Master of Science in Industrial Economics
June 2021

Customer churn prediction using machine learning

A study in the B2B subscription based service context

Benjamin Ghaffari & Yasin Osman

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Engineering at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Master of Science in Industrial Economics. The thesis is equivalent to 20 weeks of full-time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):

Benjamin Ghaffari
E-mail: Begh16@student.bth.se

Yasin Osman
E-mail: Yaos16@student.bth.se

University advisor:
Shahiduzzaman Quoreshi
Department of Industrial Economics

Faculty of Engineering
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone: +46 455 38 50 00
Fax: +46 455 38 50 57

ACKNOWLEDGMENTS

We would like to express our gratitude to our supervisors at Blekinge Tekniska Högskola and Fortnox AB. Shahiduzzaman Quoreshi, we want to thank you for your guidance, support, and contribution to our study. To our external supervisor, Johan Hagelbäck, we are eternally grateful for your support and wisdom. Your willingness to contribute to our study with your unique knowledge has been a key factor in our success, please accept our deepest thanks.

ABSTRACT

The rapid growth of technological infrastructure has changed the way companies do business. Subscription based services are one of the outcomes of the ongoing digitalization, and with more and more products and services to choose from, customer churning has become a major problem and a threat to all firms. We propose a machine learning based churn prediction model for a subscription based service provider, within the domain of financial administration in the business-to-business (B2B) context. The aim of our study is to contribute knowledge within the field of churn prediction. For the proposed model, we compare two ensemble learners, XGBoost and Random Forest, with a single base learner, Naïve Bayes. The study follows the guidelines of the design science methodology, where we used the machine learning process to iteratively build and evaluate the generated model, using the metrics, accuracy, precision, recall, and F1-score. The data has been collected from a subscription-based service provider, within the financial administration sector. Since the used dataset is imbalanced with a majority of non-churners, we evaluated three different sampling methods, that is, SMOTE, SMOTEENN and RandomUnderSampler, in order to balance the dataset. From the results of our study, we conclude that machine learning is a useful approach for prediction of customer churning. In addition, our results show that ensemble learners perform better than single base learners and that a balanced training dataset is expected to improve the performance of the classifiers.

Keywords: *Customer churning, Machine Learning, business-to-business, subscription-based companies.*

SAMMANFATTNING

Den snabba tekniska utvecklingen har förändrat hur företag bedriver sina verksamheter. Prenumerationsbaserade tjänster är en av följderna av den pågående digitaliseringen, och med allt fler produkter och tjänster att välja mellan har kundavbrott (eng: customer churning) blivit både ett problem och hot för företag. I vår studie, presenterar vi en maskininlärningsmodell för att prediktera kundavbrott inom sektorn för ekonomisk administration i ett business-to-business sammanhang. Syftet med vår studie är att bidra med kunskap relaterat till prediktering av kundavbrott inom business-to-business. För den föreslagna prediktionsmodellen jämför vi två olika ensemble learners, XGBoost och Random Forest, med Naïve Bayes som är en single base learner. Studien följer riktlinjerna för design science metodiken, där vi använde maskininlärningsprocessen för att iterativt bygga och utvärdera den genererade modellen med användning av olika utvärderingsmått. Vi har använt data från en prenumerationsbaserad tjänsteleverantör inom sektorn för ekonomisk administration. Då den använda datamängden är obalanserad med en minoritet av kunder som sagt upp sina tjänster, utvärderade vi tre olika metoder för balansering av datamängden, det vill säga, SMOTE, SMOTEENN och RandomUnderSampler. Från resultaten av vår studie drar vi slutsatsen att maskininlärning är ett användbart verktyg för att prediktera kundavbrott. Våra resultat visar även att ensemble learners presterar generellt bättre än single base learners samt att en balanserad datamängd förväntas leda till förbättrad algoritm prestanda.

Nyckelord: *Customer churning, Machine Learning, business-to-business, Prenumerationsbaserade tjänsteleverantörer.*

TABLE OF CONTENTS

ACKNOWLEDGMENTS III

ABSTRACT IV

SAMMANFATTNING V

TABLE OF CONTENTS VI

LIST OF FIGURES VII

LIST OF TABLES VIII

1 INTRODUCTION 1

1.1 PROBLEM FORMULATION 2
1.2 PURPOSE 3
1.3 RESEARCH QUESTION 4
1.4 DELIMITATIONS 4
1.5 OUTLINE 5

2 THEORETICAL CONCEPTS 6

2.1 CUSTOMER RELATIONSHIP MANAGEMENT (CRM) 6
2.2 CUSTOMER CHURN MANAGEMENT 6
2.3 MACHINE LEARNING 7
2.3.1 Supervised Learning 7
2.3.2 Ensemble Learning 7
2.3.3 Random Forest classifier 9
2.3.4 XGBoost classifier 10
2.3.5 Naïve Bayes classifier 10
2.3.6 Imbalanced Learning 10
2.4 EVALUATION SCORE 11

3 RELATED WORK 13

3.1 TIME TO CHURN PREDICTION 13
3.2 BINARY CHURN PREDICTION IN THE DOMAIN OF B2C 14
3.3 BINARY CHURN PREDICTION IN THE DOMAIN OF B2B 16

4 RESEARCH METHOD 17

4.1 LITERATURE STUDY 17
4.2 DESIGN SCIENCE 17
4.2.1 The Machine Learning Process 18
4.2.2 Implementation 25

5 RESULTS & ANALYSIS 26

5.1 COMPARISON OF CLASSIFIERS 30

6 DISCUSSION 31

6.1 PERFORMANCE OF CHURN PREDICTION MODEL 31
6.2 VALIDITY THREATS 33
6.3 IMPLICATIONS 34

7 CONCLUSION AND FUTURE WORK 36

7.1 FUTURE WORK 36

8 REFERENCES 2

LIST OF FIGURES

FIGURE 1. VISUALIZATION OF THE BAGGING APPROACH. 8

FIGURE 2. VISUALIZATION OF THE BOOSTING APPROACH. 9

FIGURE 3. RANDOM FOREST CLASSIFIER. 9

FIGURE 4. CONFUSION MATRIX FOR BINARY CLASSIFICATION. 11

FIGURE 5. THE MACHINE LEARNING PROCESS. 18

FIGURE 6. DISTRIBUTION OF NON-CHURNERS AND CHURNERS IN THE DATASET. 19

FIGURE 7. DISTRIBUTION OF NUMBER OF EMPLOYEES IN THE DATASET. 20

FIGURE 8. DISTRIBUTION OF INDUSTRY CODES IN THE DATASET 20

FIGURE 9. DISTRIBUTION OF CORPORATE FORMS IN THE DATASET. 21

FIGURE 10. DISTRIBUTION OF THE TYPE OF CUSTOMER IN THE DATASET. 21

FIGURE 11. DISTRIBUTION OF NUMBER OF INVOICES CREATED PER MONTH IN THE DATASET 21

FIGURE 12. DISTRIBUTION OF NUMBER OF SUBSCRIBED LICENSES IN THE DATASET. 22

FIGURE 13. DISTRIBUTION OF SUBSCRIPTION START YEAR FOR THE CUSTOMERS IN THE DATASET. 22

FIGURE 14. AN EXAMPLE OF A BINARY REPRESENTATION OF A CATEGORICAL VARIABLE CONSISTING OF FOUR DIFFERENT VALUES. 23

LIST OF TABLES

TABLE 1. OVERALL PERFORMANCE OF THE THREE CONSIDERED CLASSIFIERS 27

TABLE 2. CONFUSION MATRIXES AND THE RELATED METRICS FOR THE NAÏVE BAYES CLASSIFIER 28

TABLE 3. CONFUSION MATRIXES AND THE RELATED METRICS FOR THE RANDOM FOREST CLASSIFIER. 29

TABLE 4. CONFUSION MATRIXES AND THE RELATED METRICS FOR THE XGBOOST CLASSIFIER 30

1 INTRODUCTION

The development and digitalization of the world has led to new ways of doing business and companies all over the globe have been forced to adapt [1]. Subscription based services are one of the outcomes of the explosive digitalization that has taken the world by storm and with this comes both possibilities and challenges that require modern day solutions [2]. Digitalization has not only changed the way business is conducted but the abundance of information available has also led to consumers facing a higher supply of subscription-based services. This can be viewed as a challenge for companies since retaining customers can potentially become more difficult. Digitalization within companies can lead to a decrease in labor costs, an increase in efficiency and a better overview of the company's operations within the organization [1]. All of this is essential for staying competitive, and to gain an edge over other companies.

As information technology is a growing trend, the available amount of data and information has increased significantly during the past years. This rapid growth has enabled storage and processing of great amounts of data while increasing the necessity of automatically finding valuable information and creating knowledge [3]. With meaningful information extracted from the stored data, firms can make appropriate decisions in order to grow the business. With this growth, the use of data mining techniques and machine learning has increased, due to its ability of handling and analyzing great amounts of data [4]. The digitalization has also brought forward an ongoing trend to improve current data processing activities as a part of customer relationship management (CRM) strategies. The idea of knowledge management and customer relationship management has lately obtained more attention in the subscription-based business model and the concepts focus on distribution of resources to activities that are customer-centric, to be able to increase competitive advantages [5, 6]. Customer knowledge is the knowledge that businesses can obtain through interaction with their customers [7]. Customer relationship management systems are systems that support interaction between businesses and customers with the objective of collecting, storing and analyzing data to get an overview of their customers [7]. Such systems have evolved through the past years and by using technology and different data analysis tools, enterprises can find patterns in customer’s behavior, which would be almost impossible to discover manually [3]. Such patterns could vary from a customer's purchasing behavior or patterns related to customer churning. In a subscription-based business model, a fundamental part of success is to minimize the rate of customers ending their subscriptions, in other words, to minimize churn [3].

Customer churning refers to the action of when a customer chooses to abandon their service provider [8]. The term is relatively new and has gained more relevance with the emergence of online services. Firms across the globe recognize customer churning as a great loss since they have already invested in attracting these customers. This is one of the major reasons that customer retention is beneficial for a firm. Customers can churn for many reasons and it is hard to pinpoint a general reason for churning. The availability of information has given consumers a bargaining power, and nowadays customers can easily find the service provider, which provides the same product with a more satisfying deal [9]. To manage this, firms invest in customer churn prediction, which means that companies try to predict which of their customers will churn, so that they can apply preventative measures. These preventive measures could differ depending on the reason a customer might churn, and could be for example, offering a lower price or including an extra service. As mentioned earlier, analyzing customer behavior serves as the basis for predicting customers who might churn, which is important for many reasons. One reason is that for companies who rely on subscription-based income, it can make a big difference on whether they can keep a steady income level or if they need to make changes to their services to keep customers. Another reason is that, compared to retaining customers, attracting new ones is costlier and firms can save money by retaining their existing customer base [10].

1.1 Problem Formulation

The subscription-based business model is continuously growing due to digitalization and offers companies an innovative way of conducting their business [2]. At the same time, more and more services are being digitalized and data has become much easier to collect, store, and process [6]. There is an abundance of different service providers to choose from, which has increased competition and made it more difficult to retain customers, in this modern-day service market [2]. Due to the availability of data and substitutes, subscription-based businesses must adapt by focusing more on Customer Relationship Management, specifically customer churn management [6]. According to Blank & Hermansson [2] the key to success within a subscription-based business is to keep a low churn rate, which is defined as the number of customers leaving their service provider during a given period of time [2]. As a service provider, there is a greater chance of selling to an existing customer rather than a completely new one. This can be highlighted by the cost of attracting new customers. According to Verbeke et al. [11], attracting new customers can cost somewhere between five to six times more relative to retaining customers, which states the importance of preventing churn. Apart from the cost of attracting and retaining customers, the net return on investments (ROI) for strategies related to retention is in general higher than for acquisition, which may increase revenue of customers who continue transacting [12]. The importance of preventing churn is greater in business-to-business (B2B) context where each customer makes more frequent purchases with higher transactional value [13]. Due to this high transactional value that B2B customers stand for, retention in this context could be highly financially rewarded [12].

Churn prediction and prevention can be very beneficial for a firm's reputation and according to Amin et al. [14], a better reputation increases revenue. Since profits are derived from customers, a decrease in the churn rate has a significant impact on the firm's profits [15]. Customer relationship management can lead to a better and more valuable relationship to the customers and create some sort of loyalty towards the company, which could increase revenue streams over time [16].

Customer retention strategies used to minimize churn, is primarily by offering retention incentives to reduce the churn rate. Considering that such strategies are not free and the fact that all customers do not have the incentives to churn, firms should not be focusing on the entire customer base while developing retention strategies. Rather, identify which customers are more likely to churn and target those customers with retention incentives and avoid unnecessary use of resources and costs [17]. Identifying churn is commonly defined as a prediction problem, and in recent years, machine learning has been utilized in many fields of research, and it is proven to be useful to many complex problems [18]. Customer churn prediction is not an exception and in the domain of business-to-customer (B2C), the use of machine learning has been highlighted due to its accurate prediction characteristics [19]. According to Cunningham [20], traditional statistical techniques are often used to extract information from data. However, such techniques require an expertise that few domain experts have. Furthermore, traditional statistical techniques are typically time consuming and machine learning is developed in order to automate the information extraction process and make statistical analysis more effective. It should, however, be noted that it is not possible to completely separate machine learning from traditional statistical methods as most machine learning algorithms integrate statistical methods.

Previous literature has thoroughly investigated the problem of churn prediction related to subscription-based services in the context of B2C. However, in the context of B2B, churn prediction is not widely researched. Previous studies have highlighted the importance of predicting churn in B2B but only a few numbers of techniques and approaches for churn prediction are proposed. The use of machine learning in order to predict churning in the context of B2B is underdeveloped as well as churn prediction within subscription based services, and our research aims to fill the gap and propose different techniques and approaches for churn prediction, specifically in the B2B subscription-based service context.

1.2 Purpose

As mentioned above, due to the digitalization and the access to information in large proportions, companies are under constant threats of customer churning, in particular subscription-based service providers that rely on such incomes. Unlike the business-to-customer domain, business-to-business characteristics, such as higher transactional value of each customer, constitutes a greater negative impact on the revenue stream when customers churn [20]. Hence, it is important to study churn prediction and to support companies in this problem. The purpose of our study is to contribute knowledge in the field of churn prediction in a B2B subscription based service context.

