition strategy and integrate them into their CRM systems.

1.4 Research Questions

To achieve the purpose of this study, the following two research questions have been proposed:

1. How can ML classification systems be used to optimize the customer acquisition process of insurance brokers in the SME market?

2. How could an ML-based classification system be integrated for use with an existing CRM system?

These research questions are designed to explore the practical application and potential benefits of developing and implementing an ML customer classification system in the context of the corporate insurance industry. By integrating the system into the acquisition strategy of insurance brokers and their CRM systems, the study will provide insights into how these tools can attract and retain new customers and improve customer satisfaction. The findings of this study aim to have positive implications for corporate insurance brokers and contribute to the literature for future researchers and practitioners interested in customer classification systems.

1.5 Delimitations

In this research, the following delimitations will be made:

Geographic delimitations

The research will focus on small to medium-sized businesses in Sweden, and the findings may not be applicable to other countries or regions.

Industry delimitations

The research will focus on the corporate insurance industry, and the findings may not apply to other insurance types.

Time delimitations

The research will be conducted over a period of six months, and the findings may not reflect the long-term effects of integrating customer classification and recommendation systems in the insurance industry.

Data & Technical delimitations

The research will use data provided by the insurance broker company partner and available data from public sources on the web, no other data sources will be analyzed. The data analysis methods are limited to implementations in Python.

Scope delimitations

The research will only focus on integrating the customer classification system into the acquisition strategy of insurance brokers, and other aspects of the insurance industry, such as claims processing and underwriting, will not be analyzed.

By acknowledging these delimitations, the scope and limitations of the research are clearly defined, and the findings can be appropriately interpreted within these boundaries.

1.6 Outline

The paper will begin with introducing the theoretical framework (Chapter 2) that will be applied, following a literature study (Chapter 3) into relevant research in the area. A methods section will then be presented (Chapter 4), outlining the research process as well as the methods used. After this is a presentation of the results (Chapter 5), followed by a discussion (Chapter 6). Finally, a conclusion (Chapter 7) with recommendations for future research.

Chapter 2

Theoretical Framework

The following section introduces the theoretical framework that will direct the research. The study's theoretical framework is based on the Theory of Customer Relationship Management (CRM), which provides a comprehensive understanding of the factors necessary for effective customer relationship management that is crucial for the acquisition and retention of customers.

2.1 Framework Introduction

A CRM system is described as an integrated way of managing customer relationships in order to improve customer relationships and customer retention (I. J. Chen and Popovich 2003). It serves various purposes, including assisting in marketing and customer acquisition by utilizing customer data across several functions to identify and focus on the most suitable customers (Xu et al. 2002). CRM systems' role in customer acquisition is in focus in this research.

This study aims to investigate how ML classification systems can be used to optimize the customer acquisition process of insurance brokers in the SME market (RQ1) and how such a classification system can be integrated into a CRM system (RQ2). To be able to answer these questions, the CRM framework has to be applied from the beginning. Regarding the first research question, it is important that an ML classification system could work together with the data from a company's CRM system in order to bring value to the organization. Thus, if the organization can integrate the ML classification systems into its CRM system, it can provide them with valuable insights and assist them

in making better decisions in terms of customer acquisition. Additionally, by utilizing the CRM framework, companies can effectively manage customer data and identify potential customers, which is vital for successful customer acquisition. Therefore, in this study, the CRM framework serves as a theoretical foundation for investigating how ML classification systems can be integrated into the customer acquisition process of insurance brokers in the SME market.

CRM systems have, in previous research, proven applicable in similar cases. For example, Müller and Te (2017) discovered that integrating a random forest model into an insurance CRM system within the insurance industry could prove economically advantageous by focusing on motor insurance policies. The model enables insurance companies to increase their insurance premiums efficiently by targeting only the most promising customers. The study, together with several others mentioned in Chapter 3 incentivizes the application of this framework.

2.2 A Framework for Customer Relationship Management (CRM)

CRM is a business framework for companies that want to build and maintain strong customer relationships. A well-designed CRM system can help companies improve customer classification, increase customer loyalty, and ultimately drive business growth (Winer 2001). Since this research paper aims to improve the customer acquisition process by classifying SME customers, it is closely related to conducting and improving a CRM system for the partner company. CRM focuses on acquiring and retaining customers, and this research mainly focuses on the framework's acquisition side. Thus, in order to provide direction for the research project, the CRM framework is used as a reference. This section explores the key pillars of a successful CRM strategy based on Winer's article "A Framework for Customer Relationship Management" (2001).

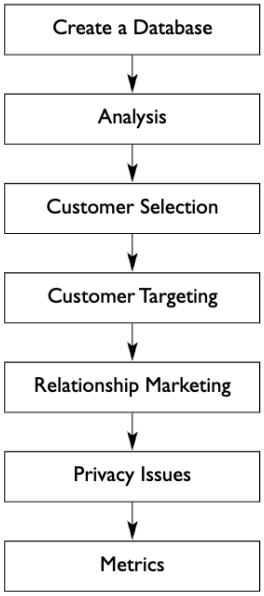


Figure 2.2.1: CRM Model (Winer 2001, p. 91)

2.2.1 Create a Database

The first pillar of a successful CRM program is creating a database of customer activity. As Winer notes, "A necessary first step to a complete CRM solution is the construction of a customer database or information file" (Winer 2001, p. 91). This database should contain information on customer contact details, purchasing history, descriptive data, and customer response data. By collecting this data, companies can gain valuable insights into their customers' behavior, preferences, and needs. The data used for customer segmentation is descriptive data, and thus the most important to collect for this research where customers are classified.

2.2.2 Analyzing the Data

The second pillar of a successful CRM program is database analysis. Companies can use various analytical techniques, such as data mining, clustering, and predictive modeling, to gain insights into customer behavior. Companies can identify high-value customers, create customer segments, and develop customer profiles by analyzing the database. The customer data analysis has gone from being group-focused to more individualized. Winer states that "... there is increased attention being paid to understanding each 'row' of the database—that is, understanding each customer and what he or she can deliver to the company in terms of profits ..." (Winer 2001, p. 94). Consequently, the term "lifetime customer value" (LCV) emphasizes the need to analyze each customer in the database regarding their present and future profitability for the firm. One example of profits from this analysis is reducing customer acquisition costs. (Winer 2001).

2.2.3 Customer Selection

After analyzing the customer database, the next step is determining which customers to target. Segmentation-type analyses are performed to identify the most desired segments of customers based on factors such as purchasing rates and brand loyalty. The descriptor variables for these segments provide information for deploying marketing tools. They can be matched with commercially available databases of names to find additional customers matching the profiles of those chosen from the database. Individual customer-based profitability analysis, such as LCV, can determine which customers to focus on by choosing what is profitable or projected to be or imposing a Return on Investment obstacle. The goal is to separate customers that will provide the most long-term profits

from those that are currently hurting profits. (Winer 2001).

The following steps of the framework are not considered in this paper, but are explained here as they are included in the original framework:

2.2.4 Targeting the Customers

While mass marketing approaches like television, radio, and print advertising are great for generating awareness, they are not well-suited for CRM. Targeted direct marketing methods such as telemarketing, direct mail, and direct sales are better suited for engaging specific customers. Personalized email marketing has emerged as a popular and effective tool for CRM, where customers have to choose to receive messages. (Winer 2001).

In this research, the most important customer targeting approach is direct sales since the brokers of the partner company directly approach prospective customers by cold calling.

2.2.5 Relationship Programs

The fourth pillar of a successful CRM program is building relationships with targeted customers. Relationship programs aim to deliver a higher customer satisfaction level than competitors. Customer service, loyalty/frequency programs, customization, reward programs, and community building are all essential components of CRM. Companies can increase customer loyalty and drive business growth by building customer relationships. (Winer 2001).

This pillar will be less focused on in this research since this is more relevant to customer retention than customer acquisition.

2.2.6 Privacy Issues

The fifth pillar of a successful CRM program is addressing privacy issues. CRM relies on customer data analysis for better marketing and relationship building. However, a trade-off exists between personalized service and the amount of personal information required to achieve this. (Winer 2001). With the rise of data mining algorithms to extract and analyze data from online platforms, consumers and advocacy groups are concerned about how much personal information is stored in databases and how it is used. General Data Protection Regulation (GDPR) is a piece of legislation that aims to protect consumers'

right to their data (EU 2016) and is a principle that must be followed in order to respect the customers' privacy.

Since this research will build its database on corporate customers, some of the information, such as financials, will be public. However, all information about the companies' insurance, such as their current premium, has to be handled as sensitive.

2.2.7 Metrics

The sixth and final pillar of a successful CRM program is the establishment of metrics. The focus on CRM requires new metrics to measure the success of products and services. Traditional financial and market-based indicators are still important, but there is now increased emphasis on customer-centric measures. These include customer acquisition costs, conversion rates, retention/churn rates, same-customer sales rates, loyalty measures, and customer share. These measures require better acquisition and processing of internal data to evaluate the company's performance at the customer level. (Winer 2001).

Due to the time limits of this research metrics such as how the classification system affects customer acquisition rates and costs could not yet be evaluated since the system then had to be tested in action.

2.3 More Recent Research

In a more recent study from 2019, CRM is described as "... a system which consists of interrelated set of components which are ambitious to improving the relationship with existing and new potential buyers with different numerous strategic advantages to the business firm." (Tigari 2019, p. 555). Further, the components of CRM are described as (1) the customer or potential buyer, (2) the relationship, which refers to the connection or association between the customer and firm, and (3) management, which is responsible for identifying, creating, attracting, developing and retaining customers (Tigari 2019).

The same study (Tigari 2019) explains that organizations can employ several types of customer relationship management (CRM) to enhance their relationships with customers. One of these is *analytical CRM*, which involves identifying target customers, collecting feedback from them, and analyzing customer information to create and deliver value to customers. Another is *strategic CRM*, which focuses on a contingency approach to a

customer-centric business strategy supported by all departments in the organization. *Operational CRM* is a modernized business process supported by information and communication technology, encompassing sales, marketing, and service automation. Finally, *collaborative CRM* involves strategically sharing customer information within an organization's network to focus on targeted customers, sometimes called strategic customer relationship management. The two CRMs that will be mostly focused on in this research are analytical CRM and operational CRM. This research aims to identify the target customer and automate parts of the current sales process.

Chapter 3

Literature review

This chapter includes an overview of the current literature regarding this thesis, divided into two parts. The first part discusses research on CRM systems, customer acquisition, and the benefits of customer classification. The second part concerns research on recommendation systems and their applications in the insurance industry, as well as the current digitalization of the insurance sector. Lastly, the third part includes literature on AI and its implications for businesses.

3.1 Customer Relationship Management

3.1.1 The Customer Acquisition Problem

One of the challenges companies face to grow and maintain their competitiveness against rival companies is acquiring new customers. It is a critical aspect of any company's growth and sustainability, as acquiring new customers helps to increase revenue, market share, and brand recognition. By understanding the value of customer relationships, firms can allocate their resources more effectively and make better strategic decisions. Customer acquisition, retention, and LCV are key metrics that are being studied exhaustively (McCarthy and Fader 2018).

As argued by S. Gupta, D. R. Lehmann, and Stuart (2004), a firm's customer base is an important asset, and the value of a firm's customer base can be used as a way to measure a firm's financial value, as well as its projected future earnings. They further state that customer acquisition costs are much higher than customer retention costs, underlining the importance of retention over acquisition. However, in order to

grow its customer base, a firm must acquire new customers. Further, Thomas (2001) argues that customer acquisition and retention are not independent processes and that acquiring new customers can positively impact customer retention rates. This is also found by Min et al. (2016), who found that a firm's market position was more sensitive to acquisition costs rather than retention costs. Another study highlighting this correlation is Nijssen, Guenzi, and van der Borgh (2017), which emphasizes the importance of developing ambidexterity in sales organizations to manage customer acquisition and retention processes. The authors identify several sales capabilities that facilitate this and positively impact sales organization ambidexterity, including incentive management, cross-functional cooperation, sales training, and customer prioritization. Nijssen, Guenzi, and van der Borgh (2017) finds a positive correlation between high levels and aligned acquisition and retention capabilities with superior organic growth, but profit growth is only achieved if acquisition capabilities are high.

Rust, Lemon, and Zeithaml (2004) further discusses the cost implied with acquiring new customers through marketing and the importance of maximizing the return of that cost for firm profitability. As a firm's customer base can be used as a measure to determine financial value, S. Gupta, D. R. Lehmann, and Stuart (2004) argue, however, that this cost should be seen as an investment and that reducing acquisition costs may not be an effective way for firms to increase their value. There is, therefore, a need for companies to both invest in customer acquisition and to simultaneously keep this process cost-efficient. Y. Chen and Hu (2005) furthermore finds that many companies disproportionately prioritize customer retention over customer acquisition due to retention activities typically having lower costs and that many companies do not know the cost of customer acquisition at all. This is a problem further researched by Reinartz, Thomas, and V. Kumar (2005), who proposes a model for maximizing acquisition-retention efforts. Further, salespeople are often overwhelmed with disorganized information on prospective customers, which further makes the acquisition process difficult (D'Haen and Van den Poel 2013).

3.1.2 Understanding Customer Needs

The importance of understanding customer needs has already been well established within the research, and how companies can improve their business practices by listening to their customer needs (Griffin and Hauser 1993; Joshi and S. Sharma 2004). Further, understanding their customers can help companies with customer retention rates (Bolton

1998) as well as marketing initiatives (King and Burgess 2008). Neslin et al. (2006) argues the importance of Customer Data Integration and achieving a complete view of the customer by analyzing this data for improved customer acquisition, retention, and firm performance.

Further, understanding customers and their needs is a way for companies to better manage their relationships, which is vital for successful acquisition and retention rates (D'Haen and Van den Poel 2013). This is also supported by Gattermann-Itschert and Thonemann (2022), who found that knowledge-driven customer relationships positively impacted customer churn rates. An improved understanding of a company's customer base can also help create a better customer journey (Lemon and Verhoef 2016). A paper by Tomczyk, Doligalski, and Zaborek (2016) looking at the Polish insurance market found that customer analysis benefited firm performance through a deeper understanding of customer needs.

In a B2B context, understanding customer needs is crucial for businesses to develop solutions that cater to their requirements. In complex B2B markets, firms embed external products and services to serve customer needs better, leading to triadic relationships between suppliers, buyers, and customers. These relationships drive innovation and the development of new solutions. A study by Fletcher-Chen, A. Sharma, and Rangarajan (2022) focused on understanding the interaction and innovation processes in triadic relationships and identified four key outcomes: product enhancement, service refinements, e-resource integration, and complementary synergy. The study also found that deep triadic relationships lead to radical innovation and highlights the importance of conflict as an antecedent to cooperation and the development of these relationships. Therefore, understanding customer needs is essential for firms to develop and sustain triadic relationships, leading to innovation and superior solutions. (Fletcher-Chen, A. Sharma, and Rangarajan 2022)

3.1.3 Customer Classification

Customer classification and segmentation have been studied as a way for companies to gain insight into and understand their customer base in order to make more informed decisions regarding their customer base (Rahim et al. 2021) as well as customer churn prediction (Baghla and G. Gupta 2022). This section looks deeper into the applications of customer classification as a method for improved customer insight as well as its benefits

in marketing and customer acquisition.

Applications of Classification Algorithms

Classification algorithms are widely used in various fields to categorize data into different groups or classes. They are used in image recognition to classify images into different categories, such as objects and people (Krizhevsky, Sutskever, and Hinton 2017). In natural language processing (NLP), classification algorithms are used for text classification, including spam detection (Metsis, Androutsopoulos, and Paliouras 2006), sentiment analysis (Pang, L. Lee, and Vaithyanathan 2002), and topic modeling (Blei, Ng, and Jordan 2003). In medical diagnosis, classification algorithms are used to diagnose diseases or predict the pathology outcome based on medical data (Vanneschi et al. 2011). In finance and banking, classification algorithms are used for fraud detection (Bhattacharyya et al. 2011), credit scoring (Fitzpatrick and Mues 2016), and risk management (Galindo and Tamayo 2000). In manufacturing, classification algorithms are used for predictive maintenance to predict when a machine is likely to fail (Jardine, D. Lin, and Banjevic 2006). In marketing and social media analysis, classification algorithms are used for sentiment analysis to classify text into positive, negative, or neutral sentiments (B. Liu and B. Liu 2011). In recommendation systems, classification algorithms are used to predict user preferences and recommend products or services based on their past behavior and preferences (Fayyaz et al. 2020).

Overall, ML and classification algorithms have been widely exploited. In this thesis, the application area in focus is customer segmentation. This is also an important application area for classification algorithms as they allow businesses to segment customers into different groups based on their behavior, demographics, and preferences (Guo, F. Liu, and X. Zhang 2021). The current literature on the field of ML in customer classification is considered in the next section.

Customer Classification & Data Mining

Customer classification using data mining techniques has proven to be useful in a wide range of industries. Studies show that it can bring value to different businesses ranging from e-commerce to car leasing. A study by E. Kim, W. Kim, and Y. Lee (2003), using classifiers to predict e-commerce customers' purchase behavior to improve directed marketing, shows that combining classifiers gives the best performance. This study's customer data included ten demographic features and five transactional features. The

best combination algorithms were Majority vote (2% error rate) and Genetic Algorithm (GA) (2.3% error rate). Further, Sadikin and Alfiandi (2018) explored the data mining techniques DT C4.5 and Naive Bayes (NB) for classifying the potential risk of customer candidates for car leasing in order to decrease stalled credit payments. The customer attributes used in the study were: salary, age, marital status, other installments, and worthiness. The results showed that the C4.5 algorithm outperformed NB on accuracy and that salary was the most influential data point.

Customer classification has also been explored in the financial sector. A study by C. Lin and Zheng (2022) researched the classification of customers for financial products with the purpose of helping financial institutions with customer acquisition to increase sales. The models used were the Decision Tree (DT) C5.0 algorithm, the NB classification algorithm, the Binary Logit model, and five combination models. Results showed that the best single model was NB, with an accuracy of 89.68%, and the best combination model was the Arithmetic Average Weighted model, with an accuracy of 89.94%. Another study (Marinakos and Daskalaki 2017) classified banking customers that would be the most likely buyers given a term deposit marketing campaign, also considering the imbalanced classification problem. The purpose was to improve direct marketing and thereby save resources by finding the best methods for practitioners regarding profitability and effectiveness. Results showed that K-Nearest Neighbor trained on under-sampled data with the cluster-based technique was the most profitable. In contrast, LR and secondarily Linear Discriminant on SMOTE over-sampled data proved the most effective.

In the insurance industry, customer classification has also proven useful, for example, due to the value of being able to target the customers most likely to purchase insurance products. A paper by Chiu (2002) tested real cases by one worldwide insurance direct marketing company, Taiwan branch, using a GA system to predict customer purchasing behavior. Results show that mining customer purchasing data is relevant. Using the GA's rapid search strengths, the system can determine the characteristics of a customer most likely and unlikely to buy an insurance product. Another study by Müller and Te (2017) investigated how changes to the company's insured revenue can be estimated using a Random Forest (RF) classification. Focusing on small– to medium–sized companies, the purpose was to be able to selectively contact the most promising companies to do business with by integrating the model into an insurance CRM system and thereby offering traditional insurance companies a practical and cost-effective strategy to raise insurance premiums. The results show that a business customer's motor vehicle insurance

policy can be used to classify shrinking, stable, and growing businesses. Müller and Te (2017) points out that there is a research gap in the use of non-financial information to estimate small to medium businesses' revenue change. They also mention that there is lacking research on using company data to classify customers by their performance. Further suggestions from the study include using other classification algorithms, such as Support Vector Machines (SVMs).

Classifying Prospective Customers

Classifying customers for insurance products is relevant for several reasons. Insurance companies use customer classification to assess each customer's risk level to determine the appropriate premium to charge (Müller and Te 2017). Insurance companies can also use customer classification to customize their sales and marketing to specific customer segments to attract and retain customers (Chiu 2002).

Classifying Corporate Customers

There is a limited amount of research in the field of classifying corporate customers. Even though customer classification is exhaustively studied in general, this is mostly in a B2C context and not B2B. However, there are a few articles on classifying corporate customers. In one study (Makinde et al. 2020), customers of a building materials manufacturing company were classified using GA-based data mining to improve seller predictive measures. The algorithm was able to classify customers into two groups: *repeat customers* and *shop-and-go customers*. The results showed that the proposed GA model efficiently distributes resources to the most profitable customer group with better accuracy compared to conventional algorithms such as C5.0, K-means, and SVMs. Also, the model provided a CRM environment that could be easily used by multiple sellers to maximize business performance (Makinde et al. 2020).

Classifying B2B customers is a critical component of marketing strategy, as it enables businesses to develop targeted marketing efforts that cater to the specific needs of different customer segments. One approach to classifying B2B customers is through market segmentation, which involves dividing the market into distinct segments based on customer characteristics, needs, and behaviors. Market segmentation can be challenging in the B2B sector due to the diversity of customer needs and preferences. Simkin (2008) suggests that sectorisation, which involves dividing the market into different sectors based on industry or customer characteristics, can help businesses achieve effective market

segmentation. This approach can aid in identifying key customer needs, developing tailored products and services, and building long-term relationships with customers. While there are challenges to consider, sectorisation is a useful tool for classifying B2B customers and enhancing marketing efforts (Simkin 2008). A paper by D'Haen and Van den Poel (2013) attempted to ease the customer acquisition process for B2B sales representatives by proposing a model to categorize and classify prospective customers. Their model generated more refined leads; however, they could not test it in a real-life environment and thus suggested further research on the topic.

Corporate Customers in the Insurance Industry

Further, there is a gap in the literature when it comes to classifying corporate customers in the insurance industry. Although ML models have been applied and researched in the insurance domain, they have mostly been for individual customers, not corporate ones. Thus, the published work available in this field pertains to the utilization of ML for predictive modeling on a policy level for individuals and is often connected to life insurance products. However, some research has been done on other types of policies that can be applied to both individuals and corporations, such as property and casualty insurance. In their research, Blier-Wong et al. (2021) conducted a review of current literature exploring the use of ML models for rate-making and reserving in property and casualty insurance. They analyzed 77 publications from 2015 to August 2020, noting a rise in interest in the topic, especially after 2017. Their comprehensive overview found significant variations among insurance policyholders, and ML models can effectively capture these differences. As a result, these models can assist in calculating premiums that accurately reflect individual risk.

Considering business insurance, the most common risk insurance is general liability insurance, giving companies coverage for harm done to third parties as a result of their operations. These possible losses include bodily harm, property damage, finished goods produced by the policyholder, as well as personal injuries, such as slander or defamation (Henry 2016). The pricing of general liability insurance is predominantly determined by two factors. The initial component is a merit rating that considers the industry's risk average based on class rates, which is subsequently modified upward or downward according to the policyholder's individual loss experience (Müller and Te 2017). The second component involves the insurer's exposure to risk, expressed as a quantity roughly proportional to the risk posed by either an individual policyholder or a group of

policyholders (Müller and Te 2017). Thus, classifying business customers for liability insurance will likely also depend on factors such as industry and the policyholder's risk.

3.1.4 CRM Systems

Another field aimed at increased understanding of a company's customer base, as well as the importance of firm-wide data integration for managing customers, is Customer Relationship Management (CRM). I. J. Chen and Popovich (2003) describes CRM as an integrated way of managing customer relationships in order to improve customer relationships and customer retention. Further, Tohidi and Jabbari (2012) motivates the necessity of using CRM systems for managing customer relations as the complexity of the organization as well as the processes of modern companies increase with the evolution of technology. Among many things, CRM is useful in marketing and customer acquisition, where it can help identify and target the best customers based on customer data (Xu et al. 2002). Ahearne, Hughes, and Schillewaert (2007) also found that IT-driven CRM solutions positively impacted the performance of sales staff.

