CHURN RATE PREDICTION IN THE HEALTHCARE INDUSTRY - A COMPARATIVE STUDY OF MACHINE LEARNING MODELS

MADALINA SECELEANU

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

410556

COMMITTEE

dr. Giacomo Spigler
dr. Emmanuel Keuleers

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

WORD COUNT: 7977

DATE

December 3, 2021

ACKNOWLEDGMENTS

Dear Reader,

Thank you for taking the time to read this thesis on churn prediction in the healthcare industry! Firstly, I would like to thank my thesis supervisor Giacomo Spigler for all the help during the thesis process. I would also like to thank Ivo Willems for inspiring me to write this paper on such an interesting topic. Finally, I would like to thank my family, Bianca Seceleanu, Giovanni Tokarski and Antonio-Stefan Chiriac for supporting and encouraging me in the toughest times.

I hope you enjoy reading my thesis,
Madalina Seceleanu

CHURN RATE PREDICTION IN THE HEALTHCARE INDUSTRY - A COMPARATIVE STUDY OF MACHINE LEARNING MODELS

MADALINA SECELEANU

Abstract

This thesis examines the prediction power of four machine learning models (Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Decision Trees) in determining the churn rate of a healthcare service provider. Churn rate prediction is an important exercise, that tries to determine how many current customers might leave the company and switch to another service provider. This endeavor ensures customer retention and increases customer loyalty. This paper starts by exploring the current literature on the subject, extracting valuable information, and implementing them on the given dataset. Due to class imbalance, the thesis explores two methods of data-sampling: random under-sampling and over-sampling using SMOTE. The machine learning models are then trained on both sets, and a comparison is made between the different datasets and the performance of each classifier applied to them. After running the models, it was found that the best-performing model was Support Vector Machine trained on the SMOTE dataset. The models were evaluated according to the F-score, AUC-ROC score, and accuracy. This study is one of the few done on the healthcare industry, and research should progress in this field, due to the many benefits for both service providers and customers.

1 DATA SOURCE/CODE/ETHICS STATEMENT

The author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is publicly available: https://github.com/MadalinaSe/ThesisCode.git.

2 INTRODUCTION

The goal of this study is to compare the performance of four machine learning algorithms using data sampling techniques to determine the churn rate of a healthcare equipment service provider. The churn rate is a metric that shows what percentage of customers choose to discontinue their subscriptions with a certain service provider or to not renew their contract once it expires OED (n.d.). The paper will analyze customer behavior based on features extracted from previous purchases to determine what percentage of the contracts will be renewed or extended in the next period. The analysis will be performed on a dataset provided by a Dutch company that has asked to remain anonymous, and shall henceforth be referred to as Company A.

Customer retention is an objective for any business that wants to focus on sustainable and predictable growth. Building a strong relationship with customers and understanding their needs ensures profitability, as well as higher satisfaction levels and loyalty. From a revenue standpoint, customer retention is at least five times less costly than acquiring a new customer CIM (n.d.). Moreover, by investing in the analysis of customer behavior and customizing the commercial offers to their needs, a service provider gains a competitive advantage over their competitors Nguyen, Sherif, and Newby (2007). Since the early '90s, companies have been creating Customer Relationship Management (CRM) systems, which enable them to acquire new customers, provide better service to their existing customer base, and design targeted marketing campaigns to unlock novel revenue streams Buttle (2008). CRM is a combination of business approaches and processes which allow businesses to focus on the customer instead of their products. As part of the CRM, some service providers have developed technological solutions, such as employing data mining to determine customer behavior. Securing customers from churning requires identifying which of them are at risk to cancel their contracts and developing proactive marketing campaigns especially targeted at churners Zhang, Zhu, Xu, and Wan (2012). Churn prediction enables companies to have some foreseeability of future trends and make decisions based on historical records.

Within the healthcare industry, a strong relationship with the customer is crucial to build a foundation of trust and ensure patients can access the equipment safely and effectively. As opposed to other industries, the healthcare domain carries more responsibility when it comes to delivering quality service in a very short period. In a hospital, for instance, some of the most important types of equipment are within the imaging department. The downtime of diagnostic imaging alone can produce almost $300,000 of lost revenue yearly in a medium-sized hospital Becker, Goldszal, Gronlund-

Jacob, and Epstein (2015). For a service provider in the healthcare industry, such as the one presented in this study, the main customers are the medical facilities that own such types of equipment. Being able to understand their needs and problems, and proactively acting to solve imminent issues builds a trust system that ensures the continuation of existing contracts. Churn prediction plays a crucial role in flagging contracts that are at risk of being discontinued, allowing the service provider to investigate the possible cause, and to open a channel of communication with its customer.

The process of predicting the churn rate combines business knowledge, data mining techniques, and machine learning algorithms. Firstly, understanding the business means being able to select features that are important to the models and drawing the correct conclusions from the data analysis. The data mining component allows us to extract knowledge from the data, visualize it and prepare the dataset to be fed to the machine algorithms. The prediction model uses a classification process that categorizes the training set in two classes: Churn and Not Churn. Ultimately, this model learns to classify unlabeled data using pattern recognition and statistics.

Churn rate prediction using machine learning algorithms and data mining techniques has been previously researched in industries such as "telecommunication, banking, retail, and cloud service subscriptions" Sabbeh (2018). The first way in which this paper distinguishes itself from the existing literature is through the industry, as healthcare has rarely been included in churn prediction projects. Some of the previously published works have approached churn rate prediction by evaluating only one classifier Adwan, Faris, Jaradat, Harfoushi, and Ghatasheh (2014); Kwon, Kim, An, and Park (2021); Rodan, Faris, Alsakran, and Ah-Kadi (2014). A common obstacle in churn prediction is the imbalanced distribution of classes, i.e., the churning class is much smaller than the non-churning one. Left unresolved, the difference in class sizes might bias the algorithm and create an illusion of well-performing models. To tackle this issue, some studies have pursued a combination of machine learning algorithms, while others have used sampling techniques Brandusoiu, Toderean, and Beleiu (2016); Hudaib, Dannoun, Harfoushi, Obiedat, and Faris (2015). This paper will use several machine learning models and compare their performance in relation to two sampling techniques. It will combine the comparative analysis of different algorithms to establish which one would be more suited for the given dataset and explore data sampling techniques to overcome the class imbalance issue of churn prediction.

The main research question this paper will pursue is:

> *How well can we predict the churn rate of a healthcare service provider by building a model using several machine learning algorithms and data mining techniques trained on historical contract data?*

This research question will be answered with the help of three sub-questions:

RQ1 *How does the performance of the machine learning algorithms change between different data sampling methods?*

RQ2 *Which of the evaluation metrics should we use to determine if the models are performing well? Should we consider F1-score rather than accuracy?*

RQ3 *Which of the listed machine learning algorithms will be the best performing one given the structure of the dataset and the selected features against the baseline model?*

The rest of the paper is organized as follows: Section 3 covers the most relevant papers in the existing literature on churn prediction. Section 4 explains methods used to predict the churn rate, while Section 5 describes the experimental setup. Section 6 shows the results of the experiments. Section 7 discusses the results in relation to the research questions, and finally, the conclusion is presented in Section 8.

3 RELATED WORK

Churn rate prediction using machine learning and data mining algorithms has been previously researched, mostly concerning industries such as telecommunication, retail, and banking. The existing literature has dealt with this research question by either using one classifier to predict the churn rate, employing several different models and comparing the accuracy of each of the models or combining different techniques such as sampling or clustering with predictive models. Another common research topic concerning churn prediction is overcoming the class imbalance issue. The health care field has not been often considered in churn rate prediction literature. Therefore, the review below will consider the services industry altogether to gain insight on methods and best-case practices in customer retention models.