In this study we employ different machine learning techniques, in order to investigate the effects on the performance of churn prediction models in the B2B subscription-based service context. In particular, the aim is to propose and evaluate different approaches and techniques of how machine learning can be used to predict churn; hence creating value for businesses offering subscription-based services. The objectives of this study are to: i) construct a binary churn prediction model, ii) compare different types of algorithms for the built model, and iii) study the effects of balancing the training dataset for the considered model.

1.3 Research Question

As discussed above, the literature clearly shows that machine learning can be used to predict customer churning in the B2C context [10, 21, 22, 23]. It also indicates that machine learning can be used to predict customer churn in the B2B context, see for example [9, 13]. We therefore focus on the question of how rather than if machine learning can be used to predict customer churning in the B2B context.

In order to fulfill the objectives of this study, we have formulated the following research question:

How can customer churning be predicted in the domain of business-to-business using machine learning?

In order to address this research question, we study different aspects. Of particular importance is to determine an appropriate model for binary churn prediction, to study which algorithm is most appropriate, and to study data processing issues including balancing of the dataset. This is of particular importance in customer churn prediction, where non-churners are typically dominating.

1.4 Delimitations

This research was conducted in collaboration with a company that works within the financial administration sector. This study will be limited to B2B enterprises for multiple reasons, one being that the company being used for this study only deals with other companies. Another reason would be that customer churning within B2B is less studied compared to B2C. Due to the time limit and complexity, three different supervised learning algorithms, based on previous literature, were chosen for comparison and evaluation. In a perfect world there are a lot of different algorithms and techniques that can be used for predicting customer churning and using a number of these algorithms means that they are not all utilized, which is why this is considered a delimitation. Due to limitations in raw data, attrition rate will not be taken into account. Predicting customer churning can be used to lower costs and improve a firm's revenue streams, however this study will not cover such economical aspects.

1.5 Outline

This outline gives an overview of the main points in our thesis

Theoretical Concepts
This section presents theoretical concepts related to churn prediction and machine learning. The presented concepts are relevant for further understanding this study.

Related Work
This section gives an insight to this field of study and presents previously used approaches for churn prediction. Related work is based on relevant research conducted for the reason of investigating customer churn prediction.

Research Method
Research method defines the used methodology and gives a complete overview of the methods and the data used for this study.

Results & Analysis
In this section the results of this study are presented, and the performance of the different classifiers is highlighted.

Discussion
In this section we discuss our findings and their relevance to our research question.

Conclusion & Future Work
This section concludes our study and highlights the most important aspects of the result. Suggestion on future work is also given.

2 THEORETICAL CONCEPTS

In this section we present a couple of theoretical concepts in order to broaden the existing knowledge of the reader. The understanding of the presented concepts below is essential to fully understand the study.

2.1 Customer relationship management (CRM)

Customer relationship management (CRM) can be viewed as a strategy of a company, a strategy with the aim of reducing costs, improving profitability, and improving customer relations by offering the right product or service to the right customer [24]. CRM is often associated with customer knowledge since it is a major aspect of CRM [7]. CRM can be done in different ways, meaning that the method or process used by companies for CRM differs [24, 25]. However, the main principle of CRM stays the same, which is to attract customers, learn about them, find the most suitable way of serving them, and then use this knowledge to retain them [24]. This is done through something called CRM systems which is the technological part of it. The purpose of these systems is to enable communication with the customer but to also store and analyze customer data in order to get an overview of the firm's customer base [25].

2.2 Customer churn management

Customer churn management is the term used to define customer attrition. It is the concept of identifying which customers might churn and has become a core strategy to survive within an industry [17]. Customer churn management can be used to identify potential churners and target those customers with proactive marketing campaigns with retention incentives [26].

According to Hadden et al. [17], customers who churn can be divided into voluntary and non-voluntary churn. Non-voluntary churners are the ones that had their service removed by the company, which are the easiest churners to identify. Voluntary churn is unlike, non-voluntary churn, harder to identify and occurs when the customer makes a decision to abandon the service provider. Such churn could occur because of changes in circumstances, such as a customer's financial situation, in other words incidental churn. Other reasons for voluntary churn could be the customer's intentions of changing service provider to a competitor, which is called deliberate churn.

Hung et al. [26] state that churn management is considered a part of customer relationship management and from a business intelligence perspective, churn management includes two tasks. The first is to predict customers who might churn, and the second task is to identify possible retention strategies from an organization's point of view.

2.3 Machine Learning

Machine learning is a part of artificial intelligence and thereby a part of data science, which focuses on developing artifacts in the form of algorithms that learn from data and experience [27] [28]. By training algorithms on data, they can improve their decision-making and the accuracy of their predictions over time [27]. Machine learning is a frequently used approach considering automation of a variety of tasks and is categorized in different types of learning based on how the algorithms learn to become better in form of accuracy. These types of learning are usually categorized as the following: Supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning and based on the scope of the problem and the type of data used, one algorithm could be more suitable than the others [29, 30].

According to Simeone [29] and Mehta et al. [28], supervised learning is often described as learning by examples, meaning that the algorithm uses labeled data for training. Unlike supervised learning, unsupervised learning is a technique where the algorithms do not learn from a supervisor, meaning that the dataset is unlabeled. The algorithms need to find their own patterns from the input. Semi-supervised learning combines the above mentioned techniques but the algorithms use more of the unlabeled dataset than the labeled for training. Reinforcement learning is a technique, where the algorithm learns by employing a reward and punishment system to reach an end goal.

2.3.1 Supervised Learning

Supervised learning is considered a sub-category of machine learning, which is frequently used within prediction problems [27]. Supervised learning can be described as learning by examples, which means that the algorithms are trained by a labeled dataset [31]. A labeled dataset is a dataset, which consists of both inputs and outputs, and the main idea is to train the algorithm by mapping these inputs and outputs from the dataset [31]. This type of learning works by finding patterns in the dataset in order to identify outputs of unseen instances [32]. Supervised learning can be further divided into regression and classification [33]. A regression problem is where the output is a continuous value, such as the days until a potential customer might churn. Classification is on the other hand used when the output is categorical, for instance churn or not churn [33].

2.3.2 Ensemble Learning

Ensemble learning is a common machine learning technique used to combine different learning models, often referred to as base learners and combine their output into one classifier [34]. Ensemble learning is often defined as a combination of weak machine learning models, which are implemented together as one stronger model and make very accurate predictions [20, 35]. Weak learners, also called base learners, are models typically just better than a random guess.

Ensemble learning is rapidly becoming a standard choice of machine learning models in the world of data science with the aim of boosting the accuracy of predictions [35]. According to

Van Wezel & Potharst [36], ensemble learning models give more accurate predictions than individual models do. Two techniques within ensemble learning are bagging and boosting, which both generally increase the predictive performance over a single machine learning model [20].

Bagging, also known as bootstrap aggregation, is an ensemble learning technique, which creates training subsets of the original training set. This procedure is done repeatedly, and the model is trained on every new generated subset. The final prediction is done through combining all the individual outputs either through majority voting in classification problems or averaging in the case of regression, see Figure 1 [20, 37, 38]. Bagging is an effective technique when it comes to reducing variance, which is a common problem with classifiers such as decision trees [20, 38].

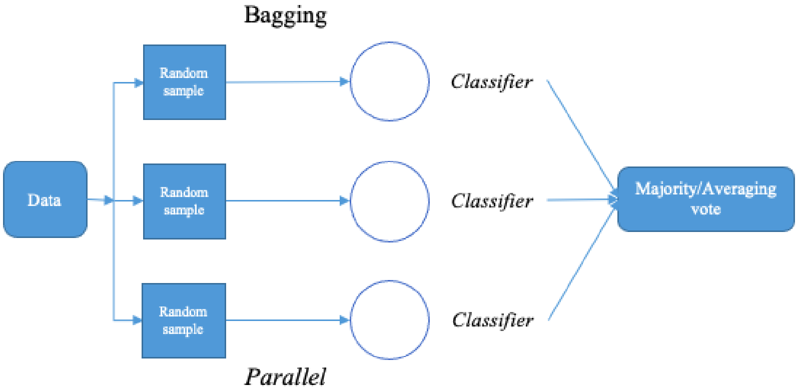


Figure 1. Visualization of the bagging approach.

Boosting is an ensemble learning technique, which seeks to combine weak learners into one strong learner with improved prediction accuracy. Boosting works by combining the weak learners in sequences, where each is influenced by its predecessor based on the weakness and the performance of the model, see Figure 2. The correctly predicted outcomes are given a lower weight and the misclassified ones are given a greater weight in the next sequence. Just like bagging, boosting increases the predictive performance of a model but in this case, in terms of reducing bias [20].

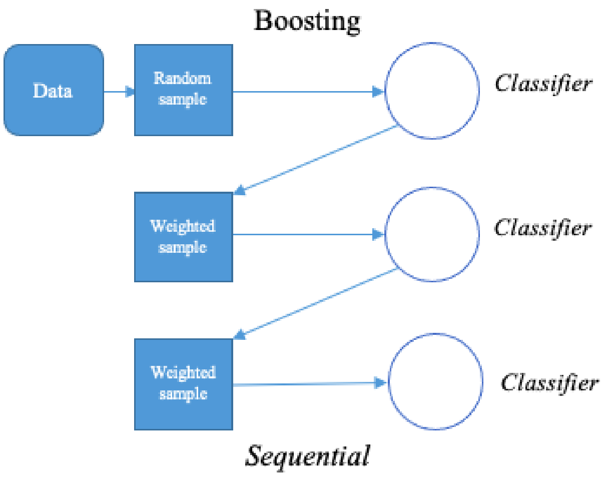


Figure 2. Visualization of the boosting approach.

2.3.3 Random Forest classifier

The Random Forest classifier is an ensemble learner, which is considered to be a supervised learning algorithm. The Random Forest classifier utilizes the bagging approach consisting of multiple classifiers, see Figure 1, where the base learners are decision trees. Each decision tree takes a random sample of the original training dataset with replacement, which means that some data is used more than once in training. However, all features are not utilized for every decision tree. In the Random Forest classifier, the division of the nodes in the trees are based on the best split of a random subset of features rather than using the best split considering all features. This reduces the correlation between the trees and decreases the generalization error. The diversity of each tree is increased in Random Forest due to the fact that different subsets of the training data is used for each tree, which leads to the classifier being more stable and robust towards noise and overtraining. Each individual decision tree in the Random Forest contributes to the final output by majority voting. The class with the most votes is chosen as the final output, see Figure 3 [39]. [40]

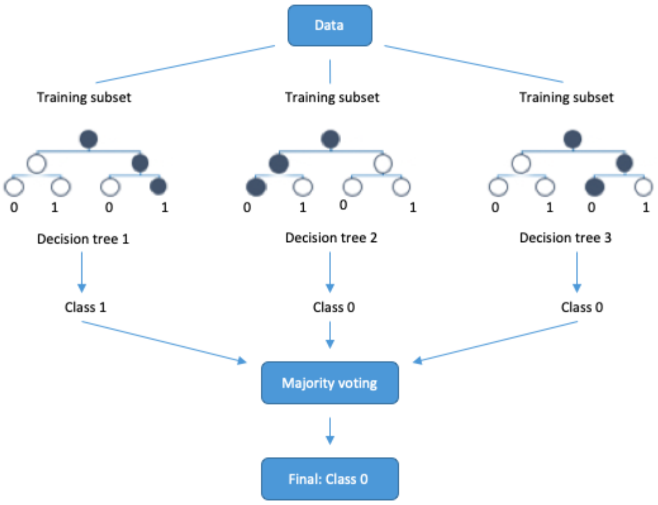


Figure 3. Random Forest classifier.

2.3.4 XGBoost classifier

XGBoost or Extreme Gradient Boosting, is a decision-tree based ensemble learner using the boosting technique, see Figure 2, which is designed to achieve great speed and optimal performance [39, 41]. XGBoost works by building decision trees in sequences, one after another, where the next built decision tree focuses on the errors and weaknesses from the previous one and improves the prediction performance with those weaknesses in consideration [39]. In each iteration or sequence the tree structure learns from the previous tree's outcome and residuals. Residuals is the difference between the real value and the predictive value [42]. By combining these sets of weak learners, starting with one base learner, XGBoost creates a strong classifier in order to make more accurate predictions. XGBoost's great cache optimization comes with both pros and cons. The advantage is that XGBoost can predict outputs with a high accuracy, but the drawback is that it needs more training time [41]. Unlike most tree learning algorithms, the XGBoost classifier finds the best split amongst trees and is able to handle datasets with missing values accurately [41]. This is why it is considered a good model when it comes to performance in terms of its high prediction accuracy.

2.3.5 Naïve Bayes classifier

Naïve Bayes classifier is a supervised learning algorithm that uses the Bayes theorem to predict certain outcomes [43]. Bayes theorem determines the conditional probability of an event, which is the occurrence of said event based on previous outcomes [44]. The formula for Bayes theorem is:

$$P(A|B) = \frac{\boldsymbol{P(A) * P(B|A)}}{\boldsymbol{P(B)}}$$

In order to predict the occurrence of an event, the Bayes formula multiplies the prior probability with the probability of it occurring, divided by the total of each possible prediction. The word "naïve" from Naïve Bayes comes from the assumption that the theorem assumes that all variables are independent from each other, which may not reflect reality [44].

2.3.6 Imbalanced Learning

Imbalanced learning could be viewed as an application of learning from imbalanced data, where the aim is to provide a balanced dataset [45]. When the dataset is dominated by one sample class, it is considered to be imbalanced, meaning that the data is mostly represented by one class [46]. According to He & Garcia [45], previous studies have shown that a balanced dataset could improve the overall performance of the classifier. There are different ways to handle imbalanced datasets and adjust the distribution of classes in a dataset, the most common ones are over- and undersampling [46].