CRM Systems & Big Data

The field of BD and its use in business have also been studied in regard to CRM (Zerbino et al. 2018). As CRM systems are already data-driven, the application of BD within CRM seems like a natural progression. However, many companies struggle to develop analytical capabilities for integrating the information from these data sources into their CRM decisions (Phillips-Wren and Hoskisson 2015). Anshari et al. (2019) further describes the new wave of BD into CRM systems and finds that it further assists in customer profiling, allowing for more precise and aggressive forms of marketing. Previous to the evolution of BD, CRM has always been data-focused, as described by Ngai, Xiu, and Chau (2009) in a literature review reviewing 87 articles in regard to CRM and Data Mining. They found that customer retention and one-to-one marketing programs were the most common application of CRM and data mining techniques.

CRM & AI

AI is another emerging technology that has found its use within CRM. Chatterjee, Rana, et al. (2021) found that organizations successfully implementing AI-CRM considerably improved their B2B engagement process. These findings are supported by Bag et

al. (2021), which similarly found that BD and AI had a positive impact on B2B marketing decisions. V. Kumar et al. (2019) also found that AI would be useful in direct, personalized marketing due to its ability to process individual customer information. Further, Paschen, J. Kietzmann, and T. C. Kietzmann (2019) found that many managers were eager to implement AI in their B2B marketing decisions. However, they were unclear about how the systems worked and, thus, how to implement them properly. In another article Libai et al. (2020) examine the impact of AI on CRM, specifically focusing on customer acquisition, development, and retention. They identify the capabilities of AI that will transform traditional CRM into AI-CRM, such as the ability to predict CLV. The authors argue that this increased predictive power will lead to greater customer prioritization and potentially discriminatory practices in the market. The article also highlights the challenges regulators may face in response to these developments. Chatterjee, Chaudhuri, and Vrontis (2022) have explored the B2B-CRM-AI context further, finding that AI-CRM had a beneficial impact on B2B relationship management and firm performance.

Designing CRM Systems

Xu et al. (2002) writes that CRM will improve many company processes, amongst other marketing, customer loyalty, and the efficiency of internal processes, when implemented correctly. However, not being implemented correctly will instead result in a cost for the company without the benefits. King and Burgess (2008) similarly finds that companies struggle to implement CRM systems and suggests a framework to make CRM more easily adoptable. A few critical success factors for the successful implementation of CRM were researched by Croteau and Li (2003), who found that CRM initiatives were more likely to be successful in organizations with adequate top management support and accurate knowledge management capabilities. Wilson, Daniel, and McDonald (2002) had similar findings but added that commitment is needed across numerous functions. Chatterjee, Chaudhuri, and Vrontis (2022) have studied the implementation of AI-CRM in B2B firms and also found that individual skills and capabilities of firm employees had an impact on the success of the implementation. Rahman et al. (2023) had similar findings, where they concluded that a firm's technology readiness had a positive relationship with AI-CRM capability.

CRM integrated with Customer Classification

Integrating customer classification with CRM systems can be a way to utilize the benefits of both customer classification and CRM systems together. As CRM is partly based on integrating customer data into company processes, it can improve customer segmentation and classification by utilizing the available data. This allows for more effective targeting of specific groups and can result in increased customer engagement and loyalty, as well as optimization of sales and marketing efforts. A study by Alsaç, Çolak, and Keskin (2017) applied classification algorithms to a Turkish company in the health sector, classifying customers by product groups and risk factors with the algorithms NB, DT, and K-Nearest Neighbour. The results could be used to improve the marketing strategy and develop a more effective sales campaign for the company.

Within the field of insurance, Müller and Te (2017) found that the integration of their RF model into an insurance CRM system could be economically beneficial for insurance companies to increase insurance premiums in an efficient way by only contacting the most promising customers. Further, a paper by Y. Chen and Hu (2005) used data mining technology in CRM systems with the aim of proposing the data mining model for customer value and customer classification. In the classification of insurance customers, four different dimensions were suggested: Vital statistics (age, gender, education, career, etc.), Customer value (high, low), Customer credit (income, career, payment history, etc.), and Customer satisfaction (expectations, perceptions).

3.1.5 CRM in the B2B Context

Traditional B2B Relationship Management

B2B relationships slightly differ from their consumer counterparts and has thus been researched as a separate field. Webster (1992) defined different stages of B2B business relationships, emphasizing the role of trust in building successful B2B relations. These findings are supported by Ganesan (1994), who further finds that dependence plays a role in B2B relationships. As the market becomes increasingly competitive, the dependence of the customer on the vendor decreases.

B2B Marketing & AI

Looking specifically to the B2B sector, AI has been frequently used in B2B marketing both with and without the use of CRM systems. Mikalef, Conboy, and Krogstie (2021)

found that the use of AI can enable an organization's dynamic capabilities for B2B marketing. For example, the use of AI led to an improvement in insight, as well as the development of new marketing approaches. The study also found, however, that some of the benefits of using AI in B2B marketing were stunted by privacy issues. The use of complex algorithms and the consequent data collection made some of the larger, most important customers uneasy, resulting in a distrust of the process. A literature review by J. R. Saura, Ribeiro-Soriano, and Palacios-Marqués (2021) found that ML and data mining were some of the most common CRM techniques for AI within B2B strategies. They found numerous benefits, among others, an improvement in sales and sales predictions. They further conclude that while AI is a popular tool in B2B relationship management, knowledge about the technical processes is necessary for a successful implementation. The authors also recommend further research into possible future uses of AI-Based CRM in B2B, such as the establishment of experiments and tests using AI to improve processes (J. R. Saura, Ribeiro-Soriano, and Palacios-Marqués 2021). Further, a paper by D'Haen and Van den Poel (2013) describes the problem that sales staff often face in B2B marketing of being overwhelmed with potential customers, and proposes a combined model featuring DTs and Neural Networks to help with the acquisition process. They recommend further research with similar types of models.

3.2 Recommendation Systems

3.2.1 Benefits of Recommendation Systems

Recommendation systems have become increasingly important in recent years, particularly in the e-commerce and entertainment industries. Personalized recommendation systems have been found to increase revenue and sales volume for online retailers (Behera et al. 2020). With the continued growth of online retailing, businesses face the problem of increasingly disloyal customers. As online shopping brings lower switching costs and less personalization, consumers become less loyal to a business which is detrimental to customer retention (Hallikainen et al. 2022). Recommendation systems such as recommendation agents (RA), are a way to combat this by increasing personalization in the online shopping experience (Hallikainen et al. 2022). Moreover, a study by Komiak and Benbasat (2006) found that RAs could increase trust, which was affected by the RA's perceived level of personalization. A more personalized experience led to a greater increase in emotional and cognitive trust. However, another type of

recommendation system that differs from RAs - dynamic pricing, where consumers are given personalized prices, was found instead to have a negative effect on trust (Komiak and Benbasat 2006). RAs have also been found to be useful tools assisting online shoppers in making decisions, allowing customers to make better decisions in less time (Häubl and Trifts 2000). Guha et al. (2021) also predicts the use of AI in online retailing, stating that it will improve customer service, personalization, as well as high-value product recommendations and customer data analysis.

3.2.2 Digitalization of Insurance Advisors

In a paper analyzing the digitalization of the insurance industry, including advisory services, the authors found that the market has seen a more rapid evolution of digital services since the Covid-19 pandemic. They note that "Over 80% of insurers believe that the future of the insurance market belongs to those organizations that will make significant investments in the area of innovation and digitization" (Pauch and Bera 2022, p. 1679). However, they also found that the digitalization of the insurance sector is in its initial stages. Moreover, they note that AI and BD are some of the solutions being tested within the sector and that the development of these technologies and digitalization will be necessary for companies to stay competitive. These services were, amongst others, useful in taking over market shares from other companies. (Pauch and Bera 2022).

Another paper by Pisoni and Díaz-Rodríguez (2023) examined the use of AI in building digital advisors for insurance and found that given the technological expertise, digital advisors were a powerful tool that could help insurance companies. Saxena and R. Kumar (2022) further found that customers nowadays are so used to digital services that they no longer feel the need to communicate with a person when making a purchase, and the insurance industry will likely be more and more affected by this in the future. The development of AI advisors will allow customers to make their own insurance decisions and purchases online. They further state that financial technology startups within the insurance sector, the so-called "insurtechs", may disrupt the entire industry. The startups are seeing a surge in investments rivaling that of other fintech companies. The insurtechs are able to improve their services by the use of AI and ML and are adapted to customers' current demands for the digitalization of services. The authors recommend insurance companies invest in AI and other types of digital technologies for the future.

3.2.3 Recommendation Systems in The Insurance Industry

In a paper by Wong et al. (2020), it is highlighted that the insurance sector has historically used human agents to analyze data and provide product recommendations, as discussed in the previous section. However, in recent years technology has been leveraged to make insurance processes easier, leading to decreased costs and more satisfied customers. (Wong et al. 2020) In their paper, Wong et al. (2020) researched product recommendation systems in partnership with a traditional life insurance company with satisfying results. The customers could be segmented into different customer classes, which highly depended on education level, occupation, age, and buying patterns. Collaborative Filtering (CF) algorithms were used, which is a class of behavior-based approaches that use previous purchase behavior to make product recommendations. In their discussion, Wong et al. (2020) highlights that the weakness of the CF algorithm is that it cannot make product recommendations for new customers due to the usage of previous customer purchasing data. This is also something that is suggested for future research.

In another study (Qazi et al. 2017) applied Bayesian Networks to deploy an insurance recommendation system for both new and existing customers. The use of Bayesian Networks was due to the small number of products to recommend and the frequency of missing data. The data used came from four insurance policies: auto, property, umbrella, and life. The purpose of the recommendation system was to give insurance agents a tool for better product recommendations making customers adequately covered for their needs. The research was tested in an action research format by agents who gave feedback, such as it helping them increase premiums, giving better attention to customer detail, and it being a valuable teaching tool. (Qazi et al. 2017) Even though the choice of Bayesian Networks was to be able to make recommendations to new customers, this was not adopted and instead given as a suggestion for future research.

From the previous research on product recommendations, there are several examples of recommending life insurance products. However, there is a limited amount of how to make recommendations to businesses instead of individuals, which is something aimed to investigate in this paper. Also, several papers suggest recommending products to new customers (Wong et al. 2020; Qazi et al. 2017), which also will be explored in this study.

Machine Learning in Classification Systems

ML has found many uses within customer classification and segmentation. In a paper by Abedin et al. (2023), an RF classifier was used to predict the behavior of bank customers. The paper also used a method for feature selection and achieved an accuracy of 75.85% for the tested data set. Another paper by Amin et al. (2019) classifying customer churn prediction used a NB classifier. In an attempt to classify insurance fraud prediction, Aslam et al. (2022) used three different classification algorithms - Logistic Regression (LR), SVM, and Naïve Bayes. The LR classifier yielded the highest f-score.

A study by Rusli, Zulkifle, and Ramli (2023) compared ML classification models for analyzing customer behavior using a data set from an in-vehicle coupon recommendation system. The data set included demographic and environmental factors, such as driving destinations, age, current time, and weather. The six models evaluated were Bayesian Network, Naïve Bayes, Instance-Based Learning with Parameter-K (Lazy-IBK), Tree J48, RF, and RandomTree. The study assessed the model performance based on accuracy, precision, processing time, recall, and F-measure. The findings revealed that Naïve Bayes and Lazy-IBK consumed the least amount of prediction time, although with the lowest accuracy. RandomTree had the highest processing time, whereas RF provided the highest accuracy, precision, recall, and F-measure values.

Yet another paper (Jain, Yadav, and Manoov 2021) classifying customer churn tested four different classifiers, LR, RF, SVM, and XGBoost. The authors found that for the three different sectors studied, banking, telecom, and IT, three separate classifiers had the highest accuracy for their respective domains. Depending on the data set and sector, different classification models perform differently. This is further supported by an empirical comparison of different training algorithms by Caruana and Niculescu-Mizil (2006, pp. 167–168), who found RF to perform well together with neural nets. However, the authors note, ”Even the best models sometimes perform poorly, and models with poor average performance occasionally perform exceptionally well.”

3.2.4 Data Used for Creating Recommendation Systems

In the case study of this paper, combined business insurance will be the recommended product. This product includes several types of policies, such as liability insurance, property insurance, interruption insurance, and others. Limited research exists into what data parameters previously have been used to create recommendation systems for similar

insurances. However, a paper by (Qazi et al. 2017) looked at the parameters used to determine different types of insurance, among therein property insurance. The data to determine property insurance included information about the insurance, the coverage, the options and endorsements, the characteristics of the home and household, customer demographics, sewer and wildfire risk, and geological information.

3.3 AI

3.3.1 Leveraging AI for Business

Several research papers have explored the implications of leveraging AI in the workplace, providing valuable insights into its potential benefits and challenges. One such study by Shollo et al. (2022) emphasizes the importance of creating tangible business value through AI projects. Organizations can harness ML algorithms to enhance their operations, improve decision-making processes, and optimize overall performance by understanding different ML value-creation mechanisms and the necessary conditions for successful implementation.

Asatiani et al. (2021) introduce the concept of sociotechnical envelopment as a means to implement inscrutable AI models, such as neural networks, in a manner that ensures accountability and safety. The authors highlight the significance of striking a balance between the performance benefits of flexible AI models and the risks associated with their lack of explainability. This is achieved through establishing clear boundaries, meticulous training data curation, and effective input and output sources management. Sociotechnical envelopment offers organizations a framework to navigate the complexities of implementing inscrutable AI models while maintaining control and mitigating potential risks.

Additionally, Waardenburg, Huysman, and Sergeeva (2022) presents a study that delves into the translation of opaque algorithmic predictions by knowledge brokers. The research sheds light on brokers' work's dynamic and influential nature in interpreting and curating algorithmic outcomes. The authors highlight the importance of understanding the role of knowledge brokers in implementing AI/ML algorithms. Brokers act as intermediaries, bridging the gap between technical experts and end-users, and play a pivotal role in navigating complex scenarios that require human judgment and decision-making. Recognizing the challenges and opportunities associated with algorithmic brokerage

provides insights into the implementation process and underscores the significance of human expertise in conjunction with AI systems.

Chapter 4

Methodology

This chapter explains the methodology used in the case study. It first presents the research setting, followed by the research design and process. Finally, the chapter delves into the quality of the research

4.1 Research Setting

The research setting involved a partnership between the authors and a Swedish insurance broker in conducting a case study on how ML-based customer classification can be applied in the business insurance industry to acquire new customers (RQ1). The classification implied determining which customers were suitable or not for a type of combined business insurance product by classifying them into two groups, yes or no, using quantitative methods. Simultaneously, the study also investigated the benefits of integrating ML classification into the partner company's CRM system (RQ2). The purpose was to investigate the current customer acquisition process and CRM system using qualitative methods, allowing the classification system to be integrated into the existing CRM and customer acquisition processes.

The authors worked closely with the partner company to collect data on potential customers. ML algorithms were used to classify the customer base into two different segments based on their characteristics to improve the customer acquisition process's efficiency and increase the insurance company's proposed customer acceptance ratio. The study investigated the potential establishment of the technology-driven customer classification process by integrating it into the CRM system. The original customer

acquisition process involved manual customer selection by individual brokers and followed a non-standardized process. The product featured as the subject in this study was a combined company insurance tailored for small to medium-sized businesses that included several different types of insurance, such as property or liability insurance coverage.

4.2 Research Design

The study's objective was to investigate the possibility of a digital solution for customer segmentation at the partner company and also to study the actual implementation of the solution and how it fits in with the company's existing processes and capabilities. Therefore, the study was designed to monitor the digital solution's implementation as it developed continuously. This was possible as the researchers implemented the classification system themselves and could thus adapt it to the needs and conditions present at the partner company. As the study was dual-focused, it was necessary to address both RQs. Consequently, the research methods were selected to ensure that both research questions were covered, which meant a mix of qualitative and quantitative methods.

An abductive research approach was applied due to the lack of research into the specific area. An abductive technique creates a hypothesis for a phenomenon based on sparse or ambiguous data, which is then tested and improved through empirical observation and analysis (Saunders, Lewis, and Thornhill 2015). The authors began by formulating a theory grounded in closely related prior research found during the initial literature study regarding CRM systems, specifically analytical CRM used in customer acquisition theory. The theory then continued to develop as the research project progressed, following the abductive research process described by Saunders, Lewis, and Thornhill (2015).

A concurrent mixed methods design was used to answer both research questions, meaning quantitative data methods were used in parallel with qualitative methods, as described by Saunders, Lewis, and Thornhill (2015). As mentioned, the study both focused on determining the effectiveness of the use of an ML classifier in the customer acquisition process (RQ1) as well as determining how this classifier fits into and develops the company's current CRM system (RQ2), and as such a mixed method was used as quantitative methods were used for RQ1 and qualitative methods was used to answer RQ2. The quantitative method for answering RQ1 was developed following the first steps of the CRM framework. The qualitative method for answering RQ2 was developed

following Saunders, Lewis, and Thornhill (2015), who recommends interviews as a suitable exploratory qualitative method.

A case study research strategy was followed, where the partner company was used as a single case, as the goal was to implement and study the potential for a technological solution in the specific setting of the partner company. As mentioned by Saunders, Lewis, and Thornhill (2015), a single case study strategy is suitable when exploring a phenomenon in a real-life setting at a specific company.

4.3 Research Process

In this section, the process of the research is presented. The process follows the theoretical framework presented in Chapter 2, with a slight modification as some steps were excluded due to not being suitable in this particular case. The exclusion criteria were based on information gathered from the pre-study interviews charting the current CRM system, and the steps were not included as they did not fit into the part of the current system where the classifier would be implemented. The excluded steps were *Customer Targeting*, *Relationship Marketing*, *Privacy Issues*, and *Metrics*.

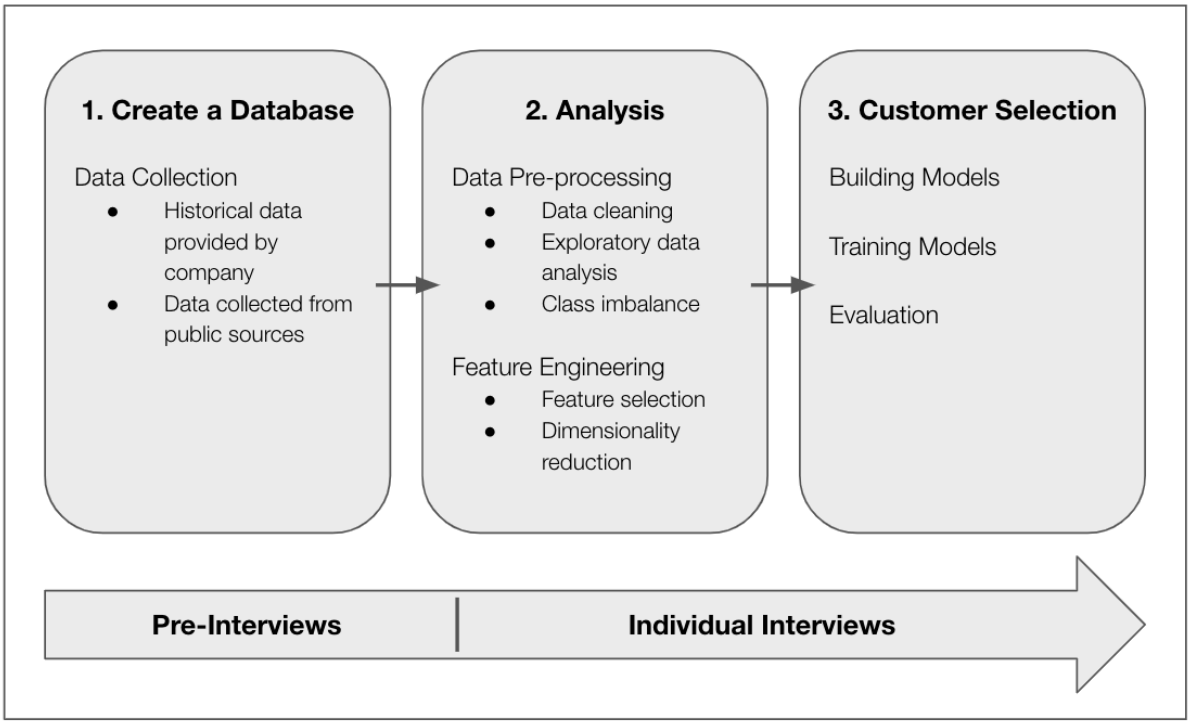


Figure 4.3.1: Research Process Overview (Based on the CRM Framework)

4.3.1 Create a Database

The data utilized in this report consist of a combination of primary and secondary sources. These were the two types of data used in the research project, where primary data were collected directly from the source, and secondary data were collected by someone else and was, therefore, already available (Saunders, Lewis, and Thornhill 2015). The primary data were provided by the company itself, while the secondary data were obtained from publicly available sources. The primary and secondary data were used to provide a comprehensive analysis of the prospective customers' financial performance, providing insights from the primary data and a broader perspective from the secondary data.

Primary Data

The primary data were a list of the customers previously involved in the customer acquisition process. The variables used from this data set were company identification numbers and a variable on whether or not they had received an offer from the insurance company. The company ID variable was only used to find and collect other variables, and the variable representing whether or not a customer had received an offer was used as the dependent variable in the study.

Secondary Data

The primary data set was subsequently completed by collecting data from a public data source hosting company information. The data were collected for each company listed in the primary data set, and due to it not being processed, it was considered *raw data* (Saunders, Lewis, and Thornhill 2015). Secondary data can be numeric and non-numeric (Saunders, Lewis, and Thornhill 2015); both forms were included in this study. The purpose of the data set extension was to include additional information. Since no studies were found on what variables were useful when classifying companies for corporate insurance, all available variables were collected. Later, different feature selection methods were used to find which variables had the most impact on the outcome of the dependent variable. The data collected was from 2021. A comparison was made between 2020-2022, and 2021 was the year with the least missing data points.

Variables

The variables collected were company information and financial metrics. Company information variables were, for example, company registration year, industry code, and number of employees. The financial variables were, for example, assets, equity, liabilities, and operating profit. The data set consisted of 32 variables; the full list of variables can be found in Appendix A.

4.3.2 Analysis

According to Saunders, Lewis, and Thornhill (2015), for a quantitative study to be straightforward and valuable, it is necessary to have prepared the data accordingly and know when to use specific graphic and statistical techniques. The data pre-processing steps included data cleaning, exploratory data analysis, feature engineering, and feature selection.

Data Cleaning

The initial sample size of the data set provided by the partner company was 6212 data points, where each data point corresponded to one company. The data set contained a few NaN (Not a Number) values, whereas the publicly available data had missing data points. This was handled by dropping all rows containing NaN values. As a consequence, the cleaned data set had varying amounts of data points depending on the combination of features. Feature combinations 1, 2, 3 & 4 had 4305, 4305, 5738 & 3873 data points, respectively. The data collected were from 2021.

Exploratory Data Analysis

In the initial quantitative analysis phase, exploratory data analysis was conducted. Exploring and visualizing the data is crucial before diving into statistical modeling or drawing conclusions because it helps to understand the underlying relationships within the data. It also enables identifying any outliers, which could significantly impact the accuracy of the results. (Saunders, Lewis, and Thornhill 2015). Some of the exploratory analysis can be found in Appendix B.

Further, since this case study partly aimed to explore which variables are important when classifying corporate customers, something that has not been researched to the knowledge of this study, exploring the data is highly necessary. Conducting Exploratory

Data Analysis provides the opportunity to adapt to new discoveries within the data and introduce analysis that was not initially planned. (Saunders, Lewis, and Thornhill 2015)

Feature Engineering

Feature engineering involves selecting, extracting, and transforming features from raw data to create a set of features suitable for ML algorithms (Verdonck et al. 2021).