3.1 *Single classifier algorithms*

A very common approach to the churn prediction problem in literature has been implementing single classifier algorithms. The choice of models is very diverse, depending on the nature of the dataset, the industry, and the aim of the papers. Most recently, Kwon et al. (2021) used neural networks (NN) to predict the churn rate of a digital health care app in Korea. The app provides customers with personalized coaching on

alimentation, exercise, and well-being. In their study, the authors had to implement topic modeling, as the database consisted of text messages which had to be vectorized. The neural networks model proved to be highly effective in predicting churning customers of the health care app. Likewise, Sharma and Panigrahi (2011) used NN to predict the customer churn in their study, resulting in an accuracy of prediction of 92%. SVM has been shown to perform well when the hyperparameter is well-tuned using a grid search and cross-validation Rodan et al. (2014).

While some papers focus on researching one model, others decided to use a comparative approach. Brandusoiu et al. (2016) applied three machine learning models on a dataset concerning prepaid mobile services. They compared the performances of Neural Networks, SVM, and BN, and found that SVM was the most accurate model. Although not listed in the most popular models, K-Nearest Neighbors (KNN) has also been shown to be highly effective. When comparing the KNN model against SVM and random forest, the KNN models yielded the best accuracy Sjarif, Yusof, Ya'akob, Ibrahim, and Osman (2019). Dahiya and Bhatia (2015) analyzed the performance of the logistic regression classifier against the DT one in the telecommunication industry. In this paper, DT has been shown to have a much higher accuracy than LR.

Boosting is often used in churn prediction, to enhance the performance of the machine learning models. Sabbeh (2018) proposed a study of eight machine-learning models engaged to predict the customer churn rate of a telecommunication company. The author used Logistic regression, Decision tree, Naïve Bayes, Support Vector Machine, K-nearest Neighbor, Ensemble learning (Ada Boost, Stochastic Gradient Boost, and Random Forest), Artificial neural network, and linear discriminant analysis. The paper concluded that the best performing models were random forest and Ada boost, proving that there was little difference between the performance of the models with and without boosting. Similarly, Vafeiadis, Diamantaras, Sarigiannidis, and Chatzisavvas (2015) studied the comparison between simple ML models and models with boosting, using ANN, SVM, DT, NB, and LR. The study found that the best performing algorithm was the boosted SVM. However, the boosting algorithms only improved accuracy by between 1% and 4%.

3.2 *Class imbalance problem*

The class imbalance problem is a very common issue in the churn prediction field, due to the rarity of churning customers in databases. Class imbalance needs to be addressed before applying the machine learning models as it might cause algorithm bias and incorrect evaluation metrics.

To overcome this concern, existing research has been exploring three methods: sampling solutions, algorithm-level solutions, and ensemble solutions Zhu, Baesens, and vanden Broucke (2017). The data level approach entails modifying the data distribution through oversampling or undersampling. The most common oversampling technique is the one using Synthetic Minority Oversampling Technique (SMOTE) Amin et al. (2016). To obtain improved accuracy, a further measure is using the ensemble strategy, which entails combining decisions from numerous classifiers, like RF, Boosting, and Bagging. The third category, algorithm-level solutions, focuses on improving the ability of the classification to learn on the minority class Zhu et al. (2017).

There are a lot of papers in the churn prediction literature that chose the data level approach to solve the class imbalance issue. In a study of churn rate prediction in the mobile internet industry, Gui (2017) compared three data sampling methods: oversampling, under-sampling, and SMOTE. They found that the random oversampling method used with a random forest model had the best evaluation metrics, followed very closely by the combination SMOTE and RF. Brandusoiu et al. (2016) also addressed the imbalanced classes in their paper on the churn rate prediction of customers who had purchased prepaid mobile services by randomly oversampling the training set, with great overall performance of the models. Farquad, Ravi, and Raju (2014) compared the same three random sampling techniques and found SMOTE to be the best method in combination with an SVM classifier, while He, Shi, Wan, and Zhao (2014) found that random sampling improved the performance of the SVM model.

Other studies use the ensemble solution for the class imbalance. For instance, Hudaib et al. (2015) studied the churn rate of the Jordanian Telecommunication Company. The churn rate of customers was 7.6%, which determined a very problematic class imbalance. To solve this problem, the study used k-means clustering, hierarchical clustering, and Self Organizing Maps. The paper found that the highest accuracy was scored by the combination of k-means clustering and Multilayer Perceptron Artificial Neural Networks, with an accuracy score of 97.2%, performing much better than a common simple model. Huang and Kechadi (2013) also used k-means clustering in their study and concluded that the combination of clustering and classification models yield maximum accuracy possible.

4 METHODS

This chapter will explain the methods used in this research, as well as the motivation behind the choice of models. It will also discuss the hyperparameters of the models which have been selected using the Scikit-learn

library in Python to improve the performance of the models. Finally, it will present the evaluation methods applied to determine the most accurate algorithm for churn rate prediction on the given dataset. The four machine learning algorithms used in this paper are Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Decision Tree.

4.1 *Logistic Regression*

Logistic Regression is a supervised learning algorithm used for classification problems, which functions based on a linear regression equation that produces binary outputs by predicting the probability of churn using the sigmoid function. If p is the probability that the label is positive (P(z=1), customer churned), and z is the linear input function, we define the logit function as follows:

$$logit\,(p) = log\left(\frac{p}{1-p}\right) \tag{1}$$

The predicted value z is converted into a probability value using the sigmoid function, which takes values between 0 and 1, and is defined as the inverse of the logit function and shown in equation 2 and Figure 1:

$$P\,(z) = \left(\frac{1}{1+e^{-z}}\right) \tag{2}$$

In order for the predicted values to then be turned in a categorical label ("churn" and "not-churn"), we can select a threshold value (generally 0.5), called the decision boundary which determines under which label the predicted value will be mapped. We can mathematically express this relationship as:

$$If\,p \geq 0.5, class = 1, otherwise, if\,p < 0.5, class = 0 \tag{3}$$

The linear regression model is very often used in the estimation of churn rate and performs very well Sabbeh (2018). The choice of using this model in this study comes from its prevalence in the existing literature, as well as its intuitive application and straightforward implementation. Although logistic regression has rarely been the best performing model, especially in comparative studies, it serves as a good baseline model and starting point in churn rate prediction.

4.2 *K-Nearest Neighbors*

K-Nearest Neighbors (KNN) is an algorithm that was proposed by Hodges and Fix as a non-parametric method for pattern classification Fix and

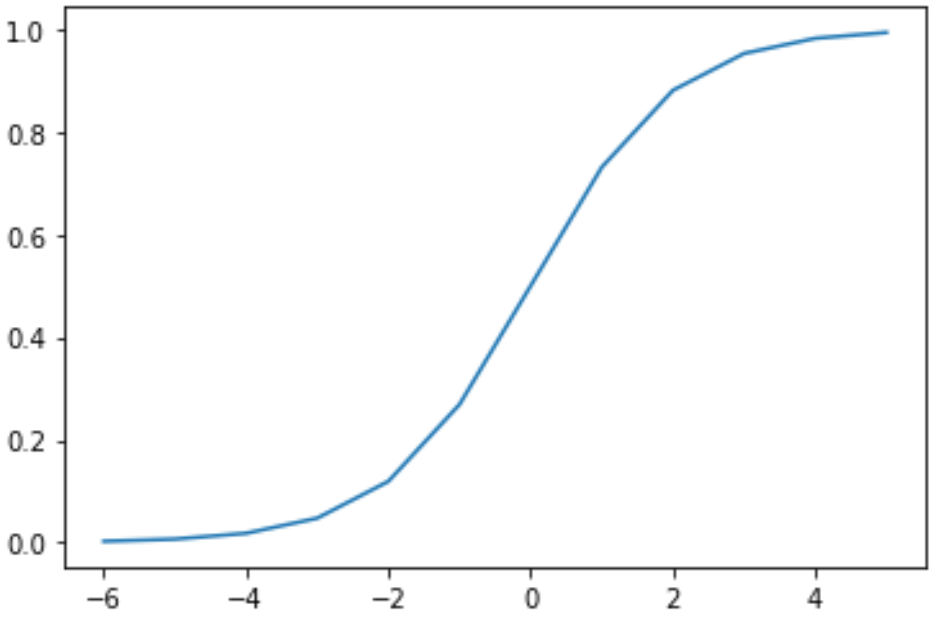

Figure 1: Sigmoid Function

Hodges (1951). KNN is often referred to as a lazy learning model, due to its instance-based approach, which is advantageous on the one hand, since it needs no prior knowledge on the distribution of the data, and disadvantageous on the other hand, as it can be computationally expensive Bhatnagar and Srivastava (2019). KNN can determine the class of a data point by looking at the classes of k closest to other points, establishing the median class, and assigning that class to the original point. The "closeness" of the k points can be determined by measuring the distance between data points. The most commonly used method to determine the distance in classification problems is the Euclidean distance, defines as:

$$Euclideandistance = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{4}$$

The value of k should be chosen based on the dataset, as small values of k might not be able to generalize well and be highly sensitive to noise, while large values of k might result in underfitting Sjarif et al. (2019). In the cases in which the issue of imbalances datasets has not been dealt with, the KNN algorithm might also misclassify cases as the largest class has a higher probability of being nearest to the instances that need to be predicted Vanhoeyveld and Martens (2018). In the existing literature that this study has covered, KNN is not always favored as the best model to predict churn rate. However, this study included KNN in its analysis due

to its intuitive nature and to further research the use of KNN in customer retention studies.

4.3 *Support Vector Machine*

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification problems. SVM was first introduced by Cortes and Vapnik (1995), and it classifies data points by finding one hyperplane in an N-dimensional that optimally separates the instances into two classes. As there a many possible hyperplanes that could be chosen, the goal of the algorithm is to find the maximum distance between data points of the two classes. The support vectors are the instances closest to the hyperplane that influence the position of the said hyperplane. The decision boundary is then decided based on the kernel function, which can be a linear, sigmoid, or a polynomial function Coussement (2008) .

SVM is an algorithm that is easily implemented by the computer but is quite difficult to comprehend by humans Farquad et al. (2014). While it has difficulties in the case of large datasets, Support Vector Machine performs very well on smaller samples that imply a pattern recognition problem Rodan et al. (2014). The main reason for choosing this model for this paper stands in its high accuracy, good computational performance, and capacity to generalize well without overfitting Rodan et al. (2014).

4.4 *Decision Trees*

Decision Trees (DTs) are non-parametric supervised learning models, which can be applied to both categorical and continuous data, and function based on a set of simple "if-then-else" decision rules. As the name suggests DTs have a structure similar to a tree and can be applied to classification and regression problems. A DT is composed of several elements known as nodes, leaves, and branches. Furthermore, nodes can be classified into root nodes, decision nodes, and leaf nodes. The root node is the highest node that splits the entire dataset into two or more sets. All the other sub-nodes which split the dataset further into sub-nodes are called decision nodes. A node that does not split any further is the leaf node or terminal node. The elements between nodes are called branches, and they represent the rules previously mentioned as "if-then-else" decisions. The most known types of decision trees are ID3 and CART (Classification and Regression Trees) Ngai, Xiu, and Chau (2009). The complex interaction of variables can lead to overfitting and poor generalization, which is why stopping rules and pruning are important steps when using DTs in classification problems Song and Lu (2015). Stopping can be done through parameters

like the minimum number of records in a leaf, while pruning is removing sub-nodes of a decision node Song and Lu (2015).

Although DTs are just as difficult to understand and interpret by humans as SVM, they are the second most researched models in churn rate prediction after Artificial Neural Networks Ngai et al. (2009). From a computational perspective, decision trees are inexpensive and often yield high accuracies with the right setting of hyperparameters chosen based on the available dataset Ngai et al. (2009).

4.5 *Class imbalance*

The problem of imbalanced data distribution is very common in churn rate prediction studies. Companies have much less data on the customers who discontinue their services than customers who decide to stay, although the former is of more interest from a business perspective Burez and den Poel (2009). Class imbalance is an issue for machine learning algorithms because it can greatly influence the prediction values, which would result in wrong conclusions. Since the "not-churn" class is larger than the "churn" class, the former has a higher weight. As will be discussed in Section 5.1, the dataset analyzed in this paper is imbalanced, with an actual churn rate of 32%. In the existing literature, there are three categories of methods used to deal with imbalanced data: data resampling, cost-sensitive learning, and ensemble techniques Nguyen et al. (2007). This paper will be examining the performance of the machine learning models on two data resampling methods: random under-sampling and Synthetic Minority Oversampling Technique (SMOTE).

Random under-sampling involves randomly removing features from the majority class until the two classes have the same number of instances. By using this method, valuable data is being discarded without even knowing what type of information it possesses, decreasing the performance of a classifier.This can result in an improbable sample that does not replicate what happens in a real-life scenario with customers of a company. However, under-sampling is computationally less expensive and yields good accuracy scores in churn rate prediction Salunkhe and Mali (2018).

SMOTE is a type of oversampling technique and it works by producing synthetic instances of the minority class to combat the imbalance in the dataset. This method selects an example in the feature space that is in the minority class and finds the k nearest neighbors to that data point. One of the neighbors is randomly chosen, and a synthetic instance is placed at a random distance between the neighbor and the initial example. This method not only does not discard important information but also created a plausible data set that mimics real-life behavior. One disadvantage to

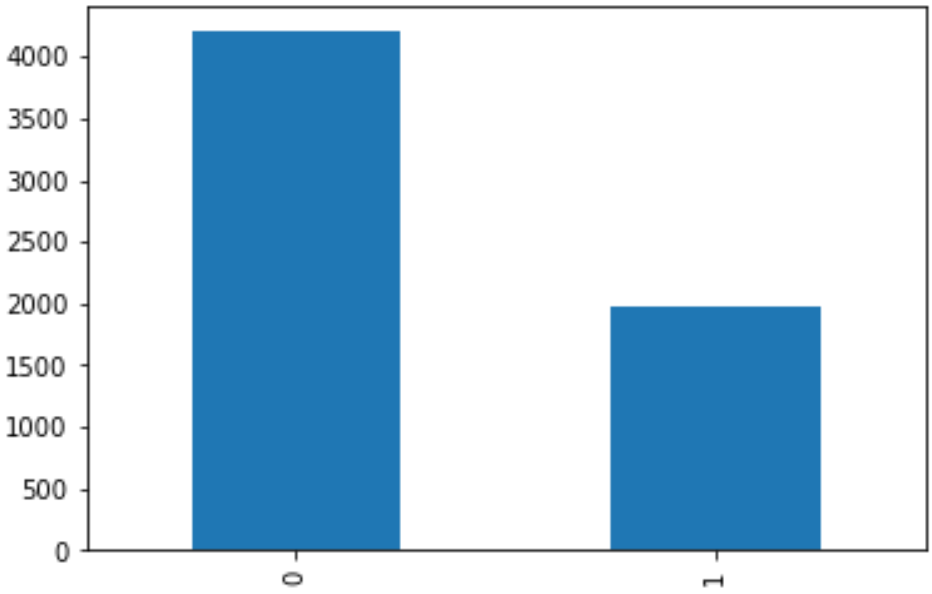

Figure 2: Class imbalance

this method is that SMOTE does not consider the possibility that the selected neighbors could be from the majority class, which would create an overlapping of classes Chawla, Bowyer, Hall, and Kegelmeyer (2002).

5 EXPERIMENTAL SETUP

5.1 *Dataset*

The data used in this paper was provided by a company, which has asked to remain anonymous. This company will be henceforth referred to as Company A. Company A manufactures medical equipment and offers service for the equipment which was installed through contracts such as warranties and service contracts. The dataset which will be analyzed by this study concerns only services, containing about 6000 service contracts signed between 2014 and 2021. The customers are not individuals, patients, or hospitals, but rather distributors and service providers. The distributors offer their services on behalf of Company A to maintain and repair the medical equipment manufactured by Company A. Therefore, the features of the dataset describe the type of distributor, their origin, their maturity from Company A's perspective, the type of contract, the type of equipment they provide service for, the tenure of the contract, the end of life of the system, and other such features. From this perspective, this paper

differentiates itself from existing literature, through the nature of the customer.