Oversampling is a sampling method, which refers to the action of increasing the minority class so that the classes become approximately equivalent. There are different ways of oversampling and random oversampling can be perceived as the most common one. Random oversampling is basically randomly replicating instances from the minority class. Synthetic Minority Oversampling Technique (SMOTE) is another oversampling approach that is frequently used. Rather than replicating random instances from the dataset, SMOTE synthetically creates instances from the minority class. From SMOTE, several oversampling methods have been created by modifying the original SMOTE such as borderline SMOTE. [46]

Undersampling refers to the action of decreasing the majority class so that the classes become approximately equivalent. There are different ways of undersampling and the most common one is RandomUnderSampler, which randomly removes instances from the majority class. [46]

SMOTEENN is a sampling method that combines the properties of over- and undersampling, in particular SMOTE and Edited Nearest Neighbor (ENN). SMOTEENN selects and connects the nearest neighbor and then cleans the data from oversampling [47].

2.4 Evaluation Score

Machine learning models are not always perfect for the given data and are in need of evaluation to see how well the model performs. The most common ways to evaluate binary classifiers are using certain metrics, which are accuracy, recall, precision and F1-score [48]. These metrics can be calculated through a confusion matrix, shown in Figure 4.

	Actual	
Predicted	True positive (TP)	False negative (FN)
	False positive (FP)	True negative (TN)

Figure 4. Confusion matrix for binary classification.

A confusion matrix is a machine learning concept, which includes information about the actual and predicted classifications, used to describe the performance of the classifier [49]. True Positives (TP) and True Negatives (TN) represent the correctly classified test instances while False Negatives (FN) and False Positives (FP) represent the incorrectly classified test instances [20].

Accuracy is a measure that shows the overall effectiveness of the classifier [48]. It is a metric showing the rate of total correctly classified instances. According to Deng et al. [49], accuracy is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \qquad \text{(Eq. 1)}$$

Precision is a measure that shows the proportion of correctly predicted positive instances. The metric shows how often the model is correct when predicting the target class, in our application, churners. According to Deng et al. [49] precision shows the accuracy of predicting a specific class and is calculated as follows:

$$\text{Precision} = \frac{TP}{TP+ FP} \qquad \text{(Eq. 2)}$$

Recall is a measure, which shows the effectiveness of the classifier to determine examples labeled as positive [48]. It shows the ability of the binary classifier to identify instances of a specific class [49]. Recall is calculated as follows:

$$\text{Recall} = \frac{TP}{TP+ FN} \qquad \text{(Eq. 3)}$$

The F-score is often used for evaluating the performance of a classifier. The F-score is a measure that takes both precision and recall into consideration and is typically defined as the harmonic mean of precision and recall. A better combined recall and precision is achieved as F-score is closer to 1. [21]

$$\text{F-score} = \frac{2 * precision * recall}{precision+recall} \qquad \text{(Eq. 4)}$$

3 RELATED WORK

Customer churn prediction has been widely researched, specifically within the domain of B2C. As mentioned above, subscription-based services have rapidly grown during the past years and they rely solely on the revenue generated from their customers. Hence, customer churning has a great impact on businesses and within the B2B context each loss of a customer directly affects the business negatively. Customers churn for various reasons and according to Gordini & Veglio [9], companies within a B2B market mostly churn for lower cost structures, their market outreach is extended, or another company is willing to cater to their specific needs, where their current service provider is not. In the last couple of years, the researcher focus has partially switched from B2C to the B2B domain. The research conducted within the field of churn prediction mostly investigates the problem in the telecom industry in the context of B2C and not within B2B. However, the relevance towards our study is that the conducted studies investigates churn in a subscription-based industry. Different approaches used for churn prediction include methods, such as survival analysis [50, 51] for time to churn prediction, as well as machine learning methods [21, 19, 22, 52, 16, 17, 23, 10] for binary churn prediction. The main difference of the used approaches is related to the problem, either identify churn/non-churn or time to churn. The most frequently used approach within customer churn prediction is machine learning.

3.1 Time to churn prediction

Junxiang [51] and Masarifoglu & Buyuklu [50], study customer churn prediction using survival analysis, which is a group of statistical methods used for studying events based on certain circumstances, such as customer churn prediction.

Using survival analysis, Masarifoglu & Buyuklu [50] emphasize that the question they answer is the time until a certain event occurs. Their model estimation is based on the survival function and the hazard function [50]. They employed the Kaplan-Meier method and the Cox model in R. Using this approach, the preliminary risks of customer churning are explained and quantified. Survival curves and hazard ratios are also presented, which they argue, allows service providers to take preventative measures against churners. Masarifoglu & Buyuklu [50] use historical telecom data acquired from a mobile operator for their study, which contains information regarding 10365 randomly selected customers. The data consists of categorical variables represented as dummy variables. These variables represent information regarding subscription type, length of subscription but also customer specific information such as gender.

The modeling process used by Junxiang [51] is done using four steps: explanatory data analysis, variable reduction, model estimation, and model validation, where the first two steps can be viewed as preparing the data for the survival analysis. Just like Masarifoglu & Buyuklu [50], the model estimation is based on the survival function and the hazard function. The author argues that the purpose of this estimation is to identify potential churn characteristics and estimate customer churn by calculating probabilities for survival. The last step in the modelling process, that is, model validation, is done by scoring predicted survival probabilities at a specified time for each customer. These survival probabilities are ranked in ascending order and they state that the customers with the lowest predicted survival probabilities are the ones that will most likely churn. Junixiang [51] ranks the survival probabilities in different deciles and compares the predicted number of churners at a specific time in each decile. The raw data used in the study consist of demographic data, customer contact data, and customer internal data, which is further divided into warehouse- and telecommunications data. The study started with a dataset containing 212 variables, which was later reduced to 115 variables. These variables are categorized into both numerical and categorical values and only 29 variables are explanatory variables, which Junixiang [51] uses in the survival analysis.

3.2 Binary churn prediction in the domain of B2C

Vafeiadis et al. [21] compare different machine learning classifiers in order to predict customer churning. The data used in this study is provided by a telecom provider and consists of 5000 samples and multiple variables. The data mostly contains usage information such as, call duration and number of texts sent. In addition, specific subscription information is also used, such as subscription period. The different classifiers compared are Artificial Neural Network (ANN), Support Vector Machine (SVN), Decision Tree (DT), Naïve Bayes, and Logistic Regression. Furthermore, they use boosting in order to compare the single learners with their boosted versions; however, Naïve Bayes and Logistic Regression are not boosted due to limitations in their parameters. In order to evaluate the performance of the different classifiers, they use precision, recall, accuracy and F-measure, which are calculated from the confusion matrix. Based on these measures they conclude that boosting significantly improves the performance in terms of both accuracy and F-measure for all of the considered classifiers. On the other hand, without the use of boosting, they suggest that Support Vector Machine is a good tool for customer churn prediction. The accuracy of all the studied classifiers range from 93 to 99 percent and the F-measure between 73 and 77 percent. For future work Vafeiades et al. [21] would like to investigate the performance of other boosting algorithms. They also suggest that more simulations with different parameters should be investigated and that the use of a larger and more detailed dataset might yield a better result.

According to Brandusoiu et al. [22], prediction tasks depend strongly on data mining techniques, since machine learning algorithms improve the performance of the prediction process. They compare different machine learning algorithms with the aim of predicting churn. The problem addressed is a classification problem with churn/non-churn as categorical variables. They compare Support Vector Machines with Bayesian Networks and Neural Networks. The result is based on a confusion matrix and a gain measure, where the overall accuracy for all the considered algorithms ranges between 99 and 100 percent. The dataset used in the study consists of usage data from a telecom company, such as how many calls or texts the subscriber has made. The data contains information regarding 3333 subscribers with 20 variables, 15 of them being continuous and five being discrete. The dependent variable is categorical, churn or non-churn. Based on their result they conclude that all three classifiers performed well on predicting churn. They further suggest that the predictive performance can be improved by applying ensemble learning structures.

Lalwani et al. [23], conduct a study related to churn prediction using a machine learning approach. They use Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and various boosting algorithms such as XGBoost and AdaBoost. To conduct a research of this magnitude, Lalwani et al. [23], use comprehensive data from a telecom provider that consists of 7000 instances. Furthermore, the dataset includes 21 features, categorized into numerical and categorical values. The dataset consists of usage information such as number of calls, but also specific customer data, such as gender. In order to address their research question, they follow six steps. The first three steps are related to data processing, feature analysis, and feature selection. In the two following steps, they develop their model based on the different algorithms. In the last step of the process, they evaluate the performance based on the metrics: precision, recall, accuracy, and F-measure, calculated from a confusion matrix. The accuracy ranges from 74 to 82 percent, recall from 74 to 82 percent, precision from 57 to 81 percent, and F-measure from 63 to 81 percent. Lalwani et al. [23] concludes that the boosted classifiers, XGBoost and AdaBoost, performed best and achieved the best accuracy. In addition, they state that machine learning is the most efficient way of dealing with predictions, which is why they consider that the future of churn prediction will revolve around machine learning.

Ullah et al. [10], investigate churn prediction using a machine learning approach. The purpose of their study is to create a general churn prediction model. They use numerous classifiers such as Random Tree, Random Forest, Naïve Bayes, Decision Stump, J48, and Logistic Regression. They apply boosting to the Decision Stump algorithm and bagging to Random Tree, which broadens the research. They use two different datasets. One is from a telecom provider and the other one is publicly available for everybody to use. The first dataset contains 64107 instances with 29 features. The second one includes 3333 instances with 16 features. All the features are in numerical form. The data is labeled as churn and non-churn and is mostly made up of usage information, but also specific subscription information such as the length of the subscription. Ullah et al. [10], evaluate the performance of the classifiers using accuracy, precision, TP rate, FP rate, recall and F-measure. Based on their results, they conclude that the Random Forest classifier outperforms the other classifiers with a Recall of 88.8 percent, precision of 89.3 percent, and a F-measure of 88.2. For future work, Ullah et al. [10] points out the relevance of exploring the behavioral patterns and how they change for customer churning, which would be done with the application of Artificial Intelligence (AI).

Rothenbuhler et al. [53], use an uncommon machine learning approach to study the churn prediction problem. They chose to use the Hidden Markov's model for their study. Hidden Markov's model is based on a stochastic process. The stochastic process is formed by a Markov's chain and since it is non-observable it is called the Hidden Markov's model. The Markov's chain is described as a discrete-time random process that takes on different values in a given state space. The value it takes is the probability of a sequence of events, where the probability depends on the value of the previous state. Due to the highly competitive nature of the mobile app market for games, Rothenbuhler et al. [53] investigates customer churning within this sector. In order to address the research question, they use customer data, which is presented in a weekly moving average. This dataset is a representation of an "active customer", that is, a customer who uses the application. Every time a customer connects to the application, the data is stored. The features are represented by a customer's weekly average connections, which are labeled activity features. Rothenbuhler et al. [53], state that churn prediction is a binary prediction problem, meaning that the goal is to predict churn or non-churn. With the use of the Hidden Markov's model, they find a clear link between the churners and their motivational state. The result is presented in the number of customers that have churned for each state and in addition, they compare their findings to other prediction models, such as Logistic Regression, Neural Network and Support Vector Machine. They conclude that their performance was quite similar in all cases.

3.3 Binary churn prediction in the domain of B2B

Within the context of B2B, churn prediction has not been researched extensively compared to the B2C context. Gordini & Veglio [9], conduct a study regarding churn prediction where they tailor the churn prediction model for a B2B e-commerce industry. They use numerous algorithms to build their churn prediction model and amongst them are Support Vector Machines (SVM), Neural Networks and Logistic regression. The dataset they use in their study is from a major online site based in Italy and contains transactional data and demographic data from customers. Their results show that SVM performs better than the other considered algorithms. Based on the results, Gordini & Veglio [9], conclude that churn prediction is crucial in B2B, and that the choice of algorithm matters greatly in the B2B context, since the result of the churn prediction models varies.

Figalist et al. [13] implements a churn prediction model within a B2B domain where the goal is to fill the gap of customer churning in the domain of B2B. Their study provides an approach that maps the data which combines stakeholders who directly and indirectly affect decision making. To get a deeper understanding of churn prediction, Figalist et al. [13] interview multiple stakeholders and analyze the data that is derived from a B2B company. To predict churn, data regarding customers or end-users is used. Based on the results Figalist et al. [13], conclude that churn prediction within B2B is often of greater significance relative to B2C since companies within B2B stand to lose a greater deal when a customer churn, compared to a customer within B2C.

4 RESEARCH METHOD

In order to address our research question, we chose to use the design science methodology [54, 55]. According to Ghauri et al. [56], a study needs to be built upon prior knowledge; hence, we also conducted a literature study in order to identify relevant information, position and justify our work, gather ideas concerning churn prediction models, and collect inputs related to the design process. We argue that design science is a natural choice based on the fact that our main objective is to propose a churn prediction model, that is, an artifact. Building an artifact is one of two focal points of design science, the second one is to evaluate the built artifact.

4.1 Literature Study

We conducted a literature study using two search engines: BTH summon and google scholar. We also used a database with a built-in search engine, Diva portal. In order to filter the results and obtain a high reliability, we have mainly used books and peer-reviewed articles published in scientific journals and conferences. As our study aims to investigate customer churn prediction, we used search words such as customer churn prediction, customer churning, customer relationship management, churn management in subscription-based services, churn prediction in B2B, and customer churning in B2B.