Categorical & Numerical variables

The data were separated into categorical and numerical variables. Categorical variables are those that represent a certain category, such as company code and year of registration. Numerical variables are those that represent a number value and can be used in computations, such as working capital turnover and number of employees.

One-hot encoding

Some ML models, such as the SVM and ANN used in this paper, can only handle numerical variables (James et al. 2021). To solve this problem, one-hot encoding was used. This involves converting the categorical variable into numerical data by creating dummy features for each value of the categorical variable and placing either 1 or 0 in that column depending on whether or not that value is present in the row (James et al. 2021).

Scaling

It is common practice to scale numerical variables for ML problems (James et al. 2021). Two scaling methods were tested and compared — MinMax and Robust scaler. The two scaling methods were chosen due to handling outliers differently. MinMax scaling is a commonly used normalization technique that rescales data from zero to one and is sensitive to outliers. MinMax scaling scales the i-th value in the data set, x_i, as:

$$x_i^{'} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{4.1}$$

(Scikit-learn Contributors 2021a). On the other hand, Robust scaling uses the median and interquartile range to rescale data and is more robust to outliers (James et al. 2021). Robust scaling scales the i-th value in the data set, x_i, as:

$$x_i^{'} = \frac{x_i - \text{median}(x)}{\text{IQR}(x)} \tag{4.2}$$

where median(x) is the median value in the data set, and IQR(x) is the interquartile range, which is the difference between the 75th and 25th percentiles of the data (Scikit-learn Contributors 2021b). By applying both MinMax scaling and Robust scaling, the study aimed to compare the effects of these two normalization techniques on the performance of subsequent analyses.

Generalizing data

To prevent overfitting, one categorical variable was modified to be less specific. The variable was a code representing an industry category. The variable had five digits, representing a more specific subgroup, but it could also be altered to represent a more general main group. The last three digits were thus dropped, converting the variable into its less specific version. To prevent overfitting, rows with very few values of this column were also dropped. Values with a frequency of 3 or less were dropped, which made up roughly 0.5% of the data set. If a variable has too many classes, it may result in overfitting, as the model might memorize the individual values of the variable instead of understanding the fundamental patterns and relationships present in the data. Overfitting occurs when the model becomes excessively complex and over-adapts to the training data, causing inadequate performance when exposed to new, unseen data (Géron 2022). The generalization was also done in order to mitigate bias that might be present in the data set due to past focus on specific industries from the current customer acquisition process.

Balancing The Data Set

The data set contained roughly 25% of one class and 75% of the other, as seen in Figure 4.3.2, resulting in an imbalanced data set. This can create a bias towards the majority class (James et al. 2021), and to combat this, the training set was balanced. The method used to balance the set was Random Oversampling, which randomly duplicates samples from the minority class until the data set is balanced. Synthetic Minority Oversampling Technique (SMOTE) was also implemented and tested. However, it was discarded as a method due to the risk of SMOTE possibly generating samples for the minority class, which would have actually been a sample belonging to the majority class.

Distribution of Data

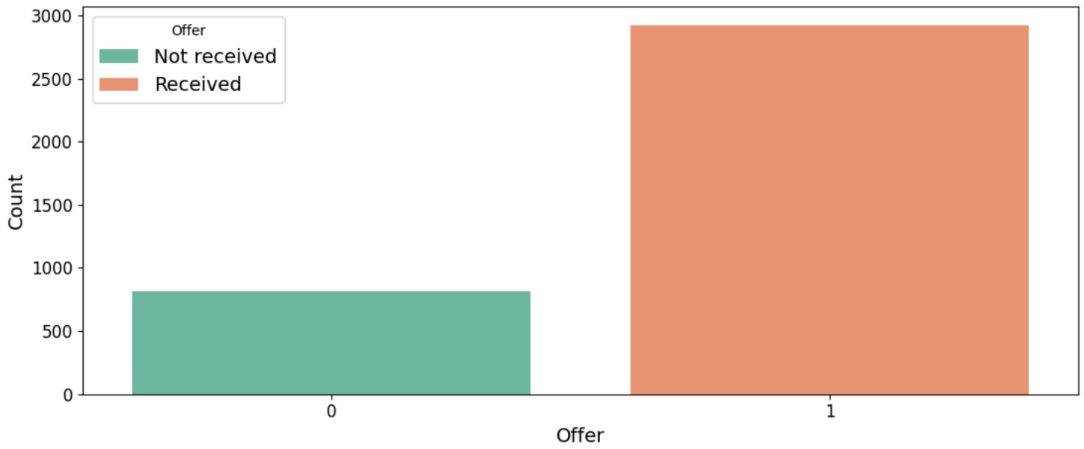


Figure 4.3.2: Visualizing Class Imbalance

The numerical data were tested for normality using the Shapiro-Wilk test, with alpha = 0.1. The results showed that none of the collected data were normally distributed. Attempts were made to normalize data using the Box-Cox transformation, which is often suggested as a remedy for asymmetry and non-linearity (LaLonde 2005). However, normality was not achieved. Transforming the data is not always a good practice as it can make the analysis more complicated, the results more difficult to interpret, and it can reduce the complexity of the model (LaLonde 2005). The data complexity, skewness, and unaccounted factors could have contributed to the difficulty of transformation (LaLonde 2005). Alternative feature selection and classification methods, suitable for non-normal data, were used instead. These methods do not require normality assumptions and can handle non-normal data. Some variable distributions can be found in Appendix B.

Feature Selection

Feature selection is critical in ML, especially in data sets with many variables. Including irrelevant variables in a model can lead to unnecessary complexity (James et al. 2021), and eliminating irrelevant features will improve the accuracy and performance of the model (R.-C. Chen et al. 2020). Therefore, selecting the most important features for the particular problem is vital. Feature selection aims to identify a given task's most relevant and informative features. There is furthermore a phenomenon called *The Curse of Dimensionality*, which describes the relationship between the number of features and the number of data points needed, meaning that the number of training samples for the model to be accurate increases with the number of used variables (Shmilovici 2005). For linear classifiers, such as the SVM used in this paper, the required number of training samples is linearly linked to the number of features. However, for the DT model applied,

the problem is even greater. It is estimated that the training data size needs to increase exponentially with the number of features. (Shmilovici 2005). Smaller models with fewer features are also easier to understand and explain (Shmilovici 2005).

Removing Multicollinearity

Firstly, the Variance inflation factor (VIF) was applied to the full feature set to check for and remove multicollinearity. Multicollinearity is a statistical phenomenon where two or more independent variables within a data set are highly correlated (Shrestha 2020). The correlation makes it difficult to estimate how each independent variable uniquely affects the dependent variable, and unresolved multicollinearity could lead to misleading interpretations of the results (Shrestha 2020). In brief, it becomes difficult to determine which independent variable is causing the observed relationship with the dependent variable. To address this issue, VIF can be used (Shrestha 2020). VIF is a statistical measure that identifies the degree of multicollinearity among the independent variables in a data set (James et al. 2021). It measures the inflation of the variance of the estimated regression coefficients due to the presence of multicollinearity. VIF is calculated for each predictor variable i as

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{4.3}$$

where R_i^2 is the coefficient of determination (R^2) for a regression model with the i-th predictor variable as the response and all other predictor variables as predictors. In this project, VIF was used to identify and remove the features with a high degree of multicollinearity, potentially leading to inaccurate results as well as variables containing redundant information. A VIF value of one indicates no correlation, and values 1–5 typically indicate the safe zone with a low correlation. However, a VIF value of 5–10 indicates a higher correlation and values ten and above an even greater degree of correlation (Shrestha 2020). VIF was applied using the stats models library in Python, and variables were subsequently dropped. VIF was reapplied until the remaining variables had a VIF value of five or lower.

Feature Selection Trade-off

As explained by O'brien (2007), removing collinearity through feature dropping via VIF has its drawbacks. Through the removal of features based on VIF, we could end up

dropping variables that were, in fact, important to the model. As the VIF threshold set in this paper does not measure perfect correlation, some information is lost (O'brien 2007). However, keeping variables with a high VIF value can mislead the results of other kinds of feature selection methods, such as the mutual class info (MCI) method and RF model applied here (Shrestha 2020).

There is thus a trade-off between removing variables with a high VIF score and then selecting variables through the feature selection methods mentioned or applying the feature selection methods without accounting for multicollinearity and skipping VIF. As the goal was to curate a very low number of features that all had high importance, the authors decided to use VIF and then feature selection methods. The risk of losing information via VIF was deemed lesser than the risk of recommending data collection for features that had a problem of correlation and thus contained much of the same information.

Feature Importance

Next, RF and MCI were employed to refine the feature selection process further to select the final features for the models. The purpose was to isolate the independent variables with the most effect on the dependent variable's outcome and exclude features containing redundant information. RF is an ensemble learning technique that utilizes multiple DTs to classify data (Breiman 2001). It can also be used to measure the importance of each feature's contribution to the data set (R.-C. Chen et al. 2020). The purpose of the RF model was further to refine the variable selection after using VIF. This RF model is not to be confused with the one used in the classification problem, as this is a separate model used solely to determine feature importance.

The feature importance score in an RF model represents each feature's relative contribution to the model's overall predictive performance. It can be calculated as the average reduction in impurity (e.g., Gini impurity, such as in this case) achieved by splitting a given feature across all DTs in the RF. The importance score can be obtained using the RF classifier as follows:

$$I(X_i) = \frac{1}{N} \sum_{t=1}^{N} \left(\frac{n_t}{n} \cdot \text{impurity}_t - \text{impurity}_{\text{left}_t} - \text{impurity}_{\text{right}_t} \right) \tag{4.4}$$

where:

- N is the total number of decision trees in the random forest,
- n_t is the number of samples in the training set that reaches node t,
- n is the total number of samples in the training set,
- impurity_t is the impurity of node t,
- $\text{impurity}_{\text{left}_t}$ is the impurity of the left child node of node t,
- $\text{impurity}_{\text{right}_t}$ is the impurity of the right child node of node t.

(Scikit-learn Contributors 2023)

MCI measures the mutual information between two random variables, which can be used to select informative features in data sets that are not normally distributed (Zong, Xia, and J. Zhang 2021). In this project, MCI was used to refine the feature selection process further and select the most informative features for the model. If X is the feature matrix and y is the target vector, the MCI score between a feature X_i and the target variable y can be computed using the method as follows (Thomas M. Cover 2005):

$$I(X_i, y) = \sum_{x_i \in X_i} \sum_{y_j \in y} P(x_i, y_j) \log \left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right) \tag{4.5}$$

where:

- x_i represents the possible values of the feature X_i,
- y_j represents the possible values of the target variable y,
- $P(x_i, y_j)$ is the joint probability of X_i and y_j,
- $P(x_i)$ is the marginal probability of X_i,
- $P(y_j)$ is the marginal probability of y_j.

Feature Combinations

In order to gain further information about the best feature combinations, the different combinations were also tested recursively for each model. This was done by testing each possible combination of features for each model and then observing which combination produced the best results. The method for evaluating feature combinations was accuracy. The tested features were the six remaining features after VIF had been applied. These variables were *Cash Liquidity*, *Number of Employees*, *Working Capital Turnover*,

Personnel Costs per Employee, *Net Turnover per Employee*, and *Net Turnover Change*. The reason for only testing the combinations with the remaining variables after VIF is that it would take too much time and computational power to test each combination of the numerical features. For example, testing each combination of 12 different features for one model would take many years with a typical processor.

In order to identify and select the best feature combinations, the features were selected through all of the explained methods. Firstly, VIF removed collinearity and allowed for testing of the remaining variables. Then, the features were tested through RF, MCI, and by recursively testing combinations via the models. These three techniques were then compared, and the features that showed the best overall results through all methods were chosen. The complete list of chosen features is available under Chapter 5.

4.3.3 Customer Selection

The literature review results showed that different ML models perform better or worse depending on the problem and data set. Notably, Jain, Yadav, and Manoov (2021) found that different models performed best depending on which sector they were applied to (banking, telecom, and IT), and a study examining different classifiers by Caruana and Niculescu-Mizil (2006, pp. 167–168) conclude that "Even the best models sometimes perform poorly, and models with poor average performance occasionally perform exceptionally well".

Based on the literature review, four models were carefully selected for implementation and testing. The decision to choose these models was based on their demonstrated success in similar classification problems. The three models, DT, RF, and SVM, were applied in previous studies mentioned in the literature review (Alsaç, Çolak, and Keskin 2017; C. Lin and Zheng 2022; Aslam et al. 2022; Jain, Yadav, and Manoov 2021). The repeated successful application of these models in similar contexts provided strong evidence for their effectiveness in handling classification problems. In addition to these established models, an ANN model was also selected for testing. This was mainly as the literature review showed promise in the implementation of ANN for AI-CRM (Bag et al. 2021) and for it being a radically different model compared to the SVM and RF (Shmilovici 2005).

One important criterion for selecting these models was their suitability for handling non-normal data. Neither model assumes normal data. For example, SVM focuses primarily

on the boundaries of the separating hyperplane and does not rely on assuming the exact shape of the data distributions (James et al. 2021). Since the data being analyzed deviated from a normal distribution, choosing models that could effectively handle such non-normality was crucial. Consequently, all four models were specifically chosen for their applicability to non-normal data, ensuring that the analysis would be accurate in the given context.

Decision Trees

The study employed a DT as the first model, which is a predictive model that utilizes a tree-like structure to make predictions. A classification tree splits the predictor space into a number of simple regions. Then it makes predictions of the observation's classes by assigning them to the most frequently occurring class among the training observations in the region to which it belongs. The tree uses a binary splitting principle and recursively partitions the data into smaller subsets based on the values of input features until a final prediction is made for each subset. The model structure comprises nodes, branches, and leaves, where each node represents a decision based on the value of an input feature, and each branch represents the possible outcomes of that decision. The tree leaves represent the final predictions of the model (James et al. 2021).

As a splitting criterion could be the classification error rate, where observations are assigned to the most common error rate class in their region and use the fraction of training observations not belonging to this class as the error measure. However, in real cases, criteria such as the Gini index or Entropy are preferred. Both these measures are used to evaluate the quality of a specific split. (James et al. 2021). In this study, the Gini index is which is calculated as:

$$G = \sum_{i=1}^{c} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.6}$$

where c is the number of classes, m is the number of regions or leaves in the DT, and $\hat{p}_{mk}$ is the proportion of training observations in the m-th region that belong to the k-th class (James et al. 2021). DTs are useful in ML due to their intuitive structure and simplicity. However, they do not tend to outperform more complex models such as SVMs. (James et al. 2021).

Random Forest

The second model used is Random forest (RF), an ML technique that uses an ensemble of DTs to make predictions. The basic idea behind RF is to create multiple DTs, each trained on a different subset of the data and a different subset of the input features. Then, they combine their predictions to improve the overall accuracy and reduce over-fitting. (Breiman 2001).

To achieve high classification accuracy, growing an ensemble of trees and letting them choose the most favored class is a common approach. The growth of each tree in the ensemble is controlled by generating random vectors. Bagging and random split selection are some of the methods used to create these vectors, where each tree is developed using random instances or features. The final result is an RF created by generating many trees and letting them vote for the most popular class. The crucial aspect of these techniques is the generation of a random vector for each tree, which is used to build the tree using the training set and the random vector. (Breiman 2001)

RF models have several advantages over individual DTs. First, they are more robust to noise and over-fitting since the individual trees are trained on different subsets of the data and input features. Second, they are more accurate since combining multiple trees reduces the variance and bias of the model. (Ali et al. 2012). However, the drawbacks of RF models are that they are computationally expensive and require a lot of training time since it builds numerous trees to combine their outputs. (Breiman 2001)

Support Vector Machine

Support Vector Machines (SVMs) are a set of ML models which can be used in classification problems by finding the best possible hyperplane which separates the classes while maximizing the distance to the nearest split data points (James et al. 2021). This is done by the simple principle that maximizing the distance minimizes the risk (Shmilovici 2005), and this hyperplane is known as the maximal margin hyperplane (James et al. 2021). Thus, the plane is only defined by the points closest to the hyperplane, called support vectors (James et al. 2021).

As only the data closest to the hyperplane defines the model, this means that SVMs can still perform well even on smaller data sets (Shmilovici 2005), which is why it was chosen for this classification problem. This also means it is resilient against observations far from the hyperplane and can handle outliers quite well (James et al. 2021). However, a

drawback of SVMs is that with complex data sets, training can take exceptionally long (Shmilovici 2005), which in this case was mitigated by the feature selection done prior to model training. Notably, training time still took several hours for the feature combination with the highest number of variables. SVMs can be used with different kernels to allow for non-linear hyperplanes (James et al. 2021). However, the kernel used in this paper was linear due to training complexity.

Given a training data set with input features X and corresponding binary class labels y, where $X = [x_1, x_2, \ldots, x_n]$ represents the input vectors and $y = [-1, 1]$ represents the class labels, the goal of a linear SVM is to find a hyperplane that separates the data points of different classes with maximum margin.

The decision function of a linear SVM can be defined as:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b \tag{4.7}$$

where:

- $f(x)$ is the decision function that predicts the class label for a new input vector x.
- $\mathbf{w}$ is the weight vector that determines the orientation of the hyperplane.
- $\mathbf{x}$ represents the input vector.
- b is the bias term.

During the training phase, the SVM algorithm aims to find the optimal values for $\mathbf{w}$ and b that maximize the margin and minimize the classification error. This is typically formulated as an optimization problem involving the minimization of a cost function subject to certain constraints.

Once the SVM is trained, the sign of $f(x)$ (positive or negative) determines the predicted class label for a new input vector x. If $f(x) \geq 0$, the predicted class label is 1; otherwise, it is -1. (James et al. 2021).

Neural Networks

Artificial neural networks (ANNs) are a type of model meant to resemble the neurological connections of the human brain (Shmilovici 2005) and are often referred to as deep learning models (James et al. 2021). They consist of interconnected nodes or neurons

arranged in layers. Each neuron receives input signals from other neurons in the previous layer and produces output signals sent to neurons in the next layer (James et al. 2021). The neurons in each layer are connected to every neuron in the adjacent layer, allowing the network to perform complex computations on the input data (James et al. 2021). This structure means that ANNs are suitable for noisy, high-dimensional, large data sets, which makes them popular for use in data mining (Shmilovici 2005), which was one of the reasons an ANN was used in this paper, as part of the purpose of this paper was to develop a classifier which could be integrated into the partner company's CRM system and the associated data set, which is expected to grow with time and further digitalization.

ANNs and SVMs have a slight problem with comprehensibility. They can be difficult to understand and are often referred to as black-box models due to their large assemblages of real-valued parameters (Shmilovici 2005; James et al. 2021). However, the ANNs have several applications within ML problems, such as image recognition and time series analysis, aside from their use in classification problems (James et al. 2021). They are currently very favored among ML models (James et al. 2021).

ANN models are quite data-hungry and require large amounts of data to train efficiently (James et al. 2021), which is both one of their strengths and weaknesses. They are particularly well-suited to tasks involving complex patterns or nonlinear relationships in the data, as they can learn these patterns through training (James et al. 2021). As data mining and BD become increasingly widespread, ANNs are useful due to their ability to handle large amounts of unstructured, noisy data.

The ANN used in this paper is a sequential neural network with three layers. The first and second layers are dense layers with 64 and 32 units, respectively. A dense layer is a fully connected layer, meaning that each neuron in this layer is connected to every neuron in the previous layer. It is often seen that the number of nodes decreases as we move deeper into the network. This is because the initial layers tend to capture low-level features, and as we go deeper, higher-level abstractions and more complex representations are learned. Reducing the number of nodes helps in reducing the computational burden and prevents overfitting (James et al. 2021).

The layers further use a ReLU activation function. Neural networks require non-linear activation functions to learn and represent complex relationships and patterns in the data. ReLU addresses this need by providing a simple yet effective non-linear transformation.

Another critical advantage of ReLU is its ability to mitigate the problem of gradient vanishing. (James et al. 2021).

The number of nodes/neurons in each layer of a neural network is a design decision that depends on various factors, including the complexity of the problem, the amount of available data, and the desired capacity of the model. After testing, it was settled upon 64 and 32.

Finally, the third layer is a dense layer with a single unit and sigmoid activation function, which is used for binary classification problems such as this one, as it outputs values between 0 and 1 (James et al. 2021). Values closer to 1 indicate a higher probability of the positive class, while values closer to 0 indicate a higher probability of the negative class.

The ReLU activation function is defined as:

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{otherwise} \end{cases} \tag{4.8}$$

where x is the input to the activation function, and $f(x)$ is the output (James et al. 2021). Further, the sigmoid function is defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4.9}$$

where x is the input to the activation function, and $f(x)$ is the output (James et al. 2021).

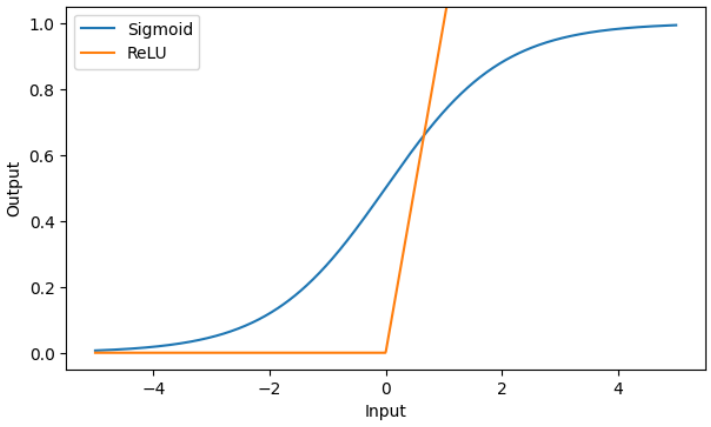


Figure 4.3.3: Activation Functions

Training

The data was split randomly into training and testing sets, with 80% reserved for training and 20% for testing. The purpose of splitting the data set into training and testing sets is to assess the model's performance on data it has not been trained on and thus not seen (James et al. 2021). The training set fits the model to the data, while the testing set evaluates the model's performance on new data. This is important because a model that performs well on the training set may not necessarily perform well on new data, which is the ultimate goal of the model (James et al. 2021). The test-training split was done prior to any feature engineering, such as scaling and one-hot encoding, which were done separately from the test and training set. The training set was, as mentioned, oversampled to balance the class distribution. However, the test set was kept in its original distribution. This was done to test the model on the actual data, and splitting the test set 50/50 between the classes would have resulted in either a small test sample or significantly fewer training samples of the minority class.

Evaluation

Evaluation is a critical step in any ML project, as it allows for measuring the performance of the models and determining how well it is able to make predictions. The evaluation metrics used in this report are Accuracy, Precision, Recall, F1-score, and the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC). Accuracy measures the percentage of correctly classified instances among all the instances in the data set. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.10}$$

where TP is the number of true positives (instances that are positive and are correctly predicted as positive), TN is the number of true negatives (instances that are negative and are correctly predicted as negative), FP is the number of false positives (instances that are negative but are incorrectly predicted as positive), and FN is the number of false negatives (instances that are positive but are incorrectly predicted as negative) (James et al. 2021).