The churn rate is determined by the status of a contract, whether it was renewed or not. The churn rate is a binary variable where 0 means the customer did not churn, and 1 means the customer churned. The dataset is standing at 32% of customers churning. As seen in Figure 2 and mentioned in Section 4.5, the dataset presents imbalanced classes with 4201 non-churners and 1979 churners. To solve this issue, this paper has tried two sampling techniques: random under-sampling (RUS) and random over-sampling using SMOTE. When using SMOTE, the over-sampling was performed after splitting the data into training and testing set. This is an important aspect because of the way SMOTE produces the synthetic data points using the KNN methodology. If the split would have been done before the train and test set creation, information from the test set could have been corrupting the test set, making the model less generalizable while maintaining a high accuracy, thereby creating the illusion that the model performs well.

5.2 *Preprocessing*

The first step in the analysis of the dataset was done in the preprocessing stage. The data was extracted in an Excel file where some feature engineering has been used. Firstly, based on the start date of the contract, the year and month were extracted to be used as variables in the prediction of the churn rate, specifically to account for seasonality. Apart from the type of the distributor and the region they conduct their business in, there was no information in the given dataset about customers. The dataset did contain a customer identification code, which was further used to provide the variable "Sales Quartile". "Sales Quartile" was calculated based on the 24-month rolling sales performance of the distributor. The variable contains four categories (Q1, Q2, Q3, and Q4), and it splits the distributor into four levels of performance in terms of sales from Company A's perspective. Therefore, a distributor in the Q4 would be a desirable partner who is capable of selling a lot of contracts and should be targeted for customer retention, while a Q1 distributor is one of the lowest-performing partners. Table 1 shows the description of all the features used in the ML modelling.

5.3 *Binary Variable Encoding*

As most of the variables in the dataset are categorical, there was a need for variable encoding in the exploratory data analysis part of the research. This study has chosen binary encoding, where categories are firstly encoded

Table 1: Description of features used in the ML models

Feature	Description
Modality	Type of medical equipment, categorical variable with 14 different categories
Market	The region where the distributor activates, categorical variable with 16 categories
Contract Type	Type of contract, categorical variables with 12 categories
Main Equipment Age Bucket in Years	The age of the contract, calculated as duration since it has been installed, categorical variable with 7 categories
Duration in Days	The age of the contract, calculated based on difference between the start date and end date of each contact, continuous variable
Sales Year	Year the service contract was sold to the distributor, categorical variable with 4 categories: 2018, 2019, 2020 and 2021
Contract Month	Month the contract was sold, categorical variable with 12 categories
Sales Value EUR	Value of the contract in euros, continuous variable
Sales Quartile	Classification of distributor, based on the value of their contract sales, categorical variable with 4 categories: Q1, Q2, Q3, Q4
Partner Type	Type of the distributor based on internal nomenclature, categorical variable with 10 categories
EOL Reached	Whether the end of life of the system installed has been reached, binary variable where 0 means end of life was not reached, and 1 is the inverse

into integers, and then converted into binary code. This method was preferred to the one-hot encoding method, due to the large number of categories for each categorical variable, therefore reducing dimensionality. After performing the binary encoding of the variables, the set of features

contains 39 variables to be fed to the machine learning algorithms, 37 of which were binary variables.

5.4 *Normalization*

Normalization was performed on the three numerical variables: "Duration in Days", "Sales Year" and "Sales Value Eur". Normalization is the method of scaling the data so that it falls in a smaller range, most commonly between 0 and 1. This step in data processing is important because it prevents the bias of the machine algorithms towards values that are on a different scale. This procedure was done in Python using MinMaxScaler, and the normalization produced values between 0 and 1.

5.5 *Hyperparameters*

In order to be able to enhance the performance of the models and compare their results in an objective manner, this paper searched and examined the hyperparameters of each of the four machine learning algorithms. Hyperparameters are settings of the models that change the configuration of the learning algorithms. Python contains the Scikit-learn library which was used to tune the parameters. To ensure fair evaluation this paper uses is the 5-fold cross-validation (CV) when tuning the models. The model splits the training set into 5 sets on every iteration, uses 4 sets to train the data and the remaining part of the data is used for validation. This process is repeated 5 times, each time with a different split. The performance is then averaged across all values computed and reported as the final result. Cross-validation helps in optimizing the machine learning algorithms for better overall performance, without losing the information by having a validation set specifically for hyperparameter tuning. Although this method can computationally expensive, it helps to not lose important data in the validation procedure.

5.6 *Evaluation metrics*

The evaluation metrics are values that quantify the performance of machine learning models. There are several choices for evaluating the quality of a classification algorithm. This paper will focus on F1-score and ROC curves, as well as its corresponding ROC-AUC score. These values are based on the confusion matrix, which contains the actual prediction capacity of a classifier. The confusion matrix contains the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) an

algorithm has predicted. Based on these values we can calculate the accuracy, precision, recall, and F1-score values as shown below in equation 5, equation 6, equation 7, and equation 8.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - score = \frac{2xPrecisionxRecall}{Precision + Recall} \tag{8}$$

The most widespread metric used for evaluation in churn rate prediction is accuracy Huang and Kechadi (2013). As seen in equation 5, accuracy shows the number of correct predictions out of the size of the entire dataset. This paper also uses F1-score to quantify performance because it considers both recall and precision, which helps in comparing multiple models. Moreover, the F-score is a better measure for the incorrectly classified instances. While this study wants to be able to report high accuracy, minimizing the wrongly predicted cases is an important approach from a business perspective. F1-score is also known to be performing well in class imbalance problems.

This study will also use the Receiver Operation Characteristic (ROC) curve, as well as the AUC (Area Under the Curve) calculated in Python by the roc-auc-score function. The ROC curve is built by plotting the true positive rate (TPR) against the false positive rate (FPR). TPR and FPR are calculated as shown below in equation 9 and equation 10:

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{TN + FP} \tag{10}$$

The advantage of using the ROC curve to compare the performance of several machine learning algorithms is that it balances the tradeoff between false positives and false negatives. In this research, it might be more helpful to have a lower number of false negatives than false positives. False positive cases would be instances that are predicted to churn, but in reality, do not churn. The outcome would simply be that Company A would give more attention to these customers, who did not have the intention to churn. The reverse, however, is the situation that we want to make sure does not occur too often. This would mean that we are not

correctly predicting the intention of some customers to churn, therefore risking losing these customers altogether.

5.7 *Software*

The data was collected from the Qlickview dashboard of Company A and exported in an Excel file. The first step in the data preparation was done in Excel. For the analysis, the programming language Python 3.0 was used in the Jupyter environment. The data processing, as well as the model implementation, have been done in Jupyter, by use of several libraries, such as Scikit-learn, Numpy, Pandas, and Matplotlib. These libraries contain easy-to-use functions for the machine learning algorithm as well as accessible ways to visualize the data. The figures presented in this paper were also realized using Python.

6 RESULTS

This chapter will contain the results of the machine learning algorithms discussed in Section4. The section will first show the performance of each of the models on the under-sampled and over-sampled sets. The models were fit on the pre-processed data, and a grid search with 5-fold cross-validation was performed to find the best hyperparameters, as discussed in Section 5.5. The F1-score and accuracy will then be presented, along with the AUC score and the ROC curves for model comparison. The chapter will then summarize the results in the final part of the chapter.