4.2 Design Science

As mentioned above, in order to answer our research question, we mainly used the design science methodology [54, 55]. Design science is a method used to build artifacts, which in turn aim to solve problems within an organization. It can be viewed as an outcome-based method that is made up of two activities and follows certain guidelines. These activities are: *building* an artifact and *evaluating* the artifact. Hevner et al. [55] propose seven guidelines for design science:

1. Design as an Artifact
2. Problem Relevance
3. Design Evaluation
4. Research Contributions
5. Research Rigor
6. Design as a Search Process
7. Communication of Research

By creating an artifact in the form of a machine learning model, the first guideline of the design science methodology is satisfied. The guideline states that the built artifact must be viable, that is, either a construct, a model, a method, or an instantiation. Based on the fact that customer churning is a growing problem in the world of business, a churn prediction model can solve problems within organizations, which corresponds to the second guideline. Churn prediction models need to be evaluated in the form of performance and efficiency, and the third guideline highlights the importance of evaluating the artifact. The fourth guideline is research contributions, which can be one of three types: i) Contributing by creating an artifact for unsolved problems, ii) improving existing knowledge within the researched field of study, and iii) contributing with methods related to building and evaluating an artifact. Since churn prediction using different machine learning approaches and techniques is not fully investigated in the domain of B2B and specifically not in the subscription based service context, our study will contribute knowledge to researchers and firms within this field, which correspond to the second contribution type, that is, improving existing knowledge. Based on previous studies, the most frequently used approach within churn prediction is machine learning, which is why we chose to build a machine learning model. In addition, customer churn prediction typically requires the analysis of large amounts of data, for which machine learning is useful. To develop our model, we decided to use the machine learning process suggested by Marsland [30], see Figure 5. It should be emphasized that the machine learning process formalizes both the construction and evaluation of the learning model, which is required according to the fifth guideline. The sixth guideline expresses that the best possible design is identified by iteratively searching through the space of possible solutions. Using Marsland's machine learning approach, we create a model and evaluate it in iterations with the aim of finding the best suitable classifier, which will represent our churn prediction model. The findings of our study, including the built artifact will be presented in a technical report in the form of this master's thesis. The thesis will be available for multiple audiences such as other researchers and organizations. According to the discussion above, our work corresponds to the guidelines of the design science methodology.

4.2.1 The Machine Learning Process

As mentioned above, we decided to use the machine learning process proposed by Marsland [30], see Figure 5. The proposed machine learning process includes data collection and preparation, feature selection, algorithm choice, parameter and model selection, training, and evaluation. We do not consider the proposed machine learning process as a linear process; there is a lot of back and forth between the steps in order to create the most optimal model, in our case a churn prediction model.

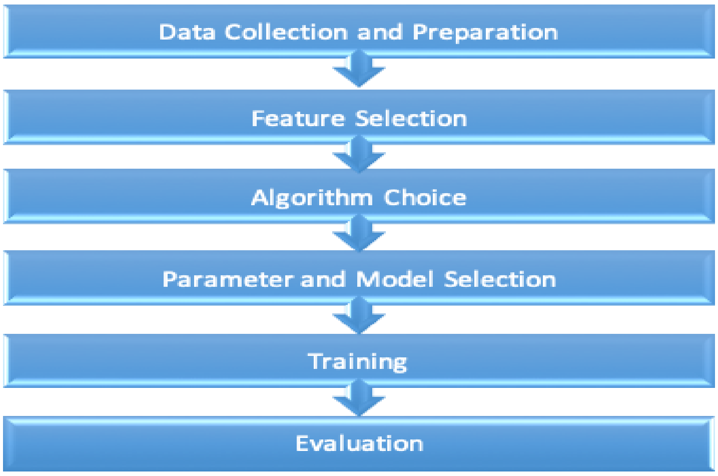

Data Collection and Preparation
Feature Selection
Algorithm Choice
Parameter and Model Selection
Training
Evaluation

Figure 5. The machine learning process proposed by Marsland [30].

4.2.1.1 Data collection and preparation

This study was performed using a dataset given by Fortnox AB, which is a company operating in the domain of B2B and provides a cloud-based platform for accounting and administration within finance. Fortnox AB provides subscription-based services in the form of licenses to different customers such as small businesses, associations, schools, and accounting firms [57]. All customer data is stored in raw data files and includes cross-sectional data of customers. Meaning that the dataset contains observations of customer specific data at a certain point. In order to use the data in our study, it was necessary to prepare the data.

We aggregated all the raw data files into one dataset including information about 96129 customers, including 94487 non-churners and 1642 churners, see Figure 6. The aggregated dataset contains Customer ID, Customer information, and class labels, that is, churn/non-churn for each customer.

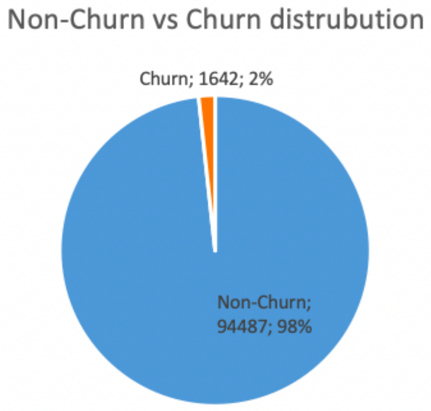


Figure 6. Distribution of non-churners and churners in the dataset.

Previous studies thoroughly investigate churn prediction in the domain of B2C, using different types of data, including customer demographics such as gender and age, customer behavior such as transactional data, customer perceptions such as overall satisfaction, and macro-environment data such as different customer experiences in life [58, 21, 10, 23]. As mentioned above, in the context of B2B less attention has been given towards churn prediction. Among others, Gordini & Veglio [9] and Figalist et al. [13], study the problem of churn prediction in the B2B domain, where they mainly use transactional data and customer demographics data to fit their classification models. All companies store different types of data and information based on their specific company activity and what is considered important to know related to their customers. In our study, we chose to use similar types of data as used in previous studies, even though the variables used partially differ.

The dataset used in the current study mainly consists of firmographic data and customer behavior data. Firmographics data, corresponding to demographics data in the B2C context, contains the number of employees, industry code, customer type, and corporate form. Customer behavior data includes the number of invoices created per month through the service platform, the number of subscribed licenses, and subscription start year. The number of employees are aggregated in intervals and represents the size of the company, see Figure 7. Industry code refers to the Swedish Standard Industrial Classification (SNI), where each customer is classified based on the activity of the business, see Figure 8. Corporate form is categorized in four different categories: corporation, sole proprietorship, general partnership, and others, see Figure 9. The customers are divided in two categories; either a direct customer or a customer through an agency, see Figure 10. The number of invoices created per month refers to the customer's usage of the service platform, see Figure 11. The dataset provides information concerning the products offered by Fortnox. These products are different licenses including accounting and invoicing, and number of licensees refers to the number of subscribed products each customer has, see Figure 12. Subscription start year refers to the year where each customer started their subscription, see Figure 13.

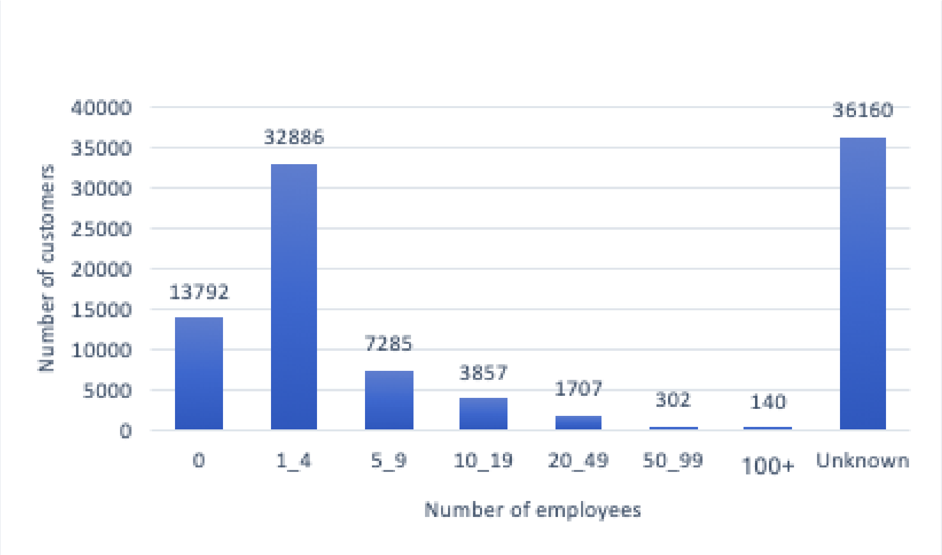


Figure 7. Distribution of number of employees in the dataset.

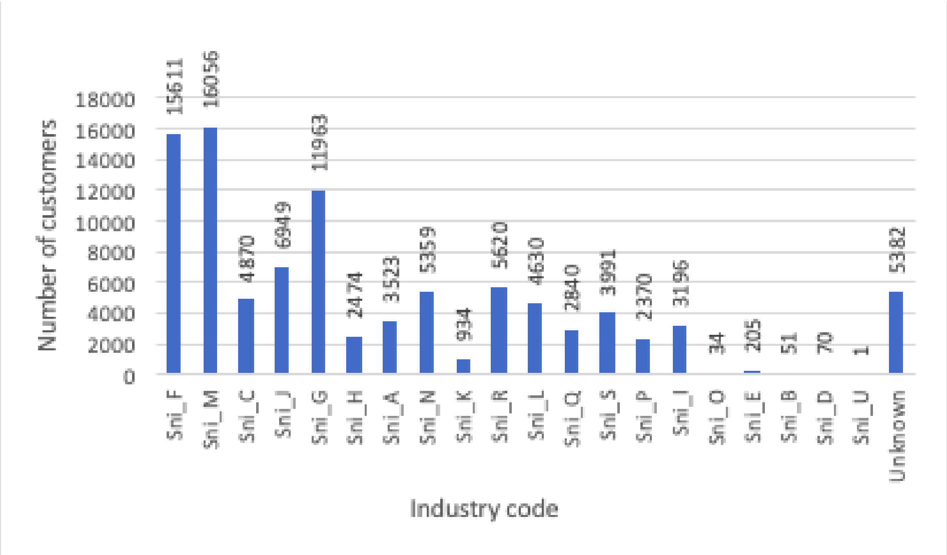


Figure 8. Distribution of industry codes in the dataset.

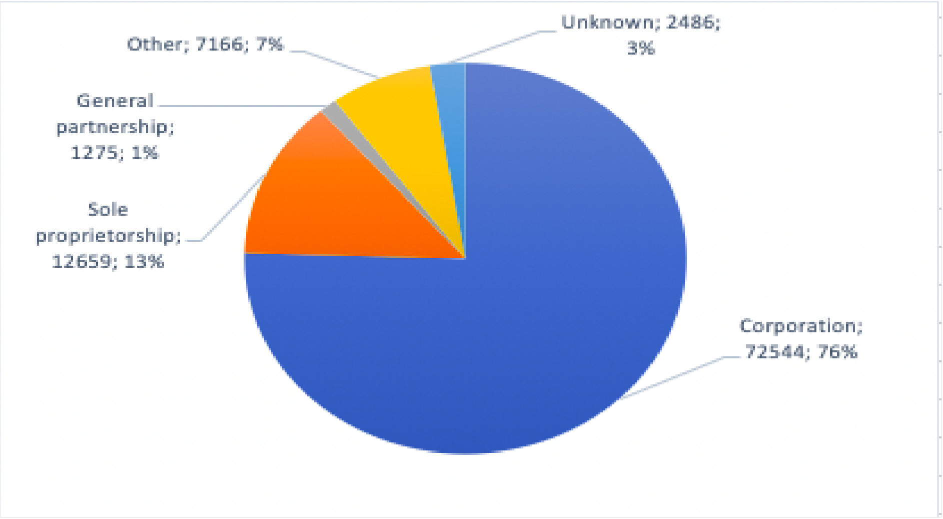


Figure 9. Distribution of corporate forms in the dataset.

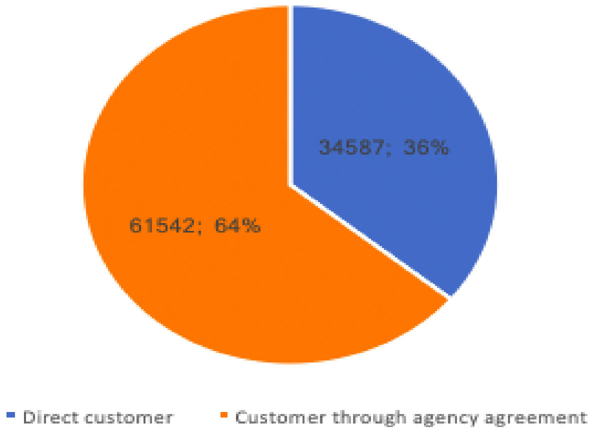


Figure 10. Distribution of the type of customer in the dataset.

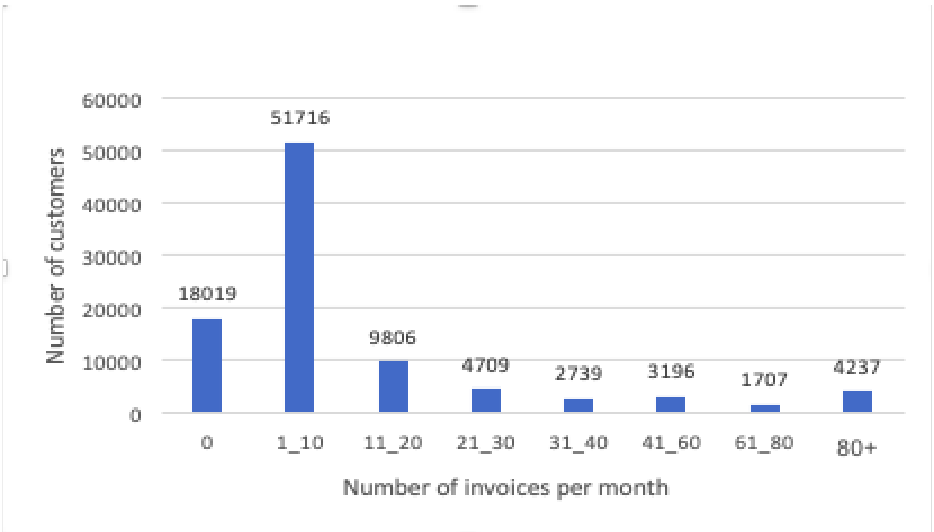


Figure 11. Distribution of number of invoices created per month in the dataset.