Precision measures the proportion of positive predictions that are correct. It is defined as the number of true positives divided by the sum of true positives and false positives. (Powers 2020)

$$Precision = \frac{TP}{TP + FP} \tag{4.11}$$

Recall, also known as sensitivity or true positive rate (TPR), measures the proportion of actual positives that are correctly identified. It is defined as the number of true positives (TP) divided by the sum of true positives and false negatives (FN). (Powers 2020)

$$Recall = TPR = \frac{TP}{TP + FN} \tag{4.12}$$

The F1-score is the harmonic mean of precision and recall. It is a balanced measure that considers both precision and recall and is often used to evaluate the overall performance of a classifier. (Powers 2020) The F1-score is defined as:

$$F1 - score = 2\frac{precision \times recall}{precision + recall} \tag{4.13}$$

The ROC curve is a graphical representation of the performance of a predictive algorithm. It plots the TPR against the false positive rate (FPR) for different threshold values of the classifier. The ROC curve visually represents the trade-off between sensitivity and specificity for a given classifier. (Handelman et al. 2019)

$$FPR = \frac{FP}{FP + TN} \tag{4.14}$$

Selecting a particular point on the ROC curve depends on the task and system-specific requirements. For instance, a higher sensitivity may sometimes be preferred, even if it results in more false positives. On the other hand, in some scenarios, a higher specificity may be more desirable, even if it comes at the expense of a lower sensitivity. Choosing a particular operating point on the ROC curve can achieve the desired balance between sensitivity and specificity for a given task. The area under the ROC curve (ROC-AUC) is a commonly used metric to compare the performance of different algorithms. The ROC-AUC provides a numerical value that summarizes the classifier's overall performance. The higher the value of the ROC-AUC, the better the algorithm's performance. (Handelman et al. 2019)

4.3.4 Interview Process

Seven interviews were conducted for this study, two during the pre-study and five during the in-depth study. Five different people were interviewed in total. Thus, some were recurrent (see Table 4.3.1). The interviewees were from three different departments. The interviewees were two Business Developers, two Analysts, and one Insurance Broker Manager. The departments were Business Development (BD), Intelligent Automation (IA), and the Brokerage (B) of the particular insurance product in this study.

Pre-Study Interviews

First, an exploratory study was conducted to gain information about the problem and the company's existing capabilities. This consisted of pre-interviews with the partner company and an exploratory literature review. The pre-interviews were held with two employees from different departments to gain an overview of the case, understand the background and context of the situation, and gain a deeper understanding of the partner company's existing capabilities and the feasibility of technical implementation. As described by Saunders, Lewis, and Thornhill (2015), unstructured interviews are useful for better understanding the context of the research problem during an exploratory study.

The interviewees belonged to two departments, *intelligent automation* and *business development*. They were chosen because both departments oversee and develop the customer acquisition process and support and develop new digital strategies and implementations. They thus also have an overview of the company's CRM system, which allowed the researchers a first overview of its current implementation.

The exploratory literature review was used to identify research gaps in the existing literature and guided the formulations of the research questions. The literature review includes all the reviewed articles from the exploratory study.

In-Depth Interviews

In-depth interviews were conducted with three separate departments to gain further insight into the current processes. The interviews aimed to understand the partner company's current acquisition strategies, customer classification systems, and CRM systems. One interview was semi-structured with a set of prepared open-ended questions, and the other two were conducted as in-depth interviews. As described by Saunders,

Lewis, and Thornhill (2015), semi-structured and in-depth interviews are suitable exploratory, qualitative methods for data gathering.

The first interviewee was the manager at the department of brokers who worked extensively with the studied insurance product. This interview (in-depth 1) aimed to learn more about the current customer selection process and how a technical solution could be integrated into it. As the brokers work with this insurance product daily and would be the main users of a digital solution, it was vital to get their perspective. Due to the location of the department being in a different city from the partner company's main offices, the interview was held digitally, which was also the reason for it being the only semi-structured one. As digital interviews sometimes can be slightly more unpredictable in the flow of the conversation, the authors decided to prepare a set of questions beforehand to keep the interview going as needed. The tasks between the interviewers were divided so that one person held the interview, and another person took notes.

The second two interviews (in-depth 2 and 3) were held with the same departments as the exploratory study, Business Development, and Intelligent Automation. The purpose of a second interview with these departments was to revisit any questions that had come up after the first interviews and also to gain new insight as the authors explored the problem further. The interviews were held in person and were, as mentioned, unstructured. The work between the interviewers was divided in the same manner as the first interview, with one person taking notes and the other conducting the interview.

Lastly, two final interviews (in-depth 4 and 5) were held to receive feedback and thoughts about the final models, how they could be applied in a real setting, and what value they could provide given their performance. The proposed integration into the CRM system was also discussed, as well as the most important features found and how these could be collected and stored to build a suitable database for the classifier. These interviews further served as an evaluation of the result of the research project.

Interview	**Dep.**	**Interviewee**	**Purpose**	**Date**	**Length**
Pre-study 1	BD	Business Developer (1)	Overview of problem, context and setting	February 8th	1h

Interview	Dep.	Interviewee	Purpose	Date	Length
Pre-study 2	IA	Business Developer (2)	Overview of problem, context and setting	February 10th	1h
In-depth 1	B	Broker Manager	Gain additional understanding of current process and end users	March 3rd	1h
In-depth 2	BD	Business Developer (1), Analyst (1)	Deeper insight into current processes and CRM systems	March 16th	1h
In-depth 3	IA	Business Developer (2), Analyst (2)	Deeper insight into current processes and CRM systems	March 16th	1h
In-depth 4	BD	Business Developer (1)	Overview of model performance and integration	May 3rd	1h
In-depth 5	IA	Business Developer (2)	Overview of model performance and integration	May 3rd	1h

Table 4.3.1: Overview of the interview process. (Departments: *Business Development [BD], Intelligent Automation [IA], and Brokers [B]*)

The interviews provided the background and context needed to understand and chart the company's current CRM system and adjacent customer acquisition process. The pre-study interviews were subsequently analyzed by reading through the notes and structuring a visual representation of the current CRM system and connected customer acquisition process, which can be seen in Figure 4.3.4.

The in-depth interviews were analyzed inspired by the qualitative data analysis method for interviews described by Saunders, Lewis, and Thornhill (2015). The real-time notes

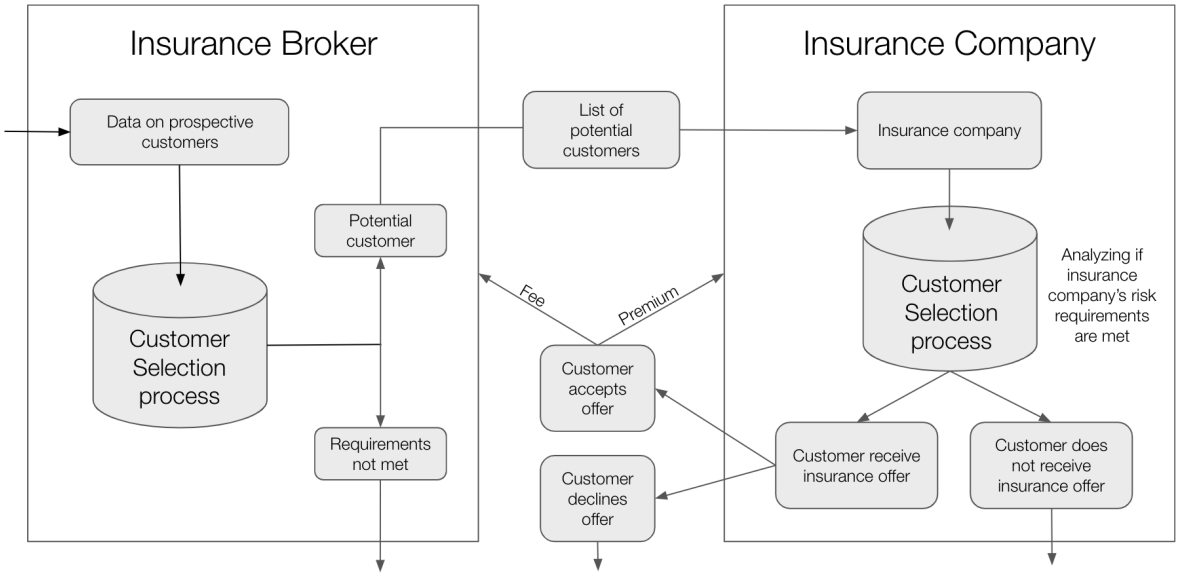


Figure 4.3.4: The Customer Acquisition Process of SME Customers

of each interview were summarized into the most relevant paraphrased statements for the particular problem, which could then be compared against the other interviews. This allowed for further summarizing into four themes that needed to be considered during the development of the classification system with respect to the CRM system. The respective statements can be found in the results section in Table 5.2.1.

4.4 Quality of Research

Several measures were considered to ensure the quality of the research, such as validity, reliability, and ethics. This was done both for the quantitative and qualitative parts of the study.

4.4.1 Validity

Validity refers to the accuracy and truthfulness of the research findings. In this study, efforts were made to ensure the validity of the data collection and analysis process. The authors worked closely with the insurance broker company to ensure that the data collected accurately reflected the target population of potential customers. The ML algorithms utilized in the study were carefully selected and implemented to ensure that the results accurately reflected the characteristics of the potential customers. To further enhance the study's validity, multiple feature selection methods were employed to ensure that the selected features accurately captured the relevant characteristics of

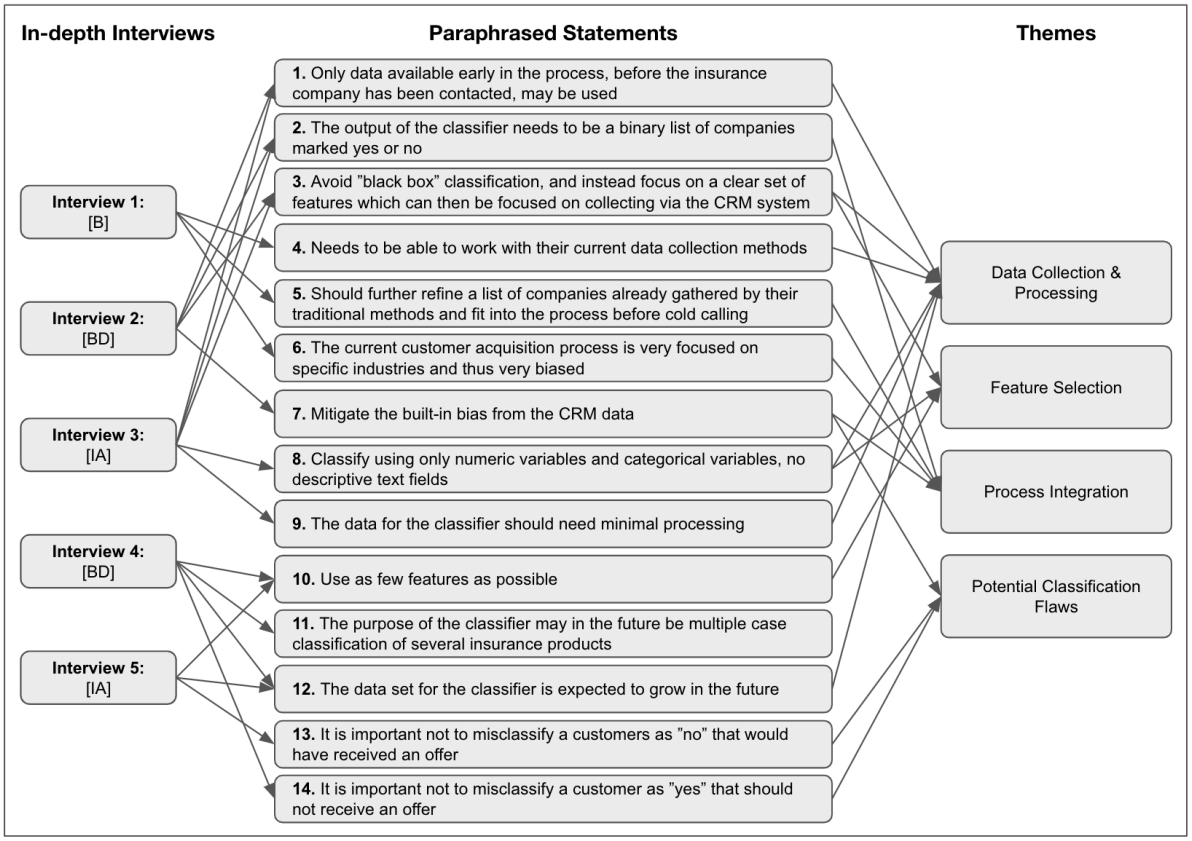


Figure 4.3.5: Interview Analysis Process

the potential customers. The test data points were randomly selected from the data set to ensure that the test data set represents the overall distribution of the data. Furthermore, the models were evaluated using various performance metrics to accurately classify potential customers into relevant segments. However, it is important to note that this particular data set of customers corresponded to one particular insurance product from one particular company. Thus, it does not reflect all SMEs or all company insurance products. Therefore, the results could vary if a similar method is used for other data sets and other products.

For the qualitative parts of the study, validity was ensured through interviews with employees from three different departments to validate their three separate perspectives to gain a complete and unbiased understanding of the process. Validity was further ensured by following established methods for qualitative analysis and interview practices. It is important to note that this is a single-case case study accounting for the specific setting of the partner company, and the results of the study are shaped by that setting.

4.4.2 Reliability

Reliability refers to the consistency and reproducibility of the research findings. In this study, efforts were made to ensure the reliability of the data collection and analysis process. The ML models were developed using standardized techniques and methods, and the same models were applied to all the potential customers in the data set. The categorical variable *Industry Code* was grouped to decrease its number of different classes in order to avoid as much overfitting as possible. Further, all different possible feature combinations were tested recursively to ensure that the results from the feature selection methods were similar.

Reliability was also assessed for the qualitative methods used. While traditional measures of reliability, such as internal consistency or test-retest reliability, may not be applicable to qualitative methods, ensuring rigor and trustworthiness is still crucial. Credibility was established through careful assessment of each of the interviewees in the study. Their role and knowledge within the company were established to ensure that they were the right fit and had the experience and knowledge to speak about their respective processes. In order to ensure dependability, all the interviews were attended by both researchers. Real-time notes were taken during each interview and kept as documentation. Each documentation was reviewed by both researchers to minimize individual bias when analyzing. Throughout the research process, the project was reviewed by the thesis supervisor as well as the peers of the researchers to ensure a reliable process.

4.4.3 Ethics

Ethical considerations were taken into account throughout the research process. The authors worked closely with the insurance broker company to ensure that the data collection and analysis process was in compliance with all relevant privacy and regulations. Further, measures were taken to prevent the ML algorithms from discriminating against any specific customer group. Also, the study strove to have a transparent ML development process and data processing for the brokers and others to understand the limits and flaws of the models.

Chapter 5

Results

In this chapter, the results of the research are presented. The feature selection results are explained in the quantitative part, followed by the selected feature combinations and the models' results. Thereafter, the results from the qualitative part of the study are presented, summarizing the most important themes from the interviews, followed by a description of how these themes contributed to the study.

5.1 Quantitative / ML Models

The study's quantitative part explored ML algorithms' potential in classifying potential customers. Four different feature combinations, two scaling techniques, and four models were tested.

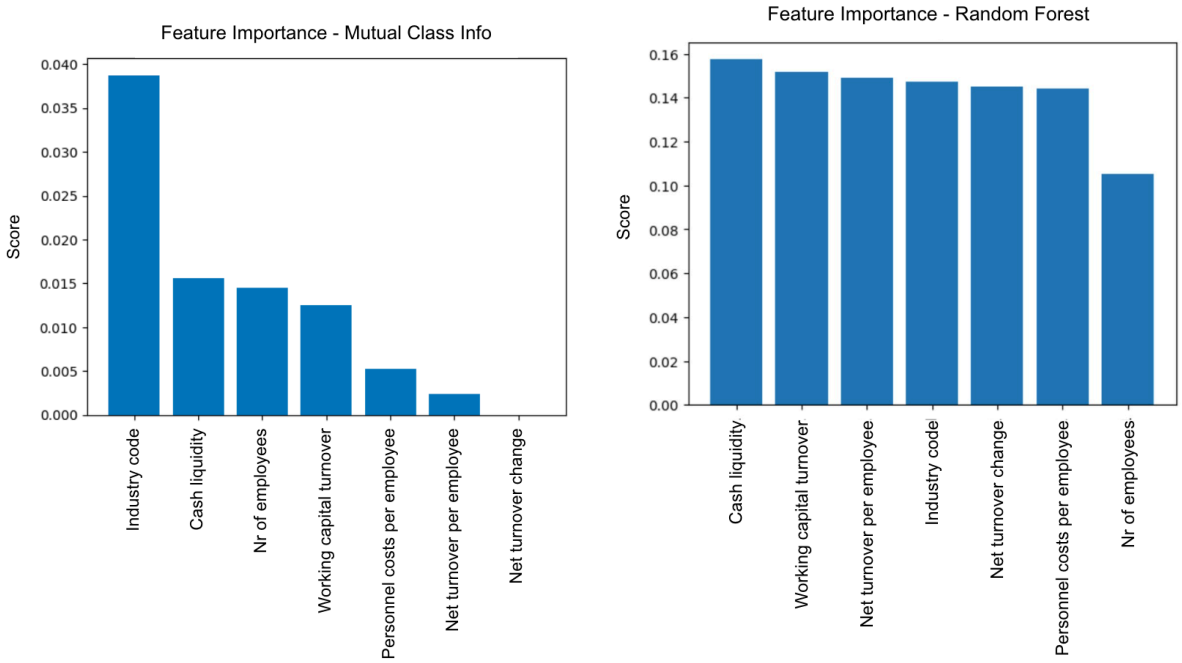
5.1.1 Feature Selection

There were many numerical variables that were highly correlated. The correlation between all variables can be seen in Appendix B. Therefore, only six numerical features remained after applying VIF to remove multicollinearity, as explained in the methodology. These variables are visible in Table 5.1.1.

The variables remaining after VIF were then evaluated with RF and MCI to find the most important features, which yielded the results shown in Figure 5.1.1. From the figure, it can be concluded that *Cash liquidity* was an important independent feature for the dependent variable due it was first in 5.1.1b and second in 5.1.1a. Also, other important variables were *Industry code*, *Working capital turnover*, *and Nr of employees*.

Table 5.1.1: Numerical Variables Remaining after Removing Multicollinearity

Numerical Variables
Cash liquidity
Nr of employees
Working capital turnover
Personnel costs per employee
Net turnover per employee
Net turnover change



(a) Mutual Class Info (b) Random Forest

Figure 5.1.1: Feature Importance Results

5.1.2 Most Important Features

A list of the most important features was curated during the feature selection process. Although the models used in this paper could handle more features than applied, a smaller set of features allows for a more informed classification process as it shows which features are responsible for the achieved result. Data collection can be expensive and time-consuming, allowing the company to focus its resources on collecting the data with the biggest impact on the customer acquisition process.

Industry Code

Industry Code emerged as one of the most valuable features in the classification of corporate customers for the insurance product. Its significance could potentially lie in its ability to provide insights into the sector in which a business operates. Different industries exhibit varying levels of risk exposure, financial stability, and regulatory compliance. By including industry code as a feature, the models could potentially capture the industry-specific reasons why a company might be a suitable candidate for the product. Industry code is further how the current customer acquisition process segments the customer base, which both speaks to its importance and demonstrates why it is useful to classify this data set as it was originally partly classified based on industry.

Cash Liquidity

Cash Liquidity appeared as another significant feature in the classification process. It measures the availability of liquid assets within a business and indicates its financial stability and ability to meet short-term obligations. Businesses with higher cash liquidity are often considered lower risk, as they have a greater capacity to withstand financial shocks and fulfill their insurance obligations. The ML models could, therefore, consider SMEs' cash liquidity to assess their financial strength and risk profiles. As the insurance company takes on all the risk of the customer, it makes it more likely that they will accept companies with greater financial stability. This is likely the reason why cash liquidity showed to be one of the most determining features.

Working Capital Turnover

Working Capital Turnover, which measures the efficiency of a company's working capital management, was also identified as an important feature. It reflects the ability of a business to utilize its current assets effectively to generate revenue. A higher working capital turnover ratio indicates efficient utilization of resources and potentially lower risk. The ML models could, therefore, potentially leverage this feature to evaluate SMEs' operational efficiency and risk profiles, enabling more accurate classification and underwriting decisions. Different companies need different amounts of working capital to operate, which might act as an identifier in the classification process. Another reason could be that companies with greater working capital turnover are seen as more financially stable in the sense that they are seen as profitable and thus able to pay their premiums on time, a factor favored by the insurance company.

Number of Employees

The number of employees within a company was found to be a valuable feature in the classification process. The size of the workforce can indicate the scale and complexity of a business's operations. This feature could result as an important one because the product only focuses on SMEs, and larger corporations are therefore not receiving an offer. The number of employees could also be correlated with factors such as revenue, market reach, and risk exposure. ML models could consider the number of employees as a proxy for the company's size and potential risk, enabling a more precise classification of SMEs into appropriate risk categories.

5.1.3 Feature Combinations

The selected feature combinations were systematically chosen based on their importance and relevance to the dependent variable, which represents whether the customer receives an insurance offer or not. The independent variables were ranked in order of importance using two different methods: MCI and RF feature importance. Based on these rankings, four selected feature combinations were determined.

Table 5.1.2: Selected Feature Combinations

Combination	Variables
1	Cash liquidity, Working capital turnover, Nr of employees
2	Cash liquidity, Working capital turnover, Nr of employees, Industry code
3	Industry code
4	All 32 variables (see Appendix A)

Combination 1

- *Cash liquidity, Working capital turnover, Nr of employees*

The MCI and RF feature importance methods consistently ranked these variables among the most important numerical features. Cash liquidity, the highest-ranked feature in the RF importance analysis, strongly influences the likelihood of receiving a corporate insurance offer. Working capital turnover and Nr of employees also exhibited significant importance, indicating their potential impact on the dependent variable.

For feature combination 1, we see similar results to feature combination 2, discussed below, which is not that surprising since they contain many of the same variables. The

results for combination 2 were slightly better, however, suggesting that the addition of the variable industry code improved the results. This is probably due to the phenomenon discussed earlier, that industry is a common way of distinguishing between businesses.

Combination 2

- *Cash liquidity, Working capital turnover, Nr of employees, Industry code*

This combination expands on Combination 1 by including the categorical variable Industry code, which was consistently ranked highly in both methods. The inclusion of the Industry code recognizes the importance of industry-specific factors in determining the likelihood of receiving a corporate insurance offer.

Among the various combinations tested, feature combination 2 stood out with the best individual performance in terms of accuracy. This is probably because it is the combination containing all the selected features and thus gave the models the most information out of feature combinations 1–3.

Combination 3

- *Industry code*

This combination focuses solely on the Industry code variable, which was ranked as the top feature by MCI. The industry in which a customer operates is likely to influence whether they receive a corporate insurance offer significantly. This is, furthermore, how the current customer acquisition process segments the customers, which is a qualitative reason why this feature combination was tested.

Interestingly, when considering the overall performance across all feature combinations, feature combination 3 had the best results. This is most likely due to the industry being a way to differentiate between companies, and some industries are more suited for this type of insurance product. This was also something that was brought up during the interviews as a likely determinant for the suitability of the product.

Combination 4

- *All 32 variables* (see Appendix A)

This combination includes all available variables and comprehensively analyzes the

relationship between independent variables and the dependent variable. It allows for a comparison with the more focused combinations to assess the importance of feature selection.

5.1.4 Models

The results of the study are presented in Table 5.1.3. The models were evaluated based on five performance metrics: accuracy, precision, recall, F1-score, and ROC-AUC.

Decision Tree

The DT model did not outperform the RF model for any combination. However, it did perform better than the SVM for feature combination 3. This is likely due to the fact that it contains only one feature and thus becomes a simplified probability problem. The best result for DT was an accuracy of 74% with feature combination 3 for both scalars.