6.1 *Logistic Regression (LR)*

The first model which we will be discussing is Logistic Regression, which will also be used as a baseline model. Table 2 shows the performance of the LR model in both the under-sampling and over-sampling situations. The model performs better on the over-sampling data set, resulting in an F-score equal to 0.74 and an AUC score of 0.82. To achieve these values, 5-fold cross-validation was performed to find the best performing model, setting the following hyperparameters: C, penalty, and solver. The logistic regression model involves an optimization problem, and the parameter called solver determines the algorithm used in the optimization. Solver has five options in Scikit-learn: “newton”, “lbfgs”, “liblinear”, “sag”, and “saga”, which were used in the cross-validation stage. The parameter penalty, also known as the regularization parameter, shrinks the coefficients of the less important features to 0, and has four options: “none”, “l1”, “l2”

and “elasticnet”. Finally, C is the inverse regularization parameter. After running the grid search, the best-performing logistic regression contains the parameters shown in Table 3.

Table 2: Logistic regression evaluation metrics after hyperparameter tuning

	Evaluation metrics		
Sampling method	F1-Score	AUC Score	Accuracy
Under-sampling	0.68	0.75	0.68
Over-sampling(SMOTE)	0.74	0.82	0.73

Table 3: Hyperparameters for LR found using CV

	Hyperparameters		
Sampling method	C	Penalty	Solver
Under-sampling	0.1	"l2"	"liblinear"
Over-sampling(SMOTE)	0.7	"l2"	"lbfgs"

6.2 *K-Nearest Neighbors (KNN)*

Table 4 presents the performance of the KNN classifier on the undersampled and oversampled data sets. As shown in the table, the model performs better in the oversampled situation, with a higher F1-score. The values of the metrics are, however, very close, therefore no significant change has been observed in the performance of KNN between different sampling techniques. In the case of KNN, the CV was done on the following parameters: neighbors, algorithm, and weights. Neighbors determine the value of k or the number of neighbors used in the calculation of distance. Algorithm is the formula used to compute neighbors, and weights determine whether nearby neighbors are given more weight than further away ones. The CV search found that the best KNN classifier used the values shown in Table 5.

6.3 *Support Vector Machine (SVM)*

The Support Vectors Machine classifier outperformed all the other ones, with a 0.91 AUC score, 0.84 F-score, and a 0.85 accuracy score on the oversampled dataset. Table 6 summarizes the comparison between the performance of the SVM model on the two different sampling methods.

Table 4: KNN evaluation metrics after hyperparameter tuning

	Evaluation metrics		
Sampling method	F1-Score	AUC Score	Accuracy
Under-sampling	0.75	0.83	0.76
Over-sampling(SMOTE)	0.79	0.85	0.8

Table 5: Hyperparameters for KNN found using CV

	Hyperparameters		
Sampling method	Neighbors	Algorithm	Weights
Under-sampling	5	"ball-tree"	"distance"
Over-sampling(SMOTE)	1	"ball-tree"	"uniform"

The hyperparameters which have been tuned for SVM are C and kernel. The kernel parameter has five options: “linear”, “poly”, “rbf”, “sigmoid”, and “precomputed”. The best parameters for the studied case found by the grid search are the same for both sampling methods, C: 2 and Kernel: “poly”.

Table 6: SVM evaluation metrics after hyperparameter tuning

	Evaluation metrics		
Sampling method	F1-Score	AUC Score	Accuracy
Under-sampling	0.74	0.82	0.77
Over-sampling(SMOTE)	0.84	0.91	0.85

6.4 *Decision Trees (DT)*

As shown in Table 7, the Decision Tree classifier performed similarly well to the SVM one, with an AUC score of 0.9, and an F-score of 0.82. This model has also been tuned using the Scikit-learn library and CV on the following hyperparameters: criterion, max-features, min-samples-leaf, max-depth and splitter. Criterion is the function the model uses to assess the quality of the split. Max-features are the number of features considered when splitting, and it was set to "none" for this analysis, and splitter is the strategy used to split at each node. Max depth is the depth of the tree, while min-samples-leaf is the parameter that determine the number of

samples required to be at a leaf node. The hyperparameters chosen by the CV for the DT classifier are summarized in Table 8.

Table 7: DT evaluation metrics after hyperparameter tuning

	Evaluation metrics		
Sampling method	F1-Score	AUC Score	Accuracy
Under-sampling	0.75	0.83	0.77
Over-sampling(SMOTE)	0.81	0.9	0.82

Table 8: Hyperparameters for DT found using CV

	Hyperparameters		
Sampling method	Criterion	Max-depth	Splitter
Under-sampling	"Entropy	14	"Random"
Over-sampling(SMOTE)	"Entropy"	14	"Best"

6.5 *Confusion matrixes*

The evaluation metrics presented in Section 6.1, Section 6.2, Section 6.3, and Section 6.4 calculated based on the confusion matrixes shown in Figure 3 and Figure 4. While the F-score and ROC curves make it easier to compare the results of the models, confusion matrixes make it possible to visualize the number of FN and FP predicted. For instance, in the under-sampling situation, it is notable that the KNN classifier had by far the least number of false negatives. That might interest a business that wants to make sure that FN is minimized. In the oversampling situation, the SVM had the smallest number of false predictions, differentiating itself from DT, although their ROC-AUC scores are similar.

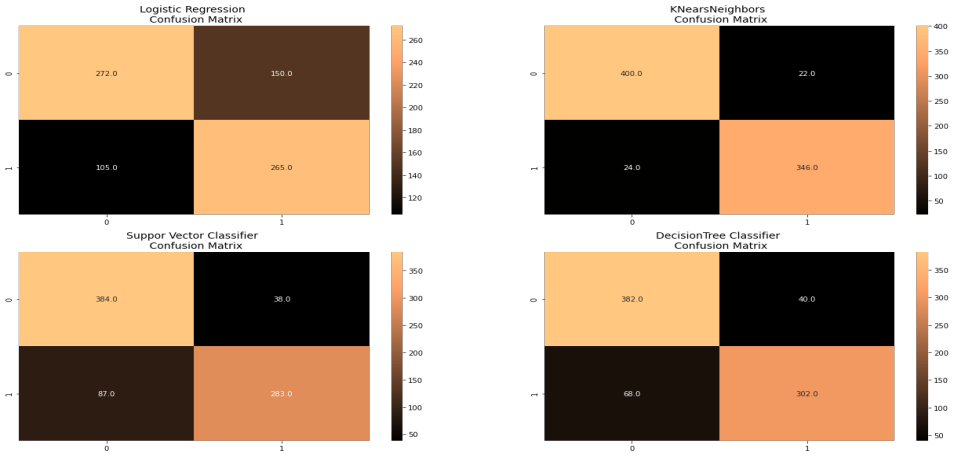

Figure 3: Confusion matrixes of the models - Under-sampling case

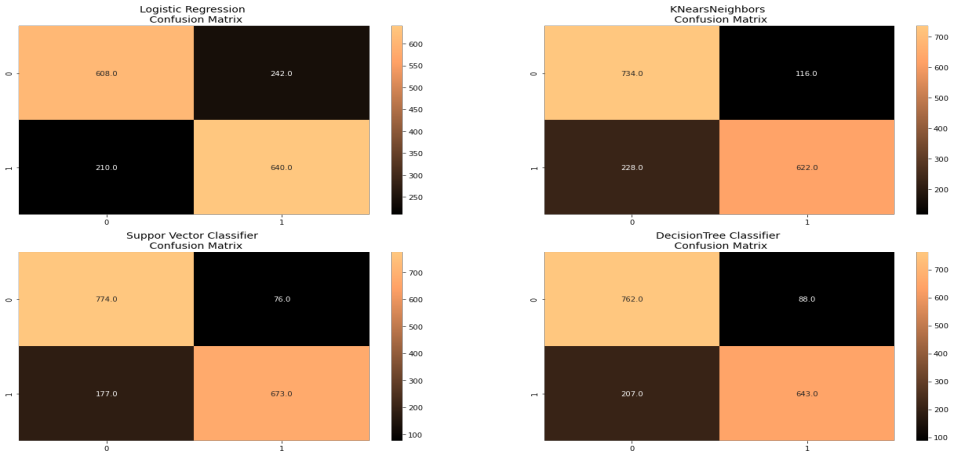

Figure 4: Confusion matrixes of the models - Oversampling case

6.6 *ROC curves*

The ROC curves of the models are presented in Figure 5 and Figure 6. Figure 5 presents the performance of the four models in the under-sampling case, alongside the random line, which is denoted by the dotted line and represent the minimum ROC score of 50%, where the model is not learning, but rather guessing. We can see that KNN, SVM and DT are very closely tying for the best ROC curve, with values for the ROC-AUC score of 0.82 for SVM and 0.83 for the DT and KNN.