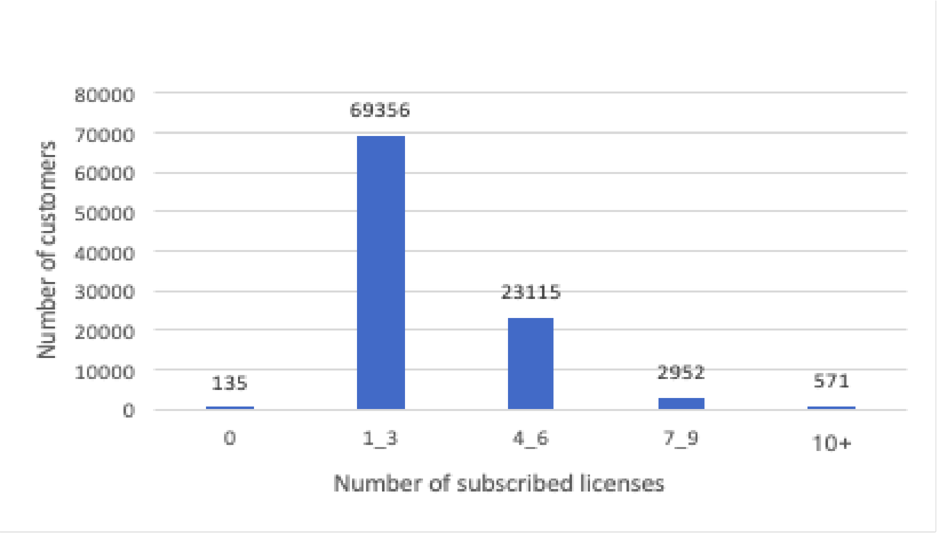


Figure 12. Distribution of number of subscribed licenses in the dataset.

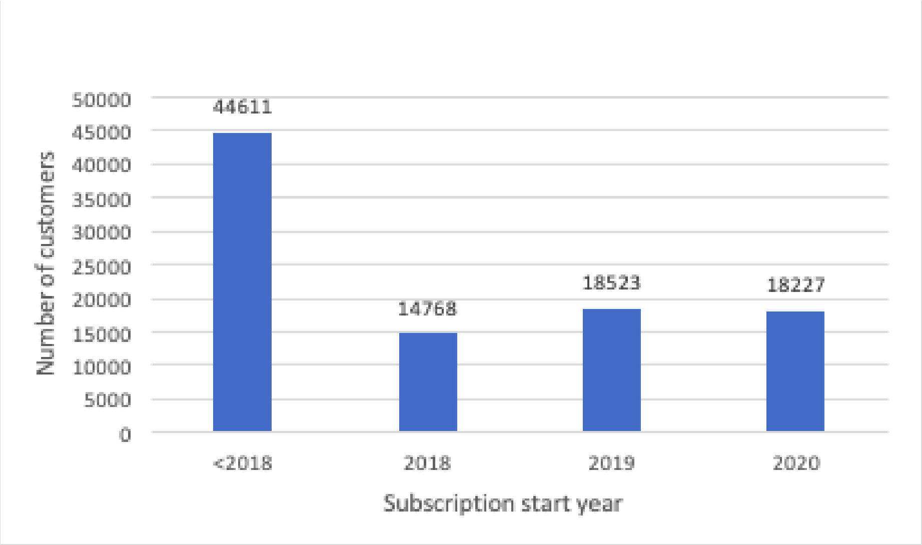
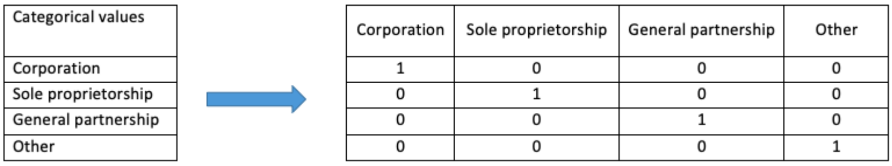


Figure 13. Distribution of subscription start year for the customers in the dataset.

In order to use the data in the model, we did some modifications to the data. Since all variables were categorical, that is, non-numerical values, such as Industry code, we chose to encode the categorical variables using one hot encoder [59]. In this case all categorical values are represented in new columns assigned with ones and zeros. In this way we present the dataset in a binary matrix including ones and zeros for a true and false representation. To illustrate this using an example, see Figure 14, let us consider a categorical variable, including four categorical values. One hot encoder transforms each of the categorical values into separate columns as features with exactly one binary value, where zero indicates non-present and one indicates present. The transformation is represented in a binary matrix where each row is a vector including ones and zeros.



Categorical values
Corporation
Sole proprietorship
General partnership
Other

Corporation	Sole proprietorship	General partnership	Other
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Figure 14. An example of a binary representation of a categorical variable consisting of four different values.

4.2.1.2 Feature Selection

Feature selection consists of determining the most impactful features for a problem. Feature selection is used in order to identify the most relevant features and is often used due to its performance enhancing properties [60]. This requires in depth knowledge of the data being used to determine which features should be included [30]. According to Tang et al., [60] too many features can result in overfitting, meaning that the model learns from what can be considered noise to the point that the result is negatively affected. This means that learning from noise complicates generalization and the ability of correctly classify unseen data instances. We decided to identify non-important features using a less complicated and straightforward approach, where the importance of the features is scored based on their prediction ability. We ranked the features based on their importance score and removed all non-important features with a score below 0.05. There exist more advanced methods for feature selection, such as recursive feature elimination (RFE) [61]. We decided not to use such methods since they are more time consuming and exceed the purpose of our study. The results we obtained were sufficiently good using feature importance. Since Naïve Bayes assumes that each feature is independent, feature selection using feature importance was not performed.

4.2.1.3 Algorithm Choice

Based on the scope of the problem, the optimal algorithm for the considered problem can vary. There are numerous algorithms to choose from when faced with a problem, for instance prediction tasks. However, not every algorithm is suitable and fits the criteria for the task at hand, which is why several algorithms are rejected. Since the problem we are trying to solve is a classification problem, all algorithms used for regression problems are rejected. Common classifiers, such as Naïve Bayes, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and Hidden Markov's Model are used in prediction problems, however previous studies show that ensemble learners are preferred for classification problems and they are rapidly becoming the standard choice among algorithms due to their performance enhancing characteristics [35, 36]. Hence, our choice of algorithms is based on previous studies. Both boosting and bagging have in previous studies shown to enhance the performance of base learners and they generate good results. We chose to use one algorithm, XGBoost, which applies boosting and one algorithm, Random Forest, which applies bagging. These algorithms are two common ensemble learners with high accuracy for churn prediction [10, 23]. In addition, we decided to compare the ensemble learners to the single classifier Naïve Bayes, which is considered being a simple probabilistic classifier, which performs well for several real-world problems including churn prediction [62].

4.2.1.4 Parameter and Model Selection

Many machine learning algorithms need to be fed with parameters such as the number of estimators. Parameter turning is the process to optimize the parameters of a machine learning algorithm and choosing the ones that obtain the best result for the desired problem. In our study, we investigate how machine learning can be used to predict customer churning within the domain of B2B; hence we considered parameter tuning to be beyond the scope of this study. In addition, we considered it to be too big of a task to accomplish. We decided to use default parameters and values for our algorithms and models.

4.2.1.5 Training

The process of training refers to using labeled data to build a model, which preferably performs well on unseen instances. The dataset should be split in a training set used for training and fitting of a model, and a test set used for evaluation. As mentioned above, supervised learning refers to learning by examples, meaning that each example contains inputs (features) and a corresponding output. To train the model, correlations are found between inputs and outputs. We decided to use supervised learning since our dataset is labeled and includes inputs with the corresponding output for each customer. We decided to randomly split the dataset into a training set and a test set with the proportions of 90:10 percent. Unfortunately, there is no distinct answer to how the split between training- and test data should be done and the choice is made from the researchers. Previous studies related to machine learning have used different splits such as 50/50, 60/40, 70/30, 80/20, and 90/10, but there is no research that suggests that one proportion preferred over another [63]. However, researchers argue that the bigger the dataset, the less test data is needed. For example, if there is data that contains 200 000 data instances, 10% can be more than enough to see if the model performs adequately but if the sample size contained 3000 data instances, 10% might be too little to validate the performance of the model and you might need to choose a larger test sample. On the other hand, there is no distinct answer to what a big or small dataset is. After consideration and a discussion with a domain expert we came to the conclusion that our dataset, including approximately 100 000 data instances, is relatively big and we chose to split our dataset with the proportions of 90:10 percent.

Many classification problems do not have a balanced dataset, meaning that the dataset includes one majority- and one minority class. Our dataset includes a majority class of non-churners. In addition to the random split, we decided to split the data so that the proportions of examples represent the same proportion as the original dataset. This means that the proportion of churners/non-churners are identical in both the training- and test set. Imbalanced datasets with more samples of one class than another is a common problem in real world applications [46]. Imbalanced data has been proven to result in significant negative effects on the performance of classifiers. There are different ways to handle imbalanced datasets, either by *oversampling* or *undersampling* [46]. Over- and undersampling are applied on the training dataset since the test dataset should be of the same characteristics as the known samples. We chose to balance the training set using both over- and undersampling methods and compare the results with the model fitted with the default training set. We chose to use the most frequently used methods, SMOTE for oversampling, RandomUnderSampler for undersampling. In addition to those sampling methods, we use SMOTEENN, a sampling method, which combines them both.

4.2.1.6 Evaluation

Evaluation is an important part of the machine learning process, where the performance of the classifier should be evaluated. In order to select the optimal machine learning algorithm different indicators are used for different types of problems [64]. In previous literature regarding churn prediction, the majority of researchers evaluate their models using accuracy, precision, recall, and F-measure, which are all calculated from the confusion matrix. We chose to evaluate the performance and efficiency of our algorithms using the above-mentioned metrics. In the case of imbalanced data, accuracy is not the most optimal metric used for evaluation since it does not always fully reflect the performance of the algorithm [65]. To handle this problem, we evaluate the classifiers based on precision, recall, and F1-score on the target class, that is Churn. However, we will focus on F1-score as it illustrates a balance between precision and recall. In our application, precision, that is, the rate of correctly classified churn instances, is less important relative to recall, which measures the model's ability to predict actual churners. Evaluation is done on unseen instances from the test set, in other words data, which the algorithm has not been trained on. This is why the data should be split into a training and a test set to be able to determine how well the model performs on examples not used during training.

4.2.2 Implementation

We chose to implement our machine learning models in Python using different libraries for data analysis, such as NumPy, Pandas, and Scikit-learn [66]. Scikit-learn is a tool used for predictive data analysis in Python with built-in functions for implementation of algorithms, training, and evaluation.

5 RESULTS & ANALYSIS

In this section, we present, for each of the three classifiers, Naïve Bayes, Random Forest, and XGBoost, the obtained results. In Table 1, we present the results as the overall performance of the three classifiers based on accuracy, see (Eq. 1), F1-score, see (Eq. 4), and the number of correct classifications. In order to fully evaluate the performance of the classifiers on the target class, that is, churn, we present a confusion matrix and the related metrics, precision, recall and F1-score, see Table 2-4. For description of the considered evaluation metrics, see Section 2.4. However, we here explain the values in the confusion matrix for our application. The confusion matrix is represented by four squares containing four values, where:

- The top left corner represents *True Positives*, that is, correctly predicted churners.
- The top right corner represents *False Positives*, that is, predicted non-churners who are actually churners.
- The bottom left corner represents *False Negatives*, that is, predicted churners who are actually non-churners.
- The bottom right corner represents *True Negative*, correctly predicted non-churners.

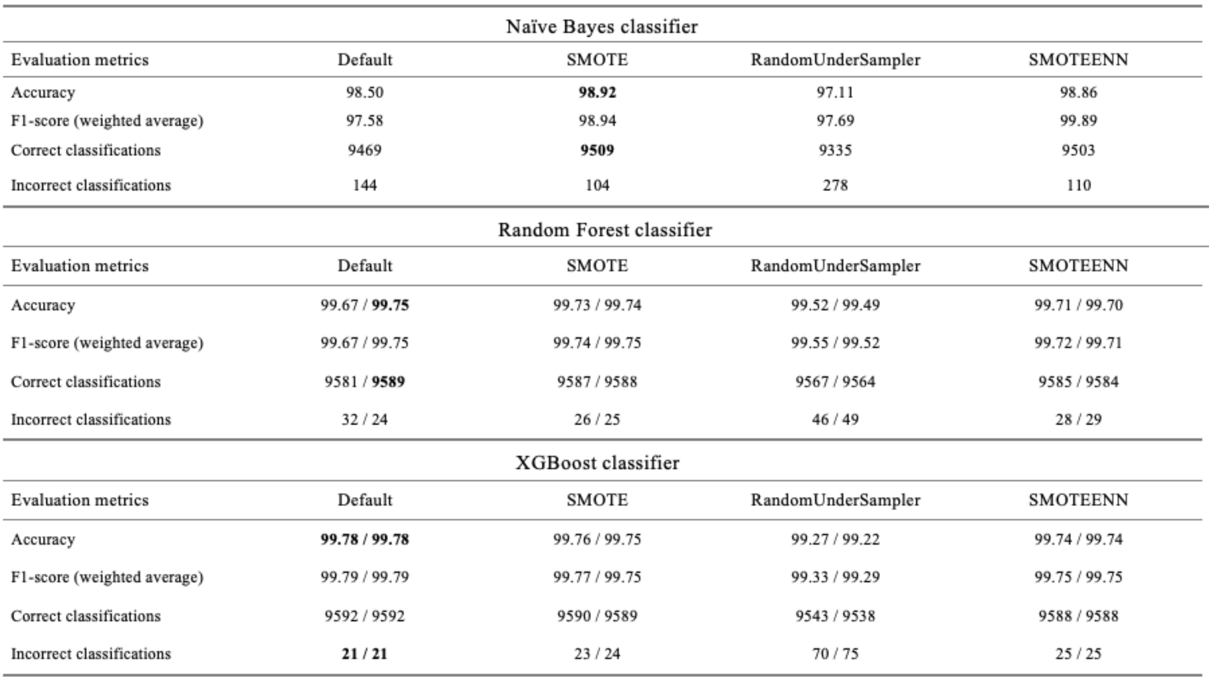
Each of the classifiers is evaluated using the default imbalanced dataset, but also using an equally distributed dataset, obtained from three different over- and undersampling methods: SMOTE, RandomUnderSampler, and SMOTEENN, see Section 2.3.6.