The DT model's poor performance was expected due to its lacking ability to handle new data points. Due to all data points tested by the model being completely new and different customers, they are difficult for the model to predict correctly. The DT model performed best on feature combination 3, which only contained the industry code as a feature. It performed less well on feature combination 2, containing the industry code and the three best numerical variables. The model performed even worse on feature combination 1, containing only the three best numerical variables. Thus, the model seemed to be confused by the numerical data. This implies that in a simple model such as a DT, the model struggles to classify companies based on numerical metrics such as liquidity, working capital, and number of employees. This could be due to a greater variation in these numerical values for customers receiving an offer and that companies are instead easier to classify based on their industry.

Random Forest

RF performed well across all feature combinations and scaling techniques. Overall, the SVM outperformed it but still proved to be better for feature combination 3. The best result for RF was an accuracy of 75% for feature combination 4 for both scalers. The RF model performed better than the DT model in all tests except for feature combination 3 (*Industry Code*), where the models performed very similarly. As mentioned in Chapter 4, RF is better at generalizing than DT and is often better at handling new data. This

is likely why the RF model can classify new customers more accurately. However, the model is still outperformed by the SVM model.

Support Vector Machine

SVM consistently produced the best accuracy and recall across all feature combinations and scaling techniques, indicating that it is the most effective model in classifying potential customers. The best result for SVM was 80%, for feature combinations 1 and 2, with MinMax scaler. The SVM model was the overall best performer, which could be attributed to their success in binary classification problems (Awad and Khanna 2015). SVMs further perform well with smaller amounts of data, which could also contribute to the effectiveness of their use in this problem. However, The drawbacks of SVMs include complex and time-consuming training on larger data sets and their tendency to perform worse on multiple class problems (Prince 2012).

Even though it performs well in this case study, the drawbacks might be an issue when the CRM system develops and yields larger data sets or the classifier needs to be extended to be used for multiple case classification, which was related to two of the statements discovered, no. 10 & 11 - *The purpose of the classifier may in the future be multiple case classification* & *The data set for the classifier is expected to grow in the future.*

Neural Network

The NN model performed worse than RF and SVM in terms of accuracy. This is most likely due to the slim number of features, as ANNs are models that tend to work better on larger data sets with more features. The ANN model was also one of the worst performers in terms of Recall. However, interestingly enough, it consistently was the best performer in terms of Precision. The best result for the ANN was an accuracy of 75% for feature combination 4 with the MinMax scaler.

One of the benefits of ANNs is their ability to handle large amounts of unstructured, noisy data (Prince 2012). They typically perform better on large data sets with many features and not as well on smaller data sets with fewer amounts of features (Prince 2012). Therefore, the very curated feature selection done in this paper is probably to blame for the slightly worse performance for this problem compared to the RF or SVM model. This phenomenon can be observed for feature combination 4, which yielded the best result and was the combination with all available features.

5.1.5 Metrics

Accuracy

Accuracy is a commonly used evaluation metric in ML that measures the overall correctness of a classifier's predictions. It represents the proportion of correctly classified instances out of the total number of instances in the data set. One of the benefits of accuracy is its intuitive interpretation, making it a convenient metric to use by, for example, users of the classification system not familiar with ML or other types of ML metrics. However, one limitation of accuracy is its sensitivity to class imbalance. In this case, where there was a 75/25 difference in the number of instances between classes in the test data, the classifier can achieve high accuracy by simply predicting the majority class most of the time. Accuracy further fails to provide insights into the type of errors made by the classifier. It treats all misclassifications equally, without distinguishing between false positives and false negatives. This was the reason for complementing accuracy with the additional metrics Precision, Recall, F1-score, and ROC-AUC.

A hypothetical classifier that predicts the majority class 100% of the time would achieve 75% accuracy on this problem whilst doing nothing beneficial and receiving a better score than DT, RF, and ANN. Therefore, it is significant that the SVM model achieves an accuracy of 80%, performing better than the hypothetical model. In order to understand the performance of the other models, which have an accuracy of 75% or below, we need to look at the other metrics to evaluate their performance beyond doing less than the hypothetical model.

Precision, Recall & F1-Score

Precision and Recall were equal or close in range for many of the model results, particularly the DT and RF model. When precision and recall are equal to each other, it means that the model is achieving a balance between these two metrics. In other words, the model makes correct positive predictions (precision) while capturing a significant portion of the actual positive instances (recall). In practice, there is often a trade-off between precision and recall. Increasing one of these metrics typically leads to a decrease in the other. For example, if a model is tuned to have high precision, it may be more conservative in labeling instances as positive, resulting in a lower recall. Conversely, if a model aims for high recall, it may be more inclusive in labeling instances as positive, potentially leading to a lower precision. We observe this phenomenon with the ANN

model, which received the highest Precision whilst often receiving the lowest Recall score, and we further observe the reverse with the SVM model, which received the highest Recall score and among the lowest precision for two of the feature combinations.

The specific balance between precision and recall depends on the specific requirements and goals of the problem at hand. For example, suppose the goal is only to have a list of companies correctly classified as yes. In that case, having high precision may be more important to avoid incorrectly classifying non-suitable customers as suitable ones. On the other hand, if the goal is to miss as few companies as possible, it may be more critical to have a high recall to avoid missing positive cases, even if it means accepting some false positives. On that note, the F1-score combines precision and recall into a single value to provide a balanced evaluation of a model's performance. The F1-score is computed as the harmonic mean of precision and recall to ensure that the F1-score remains high only if both precision and recall are high. It penalizes models with a significant difference between precision and recall, favoring models with a more balanced performance. Thus, depending on the specific requirements of the output, one of these metrics can be chosen as the most important one.

ROC-AUC

The ROC curve is a graphical representation of the performance of a classifier as the discrimination threshold is varied, while the AUC quantifies the overall performance of the classifier. This score is always between 0.5 and 1, where 1 is a perfect score, and 0.5 represents a random classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. A classifier with a higher TPR and a lower FPR across various threshold settings will have a curve closer to the plot's top-left corner, indicating better performance. The diagonal line in the ROC plot represents a random classifier. It is useful as it comprehensively evaluates the classifier's performance across various threshold settings, considering the trade-off between TPR and FPR. It does not provide insights into the specific misclassifications made by the classifier or the optimal threshold to use in a given context. Additionally, the ROC-AUC is less informative in cases where the costs of false positives and false negatives are significantly different. As this is an imbalanced data set, it is useful as an overall performance score. However, when the goal is to gain more insight into the results and performance of the classifier, metrics such as Precision and Recall might be more useful.

The DT model had the worst AUC score overall, as can be observed in the table as well as the graphs below, where the blue line is often closest to the diagonal line representing the random classifier. It also received a lower accuracy than a model which would have predicted the majority class 100% of the time, indicating that it does not perform much better than a random classifier. The ANN overall had the best AUC score, which can be observed as the green line is the closest to the top left corner for many of the graphs. The ANN received worse AUC scores for the Robust scaler, indicating it can more effectively distinguish between positive and negative cases when the data set is scaled using the MinMax scaler. It is interesting that the ANN model received the highest ROC score while receiving the lowest accuracy. This indicates that it is more effective at correctly identifying and ranking the minority class instances, even if it misclassifies some majority class instances, whereas the other models, which received a higher accuracy and instead worse AUC scores relative to the ANN, prioritize the majority class when predicting.

5.1.6 Scaling Technique

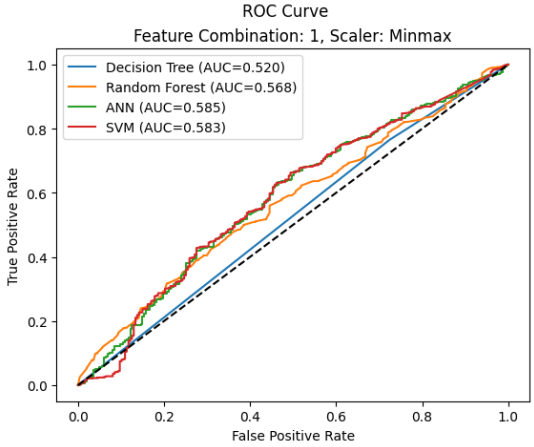
Two scaling techniques were used, MinMax scaler and Robust scaler, to normalize the feature values to a specific scale. MinMax scales the features to a range between 0 and 1, whereas the Robust scales the features to a range between the 25th and 75th percentile of the data, which makes it more robust to outliers. MinMax scaling consistently outperformed Robust scaling across all feature combinations and models, except for SVM, where both scaling techniques produced similar results. The scaling techniques' impact on the different models' performance metrics can be seen in Table 5.1.3. For example, when using feature combination 1 with MinMax scaling, the SVM model has the highest accuracy and F1 score among all the models tested. On the other hand, the NN model has very low-performance scores across all metrics. When using Robust scaling with feature combination 2, the SVM and RF models have the highest accuracy and F1 score, while the NN model still has low-performance scores.

MinMax scaling is generally a good choice when the data is well-behaved, and there are no significant outliers, whereas Robust scaling is better when the data contains outliers. However, it is important to note that the choice of scaling technique can also depend on the specific characteristics of the data and the models being used. In some cases, MinMax scaling may still perform better than robust scaling, even with outliers. Therefore, trying multiple scaling techniques and comparing their performance is always a good practice.

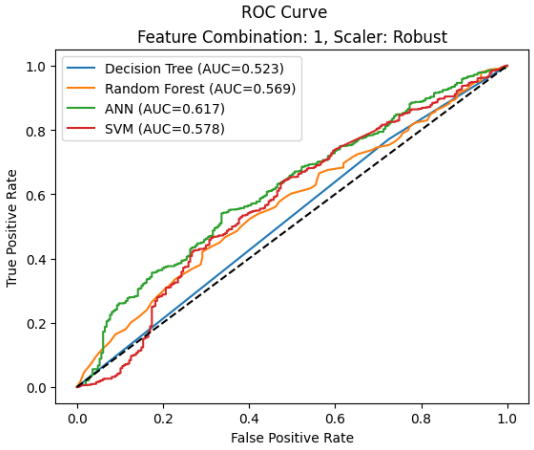
Table 5.1.3: Model performance metrics for different feature combinations, scaling techniques, and models.

Feature combination	**Scaling Technique**	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
1	MinMax	DT	0.69	0.68	0.69	0.68	0.52
1	MinMax	RF	0.71	0.68	0.72	0.69	0.57
1	MinMax	SVM	0.80	0.69	0.80	0.72	0.58
1	MinMax	NN	0.62	0.73	0.62	0.66	0.59
1	Robust	DT	0.66	0.69	0.66	0.68	0.52
1	Robust	RF	0.71	0.70	0.71	0.71	0.57
1	Robust	SVM	0.75	0.70	0.75	0.72	0.59
1	Robust	NN	0.61	0.76	0.61	0.65	0.62
2	MinMax	DT	0.68	0.70	0.68	0.69	0.56
2	MinMax	RF	0.74	0.73	0.74	0.73	0.63
2	MinMax	SVM	0.80	0.69	0.80	0.72	0.70
2	MinMax	NN	0.70	0.75	0.70	0.71	0.71
2	Robust	DT	0.69	0.71	0.69	0.70	0.57
2	Robust	RF	0.74	0.73	0.74	0.73	0.61
2	Robust	SVM	0.70	0.72	0.70	0.71	0.70
2	Robust	NN	0.71	0.74	0.71	0.72	0.64
3	MinMax	DT	0.74	0.75	0.74	0.74	0.71
3	MinMax	RF	0.74	0.75	0.74	0.74	0.71
3	MinMax	SVM	0.73	0.76	0.73	0.74	0.67
3	MinMax	NN	0.71	0.74	0.71	0.72	0.71
3	Robust	DT	0.74	0.75	0.74	0.74	0.71
3	Robust	RF	0.74	0.75	0.74	0.74	0.71
3	Robust	SVM	0.73	0.76	0.73	0.74	0.67
3	Robust	NN	0.71	0.74	0.71	0.72	0.71
4	MinMax	DT	0.67	0.65	0.67	0.66	0.55
4	MinMax	RF	0.74	0.67	0.74	0.67	0.68
4	MinMax	SVM	0.70	0.72	0.70	0.71	0.70
4	MinMax	NN	0.75	0.75	0.75	0.75	0.70
4	Robust	DT	0.66	0.64	0.66	0.65	0.52
4	Robust	RF	0.75	0.69	0.75	0.68	0.65
4	Robust	SVM	0.70	0.72	0.70	071	0.70
4	Robust	NN	0.71	0.72	0.71	0.71	0.66

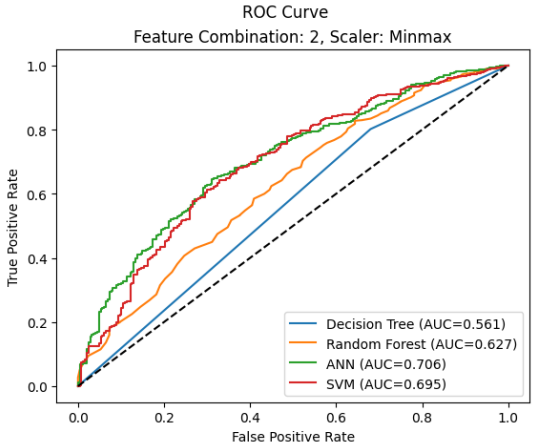
Figure 5.1.2: ROC Curves



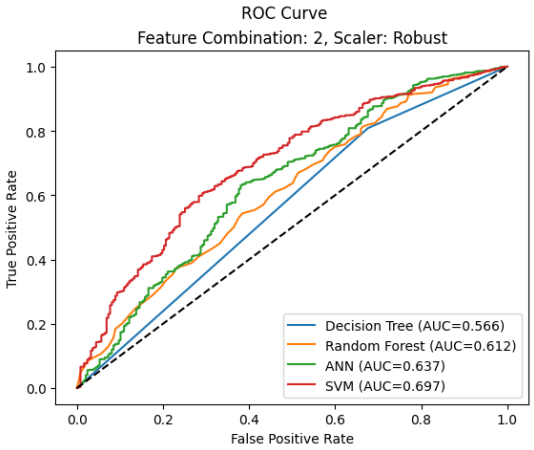
(a) Feature combination: 1, Scaler: MinMax



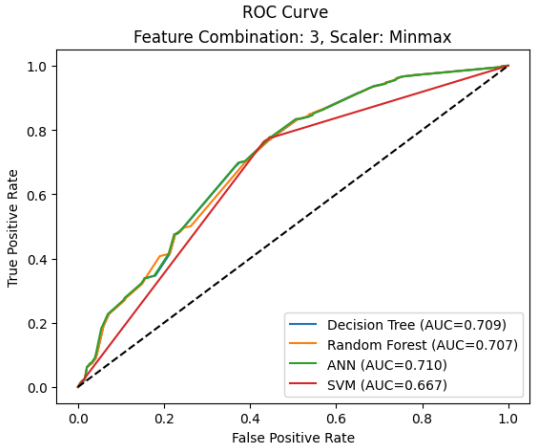
(b) Feature combination: 1, Scaler: Robust



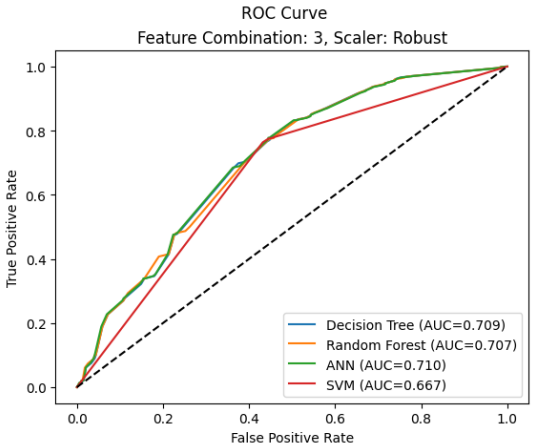
(c) Feature combination: 2, Scaler: MinMax



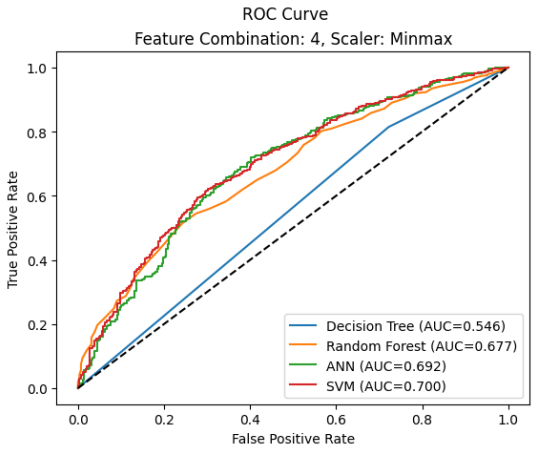
(d) Feature combination: 2, Scaler: Robust



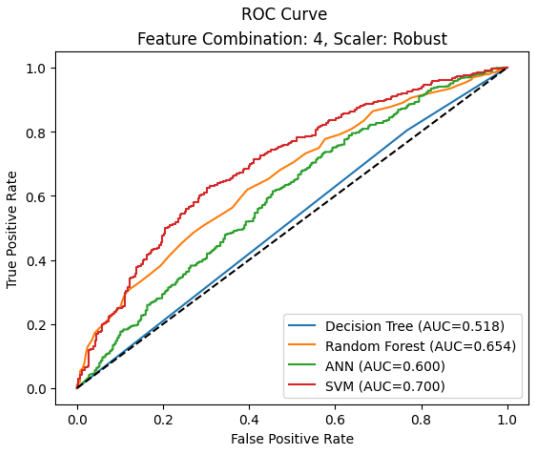
(e) Feature combination: 3, Scaler: MinMax



(f) Feature combination: 3, Scaler: Robust



(a) Feature combination: 4, Scaler: MinMax



(b) Feature combination: 4, Scaler: Robust

5.2 Qualitative / Interviews

The interviews provided valuable insights into insurance brokers' challenges in the business insurance industry and how customer classification can be integrated into their current acquisition strategy. It also provided further insight into how a digital classification implementation could be integrated into the current CRM system. Real-time notes of each interview were summarized into the most relevant paraphrased statements, which could then be summarized further into four themes as seen in Figure 4.3.5. The paraphrased statements from the interviews can be found in Table 5.2.1, and the summarizing themes can be found in Table 5.2.2

5.2.1 Themes

The resulting paraphrased statements were summarized into four themes. The statements provided different guidelines for the implementation of the classification system depending on which theme it was grouped to. As the interviews were held and statements subsequently gathered over the research process, the classification system was gradually developed according to the statements and their specified requirements.

Theme 1: Data collection

Statements 1, 3, 4, 8, and 9 are related to collecting and processing data for the classifier. For example, statement no. 1 leads to the use of publicly available data sources as they are available at any time in the process. For the intended users to be able to use the

Departments	Number	Paraphrased Statements
BD, IA	1	Only data available early in the process, before the insurance company has been contacted, may be used
BD, IA	2	The output of the classifier needs to be a binary list of companies marked yes or no
BD, IA	3	Avoid ”black box” classification, and instead focus on a clear set of features which can then be focused on collecting via the CRM system
B	4	Needs to be able to work with their current data collection methods
B	5	Should further refine a list of companies already gathered by their traditional methods and fit into the process before cold calling
B	6	The current customer acquisition process is very focused on specific industries and thus very biased
BD	7	Mitigate the built-in bias from the CRM data
IA	8	Classify using only numeric variables and categorical variables, no descriptive text fields
IA	9	The data for the classifier should need minimal processing
BD, AI	10	Use as few features as possible
BD	11	The purpose of the classifier may, in the future, be multiple case classification of several insurance products
BD, AI	12	The data set for the classifier is expected to grow in the future
IA	13	It is important not to misclassify a customer as ”no” that would have received an offer
BD	14	It is also important not to misclassify a customer as ”yes” that should not receive an offer

Table 5.2.1: Paraphrased Statements (Departments: *Business Development [BD], Intelligent Automation [IA], and Brokers [B]*)

classifier, the data collection also needs to be considered. Much of the data collection statements are related to transparency and insight, i.e., we want to know exactly which data to collect and its importance for the classification process. The interviewed business developer from in-depth 4 remarked that ”It is important for us to understand the process behind the segmentation, and to see which parameters make the classifier do its job.”. This is, in turn, connected to the feature selection process employed in this paper. The reason for employing several feature selection methods and curating the number of features to be very few was to give a more insightful data collection process that allowed for the collection of as few variables as possible. Statement no. 9 also describes the preferred data collection process, as not having to spend resources on cleaning and processing data saves time and money. This was solved by collecting publicly available data, which is already

Paraphrased Statements	Theme
1, 3, 4, 8, 9	Data Collection
3, 4, 10	Feature Selection
2, 5, 6, 7	Process Integration
7, 13, 14	Potential Flaws

Table 5.2.2: Themes

labeled and organized by company and year. It was further solved by building some data processing into the implementation so that the code for the classifier automatically generalizes the industry code variable and removes NaN values.

Theme 2: Feature selection

Statements 3, 8, and 10 are grouped as theme *Feature Selection*. From these statements, it became clear that less is more when it came to the data set and feature selection. By having a smaller set of features, the classification system is easier for the brokers to understand and thus gives them insight into what can be used as a good determinant for customer selection. A smaller set of features is also desirable as data costs money to procure, and by knowing a smaller subset of features to focus on, the data collection process also takes less time. One business developer commented that "One of the most valuable ways to develop the CRM system currently is to find the correct information to focus on, so we can spend resources where it matters the most" (In-depth 4). Statement 10, *Use as few features as possible*, can be slightly misleading, as using as few as possible would mean zero features and no classifier at all. The idea is to use the least number of features possible without a significant drop in performance and to find a good trade-off between classifier performance and the number of features. The business developer from in-depth 3 commented that "it's probably good to collect as much data as possible as we currently don't know what could be of importance, and then to narrow it down as much as possible so we can see what has the most impact". Feature combination 4 was added with the purpose of demonstrating that the complete set of features does not deliver a better result than the set of features selected through the used feature selection methods, except for the ANN model.

Theme 3: Process Integration

Statements 2, 5, 6, and 7 refer to the *Process Integration* theme. Thus, the current customer acquisition process and how the classifier needs to be adapted in order to work

as a useful tool. Theme no. 2 needs to be fulfilled as this is how the brokers work currently and explains how the output needs to be formatted in order to be useful to the broker team. This is closely related to theme no. 5, which explains at which point in the process the output should be. Themes no. 6 and 7 inform of the current process and how the classifier can be used to balance out imperfections. For example, the interviewed manager for the broker department remarked that ”We first get plenty of raw data that we have to manually go through in order to find the best candidates. The customers for this type of product expect a personal treatment, so we then have to spend time on each customer we decide on personally” (In-depth 1). This meant that the best way to assist the process was to help the brokers find the correct customers, and automate that process, and let the brokers handle the personal touch. The classification system is therefore designed to process the customer data and then let the brokers take over at the best potential customers.

Theme 4: Potential Risks and Flaws

Themes 7, 13, and 14 refer to potential flaws that can happen with the classification system. As the data set is currently biased toward the brokers' favorite industries, implementing an ML-based classification system could possibly propagate that bias further. It was thus important that the classification system was implemented to be blind to some of the bias. The most obvious bias was found in industry, as the brokers focus on companies from specific industries while ignoring others. This was part of why the industry code feature was shortened into more general categories. The case could be made for omitting industry as a feature altogether to erase the bias completely. However, as the results show, the classifier performs better with industry as a feature, and some industries are better suited for this type of product. Themes 13 and 14 show that misclassification would be a problem whether it's a no or a yes, which meant several evaluation methods looking at both FPs as well as FNs had to be used. During in-depth 4 it was said that ”bringing a non-suitable customer into the process means we would have to spend much more time on that customer, as we would need to manually put together a package deal rather than just offering them the product”, whereas during in-depth interview 5 it was stated that ”it's important not to put 'no' on a customer that would have been good, as this means we could lose revenue”. The statements are not conflicting from a business perspective, however from an ML perspective it meant that there was a need for metrics which allowed evaluation of both sides, as mentioned

earlier.