Figure 6 shows the ROC curves of the four models in the oversampling case. In this figure, we can see a big improvement of the DT and SVM classifiers (ROC-AUC score of 0.91), as opposed to the case of the under-sampled dataset. The KNN classifier performs better than the baseline

logistic regression model (ROC-AUC score of 0.85), why LR had a ROC-AUC score of 0.82.

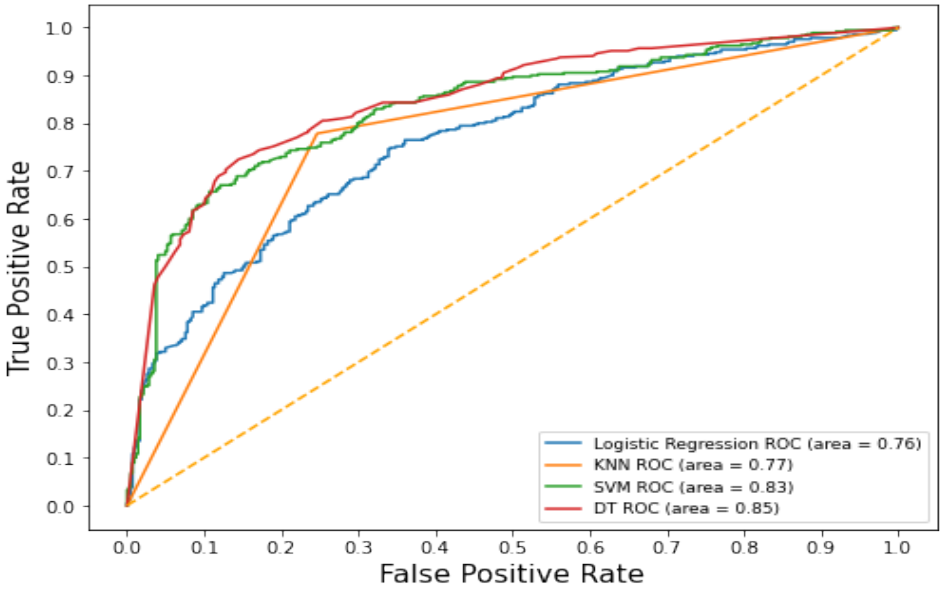

Figure 5: ROC curves for the under-sampling case

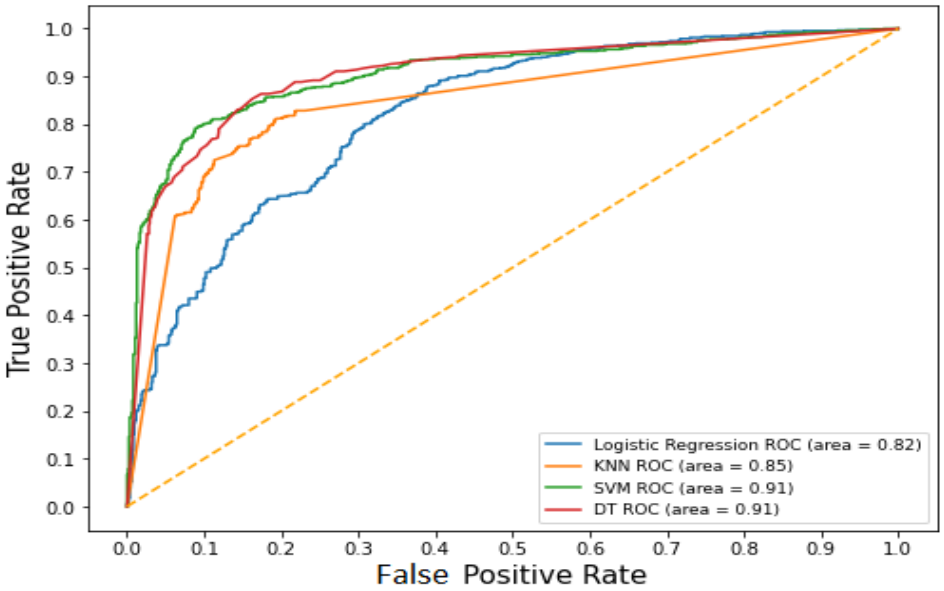

Figure 6: ROC curves for the oversampling case

7 DISCUSSION

The research question addressed in this paper is: "How well can we predict the churn rate of a healthcare service provider by building a model using several machine learning algorithms and data mining techniques trained on historical contract data?". The dataset used contains contract information of previous customers of a healthcare service provider. Three machine learning models (KNN, SVM, and DT) have been compared to the baseline logistic regression model, in the context of two sampling methods: random under-sampling and oversampling using SMOTE. Previous research had found that SVM and DT are some of the best models in churn rate prediction Brandusoiu et al. (2016); Dahiya and Bhatia (2015); Rodan et al. (2014); Sabbeh (2018); Vafeiadis et al. (2015). This paper's findings are in line with the literature, as the SVM and DT classifiers have resulted in high F-scores. To reach this result, this paper has examined three sub-questions.

When starting the analysis, this thesis discovered an issue with the dataset: class distribution. Class distribution means that the number of churners is a lot smaller than the number of non-churners. Class imbalance is a very widespread problem in churn rate prediction studies Amin et al. (2016); Farquad et al. (2014). Existing literature has approached this issue by using three methods: data-level solutions, ensemble solutions, and algorithm-level solutions Huang and Kechadi (2013). The method used in this paper is data sampling. The dataset contains 6180 entries out of which only 1979 are churners, which means that the actual churn rate is 32%. The data is not equally distributed between churners and non-churners, and if, left unresolved, could potentially cause algorithm bias. To solve this issue, this paper has used two random sampling techniques: under-sampling and oversampling using SMOTE. The choice of sampling techniques has been made based on previous studies. Random undersampling has been shown to be effective in predicting churning customers while being computationally inexpensive Salunkhe and Mali (2018) and SMOTE is one of the most popular sampling methods in churn rate prediction Zhu et al. (2017). For a correct implementation of SMOTE, the dataset was split into training and validation sets before populating with synthetic data points, as to not contaminate the two classes. The results of the analysis are in line with the literature. The values of accuracy and F1-score increased by as much as 10% when switching from under-sampling to SMOTE. This is to be expected, as in the case of under-sampling we are discarding valuable information randomly, which can be used to train the classification models. A future topic of research could be experimenting with other possible solutions for the class imbalance, and combining them with the classification algorithms.

This study set out to compare four machine learning models: logistic regression, k-nearest neighbors, support vector machine, and decision trees. For an objective evaluation of the performance of the models, and based on the literature review, logistic regression has been chosen as the baseline algorithm. The classifiers were tuned by performing 5-fold cross-validation on the hyperparameters to determine which parameters yield the highest evaluation metrics. After tuning the models, they were applied to both the under-sampled and the over-sampled datasets. The best performing algorithm was SVM implemented on the over-sampled set, which yielded an AUC-score of 0.91 and an F1-score of 0.84. The second model in terms of results was the SMOTE - DT one with an AUC-score of 0.9 and an F1-score of 0.81, followed by the under-sampling DT with an AUC-score of 0.83 and an F1-score of 0.75. These results have been reflected in the literature as well, as presented in Section 3.1 Brandusoiu et al. (2016); Rodan et al. (2014); Vafeiadis et al. (2015). KNN was expected to perform better than the LR classifier, but while in the under-sampling situation that was the case, there was no significant difference in the oversampling one. In the future, it would be interesting to experiment with neural networks to predict churn rate in the health care industry, as this is a field which had not yet been researched. An ANN model would be able to obtain excellent results, making the prediction even more accurate.