As mentioned above, in Table 1, we present the overall performance of the three different classifiers using an imbalanced training dataset but also using a balanced training dataset, transformed by three different sampling methods, SMOTE, RandomUnderSampler, and SMOTEENN. Our result indicates that all three classifiers are able to correctly classify the test instances with a high accuracy ranging between 97.1-99.8 percent. What we can see from the results is that the overall performance of the classifiers gets worse when using RandomUnderSampler as a sampling method considering overall accuracy and number of incorrect classifications. For the Naïve Bayes classifier, the accuracy ranges between 97-98.92 percent, which reflects that a high number of test instances are classified correctly. The highest possible accuracy was obtained using SMOTE as a sampling method with 9509 correct classifications, represented by an accuracy of 98.92 percent. The results for the Random Forest classifier and the XGBoost classifier are presented differently than for the Naïve Bayes classifier. Each cell in the considered tables is represented by two values. The value to the left shows the performance of the classifier without conducting feature selection and the value to the right represents the achieved results when non-important features are removed. The Random Forest classifier performed overall well with a high accuracy ranging between 99.5-99.8 percent. Our results show that feature selection slightly affected the overall performance of the Random Forest Classifier. The performance was positively affected by feature selection using the default imbalanced dataset and when using SMOTE as a sampling method. When using RandomUnderSampler and SMOTEENN the performance was negatively affected but the results were still similar. The highest number of correctly classified test instances using the

Random Forest Classifier corresponds to 9589 classifications, which reflects a 99.75 percent accuracy. This score was achieved using SMOTE as a sampling method in combination with conducting feature selection. Further, our results indicate that the XGBoost classifier obtained the highest possible accuracy of 99.78 percent, with 9592 correctly classified test instances using the default dataset. Using RandomUnderSampler as a sampling method did have a negative effect on the performance of the classifier. However, Feature selection in general and the other two sampling methods, that is, SMOTE and SMOTEENN did not have any remarkable effects on the overall performance of the XGBoost classifier.

In the case of imbalanced datasets, accuracy may not fully reflect the performance of a classifier [65]. For example, in our case where we have 98 percent non-churners, it is possible to obtain a 98 percent accuracy by predicting all examples as non-churners. To further be able to fully evaluate the performance of the classifiers on the target class, that is, churners, we present a confusion matrix and the related metrics for every classifier and for each of the over- and undersampling methods, including the default imbalanced dataset, see Table 2-4.

Table 1. Overall performance of the three considered classifiers, using the default imbalanced dataset and the three considered sampling methods. Each cell contains two values; the value to the left and to the right represents the result with and without feature selection, respectively.



Naïve Bayes classifier

Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Accuracy	98.50	**98.92**	97.11	98.86
F1-score (weighted average)	97.58	98.94	97.69	99.89
Correct classifications	9469	**9509**	9335	9503
Incorrect classifications	144	104	278	110

Random Forest classifier

Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Accuracy	99.67 / **99.75**	99.73 / 99.74	99.52 / 99.49	99.71 / 99.70
F1-score (weighted average)	99.67 / 99.75	99.74 / 99.75	99.55 / 99.52	99.72 / 99.71
Correct classifications	9581 / **9589**	9587 / 9588	9567 / 9564	9585 / 9584
Incorrect classifications	32 / 24	26 / 25	46 / 49	28 / 29

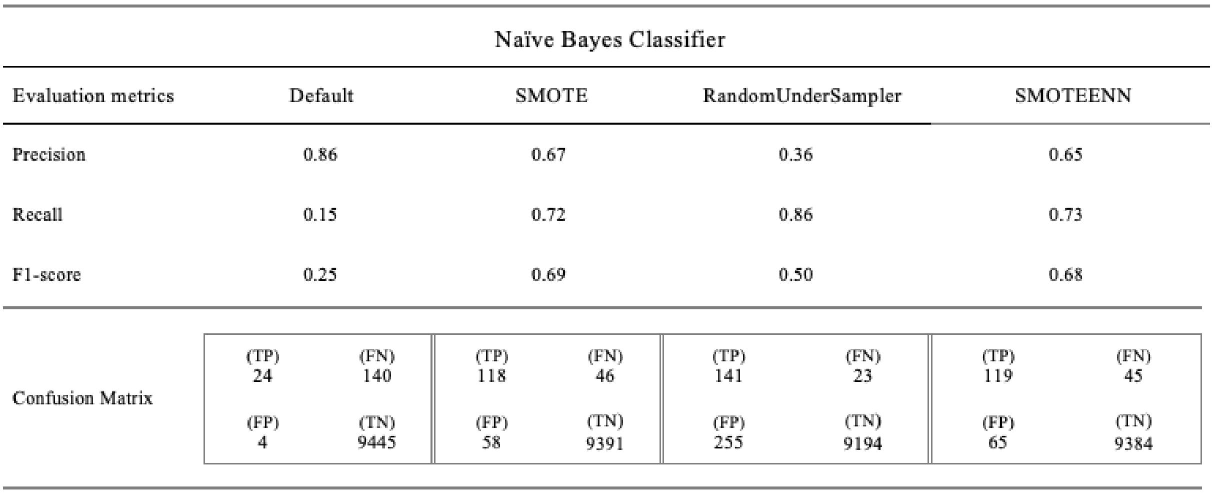
XGBoost classifier

Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Accuracy	**99.78 / 99.78**	99.76 / 99.75	99.27 / 99.22	99.74 / 99.74
F1-score (weighted average)	99.79 / 99.79	99.77 / 99.75	99.33 / 99.29	99.75 / 99.75
Correct classifications	9592 / 9592	9590 / 9589	9543 / 9538	9588 / 9588
Incorrect classifications	**21 / 21**	23 / 24	70 / 75	25 / 25

In Table 2, we present a confusion matrix and all the related metrics for the target class, churn, using the default dataset but also for each of the considered sampling methods, for the Naïve Bayes classifier. The presented confusion matrices show that the number of True Positives, that is, correctly classified churners, increase by implementing SMOTE, SMOTEENN, and RandomUnderSampler. However, the number of True Negatives, that is, correctly classified non-churners, decreases as the dataset becomes more balanced. From the values in the confusion matrix, the related metrics can be calculated. In our study, precision corresponds to how often the model is correct when it predicts churn, that is, the rate of correct churn predictions, see (Eq. 2). By maximizing the precision, True Positives are maximized, and False Positives are minimized. Recall corresponds to how often the model is able to predict actual churners, see (Eq. 3). By maximizing recall, True Positives are maximized, and the False Negatives are minimized. F1-score is a metric, which aggregates precision and recall and illustrates a balance between them, see (Eq. 4). A good F1-score represents a lower rate of False Positives and False Negatives; in our application this means, improved ability to correctly predict churners, without being concerned about False Positives, that is, non-churners classified as churners.

Based on the assumptions mentioned above, our result shows that the Naïve Bayes classifier performs far from optimal when predicting churn. Using an imbalanced dataset, the recall is 15 percent, meaning that the model is only able to predict 15 percent of the actual churners. In addition, the rate of correct churn predictions is 86 percent, which is relatively good. However, the F1-score of 25 percent indicates a poor performance on churn prediction using the imbalanced training dataset. Furthermore, our result shows that the F1-score increases and the ability to predict churn increases as the data becomes more balanced.

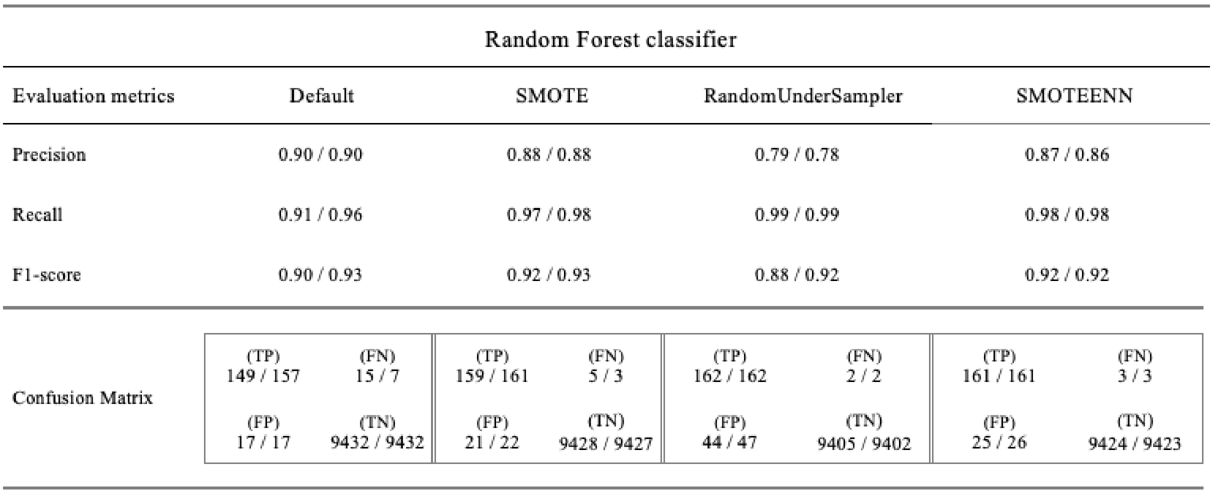
Table 2. Confusion matrixes and the related metrics for the Naïve Bayes classifier, using the default imbalanced dataset and the three considered sampling methods.



Naïve Bayes Classifier				
Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Precision	0.86	0.67	0.36	0.65
Recall	0.15	0.72	0.86	0.73
F1-score	0.25	0.69	0.50	0.68
Confusion Matrix	(TP) 24 (FN) 140 (FP) 4 (TN) 9445	(TP) 118 (FN) 46 (FP) 58 (TN) 9391	(TP) 141 (FN) 23 (FP) 255 (TN) 9194	(TP) 119 (FN) 45 (FP) 65 (TN) 9384

In Table 3, we present the obtained results while predicting churn using the Random Forest classifier and in addition to the Naïve Bayes classifier, feature selection is conducted. Each cell contains two values; the value to the left and to the right represents the result with and without feature selection, respectively. Based on the scores from the evaluation metrics it can be seen that the model performs well on predicting churn. It can be seen that a balanced training dataset increases the number of correctly classified churners, that is True Positives, but also the number of False Positives, that is non-churners predicted as churners. Feature selection did not have any remarkable effects on the results. Using the default dataset and the balanced training dataset generated by SMOTE, the performance slightly improved by removing non-important features. In the other two cases, using RandomUnderSampler and SMOTEENN the result was almost similar with and without feature selection. Overall, the Random Forest classifier scores a precision, ranging from 79 to 90 percent. The recall ranges between 91-99 percent, which indicates that the model is able to predict actual churners with a high accuracy. Since the F1-score takes both these values in consideration, The F1-score is also pretty high in all cases.

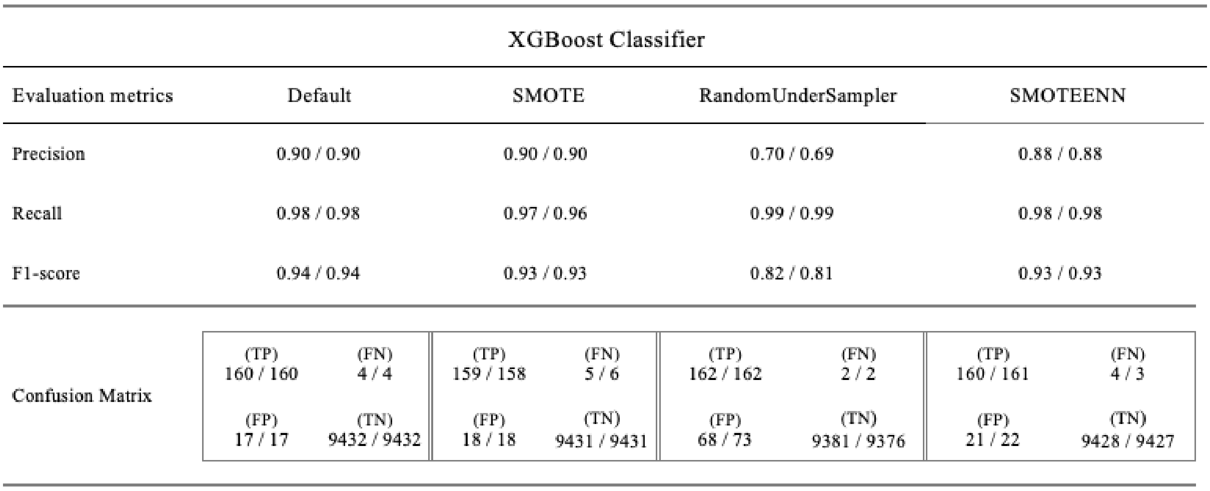
Table 3. Confusion matrixes and the related metrics for the Random Forest classifier, using the default imbalanced dataset and the three considered sampling methods. Each cell contains two values; the value to the left and to the right represents the result with and without feature selection, respectively.



Random Forest classifier				
Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Precision	0.90 / 0.90	0.88 / 0.88	0.79 / 0.78	0.87 / 0.86
Recall	0.91 / 0.96	0.97 / 0.98	0.99 / 0.99	0.98 / 0.98
F1-score	0.90 / 0.93	0.92 / 0.93	0.88 / 0.92	0.92 / 0.92
Confusion Matrix	(TP) 149 / 157 (FN) 15 / 7 (FP) 17 / 17 (TN) 9432 / 9432	(TP) 159 / 161 (FN) 5 / 3 (FP) 21 / 22 (TN) 9428 / 9427	(TP) 162 / 162 (FN) 2 / 2 (FP) 44 / 47 (TN) 9405 / 9402	(TP) 161 / 161 (FN) 3 / 3 (FP) 25 / 26 (TN) 9424 / 9423

In Table 4, we present the obtained result on the target class, churn, for the XGBoost classifier. Our results show a good churn prediction ability with a high precision, recall, and F1-score. Balanced data did not have a great impact on the result, using SMOTE and SMOTEENN. However, the number of False Positives increased significantly using RandomUnderSampler which is reflected on the precision and F1-score. Feature selection did not have any outstanding effects on the performance of the classifier. Our result states that the best churn prediction performance was achieved using the default dataset. The classifier is 90 accurate when predicting churn and is able to predict 98 percent of actual churners. The F1-score is 94 percent, indicating a good churn prediction performance.