5.2.2 Customer Process Integration

In order to answer RQ2 (*How could an ML-based classification system be integrated for use with an existing CRM system?*), the integration of the classification system had to be evaluated. Due to the limited time, an actual integration was not possible, and the authors thus developed a proposed integration of a recommended implementation, following the themes discovered during the interviews. In the end, the integration was organized following the CRM framework, and findings from the interviews suggested that the first three pillars (*Creating a Database, Analysis, and Customer Selection*) were the most important in this case.

Given statements 4 & 5, *Needs to be able to work with their current data collection methods* and *Should further refine a list of companies already gathered by their traditional methods and fit into the process before cold calling*; it is proposed that the model is integrated at the point where the brokers receive the first list of corporations to process (see Figure 5.2.1). The model can then be used to expedite the first selection process, and the brokers can then focus on the remaining, refined list of potential customers to pursue.

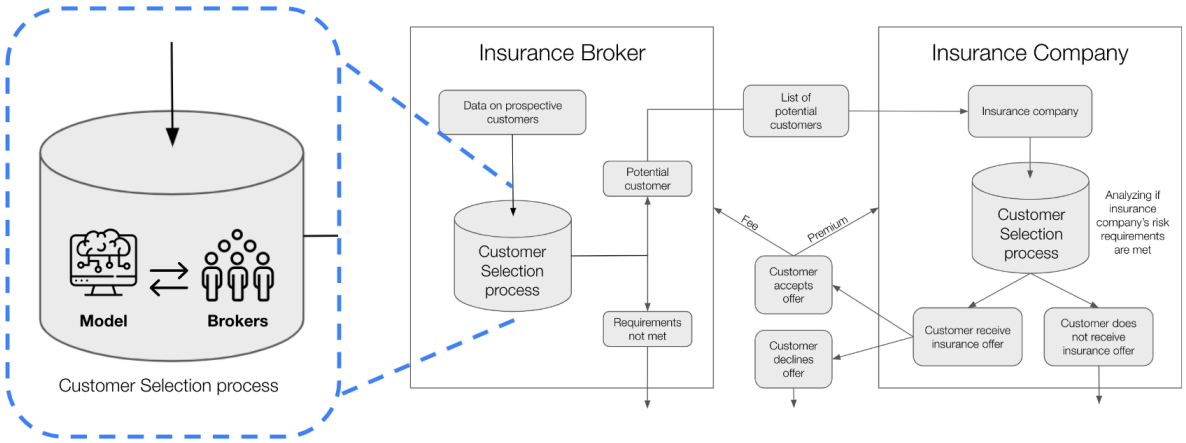


Figure 5.2.1: Model Integration into The Customer Acquisition Process

It was further found that an implementation of a digital customer selection process such as this would fit well into the existing CRM system when the classifier and system work in symbiosis to improve the data-gathering process. This insight was derived from

statement no. 3 *Avoid "black box" classification, and instead focus on a clear set of features which can then be focused on collecting via the CRM system*, and 8 *Classify using only numeric variables and categorical variables, no descriptive text fields.* Symbiosis here means that the classification system can, through feature selection and informed classification, give insight into which variables and features have the most effect on the prediction of the classification. Thus, time and resources should be focused on collecting that data. As the CRM system develops and accumulates additional data points, both the data recommended by the classifier as well as data gathered at other points of the system can be used to further improve the classifier by feeding it larger amounts of higher-quality data.

CRM Framework

The CRM framework was used to supply structure for the implementation of the classifier. Given that the first three pillars in the framework are also features that could be done by the classifier, the framework could then be used to hypothesize that a classification system could be developed to be part of a CRM system.

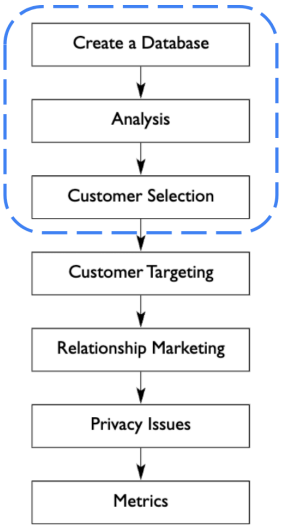


Figure 5.2.2: Classification Model Integrated into CRM System

The classification system was developed to function as the first three pillars of the framework, see Figure X. It was then proposed that the brokers take over by the "customer targeting" pillar, where they target the selected customers by the classification system by offering them the product.

Create a Database

This step refers to the action of creating a data set that can be used with the classifier. In order to adhere to the discoveries from the interviews, this data set was developed to be as refined as possible and to avoid bloated data sets.

Analysis

In this paper, a detailed description of the data analysis performed can be found in the methods section. The authors found insights into which data is more valuable when

classifying customers, which is discussed under the *feature selection* header in the Methods section.

Customer Selection

This step refers to the actual classification process. A list of available companies is processed and segmented by the classifier into the ”yes” and ”no” groups, which are then used by the brokers during the customer targeting phase.

Chapter 6

Discussion

In this section, the empirical findings from the study are presented, analyzed, and discussed. First, the most important features for classifying corporate customers in this context are discussed. Second, RQ1 and the models' performances are discussed. This is followed by a discussion of RQ2 and AI-CRM integration. Finally, the outcomes and challenges of the study are mentioned, including return on investments, misclassification, and bias.

6.1 Features for Corporate Insurance Classification

Not much previous research about variables used to classify corporate customers for insurance products was found when conducting the literature review for this study. However, some research was found on classifying private customers for insurance products. For private customers, variables such as age, gender, education, career, income, and payment history were used (Y. Chen and Hu 2005). For companies, this study found the variables *Cash Liquidity* and *Working Capital Turnover* to be two of the most important. These variables are indicative of a company's financial health and operational efficiency. Cash liquidity refers to the availability of cash or easily convertible assets, which reflects a company's ability to meet its short-term obligations (Demarzo and Berk 2014). On the other hand, working capital turnover measures how effectively a company utilizes its working capital to generate revenue. A steady income is an important variable for assessing risk in private customers because it indicates their ability to consistently

meet their financial obligations, such as insurance premium payments. Similarly, for corporate customers, higher cash liquidity implies that the company has readily available funds to cover unforeseen expenses, including insurance claims. This liquidity acts as a buffer and reduces the risk of financial distress, making the company more resilient. However, it is important to note that while these similarities can be drawn, the specific risk factors and their impact on private and corporate customers may differ. Private customer risk assessment may focus more on personal financial stability, creditworthiness, and payment history. In contrast, corporate customer risk assessment may involve considerations such as industry-specific risks, business performance indicators, and the company's overall financial health.

Continuously, it was found from the literature review that liability insurance, which is one of the most common business insurances (Henry 2016), also included in the case study insurance product, is mainly priced based on industry and policyholder risk (Müller and Te 2017). This does thus support the finding that *Industry Code* is one of the most important features when classifying corporate customers. The significance of the *Industry Code* variable in classifying corporate customers can be attributed to the varying levels of risk associated with different industries. Certain industries inherently carry higher risks due to the nature of their operations, potential liabilities, and regulatory factors. Insurance companies have to consider these industry-specific risks when determining coverage and pricing for liability insurance policies.

6.2 Model Performance

To answer RQ1, the models and their performances for this classification problem were discussed, also its implications for research by linking back to previous studies. The models implemented in this study were chosen due to many previous successful attempts where they were used for similar classification problems.

The RF model in this study achieved a maximum accuracy of 75.0%, which is consistent with the findings of other recent papers using RF for similar classification problems. For example, a recent study by Abedin et al. (2023) predicting the behavior of bank customers using an RF model achieved an accuracy of 75.85%, and another study using customer behavior data from an in-vehicle coupon recommendation system found that their RF model achieved the highest accuracy at 76.1%.

The DT model, which produced slightly worse results than the other models compared, also performed somewhat as expected. A study by C. Lin and Zheng (2022) also implemented a DT model and mentioned that "the result is not satisfactory" (C. Lin and Zheng 2022, p. 870) and used gradient boosting to improve the model's performance. A similar study implementing a model for B2B customer acquisition also utilized a DT model as part of the customer acquisition process (D'Haen and Van den Poel 2013). The study used AUC as a metric for evaluating the performance of the DT, which received a score of 0.99985. This is higher than the results in this study, where the highest AUC achieved by the DT model was 0.71. D'Haen and Van den Poel (2013) applied their model with only categorical variables, which could explain why it performed better. The best result for the DT model in this study was also achieved when used with only one categorical variable, and it performed significantly worse in the cases with numerical variables.

As for the SVM model, it achieved the highest accuracy out of all the models used in this study but with lower precision scores. These findings are congruent with those of Aslam et al. (2022), who implemented different models for insurance fraud detection. Their SVM model also produced the best results in terms of accuracy and the lowest score in terms of precision for the implemented models. Their SVM model achieved an accuracy of 94% however, compared to this study's 80%. This could be due to the difference between the classification problems, as this was done on fraud detection.

Not many studies using ANN for similar classification problems were found. However, studies have used ANNs for other types of classification problems. A scoping review of ANNs for decision-making in the health care industry found varying accuracy scores between 50% and 100% (Shahid, Rappon, and Berta 2019). A study researching early outcome prediction and risk classification in out-of-hospital cardiac arrest patients admitted to intensive care also employed a simplified ANN model with only three variables and saw an AUC score of 0.852 for that model (Johnsson et al. 2020). These findings are slightly higher than those of this study, which had the highest AUC of 0.71 for the feature combinations with few features. The feature combination with all the present features received the highest scores, supporting this study's hypothesis that the ANN model performed worse because of the low number of features. However, the findings of Johnsson et al. (2020) suggest that the quality of the features and the type of problem at hand can yield better performance for ANNs even with very few features, as they did not see much of an improvement using all of their 54 available variables as that gave an

AUC of 0.891.

Overall, the performance of the tested models suggests that depending on the data set, feature selection, and type of evaluation metric used, different models are useful for these types of classification problems. These findings are congruent with those of Jain, Yadav, and Manoov (2021), who tested different models on data sets from different industries and found that different models were useful depending on which data set was used.

6.3 AI-CRM Integration

To answer RQ2, this section discusses the implications of the results found from the qualitative part of this study. In practical terms, the results obtained from the interviews with insurance brokers provide valuable insights into the challenges they face in the business insurance industry and shed light on how customer classification can be integrated into their existing acquisition strategy. One important observation is the need to consider data collection throughout the entire process. The interviews highlighted the importance of utilizing data available early in the process before contacting the insurance company. This suggests that publicly available data sources can be leveraged effectively to enhance the classification system. This is congruent with the findings of Libai et al. (2020), who suggest that data from external sources can be used to improve the customer acquisition strategy for AI-CRM and that this will be an important capability when developing a CRM system into AI-CRM. Similarly, other studies also highlight the potential value of utilizing data in the CRM process. One example is a study in the field of direct marketing (V. Kumar et al. 2019) highlighting that the right data and AI as an analysis tool could improve the personalization of marketing.

Another practical implication is the emphasis on feature selection. The statements from the interviews indicated that a smaller set of features is preferable for the classification system. By having a concise set of features, brokers can better understand the determinants for customer selection. Moreover, a smaller feature set reduces data procurement costs and speeds up the data collection process, making it more efficient. Our quantitative model analysis further shows that a smaller set of features produces an equally satisfying result as the full set of features. This is in accordance with the findings of Tillmanns et al. (2017), where they used feature selection methods to narrow down the set of variables used from 100 to 8. The resulting model is ”surprisingly powerful in predicting new customer responses” and they further state that the benefit of the slim

feature selection is lower data procurement costs (Tillmanns et al. 2017, p. 112).

Process integration is another significant aspect revealed by the interviews. The classification system needs to be seamlessly integrated into the current customer acquisition process to ensure its usefulness for brokers. This includes formatting the output as a binary list of companies marked ”yes” or ”no” to align with the brokers' workflow. Additionally, the classifier should be adapted to refine the list of companies gathered through traditional methods before engaging in cold calling. The literature review found several studies on the benefits of integrating AI into CRM in a B2B context. For example, one study identified the capabilities AI could transform CRM to AI-CRM (Libai et al. 2020), such as the ability to predict customers' value, argued that this could lead to increased prioritization among customers. Similarly, brokers could potentially, with the help of the classifier, prioritize which customers to call as they will more likely be accepted by the insurer.

6.4 Outcomes and Challenges

As discussed previously, the goal of implementing an ML-based classification system into the current customer acquisition process at the partner company was to improve the efficiency of the customer acquisition process and to increase the likelihood that the insurance company would accept proposed customers. The expected benefit of this is an increased return on resources, and the main challenge is the problem of misclassification.

6.4.1 Increased Return on Investments

The main estimated benefit of a classification system is the potential to make the current process more efficient and effective. Improved efficiency can be achieved by reducing brokers' time on customer acquisition. The classification system automates the initial selection phase, allowing brokers to spend less time manually filtering potential customers. This streamlines the process and enables brokers to focus their efforts on customers selected by the classifier. Consequently, the customer acceptance rate from the initial contact with the insurance company is expected to increase as the classifier enhances the probability of identifying suitable candidates for the studied product. This increased efficiency and a higher acceptance rate per processed customer lead to a higher return on invested time and resources for brokers. By optimizing their efforts, brokers

can achieve more favorable outcomes while minimizing the resources expended.

From a data perspective, implementing a classification system can also yield increased returns on invested resources. The complexity of the system's data requirements is reduced by curating a smaller subset of relevant features. This, in turn, lowers the costs associated with data acquisition. As a result, resources can be allocated more effectively to collect high-value data, enhancing the overall data collection process. This approach helps identify where resources should be allocated for the highest value per invested funds, contributing to improved returns.

6.4.2 Misclassification

An implementation of a classification system faces two challenges regarding the misclassification of potential customers. Misclassification occurs when the classifier assigns a label other than the true one to a given data point. For this binary classification problem, the two kinds of misclassification are False Negative (FN) and False Positive (FP).

In this classification problem, a FP would be a customer that the classification system classified as a suitable customer when it, in fact, was not. Given that the studied insurance product in this paper is a type of package product that incorporates a collection of insurances suitable for many companies, once a suitable customer gets involved in the customer acquisition process, the process is typically quite fast. If, however, a customer who is not suitable gets brought into the process, significantly more time will have to be spent on that customer as their insurance deal will have to be negotiated manually instead of them being offered the ready package deal. This will lower the return on invested resources discussed earlier, increasing the time it takes to earn back the cost of acquiring this customer. To minimize the risk of FPs, the classifier should be optimized using Precision as a metric. Precision considers FPs when it is calculated, so it has an inverse relationship with the FP rate. A high Precision thus yields a low rate of FPs.

A FN would instead be a customer that the classification system classified as unsuitable for the product when it would have been a good candidate. The challenge here is not the waste of invested resources but rather the cost associated with missing a potential customer. Depending on the customer, this cost can be quite high in terms of missed yearly premium revenue. In order to minimize the rate of FNs, Recall should be used as

a metric. Recall instead has an inverse relationship with the rate of FNs, and a higher Recall will yield fewer FNs.

6.4.3 Bias

The interviews revealed the need to address biases in the current CRM data, ensuring an impartial and balanced classifier. Another study (Libai et al. 2020) also highlighted bias issues and the potential for customer discrimination with AI-CRM. Brokers must be mindful of this risk when integrating the system. For instance, they should avoid favoring or discriminating against specific industries based on personal preference. Implementing an ML-based system could perpetuate this bias. To mitigate it, steps should be taken to reduce bias, such as using broader industry categories instead of specific ones. However, it's important to consider that certain industries may be more suitable for the product, and the industry category feature still contributes to better classifier performance.

Chapter 7

Conclusion

In conclusion, this thesis investigated the application of ML-based classification in the business insurance industry and how the classification could potentially be integrated into a CRM system. It investigated the current state of these systems, explored their potential benefits for collecting and utilizing customer data in customer acquisition, and developed an ML-based system for the study partner company. The research also aimed to provide practical insights and recommendations for insurance brokers looking to implement these systems in their acquisition strategy and integrate them into their CRM systems.

The study's main findings are related to the use of ML models in AI-CRM integration. In regards to RQ1, the authors found that an ML-based classification system can successfully be implemented to assist in the customer acquisition process by segmenting customers based on customer data into two groups — customers suitable for the given product and customers who should not be pursued. It was found that different models are useful depending on the feature set and evaluation metric. A model should be chosen depending on the available data and the least desired type of misclassification. Further, a lower number of features still produced equally satisfying results as the full set of features, allowing for lower data procurement costs while maintaining comparable performance levels. Additionally, leveraging external data sources can be useful for gaining additional information on the target customer base. For this data set, a max accuracy of 80% was reached. Considering RQ2, the classification system was successfully integrated into the current customer acquisition process as well as the existing CRM system. This has implications for developing traditional CRM systems into AI-CRM and how to utilize ML classification with customer data. It was found that AI-CRM integration for customer

acquisition is possible through consideration of the company's specific requirements. This study found it important to consider the data collection process and current customer acquisition process integration and to estimate misclassification risks. The company-specific requirements that enable integration can be charted by studying the existing processes and the behaviors of both the end users of a system, as well as the parties responsible for managing it. In this case, the end users were the brokers and the managing parties were the business developers. It is also important to note that bias present in customer data can be propagated through the use of ML models for customer acquisition. That present bias in a CRM system should be evaluated and mitigated when implementing AI-CRM.

7.1 Theoretical Contributions

Many of the findings in this study are congruent with other similar studies. However, the research also sheds light on the unexplored field of using ML to classify corporate customers in the insurance industry and what data could be useful in this case. The study identifies the potential predictors that are most informative for accurately classifying corporate customers — a research gap that the authors wished to address. Further, the development of an ML-based system for the study partner company provides a practical demonstration of the application of customer classification techniques. This contribution validates the theoretical concepts and offers a tangible example of the practical implications and benefits of integrating customer classification into CRM systems and AI-CRM development.

7.1.1 Practical implications

To stay competitive in a rapidly changing landscape, business insurance brokers should embrace the potential of AI technologies. As highlighted by Shollo et al. (2022), it is crucial for brokers to create tangible business value through AI projects. By leveraging machine learning (ML) algorithms, brokers can enhance their operations, improve decision-making processes, and ultimately drive better client outcomes.

Responsible and accountable deployment is a key aspect of implementing AI in the insurance sector. Asatiani et al. (2021) introduce the concept of sociotechnical envelopment as a means to implement inscrutable AI models, such as neural networks, in a safe and transparent manner. Business insurance brokers should adopt this approach

by setting clear boundaries for AI systems, meticulously curating training data, and effectively managing input and output sources. This ensures that the benefits of flexible AI models are balanced with the need for explainability and regulatory compliance.

While AI plays a crucial role in decision-making, the expertise and judgment of business insurance brokers remain invaluable. As Waardenburg, Huysman, and Sergeeva (2022) emphasized, brokers should view AI as a complementary tool rather than a replacement for human expertise. Thus, it is important to consider this paper's practical implementations as a resource rather than a substitute. By combining algorithmic outputs with their domain knowledge, brokers can provide nuanced assessments, interpret algorithmic predictions, and make well-informed decisions, especially in complex scenarios.

The role of knowledge brokers also emerges as an important consideration. Business insurance firms should foster the development of individuals who can bridge the gap between technical experts and end-users. These knowledge brokers, as highlighted by Waardenburg, Huysman, and Sergeeva (2022), play a crucial role in translating algorithmic outcomes and providing valuable insights into insurance management. Nurturing knowledge brokerage capabilities within the organization enables successful implementation and adoption of AI/ML algorithms.

Continual learning and adaptation are key in the journey of AI implementation. Business insurance brokers should maintain a culture of ongoing learning to keep up with the latest AI technologies and ML value-creation mechanisms advancements. This ensures that brokers can adapt their strategies, approaches, and capabilities as the business landscape evolves.

7.2 Limitations and Future Work

One notable challenge encountered was the issue of limited data. The data set consisted of customers already chosen by the brokers and thus suffered from slight survival bias. The data used in this study is further from 2021, which might not accurately reflect the state when acquired of the companies for those signed in 2020 and previously. The companies that were dropped as a result of missing data points also did not make it into the study. Another important limitation of the study and the current research was the absence of testing the classification models in a real-world operational context. A suggestion for

future research is to test other models, for example, a Genetic Algorithm model. Further, as this study was not able to test the implementation in a real-life setting, this would also be an interesting topic for future research. Another area for future research would be to test this type of implementation in fields other than corporate insurance.

Bibliography

Abedin, Mohammad Zoynul, Hajek, Petr, Sharif, Taimur, Satu, Md. Shahriare, and Khan, Md. Imran (2023). "Modelling bank customer behaviour using feature engineering and classification techniques". In: *Research in International Business and Finance* 65, p. 101913. ISSN: 0275-5319. DOI: https://doi.org/10.1016/j.ribaf.2023.101913. URL: https://www.sciencedirect.com/science/article/pii/S0275531923000399.

Ahearne, Michael, Hughes, Douglas E., and Schillewaert, Niels (2007). "Why sales reps should welcome information technology: Measuring the impact of CRM-based IT on sales effectiveness". In: *International Journal of Research in Marketing* 24.4, pp. 336–349. ISSN: 0167-8116. DOI: https://doi.org/10.1016/j.ijresmar.2007.09.003. URL: https://www.sciencedirect.com/science/article/pii/S016781160700047X.

Ali, Jehad, Khan, Rehanullah, Ahmad, Nasir, and Maqsood, Imran (2012). "Random forests and decision trees". In: *International Journal of Computer Science Issues (IJCSI)* 9.5, p. 272.

Alsaç, Ali, Çolak, Murat, and Keskin, Gülşen Aydin (2017). "An integrated customer relationship management and Data Mining framework for customer classification and risk analysis in health sector". In: *2017 6th International Conference on Industrial Technology and Management (ICITM)*, pp. 41–46. DOI: 10.1109/ICITM.2017.7917893.

Amin, Adnan, Al-Obeidat, Feras, Shah, Babar, Adnan, Awais, Loo, Jonathan, and Anwar, Sajid (2019). "Customer churn prediction in telecommunication industry using data certainty". In: *Journal of Business Research* 94, pp. 290–301. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2018.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0148296318301231.

Anshari, Muhammad, Almunawar, Mohammad Nabil, Lim, Syamimi Ariff, and Al-Mudimigh, Abdullah (2019). "Customer relationship management and big data enabled: Personalization customization of services". In: *Applied Computing and Informatics* 15.2, pp. 94–101. ISSN: 2210-8327. DOI: https://doi.org/10.1016/j.aci.2018.05.004. URL: https://www.sciencedirect.com/science/article/pii/S2210832718300735.

Asatiani, Aleksandre, Malo, Pekka, Nagbøl, Per Rådberg, Penttinen, Esko, Rinta-Kahila, Tapani, and Salovaara, Antti (2021). "Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems". In: *Journal of the Association for Information Systems (JAIS)* 22.2, pp. 325–252.

Aslam, Faheem, Hunjra, Ahmed Imran, Ftiti, Zied, Louhichi, Wael, and Shams, Tahira (2022). "Insurance fraud detection: Evidence from artificial intelligence and machine learning". In: *Research in International Business and Finance* 62, p. 101744. ISSN: 0275-5319. DOI: https://doi.org/10.1016/j.ribaf.2022.101744. URL: https://www.sciencedirect.com/science/article/pii/S0275531922001325.

Awad, Mariette and Khanna, Rahul (2015). "Support Vector Machines for Classification". In: *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, pp. 39–66. ISBN: 978-1-4302-5990-9. DOI: 10.1007/978-1-4302-5990-9_3. URL: https://doi.org/10.1007/978-1-4302-5990-9_3.