This paper also explored the best options for the evaluation metrics. Accuracy has been widely used in churn rate literature Huang and Kechadi (2013). This however does not necessarily mean that it also gives us the most insight into how well a model is learning. This paper has chosen to explore ROC curves and F1-score as well, to make sure that the proportion of false-negative cases is minimized. From a business perspective, this strategy is valuable, as it ensures that the number of customers who churn but are predicted to stay is as small as possible. With this purpose in mind, this study also provided the confusion matrixes for the trained models, which revealed that KNN in the under-sampled set, although performing worse than both SVM and DT, had the smallest number of false negatives.

The biggest limitation this study has encountered is the amount of data available for the churn prediction exercise. The dataset consisted of 6180 entries and 10 features provided by Company A. The features, however, contained very little information on the customer, that could potentially enhance the performance of the machine learning algorithms. Due to the nature of the customers, who are firms and not individuals, it can be very difficult to get more data about them. Further research on this topic should include either more features about the customer or focus on individual customers rather than distributors. In future studies, more complex models could be used to predict the churn rate, using neural networks or hybrid

models. An ANN model would be able to obtain excellent results, making the prediction even more accurate. Current literature does not focus on the health care industry at all, although it is also heavily reliant on services like retail, banking, and telecommunication.

8 CONCLUSION

The goal of this thesis is to examine the extent to which it is possible to predict the churn rate prediction of a healthcare service provider using four machine learning models (LR, KNN, SVM, and DT) trained on historical data containing contract information of previous purchases. To reach this result, this paper has examined three sub-questions.

The first sub-question this paper considered is: "How does the performance of the machine learning algorithms change between different data sampling methods?". The result showed that SMOTE offers more data for the algorithms to learn on, which caused higher performance for all four of the ML models, as expected from the literature Farquad et al. (2014). The biggest difference was observed in the SVM classifier, which saw an increase of 10% in F1- and ROC-AUC score between its implementation on the under-sampled set and the over-sampled one. Overall, it can be concluded that over-sampling is preferred to undersampling especially in the cases of smaller datasets, like the present one.

The second sub-question explored in this paper is: "Which of the listed machine learning algorithms will be the best performing one given the structure of the dataset and the selected features against the baseline model?". The churn prediction literature has shown us that out of the four models, the best-performing ones should be SVM and DT. The models were trained on the training sets of an under-sampled dataset and an over-sampled dataset. To improve the performance of all the models, 5-fold cross-validation was performed to tune the hyperparameters. As discussed in Section 6.3, SVM implemented on the SMOTE dataset was found to have the highest F-score of 0.84 and AUC-score of 0.91. The second model, although at a statistically insignificant difference was DT (F-score 0.81 and AUC-score 0.9).

The last sub-question answered by this study is:" Which of the evaluation metrics should we use to determine if the models are performing well? Should we consider F1-score rather than accuracy?". This paper concluded that F1-score is a more important value from a strategical point of view, as it considers not only the well-predicted data points, but also the false negatives that are important for the business. ROC curves were also used in the analysis for a better comparison of the models.

In conclusion, this paper found that the most suited model for churn prediction given the dataset provided by the healthcare provided is the combination of sampling with SMOTE and implementing an SVM classifier. Given the limited dataset, SVM exceeded the expectations with values for the evaluation metrics. However, improvements can be done to the model to raise the performance of the predicting models. Further research should use more complex models, such as ANN or hybrid models. More solutions for the class imbalance problem could be explored, using ensemble techniques or algorithm-level solutions. Moreover, in future studies, a larger dataset should be included, with more information concerning the customer. The healthcare industry would be able to benefit greatly from more studies on churn rate prediction.

REFERENCES

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis. *Life Science Journal*, 75–81.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940–7957.

Becker, M., Goldszal, A., Gronlund-Jacob, J., & Epstein, R. (2015). Managing a multisite academic–private radiology practice reading environment: Impact of it downtimes on enterprise efficiency. *Journal of the American College of Radiology*, *12*(6), 630–637.

Bhatnagar, A., & Srivastava, S. (2019). A robust model for churn prediction using supervised machine learning. *IEEE 9th International Conference on Advanced Computing (IACC)*.

Brandusoiu, I., Toderean, G., & Beleiu, H. (2016). Methods for churn prediction in the pre-paid mobile telecommunications industry. *2016 International Conference on Communications (COMM)*.

Burez, J., & den Poel, D. V. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626–4636.

Buttle, F. (2008). Making sense of customer relationship management. *Customer relationship management*.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

CIM. (n.d.). *Cost of customer acquisition versus customer retention.* Retrieved from https://www.cim.co.uk/membership/marketing-library/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(20), 273–297.

Coussement, K. D. V. d. P. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327.

Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 1–6.

Farquad, M., Ravi, V., & Raju, S. (2014). Churn prediction using comprehensible support vector machine: An analytical crm application. *Applied Soft Computing*, *19*, 31–40.

Fix, E., & Hodges, J. (1951). Discriminatory analysis, nonparametric

discrimination: Consistency properties. *Technical Report*, *4*.

Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. . *Artificial Intelligence Research*, *6*(2), 93.

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on svm model. *Procedia Computer Science*, *31*, 423–430.

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*, *40*(14), 5635–6547.

Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid data mining models for predicting customer churn. *International Journal of Communications, Network and System Sciences*, *8*(5), 91-96.

Kwon, H., Kim, H., An, J., & Park, Y. (2021). Lifelog data-based prediction model of digital health care app customer churn: Retrospective observational study. *Journal of Medical Internet Research*, *23*(1).

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602.

Nguyen, T., Sherif, J., & Newby, M. (2007). Strategies for successful crm implementation. *Information Management & Computer Security*, *15*, 102–115.

OED. (n.d.). *Churn rate.* Retrieved from https://www.oxfordreference.com/view/10.1093/oi/authority.2011080309561214

Rodan, a., Faris, H., Alsakran, J., & Ah-Kadi, O. (2014). A support vector machine approach for churn prediction in telecom industry. *International Journal on Information*, *17*, 3961–3970.

Sabbeh, S. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, *9*(2), 630–637.

Salunkhe, U., & Mali, S. (2018). A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to undersampling. *International Journal of Intelligent Systems and Applications*, *10*(5), 71–81.

Sharma, A., & Panigrahi, P. K. (2011). A neural network based approach for predicting customer churn in cellular network services. *International Journal of Computer Applications*, *27*(11), 26–31.

Sjarif, N., Yusof, M., Ya'akob, D., Ibrahim, S., & Osman, M. (2019). A customer churn prediction using pearson correlation function and k nearest neighbor algorithm for telecommunication industry. *Int. J. Advance Soft Compu. Appl*, *11*(2).

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2).

Vafeiadis, T., Diamantaras, K., Sarigiannidis, G., & Chatzisavvas, K. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9.

Vanhoeyveld, J., & Martens, D. (2018). Imbalanced classification in sparse and large behaviour datasets. *Data Mining and Knowledge Discovery*, 25–82.

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting customer churn through interpersonal influence. *Knowledge-Based Systems*, *28*, 97–104.

Zhu, B., Baesens, B., & vanden Broucke, S. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, 84–99.

