Customer Churn Prediction in Computer Security Sofware

Dang Van Quynh

Master's Thesis

Degree Programme

2019

MASTER'S THESIS	
Arcada	
Degree Programme:	Big Data Analytics
Identification number:	23274
Author:	Dang Van Quynh
Title:	Customer Churn Prediction in Computer Security Software
Supervisor (Arcada) :	Anton Akusok
Instructor (F-Secure):	Matti Aksela
Commissioned by:	

Abstract:

One of the most valuable assets of any company is its customer; hence it is vital for a company to focus on customer retention strategy by its advantage over customer acquisition. Customer churn prediction is a tool for increasing customer retention by identifying potential churners prior to them leaving. There are many studies on churn prediction over the past decade on various sectors but still lacking churn prediction study on the computer software industry, especially, in the context of the combination between telecommunication and security software. The paper attempted to find a solution to this problem by applying state-of-the-art machine learning models to real-world data. The results show the best model performs 4.6 times better than random as well as correctly identify 100 percent of churn in the top five percent of ranked customers. The results also demonstrate the importance of handling the imbalanced class issue in predicting customer churn.

Keywords:	Customer churn prediction, customer retention, computer software, security.
Number of pages:	50
Language:	English
Date of acceptance:	

CONTENT

Figures

Tables

FOREWORD

I would like to express my gratitude to my supervisor, Dr. Anton Akusok, for his support and guidance throughout this thesis. I would also like to express my sincere gratitude to Dr. Matti Aksela for this thesis opportunity as well as for giving me many ideas and invaluable help for my thesis. Lastly, I would like to thank all my colleagues at F-Secure Corporation, especially Myriam Munezero, Jouni Kallunki, Christine Bejerasco, Urmas Rahu for supporting and creating a fantastic work atmosphere.

1 INTRODUCTION

1.1 Background

Every company, of course, want to be successful and remain in business. Especially for-profit companies, profitability is critical to their long-term survivability. One way for companies to increase their profits is focused on increasing the number of customers. To do that, companies can focus on new customer acquisition or focus on customer retention.

Customer acquisition refers to gaining new customers or finding customers that have not had companies' products or services and persuading those consumers to purchase the products or services. The difficulty with this approach is that many markets are saturated, which means not many people out there with no customer account with some companies as if they are not with the company, they are likely with its competitors. Therefore, taking these people away from their current providers or bringing them as your brand new customers are nearly impossible or very costly. The cost of companies pay to acquire new customers is often high, around five times to twenty times (Longo 2016, Gur Ali and Ariturk 2014) comparing the cost of keeping current customers.

Customer retention refers to keeping current customers, which has various benefits that help increase companies' profitability as well as plays a key to the business' success. The first benefit of customer retention is low cost as the cost of retaining old customers is much cheaper than the one of acquisition. There also are loyalty benefits associated with existing customers. Loyal customers tend to spend more and a more extensive range of products and services than new customers. The more extended customers stay with a company, the less likely they are to leave the company. Besides, a company is expected to benefit from the most cost-effective advertising, namely word-of-mouth, from its loyal, happy customers. These retained customers can also provide valuable feedback for the company as the one who makes frequent purchases often know which area of the product or service could be improved. In general, companies should be making dedicated efforts to retain their existing customers. To maintain current customers, companies need to know churn and focus on customer churn prediction.

1.2 The practical problem

The security computer software industry includes antivirus, identity and access management, encryption, intrusion detection, and other security software. With the trend of the data explosion, a rise in internet-enabled solutions and cloud services, a shifted toward tablets and smartphones of consumers, the increase in cyber-attacks is expected, which further stimulate demand on a security software product. Hence, the industry is characterized by strong market growth, and low product differentiation, make the competitive rivalry intensive among industry players.

F-Secure Corporation has three types of customers: home customers, business customers, and operators customers. Among those, operators customers play an important role in generating profit for the company. F-Secure operators' customers mostly are network operators around the world. The telecom market has reached a saturation point that operators will find it hard to attract new customers as each new customers must likely be won over from the competitors. Also, customers can easily switch over from one operator to another. Therefore, understanding churn and customer retention become vital.

Churn is the term to indicate that a customer is just leaving a company by stopping transacting with the company or canceling service. The challenging problem of a customer churning prediction differs depending on the industry that is looking at. There are many studies on churn prediction over the past decade on various sectors such as telecommunications, banking and insurance, retail market, etc. but still lacking churn prediction study on the security software industry, especially, in the context of the combination between telecommunication and security software industry. This study is an attempt to find a solution to this type of problem.