Bag, Surajit, Gupta, Shivam, Kumar, Ajay, and Sivarajah, Uthayasankar (2021). "An integrated artificial intelligence framework for knowledge creation and B2B marketing rational decision making for improving firm performance". In: *Industrial Marketing Management* 92, pp. 178–189. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2020.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0019850120309044.

Baghla, Seema and Gupta, Gaurav (2022). "Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce". In: *Microprocessors and Microsystems* 94, p. 104680. ISSN: 0141-9331. DOI: https://doi.org/10.1016/j.micpro.2022.104680. URL: https://www.sciencedirect.com/science/article/pii/S0141933122002101.

Behera, Rajat Kumar, Gunasekaran, Angappa, Gupta, Shivam, Kamboj, Shampy, and Bala, Pradip Kumar (2020). "Personalized digital marketing recommender engine".

In: *Journal of Retailing and Consumer Services* 53, p. 101799. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2019.03.026. URL: https://www.sciencedirect.com/science/article/pii/S0969698918307987.

Bhattacharyya, Siddhartha, Jha, Sanjeev, Tharakunnel, Kurian, and Westland, J Christopher (2011). "Data mining for credit card fraud: A comparative study". In: *Decision support systems* 50.3, pp. 602–613.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Blier-Wong, Christopher, Cossette, Hélène, Lamontagne, Luc, and Marceau, Etienne (2021). "Machine Learning in Pamp;C Insurance: A Review for Pricing and Reserving". In: *Risks* 9.1. ISSN: 2227-9091. DOI: 10.3390/risks9010004. URL: https://www.mdpi.com/2227-9091/9/1/4.

Bolton, Ruth N. (1998). "A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction". In: *Marketing Science* 17.1, pp. 45–65. ISSN: 07322399, 1526548X. URL: http://www.jstor.org/stable/193196 (visited on 02/28/2023).

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Bryzgalov, DV and Tsyganov, AA (2022). "Consumer Limitations on the Digitalization of the Insurance Market and Ways to Overcome Them." In: *Studies on Russian economic development* 33.5. ISSN: 1422-8890. DOI: 10.1134/S1075700722050057. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9491652/#Fn1.

Caruana, Rich and Niculescu-Mizil, Alexandru (2006). "An Empirical Comparison of Supervised Learning Algorithms". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 161–168. ISBN: 1595933832. DOI: 10.1145/1143844.1143865. URL: https://doi.org/10.1145/1143844.1143865.

Chatterjee, Sheshadri, Chaudhuri, Ranjan, and Vrontis, Demetris (2022). "AI and digitalization in relationship management: Impact of adopting AI-embedded CRM system". In: *Journal of Business Research* 150, pp. 437–450. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2022.06.033. URL: https://www.sciencedirect.com/science/article/pii/S0148296322005719.

Chatterjee, Sheshadri, Chaudhuri, Ranjan, Vrontis, Demetris, and Jabeen, Fauzia (2022). "Digital transformation of organization using AI-CRM: From microfoundational perspective with leadership support". In: *Journal of Business Research* 153, pp. 46–58.

ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2022.08.019. URL: https://www.sciencedirect.com/science/article/pii/S0148296322007019.

Chatterjee, Sheshadri, Rana, Nripendra P., Tamilmani, Kuttimani, and Sharma, Anuj (2021). "The effect of AI-based CRM on organization performance and competitive advantage: An empirical analysis in the B2B context". In: *Industrial Marketing Management* 97, pp. 205–219. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2021.07.013. URL: https://www.sciencedirect.com/science/article/pii/S0019850121001425.

Chen, Injazz J. and Popovich, Karen (2003). "Understanding customer relationship management (CRM): People, process and technology". In: *Business Process Management Journal* 9.5, pp. 672–688. DOI: 10.1108/14637150310496758. URL: https://doi.org/10.1108/14637150310496758.

Chen, Rung-Ching, Dewi, Christine, Huang, Su-Wen, and Caraka, Rezzy Eko (2020). "Selecting critical features for data classification based on machine learning methods". In: *Journal of Big Data* 7.1, p. 52. ISSN: 2196-1115. DOI: 10.1186/s40537-020-00327-4. URL: https://doi.org/10.1186/s40537-020-00327-4.

Chen, Yongqiang and Hu, Leifang (2005). "Study on Data Mining Application in CRM System Based on Insurance Trade". In: *Proceedings of the 7th International Conference on Electronic Commerce*. ICEC '05. Xi'an, China: Association for Computing Machinery, pp. 839–841. ISBN: 1595931120. DOI: 10.1145/1089551.1089715. URL: https://doi-org.focus.lib.kth.se/10.1145/1089551.1089715.

Chiu, Chaochang (2002). "A case-based customer classification approach for direct marketing". In: *Expert Systems with Applications* 22.2, pp. 163–168. ISSN: 0957-4174. DOI: https://doi.org/10.1016/S0957-4174(01)00052-5. URL: https://www.sciencedirect.com/science/article/pii/S0957417401000525.

Croteau, Anne-Marie and Li, Peter (2003). "Critical Success Factors of CRM Technological Initiatives". In: *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration* 20.1, pp. 21–34. DOI: https://doi-org.focus.lib.kth.se/10.1111/j.1936-4490.2003.tb00303.x. eprint: https://onlinelibrary-wiley-com.focus.lib.kth.se/doi/pdf/10.1111/j.1936-4490.2003.tb00303.x. URL: https://onlinelibrary-wiley-com.focus.lib.kth.se/doi/abs/10.1111/j.1936-4490.2003.tb00303.x.

D'Haen, Jeroen and Van den Poel, Dirk (2013). "Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework". In:

Industrial Marketing Management 42.4. Special Issue on Applied Intelligent Systems in Business-to-Business Marketing, pp. 544–551. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2013.03.006. URL: https://www.sciencedirect.com/science/article/pii/S0019850113000564.

Demarzo, Peter and Berk, Jonathan (2014). *Corporate Finance: Global Edition*. Boston; Columbus; Indianapolis etc.: Pearson Education.

Dexe, Jacob, Franke, Ulrik, and Rad, Alexander (2021). "Transparency and insurance professionals: a study of Swedish insurance practice attitudes and future development". In: *The Geneva Papers on Risk and Insurance - Issues and Practice* 46.4. ISSN: 1468-0440. DOI: 10.1057/s41288-021-00207-9. URL: https://doi.org/10.1057/s41288-021-00207-9.

Eckert, Christian, Eckert, Johanna, and Zitzmann, Armin (2021). "The status quo of digital transformation in insurance sales: an empirical analysis of the german insurance industry." In: *Zeitschrift für die gesamte Versicherungswissenschaft* 110.2-3. ISSN: 1422-8890. DOI: 10.1007/s12297-021-00507-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8628284/.

Eling, Martin and Lehmann, Martin (2018). "The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks". In: *The Geneva Papers on Risk and Insurance - Issues and Practice* 43.3. ISSN: 1468-0440. DOI: 10.1057/s41288-017-0073-0. URL: https://doi.org/10.1057/s41288-017-0073-0.

EU (2016). *General Data Protection Regulation (GDPR)*. European Union. URL: https://gdprinfo.eu/ (visited on 05/10/2023).

Fayyaz, Zeshan, Ebrahimian, Mahsa, Nawara, Dina, Ibrahim, Ahmed, and Kashef, Rasha (2020). "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities". In: *Applied Sciences* 10.21. ISSN: 2076-3417. DOI: 10.3390/app10217748. URL: https://www.mdpi.com/2076-3417/10/21/7748.

Fitzpatrick, Trevor and Mues, Christophe (2016). "An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market". In: *European Journal of Operational Research* 249.2, pp. 427–439.

Fletcher-Chen, Chavi C.-Y., Sharma, Arun, and Rangarajan, Deva (2022). "Examining supplier, buyer, and customer triads: The critical role of conflict in interaction processes and product/service innovations". In: *Industrial Marketing Management* 107, pp. 337–352. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.

2022.10.019. URL: https://www.sciencedirect.com/science/article/pii/S0019850122002589.

Galindo, Jorge and Tamayo, Pablo (2000). "Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications". In: *Computational economics* 15, pp. 107–143.

Ganesan, Shankar (1994). "Determinants of Long-Term Orientation in Buyer-Seller Relationships". In: *Journal of Marketing* 58.2, pp. 1–19. ISSN: 00222429. URL: http://www.jstor.org/stable/1252265 (visited on 03/02/2023).

Gattermann-Itschert, Theresa and Thonemann, Ulrich W. (2022). "Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests". In: *Industrial Marketing Management* 107, pp. 134–147. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2022.09.023. URL: https://www.sciencedirect.com/science/article/pii/S0019850122002255.

Géron, Aurélien. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. eng. 3rd ed. O'Reilly Media, Inc. ISBN: 1-0981-2596-7.

Grant, Eric (2012). "The Social and Economic Value of Insurance". In: URL: https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/ga2012-the_social_and_economic_value_of_insurance.pdf.

Griffin, Abbie and Hauser, John R. (1993). "The Voice of the Customer". In: *Marketing Science* 12.1, pp. 1–27. ISSN: 07322399, 1526548X. URL: http://www.jstor.org/stable/183735 (visited on 02/28/2023).

Guha, Abhijit, Grewal, Dhruv, Kopalle, Praveen K., Haenlein, Michael, Schneider, Matthew J., Jung, Hyunseok, Moustafa, Rida, Hegde, Dinesh R., and Hawkins, Gary (2021). "How artificial intelligence will affect the future of retailing". In: *Journal of Retailing* 97.1. Re-Strategizing Retailing in a Technology Based Era, pp. 28–41. ISSN: 0022-4359. DOI: https://doi.org/10.1016/j.jretai.2021.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0022435921000051.

Guo, Weien, Liu, Fang, and Zhang, Xinyue (2021). "Research on Insurance Customer Segmentation Model and Marketing Strategy Based on Big Data and Machine Learning". In: *2021 2nd International Conference on Artificial Intelligence and Information Systems*. ICAIIS 2021. Chongqing, China: Association for Computing Machinery. ISBN: 9781450390200. DOI: 10.1145/3469213.3471326. URL: https://doi-org.focus.lib.kth.se/10.1145/3469213.3471326.

Gupta, Sunil, Lehmann, Donald R., and Stuart, Jennifer Ames (2004). "Valuing Customers". In: *Journal of Marketing Research* 41.1, pp. 7–18. DOI: 10.1509/jmkr.41.1.7.25084. eprint: https://doi.org/10.1509/jmkr.41.1.7.25084. URL: https://doi.org/10.1509/jmkr.41.1.7.25084.

Hallikainen, Heli, Luongo, Milena, Dhir, Amandeep, and Laukkanen, Tommi (2022). "Consequences of personalized product recommendations and price promotions in online grocery shopping". In: *Journal of Retailing and Consumer Services* 69, p. 103088. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2022.103088. URL: https://www.sciencedirect.com/science/article/pii/S0969698922001813.

Handelman, Guy S, Kok, Hong Kuan, Chandra, Ronil V, Razavi, Amir H, Huang, Shiwei, Brooks, Mark, Lee, Michael J, and Asadi, Hamed (2019). "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods". In: *American Journal of Roentgenology* 212.1, pp. 38–43.

Häubl, Gerald and Trifts, Valerie (2000). "Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids". In: *Marketing Science* 19.1, pp. 4–21. DOI: 10.1287/mksc.19.1.4.15178. eprint: https://doi.org/10.1287/mksc.19.1.4.15178. URL: https://doi.org/10.1287/mksc.19.1.4.15178.

Henry, O. (2016). "Commercial General Liability Insurance and Coverage: A Theoretical Review". In: 5.1, pp. 509–517.

Jain, Himani, Yadav, Garima, and Manoov, R. (2021). "Churn Prediction and Retention in Banking, Telecom and IT Sectors Using Machine Learning Techniques". In: *Advances in Machine Learning and Computational Intelligence*. Ed. by Srikanta Patnaik, Xin-She Yang, and Ishwar K. Sethi. Singapore: Springer Singapore, pp. 137–156. ISBN: 978-981-15-5243-4.

James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert (2021). "An Introduction to Statistical Learning: with Applications in R". In: New York, NY: Springer US. ISBN: 978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1_8. URL: https://doi.org/10.1007/978-1-0716-1418-1_8.

Jardine, Andrew KS, Lin, Daming, and Banjevic, Dragan (2006). "A review on machinery diagnostics and prognostics implementing condition-based maintenance". In: *Mechanical systems and signal processing* 20.7, pp. 1483–1510.

Johnsson, Johan, Björnsson, Oscar, Andersson, Per, and al., et (2020). "Artificial neural networks improve early outcome prediction and risk classification in out-of-hospital

cardiac arrest patients admitted to intensive care". In: *Critical Care* 24.1, p. 474. ISSN: 1466-609X. DOI: 10.1186/s13054-020-03103-1. URL: https://doi.org/10.1186/s13054-020-03103-1.

Joshi, Ashwin W. and Sharma, Sanjay (2004). "Customer Knowledge Development: Antecedents and Impact on New Product Performance". In: *Journal of Marketing* 68.4, pp. 47–59. DOI: 10.1509/jmkg.68.4.47.42722. eprint: https://doi.org/10.1509/jmkg.68.4.47.42722. URL: https://doi.org/10.1509/jmkg.68.4.47.42722.

Khade, Anindita A. (2016). "Performing Customer Behavior Analysis using Big Data Analytics". In: *Procedia Computer Science* 79. Proceedings of International Conference on Communication, Computing and Virtualization (ICCCV) 2016, pp. 986–992. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.03.125. URL: https://www.sciencedirect.com/science/article/pii/S1877050916002568.

Kim, Eunju, Kim, Wooju, and Lee, Yillbyung (2003). "Combination of multiple classifiers for the customer's purchase behavior prediction". In: *Decision Support Systems* 34.2. Agents and E-Commerce Business Models, pp. 167–175. ISSN: 0167-9236. DOI: https://doi.org/10.1016/S0167-9236(02)00079-9. URL: https://www.sciencedirect.com/science/article/pii/S0167923602000799.

King, Stephen F. and Burgess, Thomas F. (2008). "Understanding success and failure in customer relationship management". In: *Industrial Marketing Management* 37.4, pp. 421–431. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2007.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0019850107000430.

Komiak, Sherrie Y. X. and Benbasat, Izak (2006). "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents". In: *MIS Quarterly* 30.4, pp. 941–960. ISSN: 02767783. URL: http://www.jstor.org/stable/25148760 (visited on 03/06/2023).

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E (2017). "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Kumar, V., Rajan, Bharath, Venkatesan, Rajkumar, and Lecinski, Jim (2019). "Understanding the Role of Artificial Intelligence in Personalized Engagement Marketing". In: *California Management Review* 61.4, pp. 135–155. DOI: 10.1177/0008125619859317. eprint: https://doi.org/10.1177/0008125619859317. URL: https://doi.org/10.1177/0008125619859317.

LaLonde, Steven M (2005). "Transforming variables for normality and linearity—when, how, why and why not's". In: *SAS conference proceedings NESUG*. Vol. 18, pp. 11–14.

Lemon, Katherine N. and Verhoef, Peter C. (2016). "Understanding Customer Experience Throughout the Customer Journey". In: *Journal of Marketing* 80.6, pp. 69–96. ISSN: 00222429. URL: http://www.jstor.org/stable/44134974 (visited on 03/15/2023).

Libai, Barak, Bart, Yakov, Gensler, Sonja, Hofacker, Charles F., Kaplan, Andreas, Kötterheinrich, Kim, and Kroll, Eike Benjamin (2020). "Brave New World? On AI and the Management of Customer Relationships". In: *Journal of Interactive Marketing* 51.1, pp. 44–56. DOI: 10.1016/j.intmar.2020.04.002. URL: https://doi.org/10.1016/j.intmar.2020.04.002.

Lin, Cong and Zheng, Jinju (2022). "Financial customer classification by combined model". In: *Applied Mathematics and Nonlinear Sciences*. Cited by: 0; All Open Access, Gold Open Access. DOI: 10.2478/amns.2021.2.00198. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139425515&doi=10.2478%2famns.2021.2.00198&partnerID=40&md5=19caeebb01509d65e7683d0d546a957a.

Liu, Bing and Liu, Bing (2011). "Opinion mining and sentiment analysis". In: *Web data mining: exploring hyperlinks, contents, and usage data*, pp. 459–526.

Liu, Chi-Chun and Liao, Yi-Ping (2020). "Insurance Acquisition Costs: Capitalizing Versus Expensing". In: *Journal of Accounting, Auditing & Finance* 35.3, pp. 558–580. DOI: 10.1177/0148558X18773841. eprint: https://doi.org/10.1177/0148558X18773841. URL: https://doi.org/10.1177/0148558X18773841.

Makinde, Ayodeji Samuel, Vincent, Olufunke Rebecca, Akinwale, Adio Taofik, Oguntuase, Adebayo, and Acheme, Ijegwa David (2020). "An Improved Customer Relationship Management Model for Business-to-Business E-commerce Using Genetic-Based Data Mining Process". In: Cited by: 3. DOI: 10.1109/ICMCECS47690.2020.240875. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084956099&doi=10.1109%2fICMCECS47690.2020.240875&partnerID=40&md5=bd0fb6a6b189851d26585d0f1104f26b.

Marinakos, Georgios and Daskalaki, Sophia (2017). "Imbalanced customer classification for bank direct marketing". In: *Journal of Marketing Analytics* 5.1. Cited by: 16, pp. 14–30. DOI: 10.1057/s41270-017-0013-7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85031314713&doi=10.1057%2fs41270-017-0013-7&partnerID=40&md5=cc9eb0c6878b1c8755cdf251d71837ef.

McCarthy, Daniel M. and Fader, Peter S. (2018). "Customer-Based Corporate Valuation for Publicly Traded Noncontractual Firms". In: *Journal of Marketing Research* 55.5, pp. 617–635. DOI: 10.1177/0022243718802843. eprint: https://doi.org/10.1177/0022243718802843. URL: https://doi.org/10.1177/0022243718802843.

Metsis, Vangelis, Androutsopoulos, Ion, and Paliouras, Georgios (2006). "Spam filtering with naive bayes-which naive bayes?" In: *CEAS*. Vol. 17. Mountain View, CA, pp. 28–69.

Mikalef, Patrick, Conboy, Kieran, and Krogstie, John (2021). "Artificial intelligence as an enabler of B2B marketing: A dynamic capabilities micro-foundations approach". In: *Industrial Marketing Management* 98, pp. 80–92. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2021.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0019850121001486.

Min, Sungwook, Zhang, Xubing, Kim, Namwoon, and Srivastava, Rajendra K. (2016). "Customer Acquisition and Retention Spending: An Analytical Model and Empirical Investigation in Wireless Telecommunications Markets". In: *Journal of Marketing Research* 53.5, pp. 728–744. DOI: 10.1509/jmr.14.0170. eprint: https://doi.org/10.1509/jmr.14.0170. URL: https://doi.org/10.1509/jmr.14.0170.

Müller, Daniel and Te, Yiea-Funk (2017). "Insurance premium optimization using motor insurance policies — A business growth classification approach". In: *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4154–4158. DOI: 10.1109/BigData.2017.8258437.

Nebolsina, Elena (2021). "The impact of the Covid-19 pandemic on the business interruption insurance demand in the United States". In: *Heliyon* 7.11. DOI: 10.1016/j.heliyon.2021.e08357. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8579736/.

Neslin, Scott A., Grewal, Dhruv, Leghorn, Robert, Shankar, Venkatesh, Teerling, Marije L., Thomas, Jacquelyn S., and Verhoef, Peter C. (2006). "Challenges and Opportunities in Multichannel Customer Management". In: *Journal of Service Research* 9.2, pp. 95–112. DOI: 10.1177/1094670506293559. eprint: https://doi.org/10.1177/1094670506293559. URL: https://doi.org/10.1177/1094670506293559.

Ngai, E.W.T., Xiu, Li, and Chau, D.C.K. (2009). "Application of data mining techniques in customer relationship management: A literature review and classification". In: *Expert Systems with Applications* 36.2, Part 2, pp. 2592–2602. ISSN: 0957-4174. DOI:

https://doi.org/10.1016/j.eswa.2008.02.021. URL: https://www.sciencedirect.com/science/article/pii/S0957417408001243.

Nijssen, Edwin J., Guenzi, Paolo, and van der Borgh, Michel (2017). "Beyond the retention—acquisition trade-off: Capabilities of ambidextrous sales organizations". In: *Industrial Marketing Management* 64, pp. 1–13. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2017.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0019850117302705.

O'brien, Robert M. (2007). "A Caution Regarding Rules of Thumb for Variance Inflation Factors". In: *Quality & Quantity* 41.5, pp. 673–690. ISSN: 1573-7845. DOI: 10.1007/s11135-006-9018-6. URL: https://doi.org/10.1007/s11135-006-9018-6.

Pang, Bo, Lee, Lillian, and Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment classification using machine learning techniques". In: *arXiv preprint cs/0205070*.

Paschen, Jeannette, Kietzmann, Jan, and Kietzmann, Tim Christian (2019). "Artificial intelligence (AI) and its implications for market knowledge in B2B marketing". In: *Journal of business & industrial marketing*. DOI: https://doi.org/10.1108/JBIM-10-2018-0295.

Pauch, Dariusz and Bera, Anna (2022). "Digitization in the insurance sector – challenges in the face of the Covid-19 pandemic". In: *Procedia Computer Science* 207. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022, pp. 1677–1684. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.09.225. URL: https://www.sciencedirect.com/science/article/pii/S1877050922011097.

Phillips-Wren, Gloria and Hoskisson, Angela (2015). "An analytical journey towards big data". In: *Journal of Decision Systems* 24.1, pp. 87–102. DOI: 10.1080/12460125.2015.994333. eprint: https://doi.org/10.1080/12460125.2015.994333. URL: https://doi.org/10.1080/12460125.2015.994333.

Pisoni, Galena and Díaz-Rodríguez, Natalia (2023). "Responsible and human centric AI-based insurance advisors". In: *Information Processing Management* 60.3, p. 103273. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2023.103273. URL: https://www.sciencedirect.com/science/article/pii/S0306457323000109.

Powers, David M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv: 2010.16061 [cs.LG].

Prince, Simon JD (2012). *Computer vision: models, learning, and inference*. Cambridge University Press.

Przybytniowski, Jarosław Wenancjusz, Borkowski, Stanisław, Pawlik, Andrzej, and Garasyim, Petro (2022). "The Risk of the COVID-19 Pandemic and Its Influence on the Business Insurance Market in the Medium- and Long-Term Horizon". In: *Risks* 10.5. ISSN: 2227-9091. DOI: 10.3390/risks10050100. URL: https://www.mdpi.com/2227-9091/10/5/100.

Qazi, Maleeha, Fung, Glenn M., Meissner, Katie J., and Fontes, Eduardo R. (2017). "An Insurance Recommendation System Using Bayesian Networks". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. Como, Italy: Association for Computing Machinery, pp. 274–278. ISBN: 9781450346528. DOI: 10.1145/3109859.3109907. URL: https://doi-org.focus.lib.kth.se/10.1145/3109859.3109907.

Rahim, Mussadiq Abdul, Mushafiq, Muhammad, Khan, Salabat, and Arain, Zulfiqar Ali (2021). "RFM-based repurchase behavior for customer classification and segmentation". In: *Journal of Retailing and Consumer Services* 61, p. 102566. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2021.102566. URL: https://www.sciencedirect.com/science/article/pii/S0969698921001326.