Table 4. Confusion matrixes and the related metrics for the XGBoost classifier, using the default imbalanced dataset and the three considered sampling methods. Each cell contains two values; the value to the left and to the right represents the result with and without feature selection, respectively.



XGBoost Classifier				
Evaluation metrics	Default	SMOTE	RandomUnderSampler	SMOTEENN
Precision	0.90 / 0.90	0.90 / 0.90	0.70 / 0.69	0.88 / 0.88
Recall	0.98 / 0.98	0.97 / 0.96	0.99 / 0.99	0.98 / 0.98
F1-score	0.94 / 0.94	0.93 / 0.93	0.82 / 0.81	0.93 / 0.93
Confusion Matrix	(TP) 160 / 160 (FN) 4 / 4 (FP) 17 / 17 (TN) 9432 / 9432	(TP) 159 / 158 (FN) 5 / 6 (FP) 18 / 18 (TN) 9431 / 9431	(TP) 162 / 162 (FN) 2 / 2 (FP) 68 / 73 (TN) 9381 / 9376	(TP) 160 / 161 (FN) 4 / 3 (FP) 21 / 22 (TN) 9428 / 9427

5.1 Comparison of classifiers

In Table 1, we present the overall performance of the three classifiers, Naïve Bayes, Random Forest, and XGBoost in order to get an overview of the models' ability to correctly classify the test instances. Our result indicates that all three classifiers are able to correctly classify the test instances with a high overall accuracy. XGBoost performed better than the other classifiers with an achieved overall accuracy of 99.78 percent with only 21 incorrect classifications. This assumption is based on the fact that the best performance of the Random Forest Classifier was achieved with only 24 incorrect classifications and the Naïve Bayes classifier with 104 incorrect classified test instances.

As mentioned above, accuracy could be a misleading metric in the case of an imbalanced dataset, since it does not tell much on how the classifiers perform on specific classes. However, in our application it is more important to predict actual churners with a high accuracy and minimizing False Negatives rather than minimizing the number of False Positives, that is, non-churners classified as churners. In Table 2-4 we present the performance of the classifiers for the target class, that is churners.

In our study, the focal point is to predict churners with a good accuracy. By analyzing the results, Naïve Bayes does not perform well on predicting churners, in particular with the default imbalanced dataset, the results were far from adequate. The Random Forest classifier and the XGBoost classifier perform remarkably better on churn prediction than the Naïve Bayes classifier does. The XGBoost classifier and the Random Forest Classifier perform similarly. The Random Forest classifier performed best using SMOTE as a sampling method in combination with conducted feature selection. The precision of the classifier is 88 percent, recall is 98 percent, and F1-score is 93 percent. Based on the highest achieved F1-score of 94 percent and a recall of 98 percent, our results indicate that the XGBoost classifier is the best performing classifier.

6 DISCUSSION

The objectives of this study was to construct a binary churn prediction model and compare different types of algorithms for the built model. Furthermore, we have studied the effects of balancing the training dataset for the considered model. For our churn prediction model, we built and analyzed classifiers using three different algorithms, that is, XGBoost, Random Forest, and Naïve Bayes. For each algorithm, we used three different sampling methods: SMOTE, RandomUnderSampler, and SMOTEENN, in order to investigate whether a balanced dataset would improve our models' performance on customer churn prediction.

6.1 Performance of Churn prediction model

Model evaluation is a big part of Design Science and with a great variety of metrics and measures, the choice of which metric(s) to use is not always obvious, and it depends on the considered application. Our results show that for the considered classifiers, Naïve Bayes, XGBoost, and Random Forest, the overall accuracy ranges between 97.0 and 99.8 percent, indicating an overall good performance. On the other hand, one can question what a good and suitable score of accuracy actually is. Accuracy measures the classifiers ability of correctly classifying all the test instances, in our case including both non-churners and churners. But what is actually a good score? This is a common question when evaluating machine learning algorithms, and unfortunately there is no distinct answer to this question. We argue that a good score of accuracy depends on the considered application and what prediction problem is investigated. 90 percent accuracy in one application may not be as good as 90 percent in another application. For example, for email spam detection, 90 percent accuracy is not bad, meaning that only 10 percent of the emails are wrongly classified. In medical predictions, regarding potential patients in a certain risk group, 10 percent wrong classifications are far from acceptable. It should be emphasized that accuracy alone is not a sufficient metric used for evaluation [67], which is also shown by the above examples. In particular, when the considered dataset is imbalanced to a high degree, accuracy can be a misleading metric. In our case, the used dataset consists of 98 percent non-churners and 2 percent churners, which in fact could give us a 98 percent accuracy by predicting all test instances as non-churners. If this would be the case, the result is far from acceptable and a 98 percent accuracy would not be good. Based on this reasoning, accuracy gives in our case a skewed view of the actual performance of the classifiers. Based on the above reasoning, we can therefore not assume that the performance of the classifier is good by only looking at accuracy. Alternative evaluation metrics, which are commonly used in the literature are precision, recall, and F1-score, all indicating the performance of the classifier on a specific target class (see also Section 2.4). In our case the target class is churn, and our results based on the above-mentioned alternative metrics show that, except for XGBoost and Random forest, Naïve Bayes with the best obtained precision of 67 percent, recall of 72 percent, and F1-score of 69 percent is actually not good at predicting churners in our application. This further proves our point stated above that accuracy can be a misleading metric.

Precision, which corresponds to how often the model is able to correctly predict the target class, is a metric to favor when the cost of False Positives is high. A typical application where precision is to favor is email spam detection, where the cost of losing a valuable email might be higher than receiving one additional spam email. On the other hand, in medical predictions, recall is the preferable metric, since the cost of False Negatives is typically much greater than False Positives; there is no room for misclassifying a patient with a certain severe diagnosis as healthy. In our application of customer churning in the B2B subscription-based context, where each purchase typically has a higher transactional value compared to B2C [13], we consider recall being more important than precision, since customers who churn directly affect the revenue stream in a negative manner. Based on the above reasoning, we argue that the cost of False Negatives, that is, churners who are classified as non-churners, is greater than the cost of targeted marketing towards False Positives, that is, non-churners classified as churners. In addition, the non-churners that are misclassified as churners, may tend to churn within the nearest future and preventive actions directed towards these misclassified non-churners might not be a wasted effort. One can argue that classifying all test instances as churners would increase the recall to 100 percent, but the precision would be considerably low. An extreme example of this, from our dataset, would lead to 98 percent of the test instances represented as False Positives, which does not reflect the churn predictive ability of the model. The purpose of a churn prediction model is to accurately predict churners and avoid wasting money and unnecessary costs of marketing towards all customers and customers who do not have intentions to churn [68]. Hence, the precision cannot be too low even though the aim is to improve recall. Thus, it is also important to consider the F1-score, which represents the harmonic mean of both precision and recall. The F1-score shows how the algorithm is able to tradeoff between recall and precision in the generated classifier.

An important finding is that the ensemble learners, XGBoost and Random Forest, performed better than the single base learner Naïve Bayes. In previous studies, regarding prediction in general, all three classifiers have proven to perform well in the real-world context. However, ensemble learners, due to their ability of combining single weak learners, have in general proved to be better than single base learners [34, 35, 36]. Churn prediction has been widely researched within the telecom industry, where the generated classifiers have been compared based on their churn prediction ability, see, for example Vafeiadis et al. [21] and Ullah et al. [10]. These studies show that ensemble learners, using bagging and boosting, perform better than single base learners when predicting customer churning. Our results are consistent with the findings of Vafeiadis et al. and Ullah et al., as XGBoost (using a boosting approach) and Random Forest (using a bagging approach) performed better than the single learner, Naïve Bayes. Bagging, in its simplest form, uses majority voting from several independent decision trees for prediction. We argue that when it comes to taking a decision, collective decision making is usually better than an individual making the same decision, in other words, many heads are better than one. Boosting can be described as learning by mistakes, where each tree is built sequentially based on the predecessor's error. We argue that when making a decision, it is better to do it based on experiences from previous mistakes rather than making a decision for the first time.

Regarding balanced and imbalanced training datasets and the effects on the results, previous studies show that the use of a balanced training dataset is expected to improve the performance of a classifier [45, 69]. He & Garcia [45] and Mujalli et al. [69] argue that in the case of an imbalanced dataset, the classifier could be biased due to the dominant effect of the majority class. In our study, we used three different sampling methods and compared the results to the default imbalanced dataset in order to investigate the impacts of having a minority target class relative to having an almost equally distributed dataset. One common finding, using Random Forest and Naïve Bayes, is that sampling methods used for balancing the training dataset improve the recall and decrease the precision. This means that the number of False Positives increase as the number of False Negatives decrease. We believe that there is a risk associated with the use of sampling methods, which is for example, that the more churn instances synthetically created, the higher the likelihood of these instances being similar to non-churn instances. This could affect the ability of correctly predicting the majority class in a negative manner. Due to the equal distribution of classes and more similar instances of churners and non-churners, classification becomes more complicated and more non-churners might be classified as churners. Another finding related to imbalanced learning was that sampling using RandomUnderSampler increased the number of False Positives far more than the other sampling methods. One reason for this could be that RandomUnderSampler might remove vital information from the majority class, in our case, non-churners, making it more complicated to correctly classify non-churners.

In previous studies, specifically in the domain of B2C, churn prediction is proven to be done with an overall high accuracy [10, 21, 22, 23]. In section 3.2 we present a few studies regarding churn prediction in the B2C context and the results of these studies show that the accuracy, precision, recall and F-measure scores above 80 percent, in some cases even over 90 percent indicating that churn prediction in general can be done accurately. In our study our churn prediction models' based on the best obtained results for each of the three considered classifiers, show an overall accuracy ranging from 98.2 to 99.78 percent, recall ranging from 86 to 99 percent, precision ranging from 67 to 90 percent, and F-measure ranging from 69 to 94 percent, see section 5. If we were to put our results in comparison with studies conducted in the B2C context we would like to argue that one could expect to obtain a relatively high performance model.

6.2 Validity threats

In all studies there are validity threats to various degrees [56]. As mentioned above, our dataset is imbalanced and only two percent of the data instances are churners. This could be a threat against validity since the majority class could have a dominant effect on the model. For example, if all examples (both actual churners and actual non-churners) were classified as non-churners, we would achieve an accuracy of 98 percent, but the model would still not perform well on predicting churners. However, our results show that our churn prediction model has proven to perform well on predicting churners. In the case of Random Forest and Naïve Bayes, it was necessary to apply balancing. Another threat to validity could be bias, introduced by the researcher, in how the features are extracted from the data. For example, during the data preprocessing step, there could be occasions where the researcher misunderstands the data and categorizes certain features incorrectly. In order to prevent such bias, we chose to discuss each feature and the data with our industrial supervisor, hence allowing us to fully understand the dataset. According to Ghauri et al. [56], external threats to validity are whether the findings from our study are generalizable or not. This is hard to say since we did not use our model using another company's data, meaning that we cannot say whether the results of our study are generalizable or not. On the other hand, since the problem related to churn prediction and our approach of predicting churn is common in previous studies in the B2C context, we believe that most companies involved in B2B facing the same problem can find our study useful. Another aspect to keep in mind, which can affect the generalizability of the model, is the availability of data. In our case we used a dataset from a company within the financial administration sector. Every company has their own segments of customers and all companies acquire and store different types of information based on the company activity and what information that is important to the company. Certain information in one business sector such as financial administration may be less important and relevant in for example the car industry. As mentioned before there are different reasons for a customer to churn and it does not necessarily need to be the same reasons in all sectors. Although previous studies have investigated churn prediction in the context of B2C and B2B there is no distinct answer to what exact input variables to use; hence, it is important to train the model using the specific company data.

6.3 Implications

Due to the digitalization and the broad variety of products and services offered to customers, churn management has become more important than ever. As mentioned in Section 2.2, from a business perspective, churn management as a part of customer relationship management includes two tasks. The first task is to predict which customer might churn and the second task is to determine retention strategies, such as customer loyalty programs, in order to retain those customers. Retention strategies are developed with the purpose of increasing long-time profitability. If such strategies are successful, the need for seeking new customers decreases and allows firms to focus on existing customers and building relationships with them. Long-term customers tend to be less sensitive towards churning and competitive offers and generate bigger profit margins [11]. These long-term customers may also provide new customers through referrals if they are satisfied, which in fact increases the revenue stream. Retention is an important task for firms to accomplish since for instance losing a customer leads to reduced sales but also increases the needs of attracting completely new customers [58]. The cost of acquiring new customers is five to six times the cost of retaining customers, which in fact is a direct loss of money [11, 58]. Apart from the cost of attracting and retaining customers, the net return on investments related to retention is higher than for acquisition, which may lead to an increase of revenue [12, 58]. Previous studies have also shown that current customers contribute more to the overall revenue than new customers do, since the probability of selling to an existing customer is higher than for new customers but also that loyal customers tend to reject offers from other service providers [12, 58]. From this viewing point we argue for the necessity of companies investing in churn management. According to Mithas et al. [70], applying CRM strategies, increases firm's knowledge regarding their customer base, which can be used to improve the degree of customer satisfaction and further the profitability of the firm. Based on such customer knowledge, loyalty programs can be used in order to improve customer satisfaction and increase the level of customer retention [70]. As mentioned before, retention strategies as a part of churn management are not free and the purpose of churn prediction is to minimize unnecessary cost by identifying potential churners and target those customers with most incentives to churn rather than the whole customer base. Not all customers have the incentives to churn so applying preventative measures on those is a direct loss of money [17]. On the other hand, this implies the importance of an accurate churn prediction model. If the model is not accurate, all resources put towards retaining the customers can be viewed as a potential waste, which contradicts the aim of CRM systems [71].