APPENDIX: ROC CURVES OF MODELS BEFORE HYPERPARAMETER TUNING

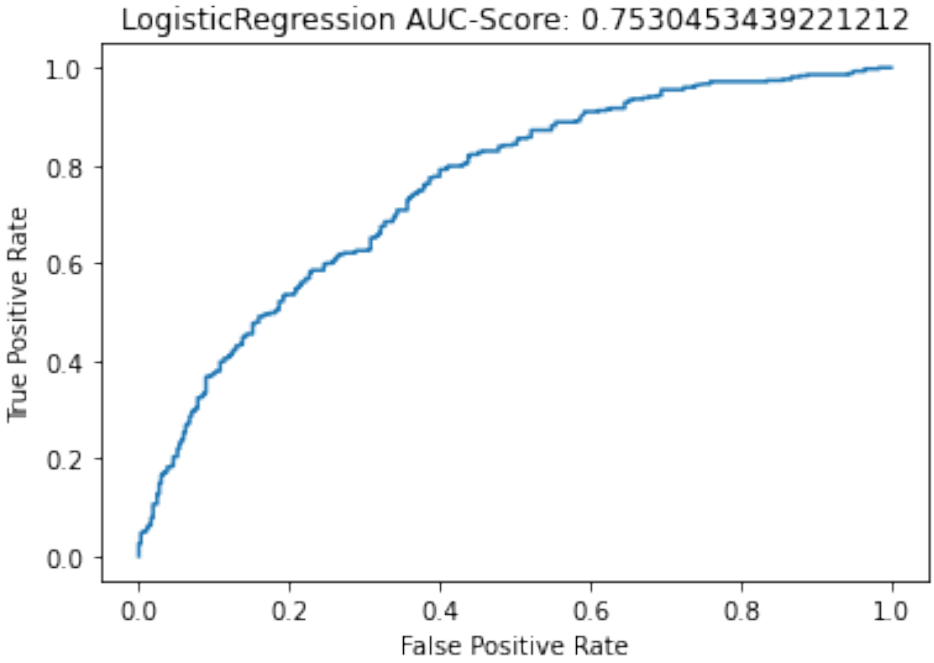

Figure 7: Logistic regression before CV

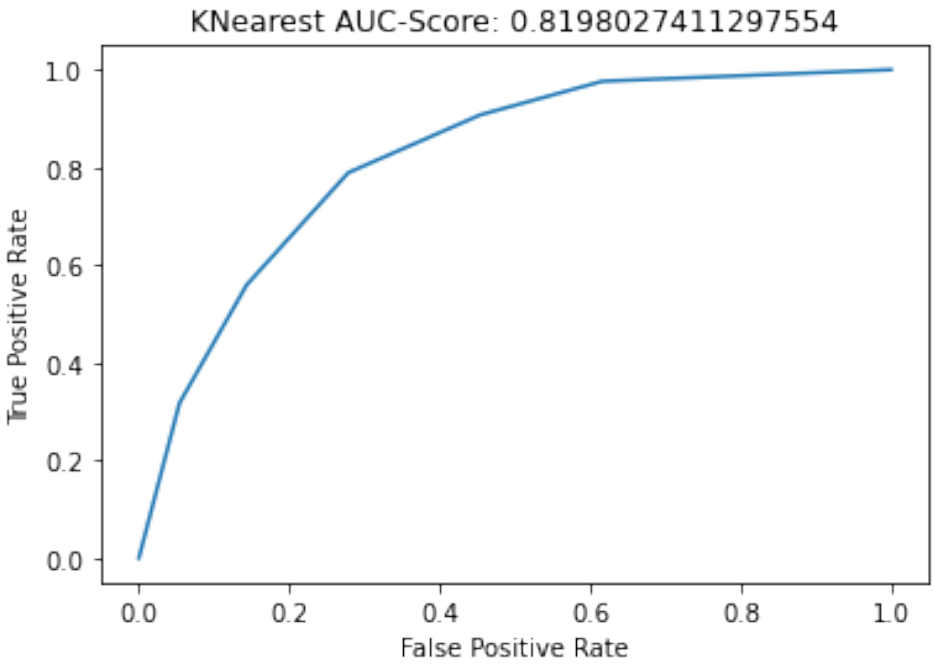

Figure 8: KNN before CV

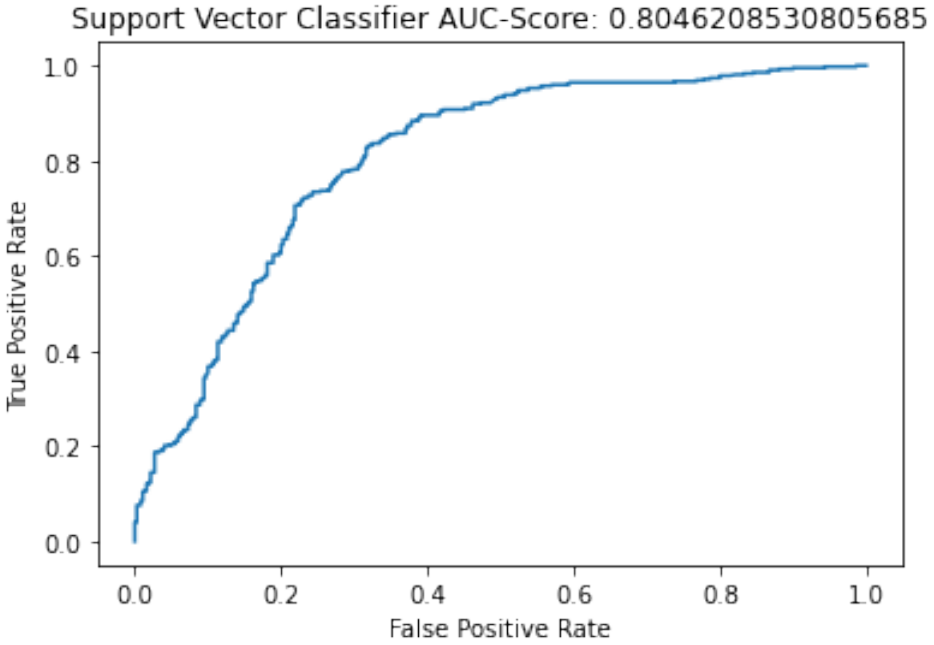

Figure 9: SVM before CV

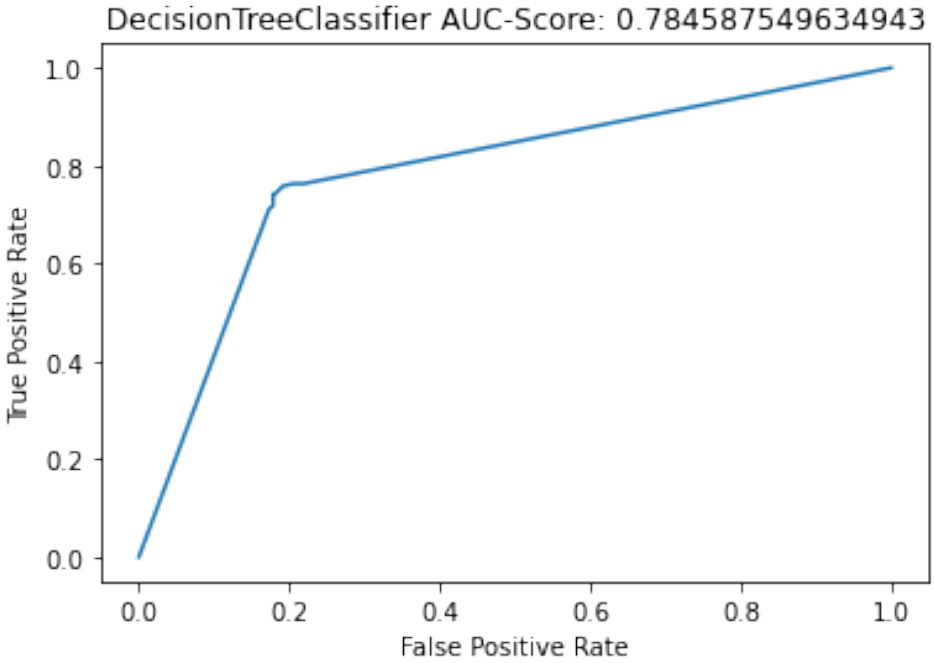

Figure 10: DT before CV