1.3 Purpose

The aim of the thesis is using state-of-the-art machine learning techniques to identify customers who are about to churn. The thesis will study two perspectives of churn. The first one is the company perspective, which includes customer service, product quality, competitive price, etc., all things that a company can do to reduce churning rate. By identify churn determination of the company, we can make an improvement, which is appealing to customers to keep more of them. However, physical or resource limitations are considered. The second one is the customer perspective, which includes customer attributes such as location and behavioral traits: how frequently customers use the service, the amount of money that customers generate, etc. The study aims to identify early churn signals and recognize customers with an increased likelihood of churning. As a result, common attributes in churned is understand as well as identified churn candidates are managed proactively.

1.4 Scope

The scope of the study includes creating and training a machine learning model to predict customer churn. More specifically, the thesis uses some of the well-known models used in churn prediction. The implemented models are Logistic Regression, Decision Trees and Random Forest, and more detail about these models are described in the method section.

1.5 Data set

The study uses a data set from F-Secure Corporation. F-Secure Corporation is a cybersecurity company, which offer advanced threat protection, attack surface enumeration, and professional services that include security assessments, risk management, and software security.

1.6 Definitions

Table 1. Definition of Churn

Term	**Definition**
Trial churn	A subscriber who did not purchase after the trial period had ended.
Predefined paid churn	A subscriber who had paid subscription of X months and did not renew after it expired.
Continuous paid churn	A subscriber who terminated their continuous paid subscription.

2 LITERATURE REVIEW

2.1 Introduction

Customer churn prediction, one of the most prominent research topics for wide range of industries, is a common topic for both academic and practical; many practitioners and researchers spent years studying customer churn prediction in various domains

At any domain, customer churn prediction has followed this process

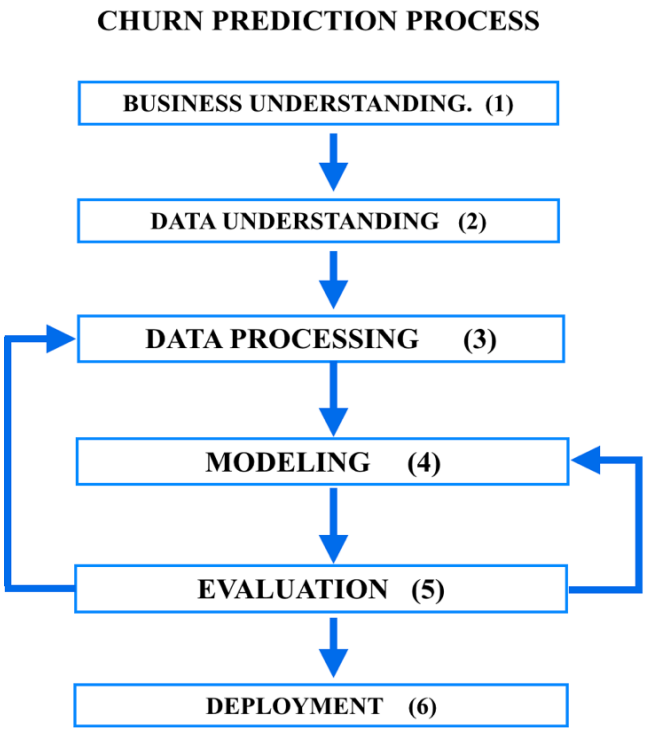


Figure 1: The standard churn prediction process

In the first step, the business should be clear understanding and defined, then it comes to the data understanding, after that it comes to the most time-consuming step: data processing. Later the predicting model has to be built. The evaluation step is necessary to evaluate the success of the model. There can be iteration between the Evaluation step and Data processing step to finding the most appropriate data processing technique such as variables selection or data reduction in order to get the best performance of the chosen predictive model. Also, there can be an iteration between the evaluation step and

modeling to tune the parameters used in the modeling. The final step is deployed the best model found from the previous step to support decision making.

This chapter mostly concentrates in the third, fourth and fifth steps; therefore, the chapter also divided into three parts. The first part discusses Data preprocessing, why it is necessary, also the techniques that previous studies have proposed. The second part discusses the popular machine learning algorithms using for modeling customer churn prediction. The third, as well as the last part, talk about evaluation metrics.

2.2 Data Processing

Data processing is a critical step because of its ultimate affection to churn prediction performance (Neslin et al. 2006). Data processing may include three steps: Data preparation, dimensionality reduction, and data reduction. Handling imbalance class issue in churn prediction is crucial during this data processing phase in order to improve the overall performance of churn prediction algorithms.