Rahman, Muhammad Sabbir, Bag, Surajit, Gupta, Shivam, and Sivarajah, Uthayasankar (2023). "Technology readiness of B2B firms and AI-based customer relationship management capability for enhancing social sustainability performance". In: *Journal of Business Research* 156, p. 113525. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2022.113525. URL: https://www.sciencedirect.com/science/article/pii/S0148296322009900.

Reinartz, Werner, Thomas, Jacquelyn S., and Kumar, V. (2005). "Balancing Acquisition and Retention Resources to Maximize Customer Profitability". In: *Journal of Marketing* 69.1, pp. 63–79. ISSN: 00222429. URL: http://www.jstor.org/stable/30162033 (visited on 03/15/2023).

Rusli, Nur Ida Aniza, Zulkifle, Farizuwana Akma, and Ramli, Intan Syaherra (2023). "A Comparative Study of Machine Learning Classification Models on Customer Behavior Data". In: *Soft Computing in Data Science*. Ed. by Marina Yusoff, Tao Hai, Murizah Kassim, Azlinah Mohamed, and Eisuke Kita. Singapore: Springer Nature Singapore, pp. 222–231. ISBN: 978-981-99-0405-1.

Rust, Roland T., Lemon, Katherine N., and Zeithaml, Valarie A. (2004). "Return on Marketing: Using Customer Equity to Focus Marketing Strategy". In: *Journal of Marketing* 68.1, pp. 109–127. DOI: 10.1509/jmkg.68.1.109.24030. eprint: https:

//doi.org/10.1509/jmkg.68.1.109.24030. URL: https://doi.org/10.1509/jmkg.68.1.109.24030.

Sadikin, Mujiono and Alfiandi, Fahri (2018). "Comparative study of classification method on customer candidate data to predict its potential risk". In: *Int. J. Electr. Comput. Eng* 8.6, pp. 4763–4771.

Satish, Laika and Yusof, Norazah (2017). "A Review: Big Data Analytics for enhanced Customer Experiences with Crowd Sourcing". In: *Procedia Computer Science* 116. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017), pp. 274–283. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.10.058. URL: https://www.sciencedirect.com/science/article/pii/S1877050917321063.

Saunders, Mark, Lewis, Philip, and Thornhill, Adrian (2015). *Research Methods for Business Students.* Vol. Seventh edition. Pearson. ISBN: 9781292016627. URL: https://search-ebscohost-com.focus.lib.kth.se/login.aspx?direct=true&db=nlebk&AN=1419381&site=ehost-live.

Saura, Irene, Frasquet, Marta, and Cervera-Taulet, Amparo (May 2009). "The value of B2B relationships". In: *Industrial Management and Data Systems* 109, pp. 593–609. DOI: 10.1108/02635570910957605.

Saura, Jose Ramon, Ribeiro-Soriano, Domingo, and Palacios-Marqués, Daniel (2021). "Setting B2B digital marketing in artificial intelligence-based CRMs: A review and directions for future research". In: *Industrial Marketing Management* 98, pp. 161–178. ISSN: 0019-8501. DOI: https://doi.org/10.1016/j.indmarman.2021.08.006. URL: https://www.sciencedirect.com/science/article/pii/S0019850121001772.

Saxena, Shikhar and Kumar, Raj (2022). "The impact on supply and demand due to recent transformation in the insurance industry". In: *Materials Today: Proceedings* 56. First International Conference on Design and Materials, pp. 3402–3408. ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.10.337. URL: https://www.sciencedirect.com/science/article/pii/S2214785321068450.

Scikit-learn Contributors (2021a). *Scikit-learn: Machine Learning in Python.* https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Accessed: May 3, 2023.

Scikit-learn Contributors (2021b). *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html. Accessed: May 3, 2023.

— (2023). *Sklearn.ensemble.randomforestclassifier*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed: 20 May 2023.

Shahid, Nida, Rappon, Tim, and Berta, Whitney (2019). "Applications of artificial neural networks in health care organizational decision-making: A scoping review". eng. In: *PLoS One* 14.2, e0212356. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0212356. URL: https://doi.org/10.1371/journal.pone.0212356.

Shmilovici, Armin (2005). "Support Vector Machines". In: *Data Mining and Knowledge Discovery Handbook*. Ed. by Oded Maimon and Lior Rokach. Boston, MA: Springer US, pp. 257–276. ISBN: 978-0-387-25465-4. DOI: 10.1007/0-387-25465-X_12. URL: https://doi.org/10.1007/0-387-25465-X_12.

Shollo, Arisa, Hopf, Konstantin, Thiess, Tiemo, and Müller, Oliver (2022). "Shifting ML value creation mechanisms: A process model of ML value creation". In: *The Journal of Strategic Information Systems* 31.3, p. 101734. ISSN: 0963-8687. DOI: https://doi.org/10.1016/j.jsis.2022.101734. URL: https://www.sciencedirect.com/science/article/pii/S0963868722000300.

Shrestha, Noora (June 2020). "Detecting Multicollinearity in Regression Analysis". In: *American Journal of Applied Mathematics and Statistics* 8, pp. 39–42. DOI: 10.12691/ajams-8-2-1. URL: https://www.researchgate.net/publication/342413955_Detecting_Multicollinearity_in_Regression_Analysis.

Simkin, Lyndon (2008). "Achieving market segmentation from B2B sectorisation". In: *Journal of Business and Industrial Marketing* 23.7. Cited by: 34, pp. 464–474. DOI: 10.1108/08858620810901220. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-51749089767&doi=10.1108%2f08858620810901220&partnerID=40&md5=65eda3dacb91c513ee57faa5370094a1.

Stoeckli, Emanuel, Dremel, Christian, and Uebernickel, Falk (2018). "Exploring characteristics and transformational capabilities of InsurTech innovations to understand insurance value creation in a digital world". In: *Electronic Markets* 28.3, pp. 287–305. ISSN: 1422-8890. DOI: 10.1007/s12525-018-0304-7. URL: https://doi.org/10.1007/s12525-018-0304-7.

Thomas, Jacquelyn S. (2001). "A Methodology for Linking Customer Acquisition to Customer Retention". In: *Journal of Marketing Research* 38.2, pp. 262–268. ISSN: 00222437. URL: http://www.jstor.org/stable/1558629 (visited on 02/25/2023).

Thomas M. Cover, Joy A. Thomas (2005). "Entropy, Relative Entropy, and Mutual Information". In: *Elements of Information Theory*. John Wiley Sons, Ltd. Chap. 2, pp. 13–55. ISBN: 9780471748823. DOI: https://doi.org/10.1002/047174882X.ch2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/047174882X.ch2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch2.

Tigari, Harish (2019). "Customer Relationship Management: A Theoretical Framework". In: pp. 555–558.

Tillmanns, Sebastian, Hofstede, Frenkel Ter, Krafft, Manfred, and Goetz, Oliver (2017). "How to Separate the Wheat from the Chaff: Improved Variable Selection for New Customer Acquisition". In: *Journal of Marketing* 81.2, pp. 99–113. ISSN: 00222429. URL: http://www.jstor.org/stable/44879028 (visited on 05/22/2023).

Tohidi, Hamid and Jabbari, Mohammad Mehdi (2012). "The Necessity of Using CRM". In: *Procedia Technology* 1. First World Conference on Innovation and Computer Sciences (INSODE 2011), pp. 514–516. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2012.02.110. URL: https://www.sciencedirect.com/science/article/pii/S2212017312001119.

Tomczyk, Przemysław, Doligalski, Tymoteusz, and Zaborek, Piotr (2016). "Does customer analysis affect firm performance? Quantitative evidence from the Polish insurance market". In: *Journal of Business Research* 69.9, pp. 3652–3658. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2016.03.026. URL: https://www.sciencedirect.com/science/article/pii/S0148296316300467.

Vanneschi, Leonardo, Farinaccio, Antonella, Mauri, Giancarlo, Antoniotti, Marco, Provero, Paolo, and Giacobini, Mario (2011). "A comparison of machine learning techniques for survival prediction in breast cancer". In: *BioData mining* 4.1, pp. 1–13.

Verdonck, Tim, Baesens, Bart, Óskarsdóttir, María, and Broucke, Seppe vanden (Aug. 2021). "Special issue on feature engineering editorial". In: *Machine Learning*. ISSN: 1573-0565. DOI: 10.1007/s10994-021-06042-2. URL: https://doi.org/10.1007/s10994-021-06042-2.

Waardenburg, Lauren, Huysman, Marleen, and Sergeeva, Anastasia V. (2022). "In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms". In: *Organization Science* 33.1, pp. 59–82. DOI: 10.1287/

orsc.2021.1544. eprint: https://doi.org/10.1287/orsc.2021.1544. URL: https://doi.org/10.1287/orsc.2021.1544.

Webster, Frederick E. (1992). "The Changing Role of Marketing in the Corporation". In: *Journal of Marketing* 56.4, pp. 1–17. ISSN: 00222429. URL: http://www.jstor.org/stable/1251983 (visited on 03/02/2023).

Wilson, Hugh, Daniel, Elizabeth, and McDonald, Malcolm (2002). "Factors for Success in Customer Relationship Management (CRM) Systems". In: *Journal of Marketing Management* 18.1-2, pp. 193–219. DOI: 10.1362/0267257022775918. eprint: https://doi.org/10.1362/0267257022775918. URL: https://doi.org/10.1362/0267257022775918.

Winer, Russell S. (2001). "A Framework for Customer Relationship Management". In: *California Management Review* 43.4, pp. 89–105. DOI: 10.2307/41166102. eprint: https://doi.org/10.2307/41166102. URL: https://doi.org/10.2307/41166102.

Wong, Wing, Lee, Yun-Tien, Ko, Shenglan, and Yau, Kwan (Dec. 2020). *Life insurance sales recommender system: December 2020*. URL: https://us.milliman.com/en/insight/life-insurance-sales-recommender-system-december-2020.

Xu, Yurong, Yen, David C., Lin, Binshan, and Chou, David C. (2002). "Adopting customer relationship management technology". In: *Industrial Management Data Systems* 102.8, pp. 442–452. DOI: 10.1108/02635570210445871. URL: https://doi.org/10.1108/02635570210445871.

Young, Louise, Wilkinson, Ian, and Smith, Andrew (2015). "A Scientometric Analysis of Publications in the Journal of Business-to-Business Marketing 1993–2014". In: *Journal of Business-to-Business Marketing* 22.1-2, pp. 111–123. DOI: 10.1080/1051712X.2015.1021591. eprint: https://doi.org/10.1080/1051712X.2015.1021591. URL: https://doi.org/10.1080/1051712X.2015.1021591.

Zeeland-van der Holst, Eveline Maria van and Henseler, Jörg (2018). "Thinking outside the box: a neuroscientific perspective on trust in B2B relationships". In: *IMP Journal* 12.1, pp. 75–110. DOI: 10.1108/IMP-03-2017-0011. URL: https://doi.org/10.1108/IMP-03-2017-0011.

Zerbino, Pierluigi, Aloini, Davide, Dulmin, Riccardo, and Mininno, Valeria (2018). "Big Data-enabled Customer Relationship Management: A holistic approach". In: *Information Processing Management* 54.5. In (Big) Data we trust: Value creation in knowledge organizations, pp. 818–846. ISSN: 0306-4573. DOI: https://doi.org/

10.1016/j.ipm.2017.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0306457317300067.

Zong, C., Xia, R., and Zhang, J. (2021). *Text Data Mining*. Springer Nature Singapore. ISBN: 9789811601002. URL: https://books.google.se/books?id=EbUvEAAAQBAJ.

Appendix A

Variables

The total data set consisted of 32 variables, and all data is from the year 2021.

Variable Index	Variable Description
v1	Registrations year
v2	Turnover
v3	Year's result
v4	Industry Code (SNI)
v5	Offer
v6	Net turnover
v7	Other turnover
v8	Operating profit (EBIT)
v9	Result after financial items
v10	Year's result
v11	Assets
v12	Subscribed equity capital
v13	Fixed assets
v14	Current assets
v15	Equity
v16	Undistributed reserves
v17	Provisions
v18	Long-term liabilities
v19	Short-term liabilities
v20	Liabilities and equity
v21	Turnover
v22	Number of employees
v23	Net turnover per employee
v24	Personnel costs per employee
v25	Operating profit (EBITDA)
v26	Net turnover change
v27	DuPont Model
v28	Profit margin
v29	Gross profit margin
v30	Working capital/turnover
v31	Solvency
v32	Cash liquidity

Appendix B

Data Exploration

A lot of data exploration was done in order to understand the data before selecting the best features and building the models. This Appendix contains both numerical and categorical data exploration.

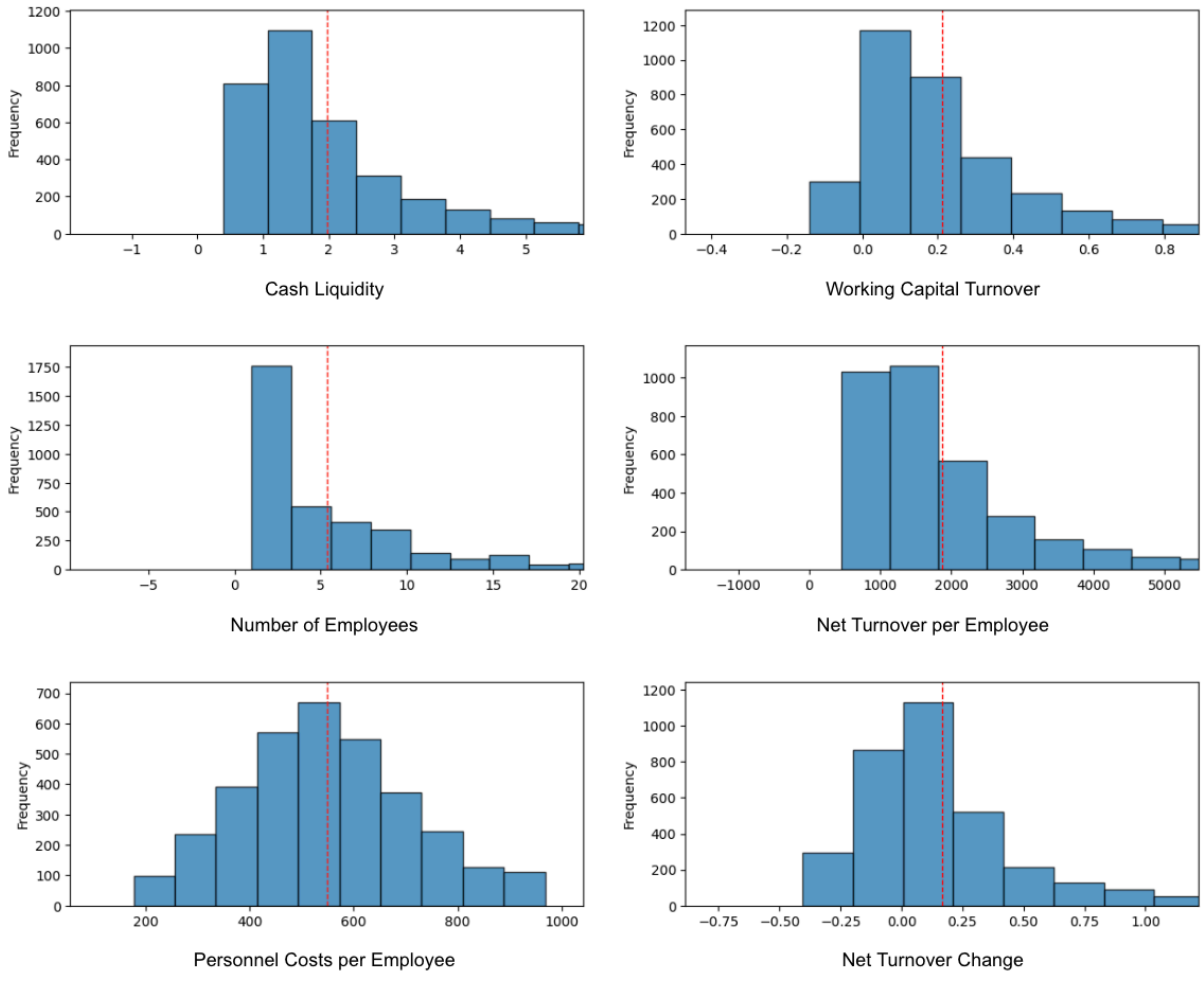


Figure B.0.1: Frequency Distributions of 6 Numerical Variables

The correlation heat map contains all 32 variables (v1-v32) visible in Appendix A, including the target variable, all numerical variables gathered from financial statements, and industry code. From the figure, it is clear that there are a lot of highly correlated variables.

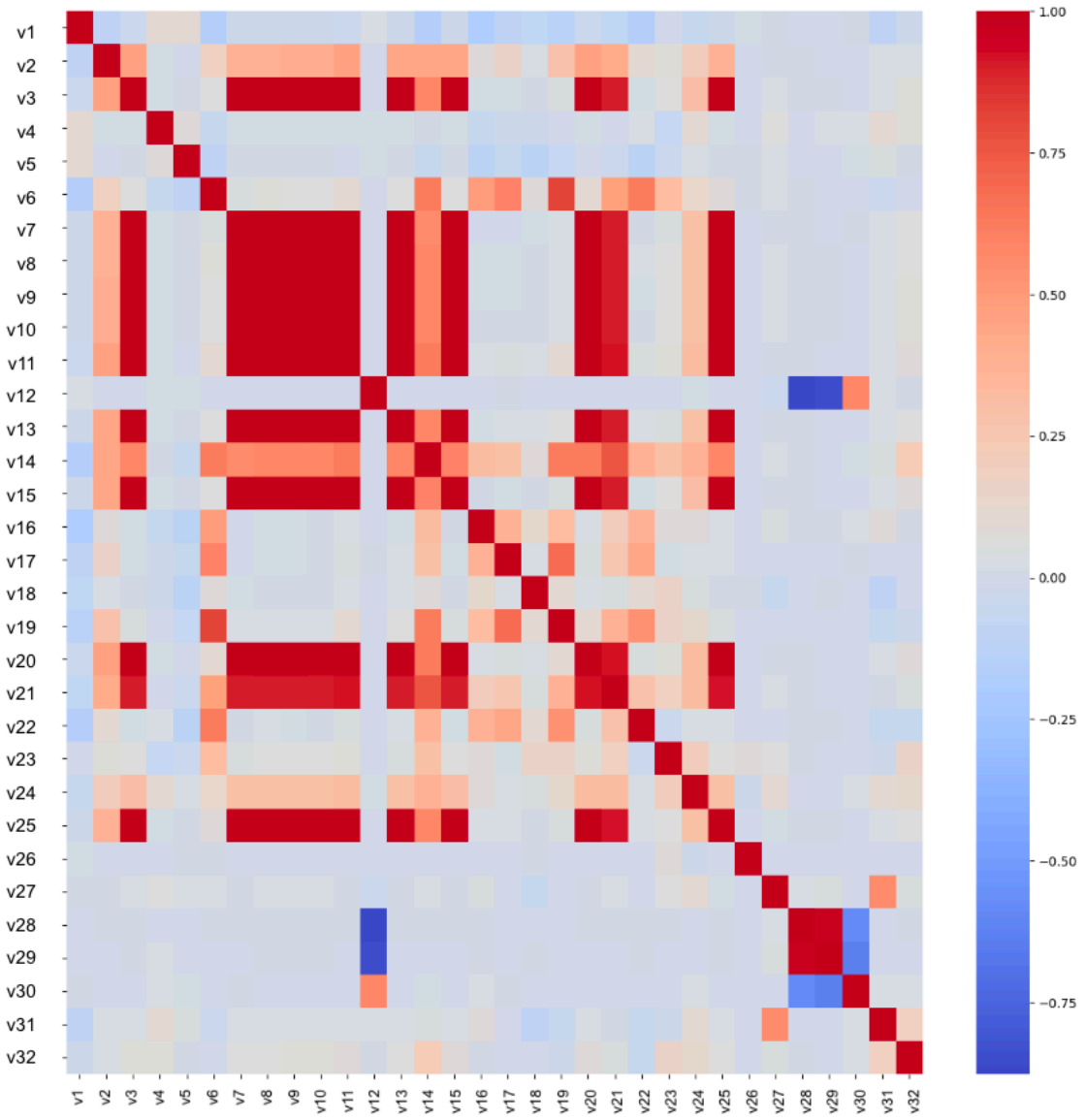


Figure B.0.2: Correlation Heat Map of all Variables

Figure B.0.3 visualizes the frequency of grouped (two numbered) SNI codes that occurred more than 20 times in the data set and how many of the companies with those codes received versus did not receive an offer for the insurance product.

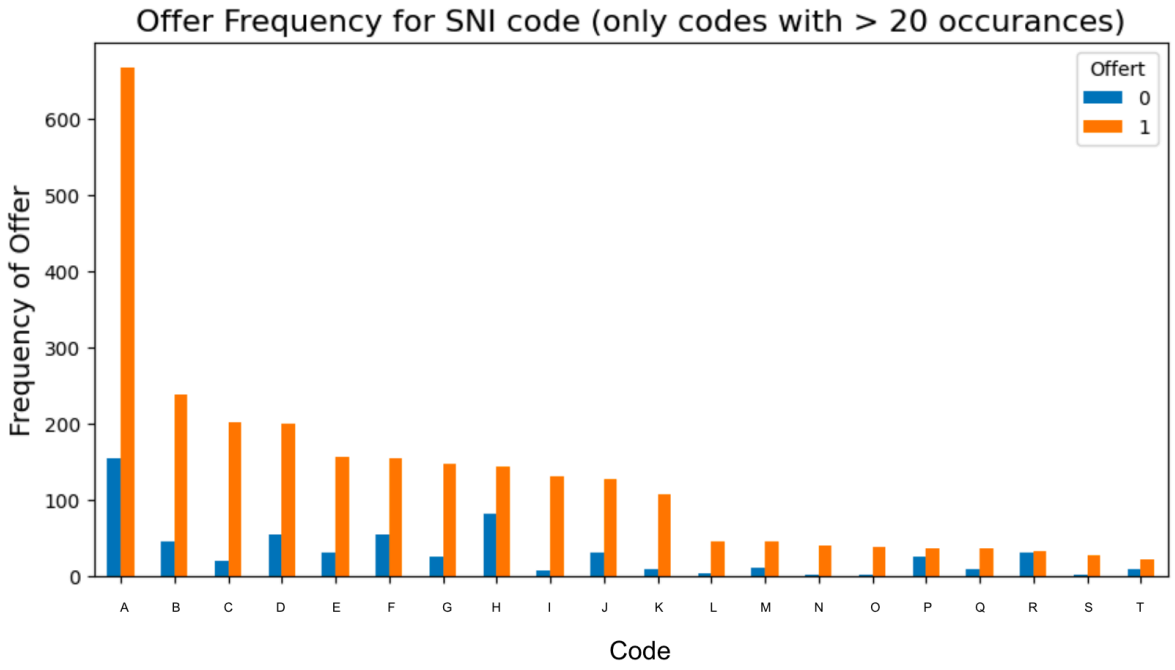


Figure B.0.3: SNI Code Frequency of Offer (*NOTE: No real SNI codes are displayed due to confidentiality*)

Appendix C

Interview Questions

These are the questions used for the one semi-structured interview. Other topics and questions arose besides these, but they were the prepared ones:

1. What does the process look like today?
2. How do you choose potential customers?
3. Which criteria or parameters do you go by?
4. For which reasons do you choose not to contact a potential customer?
5. Which tools do you currently use in the process?
6. Do you have a list of or data on the customers which got excluded earlier in the selection process and thus never made it in the current data set?
7. What challenges do you currently face is the customer acquisition process?
8. Could you talk about what is currently difficult?
9. Could you talk about what is currently straightforward?
10. What typically goes according to plan?
11. How do you get the information on the potential customers?
12. From your perspective, do you have any suggestions for how the process could be improved?