Our results indicate that our built churn prediction model is able to accurately predict churners based on the low rate of False Negatives and False Positives. This is valuable for firms since we could target actual churners with preventative measures and stop wasting money on customers with no intentions of leaving. Yet an accurate churn prediction model alone is important but it will only help companies determine potential churners. In order to bring complete value to a firm, retention strategies towards those specific customers are necessary. This applies in the context of subscription based services in both B2C and B2B. Blank & Hermansson [2], argue that subscription-based service providers need to keep a low churn rate in order to remain successful. We argue that this applies in both B2C and B2B, although the terrain is different. In all businesses, in particular subscription based businesses, customers are the focal point and since customers and their transactions constitute the incomes and profits, a lower churn rate directly impacts the firm's profitability regardless of the business model. However, even if preventative measures are applied, there are no guarantees that the customer will still not churn, since customers tend to churn due to various reasons. According to Hadden et al. [17], churners can be categorized into voluntary and non-voluntary churn. Non-voluntary churners could be customers who either had their subscription removed by the company or their subscription was cancelled by their own doing, due to some sort of mistake, such as missing a payment. Voluntary churners are customers who actively churn due to various reasons. It can be due to changes in circumstances such as financially or demographically, but also leaving for another service provider. Even though our result showed that we could predict churn with a high accuracy, we cannot distinguish between voluntary- and non-voluntary churn. However, it is still useful to know which of the customers might churn or not to increase the possibilities to even retain a few of them. As mentioned before, each loss of a customer affects the revenue stream and if there is a possibility to decrease this effect, firms should act upon this. Even a small improvement in customer retention could increase the profits and revenues significantly [11].

7 CONCLUSION AND FUTURE WORK

In the domain of B2B, customer churn management, in particular, churn prediction has become crucial for today's businesses operating as subscription-based service providers. The purpose of our study is to contribute knowledge regarding churn prediction in the domain of B2B with the objective of presenting a machine learning based churn prediction model. The main research question stated is the following:

How can customer churning be predicted in the domain of business-to-business using machine learning?

Our results indicate that customer churning can be predicted with a high accuracy using machine learning. Furthermore, from our results we can conclude that ensemble learners, using boosting and bagging, improved the performance of our churn prediction model, relative to the studied single learner Naïve Bayes, which is consistent with previous studies. In particular, the XGBoost classifier performed best based on overall accuracy, precision, recall, and F1-score. It should be emphasized that the dataset used for this study was imbalanced, and we applied different sampling methods in order to investigate how balancing the training dataset would affect the result. We conclude that balancing, using sampling methods, has a positive influence for the studied problem. Further we observe that balancing leads to reduced precision of the classifiers, that is, the rate of correctly classified churners, and improved recall of the classifiers, which corresponds to the models' ability to predict churn.

We have studied different aspects related to customer churn prediction using machine learning. By studying these aspects, which are captured in the machine learning process, we contribute to building a complete understanding on how machine learning can be used for churn prediction. Hence, we provide an answer to our research question.

We consider our findings valuable for researchers in our field of study, but also for organizations developing churn management systems as a part of their customer relationship management. In particular, our findings fill the gap in the domain of B2B, and give researchers and firms more knowledge regarding churn prediction in the subscription based service context, which is still underdeveloped.

7.1 Future Work

Our study investigates how machine learning can be used to predict customer churning in the B2B context. As mentioned above, churn prediction is one of two parts in customer churn management, and for future work, it would be interesting to investigate what features to use and how they impact churn prediction in the context of B2B. It would also be interesting to investigate, based on the variables, what measures could be taken related to retention strategies and how organizations should actively act towards customers that are predicted to churn. Another aspect, which would be interesting to further investigate, is what other methods can be used for feature selection and sampling and how they would impact the result. In addition, what other algorithms that could be used and how they could be modified in order to achieve the best result.

8 REFERENCES

[1] R. Shkurti and A. Muça, "AN ANALYSIS OF CLOUD COMPUTING AND ITS ROLE IN ACCOUNTING INDUSTRY IN ALBANIA," *Romanian-American University,* vol. 8, no. 2, pp. 219-229, 2014.

[2] C. Blank and T. Hermansson, "A Machine Learning approach to churn prediction in a subscriptionbased service," KTH, Stockholm, 2018.

[3] D. Buö and M. Kjellander, "Predicting Customer Churn at a Swedish CRM-system Company," Linköpings Universitet, Linköping, 2014.

[4] K. Mishra and R. Rani, "An inclusive survey on machine learning for CRM: a paradigm shift," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017.

[5] H. Gebert, M. Geib, L. Kolbe and W. Brenner, "Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts," *Journal of Knowledge Management,* vol. 7, no. 5, pp. 107-123, 2003.

[6] M. Sergue, "Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company," KTH, Stockholm, 2020.

[7] F. Khodakarami and Y. Chan, "Exploring the role of customer relationship management (CRM) systems in customer knowledge creation," *Information & Management,* vol. 51, pp. 27-42, 2014.

[8] D. L. Garcia, A. Nebot and A. Vellido, "Intelligent data analysis approaches to churn as a business problem: a survey," *Knowledge and Information Systems,* vol. 51, no. 3, pp. 1-56, 2017.

[9] N. Gordini and V. Veglio, "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry," *Industrial Marketing Management ,* vol. 62, pp. 100-107, 2017.

[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access,* pp. 60134-60149, 2019.

[11] W. Verbeke, K. Dejaeger, D. Martens, J. Hur and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research ,* pp. 211-229, 1 April 2012.

[12] A. T. Jahromi, S. Stakhovych and M. Ewing, "Managing B2B customer churn, retention and profitability," *Industrial Marketing Management ,* vol. 43, no. 7, pp. 1258-1268, 2014.

[13] I. Figalist, C. Elsner, J. Bosch and H. H. Olsson, "Customer Churn Prediction in B2B Contexts," Conference: International Conference on Software Business, 2020.

[14] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research,* vol. 94, pp. 290-301, 2019.

[15] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications,* vol. 36, no. 10, pp. 12547-12553, 2009.

[16] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications,* vol. 34, no. 1, pp. 313-327, 2008.

[17] J. Hadden, A. Tiwari, R. Roy and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," *Computers & Operations Research,* vol. 34, no. 10, pp. 2902-2917, October 2007.

[18] N. Singh, P. Singh and M. Gupta, "An inclusive survey on machine learning for CRM: a paradigm shift," *DECISION,* vol. 47, 19 January 2021.

[19] A. K. Ahmad, A. Jafar and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data volume 6,* vol. 6, no. 28, 2019.

[20] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, Data Mining : Practical Machine Learning Tools and Techniques, San Francisco: Elsevier Science & Technology, 2016.

[21] T. Vafeiadis, K. Diamantaras and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory,* vol. 55, pp. 1-9, June 2015.

[22] I. Brandusoiu, G. Toderean and H. Beleiu, "Methods for Churn Prediction in the Pre-paid Mobile Telecommunications Industry," *International conference on communications,* pp. 97-100, 2016.

[23] P. Lalwani, M. M. Kumar, J. Singh Chadha and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing,* pp. 1-24, 2021.

[24] S. R. Gulliver, U. B. Joshi and V. Michell, "Adapted customer relationship management implementation framework: Facilitating value creation in nursing homes," *Total Quality Management and Business Excellence,* vol. 24, pp. 9-10, October 2013.

[25] A. Payne and P. Frow, "A Strategic Framework for Customer Relationship Management," *Journal of Marketing,* vol. 69, no. 4, pp. 167-176, 2005.

[26] S.-Y. Hung, D. C. Yen and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications ,* vol. 31, no. 3, pp. 515-524, October 2006.

[27] T. Jiang, J. L. Gradus and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behavior Therapy ,* vol. 51, no. 5, pp. 675-687, 2020.

[28] P. Mehta, M. Bukov, C. H. Wang, A. G. Day, C. Richardson, C. K. Fisher and D. J. Schwab, "A high-bias,l ow-variance introduction to Machine Learning for physicists," *Physics Reports,* vol. 810, pp. 1-124, 2019.

[29] O. Simeone, "A Very Brief Introduction to Machine Learning With Applications to Communication Systems," *IEEE Transactions on Cognitive Communications and Networking,* vol. 4, no. 4, pp. 648-664, 2018.

[30] S. Marsland, Machine Learning: An Algorithmic Perspective, 2nd edition ed., Chapman and Hall/CRC, 2014.

[31] M. Mohammed, M. b. Khan and E. B. M. Bashier, Machine Learning: Algorithms and Applications, CRC Press, 2016.

[32] M. Oral, E. L. Oral and A. Aydin, "Supervised vs. unsupervised learning for construction crew productivity prediction," *Automation in Construction,* vol. 22, pp. 271-276, 2012.

[33] V. Verdhan, publisher logoSupervised Learning with Python: Concepts and Practical Implementation Using Python, Apress, 2020.

[34] Z.-H. Zhou, Ensemble Methods - Foundations and Algorithms, Taylor & Francis group, LLC, 2012.

[35] A. Kumar and M. Jain, Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases, Apress, 2020.

[36] M. Van Wezel and R. Potharst, "Improved customer choice predictions using ensemble methods," *European Journal of Operational Research,* vol. 181, no. 1, pp. 436-452, 2007.

[37] J. Karlberg and M. Axen, "Binary Classification for Predicting Customer Churn," Umeå University, Umeå, 2020.

[38] D. Windridge and R. Nagarajan, "Quantum Bootstrap Aggregation," in *International Symposium on Quantum Interaction*, 2017.

[39] B. Raja and P. Jeyakumar, "An Effective Classifier for Predicting Churn in Telecommunication," *Journal of Advanced Research in Dynamical and Control Systems,* vol. 11, no. 1, Juni 2019.

[40] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo and J. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 67, pp. 93-104, 2012.

[41] J. Pamina, B. Raja, S. SathyaBama, S. Soundarja, M. S. Struthi, S. Kiruthikas, V. J. Aiswaryadevi and G. Priyanka, "An Effective Classifier for Predicting Churn in Telecommunication," *Jour of Adv Research in Dynamical & Control Systems,* vol. 11, no. 1, pp. 1-9, 2019.

[42] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu and J. Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud," in *2018 IEEE international conference on big data and smart computing*, 2018.

[43] K. P. Murphy, "Naive Bayes classifiers," University of British Columbia, 2006.

[44] J. N. Rouder and R. D. Morey, "Teaching Bayes' Theorem: Strength of Evidence as Predictive Accuracy," *The American Statistician,* vol. 73, no. 2, pp. 186-190, 2019.

[45] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on knowledge and data engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[46] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *nternational Journal of Emerging Technology and Advanced Engineering,* pp. 42-47, 2012.

[47] A. R. Nair, "Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms," in *International Symposium on Embedded Computing and System Design (ISED)*, 2019.

[48] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management ,* vol. 45, no. 4, pp. 427-437, July 2009.

[49] X. Deng, Q. Liu, S. Mahadevan and Y. Deng, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Information Sciences,* Vols. 340-341, pp. 250-261, May 2016.

[50] M. Masarifoglu and A. H. Buyuklu , "Applying Survival Analysis to Telecom Churn Data," *American Journal of Theoretical and Applied Statistics,* vol. 8, no. 6, pp. 261-275, 23 November 2019.

[51] L. Junxiang, "Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS," Sprint Communications Company, 2002.

[52] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications,* vol. 36, no. 3, pp. 4626-4636, 2009.

[53] P. Rothenbuehler, J. Runge, F. Garcin and B. Faltings, "Hidden Markov Models for churn prediction," in *2015 SAI Intelligent Systems Conference (IntelliSys)*, 2015.

[54] S. T. March and G. F. Smith, "Design and natural science research on information technology.," *Decision Support Systems ,* vol. 15, no. 4, pp. 251-266, December 1995.

[55] A. Hevner, S. T. March, J. Park and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly,* vol. 28, no. 1, pp. 75-105, 2004.

[56] P. Ghauri, K. Gronhaug and R. Strange, Research Methods In Busniess Studies, Cambridge University Press, 2020.

[57] "Fortnox," 16 April 2021. [Online]. Available: www.fortnox.se. [Accessed 16 April 2021].

[58] D. Van den Poel and B. Larivière, "Customer attrition analysis for financial services using proportional hazard models," *European Journal of Operational Research,* vol. 157, no. 1, pp. 196-217, 16 August 2004.

[59] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," KTH, Stockholm, 2018.

[60] J. Tang, S. Alelyani and H. Li, "Feature Selection for Classification: A Review," p. 37, 2014.

[61] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 2007.

[62] S. Taheri and M. Musa, "Learning the naive Bayes classifier with optimization models," *International Journal of Applied Mathematics and Computer Science,* vol. 23, no. 4, 2014.

[63] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," University of Wisconsin–Madison, 2018.

[64] G. Hongqing, S. Peiyong, G. Wenzhong and G. Kun, "Component-based Assembling Tool and Runtime Engine for the Machine Learning Process," in *International Conference on Cloud Computing, Big Data and Blockchain*, 2018.

[65] S. Tyagi and S. Mittal, "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning," in *Proceedings of ICRIC 2019*, 2020.

[66] "Scikit Learn," [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html. [Accessed 18 April 2021].

[67] B. J. Jacksson, P. Korfiatis, Z. Akkus and T. L. Kline, "Machine Learning for Medical Imaging," *RadioGraphics,* vol. 37, no. 2, pp. 505-515, 17 February 2017.

[68] J. Burez and D. Van den Poel, "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services," *Expert Systems with Applications Volume 32, Issue 2, February 2007, Pages 277-288,* vol. 32, no. 2, pp. 277-288, February 2007.

[69] R. O. Mujalli, G. Lopez and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Analysis & Prevention ,* vol. 88, pp. 37-51, March 2916.

[70] S. Mithas, S. M. Krishnan and C. Fornell, "Why Do Customer Relationship Management Applications Affect Customer Satisfaction?," *Journal of Marketing,* vol. 69, no. 4, March 2006.

[71] A. Tamaddoni Jahromi, "Customer attrition analysis for financial services using proportional hazard models," Luleå University of Technology, 2009.

[72] M. Tamassia, W. Raffe, R. Sifa, A. Drachen, F. Zambetta and M. Hitchens, "Predicting Player Churn in Destiny: A Hidden Markov Models Approach to Predicting Player Departure in a Major Online Game," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, Santorini, Greece, , 2016.

[73] S. J. Cunningham, "Machine Learning and Statistics: A matter of perspective," University of Waikato, Hamilton, 1995.