Data Preparation

Data preparation refers to steps of preparing data in the form that can be used and support a particular machine learning algorithm. These steps consist of cleaning data, handling missing value, detecting and possibly removing outlier as well as transforming original variables. Transforming step is necessary because some algorithms are not able to handle some kind of variable; for example, Logistic regression is not able directly to handle categorical variables. Different approaches to data transforming, based on whether variables are categorical or continuous, are applied. Categorical original variables can be transformed into new discrete categorical variables to obtain optimized, possibly by using remapping with a decision tree (Zhang et al. 2003). The idea behind is that giving the same category label for homogenous customers by grouping ones belonged to the same terminal node or leaf. They represent the new variable by dummy coding if necessary. Continuous variables will be discretization by three methods such as equal frequency, equal width and decision tree-based (Fayyad and Irani 1992, Coussement et al. 2017)

Equal frequency sorts the considered variable in ascending order, then divides into a certain number of bins b, which is calculated as (Coussement et al. 2017):

$$\frac{Total\ number\ of\ customers\ in\ a\ dataset}{b}$$

Equal width sorts the continuous variable, find the minimum and the maximum of the variable, then create b bins by calculating the width Ω of bin as (Coussement et al. 2017):

$$\frac{x_{max} - x_{min}}{b}$$

with boundaries at $x_{\min + i} \times \Omega$ Where i=1,2,…,b-1 and b is a chosen parameter.

Decision tree-based relabels the original variable accounted for the relation with the dependent variable by building a decision tree and then grouping the continuous variable into different groups based on the leaf that the customers belong to (Coussement et al. 2017).

Data reduction

Data reduction or instance selection tries to reduce dataset but still maintain the integrity of the original dataset (Wilson and Martinez 2000). This step is necessary since the problem of missing data is typical in the original dataset as well as the possibility of outlier or noisy data in the original one. Data reduction also mean to discard the incorrect data or outliers - the data points that are unlikely to occur in the dataset. One technique used to perform data reduction is clustering, which is based on the distances of data points to neighboring (Ghosting et al., 2008).

Dimensionality reduction

Dimensionality reduction or variable selection tries to reduce the variable space to reduce computational expense while still maintain the informative and discriminative power of variables (Ke et al. 2014, Rainartz 2002). Some of the proposed techniques for dimensionality reduction are PCA, correlation-based variable selection heuristic, and association rule. PCA tries to transform the original variable space to the new space of lower

dimension by performing a linear mapping of independent variables to extract the maximum variance. Correlation-based variable selection heuristic tries to map independent variables based on the scores, which give the high score for the independent variable of high correlation with the dependent variable and low inter-correlation with the other independence (Hall 2000). The rule of only one variable is kept if two variables are perfectly correlated is applied in heuristic technique. Association rule tries to find the co-occurred relation between variables, which identifies the independent variables that co-occur frequently (Agrawal et al. 1993) By performing variable selection before modeling, the predictive accuracy and stability of churn models are increased (Ke et al. 2014, James et al. 2013).

Imbalance classes problem

Churn is often a rare even in a company and of great interest. As a result, the number of instances of churn class is much smaller compared to the one of non-churn, which causes imbalance classes problem in churn as a bias toward non-churn class or a majority class is exhibited for most standard algorithms (Zhu et al. 2017) Many researches proposed solution for this issue, which can be grouped into two: data-level and algorithm-level solution.

Data-level solution tries to alter the distribution of minority or majority class to rebalance them. A variety of sampling techniques are included, such as random sampling that consists of random oversampling (ROS) and random undersampling (RUS), and synthetic minority over-sampling technique (SMOTE). ROS tries to randomly replicate minority class sample by making exact copies of samples; as a result, ROS has a drawback of increasing the likelihood of overfitting (Chawla et al. 2002). On the other hand, RUS tries to eliminate the majority class sample randomly; therefore, RUS may have a drawback of degrading classifier performance (Weiss 2004). SMOTE works to randomly oversampling minority class by creating ”synthetic” samples in the way of producing a new sample based on linear interpolations between the original sample and its k minority class nearest neighbors, as a result, SMOTE does not take account majority class samples, which may lead to increase overlap between the classes (Chawla et al. 2002).

Algorithm-level solution tries to emphasize the learning of classification algorithm on the minority class. One of the popular techniques is cost-sensitive learning. The cost-sensitive method works to assign a cost for misclassification, typically, a higher cost for misclassification of minority class than for the one of the majority class. The objective of cost-sensitive learning is to minimize the total cost (Kai Ming Ting 2002).

All in all, it is critical to ensure that the format of discrete, as well as continuous variables produced after data processing, is fit for churn prediction algorithms. The quality of variables has a direct effect on the quality of algorithms' performances as "Garbage in garbage out." A typical example is the Logistic Regression algorithm has no natural means to deal with missing value and outliers in dataset distort logistic regression. Ke et al. (2014) presented that data processing helps to improve the quality, as well as the efficiency, of the churn prediction model. Coussement et al. (2017) also find performing data preparation improving the prediction model performance by up to 14.5 percent in AUC and 34 percent in the top decile lift. Finally, it is important to aware of the imbalance class issue in churn prediction.

2.3 Modeling

Researchers, from variety of industries such as telecommunication (Kim et al. 2014, Kirui et al. 2013, Jadhav and Pawar 2011), banking (Gur Ali and Ariturk 2014, Eichinger et al., Prasad and Madhavi 2012, He et al. 2014), retail market and commercial business (Khodabandehlou and Rahman 2017) as well as gambling (Coussement and De Bock 2013) and gaming (Milosevic et al. 2017, Kawale et al.) have been focused on churn prediction over the past decade. Researchers proposed various machine learning techniques for predicting customer churner such as Logistic Regression (LR), Classification and Regression Trees (CART), Generalized Additive Model (GAM), Survival Analysis, Decision Tree (DT), Alternating Decision Tree (ADT), Random Forest (RF), Bayesian Networks (BN), Näive Bayes, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Convolution Neural Network (CNN), Restricted Boltzmann Machine (RBM) and Relational Learning.

In the retail market and commercial business, Khodabandehlou and Rahman (2017) proposed three classification models Decision Tree, Support Vector Machine, and Artificial Neural Network in predicting customer churn. Alexiei, Vincent, and Nicole (2017) implemented Convolution Neural Network and Restricted Boltzmann Machine to predict churn in the grocery industry. Khodabandehlou and Rahman identified that Artificial Neural Network performed the best with the highest accuracy, and Decision Tree performed the worse. Samira and Mahmoud also implemented ensemble learners, which is combining various prediction models, and shown that using the ensemble learners method increases efficiency and accuracy of prediction.

In online gambling and the game industry, Coussement and De Bock (2013) employed the Generalized Additive Model in their studying for predicting gamblers who are likely to leave. Similarly, Milosevic et al. (2017) used Classification and Regression Trees and Generalized Additive Model as well as performed ensemble algorithm and revealed the superiority of the ensemble in prediction compared with the single models.

In the banking industry, He et al. (2014) proposed a Support Vector Machine for predicting customer churn in commercial banking. They pointed out the necessity of using a sampling method in solving the imbalance class issue, which helps enhance the prediction accuracy of the Support Vector Machine. Gur Ali and Ariturk (2014), on the other hand, employed Survival Analysis, Logistic Regression, and Decision Tree with a dynamic churn prediction framework. With this framework, three types of training data were generated and used to train models, which included Multiple period training data (MPTD) or multiple training samples from different time period per customer, Single period training data (SPTD) or only one example per customer, and the lags version of SPTD that is k lag variables for each customer. Gur Ali and Ariturk identified that regardless of the models, MPTD enhanced the predictive accuracy across prediction horizons.

Finally, in the telecommunication industry, Verbeke et al. (2014) employed both non-relational customer churn prediction models (CCP) and the relational model for predicting customers with a high propensity to churn. The five non-relational models are Alternating Decision Trees, Bagging, Random Forests, Bayesian Network, and Logisitic Regression. Incorporation with these non-relational churn prediction models, Verbeke, et

al. developed a so-called relational learning model that used social network information for churn prediction. The explanations for social network effects on churn are contacting between similar people occurs at a higher rate than dissimilar ones, the word-of-mouth effect as well as the wide existence of operators' promotional offers of reducing tariffs for intra-operator traffic. Verbeke et al. find the significant impact of social network on churn behavior of telco subscribers. The experiment results had shown that the incorporated social network models outperformed all other models and relational churn prediction models could detect different types of churners compared to non-relational churn prediction models; as a result, the combination models generated significant profit gains for the telecom operator. Verbeke et al. also emphasized the impact of handling class skewed by non-Markovian technique for relational learning that improved the predicting power of churn prediction models. Abinash and Reddy (2017) implemented a comparative study of well-known prediction models: CART, Random Forest, SVM, DT, Näive Bayes, ANN, boosting and bagging for churn prediction problem and gave a conclusion of Random Forest performed the best in all terms of accuracy, sensitivity, specificity, and error rate. Vafeiadis et al. (2015), on the other hand, performed a comparison experiment similarly and concluded the superiority performance of SVM boosted version over all other models.

2.4 Evaluation Metrics

Evaluating the performance of churn prediction models is a crucial step in churn prediction process to assess how well the model generalizes. Some of the popular evaluation metrics used by many researchers for churn prediction evaluation are accuracy (Verbeke et al. 2012), Receiver Operating Characteristics curve (ROC) and the area under the receiver operating characteristics curve (AUC) (Coussement et al. 2017, Milosevic et al. 2017, Verbeke et al. 2012), precision and recall (Coussement and De Bock 2013), F-Measure (Coussement and De Bock 2013), and Lift Metric (Coussement et al. 2017, Verbeke et al. 2012, Verbeke et al. 2014).

A popular tool that used to evaluate a classification model's performance is the confusion matrix. A confusion matrix presents information about actual and predicted classification produced by a classifier.

Table 2: A confusion matrix for a binary classifier (Hassouna et al. 2015)

		Predicted classes	
		Class=Yes/+/ Churn	Class=No/-/No-churn
Actual classes	Class=Yes/+/ Churn	TP (true positive)	FN (false negative)
	Class=No/-/ No-churn	FP (false positive)	TN (true negative)

True positives (TP): the number of customers that are actually churners and the classification model has identified them correctly as churners.

True negatives (TN): the number of customers that actual are non-churners and the classification model has identified them correctly as non-churners.

False positives (FP): the number of customers who are non-churners but the classification model incorrectly determined them as churners.

False negatives (FN): the number of customers who are churners but the classification model incorrectly determined them as non-churner.

a) Accuracy

Accuracy, the portion of the total number of correctly predicted cases, is calculated as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Accuracy is a straightforward measurement of classification performance. Although, for churn prediction, accuracy may not be the best evaluation criterion and should not be the sole performance determinant since based on the assumption of equal misclassification costs that is not the case for a classifier with skewed class issue (Verbeke et al. 2012).

b) ROC curve and AUC

Sensitivity, the fraction of real churners which are correctly identified, is calculated as follow:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Specificity, the fraction of real non-churners which are correctly identified, is calculated as follow:

$$Specificity = \frac{TN}{(TN + FP)}$$

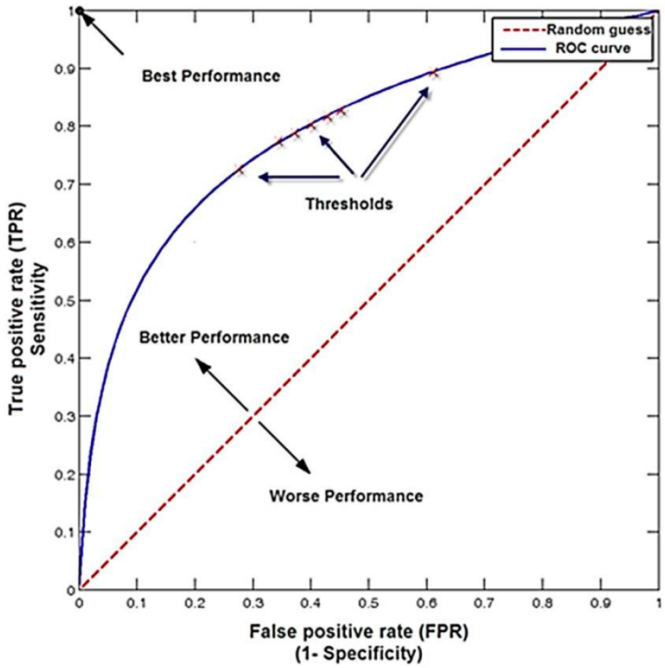


Figure 2: An example of ROC curve (Hassouna et al. 2015)

In ROC curve, the true positives rates (TPR) or Sensitivities are compared to the false positives rates (FPR) or 1 minus Specificity. It is the prevalent metric for assessing the performance of classifiers contained skewed class issue because it explains the classifier performance without any assumption of misclassification costs (Verbeke et al. 2012). AUC is an estimation of the area under the ROC curve, which means an evaluation of the overall classifier performance. All possible thresholds on the ROC curve are considered and aggregate into a single number of AUC. Therefore, from the ROC, we can easily identify which threshold is the best or the optimal threshold for the consideration model, as well as by using AUC, the overall performance of multiple classification models can easily be compared (Coussement et al. 2017, Verbeke et al. 2012).

An example of ROC curve is depicted in figure 2, in which the diagonal red line shows a ROC curve for random predictor. The model with the more passing top left of the ROC curve is the better. The AUC value ranges from 0.0 to 1.0, in which the random classification model has AUC of 0.5, and the classification model with larger AUC is the better.

c) Precision, Recall and F-Measure (Burez and Van den Poel 2009, Counssement et al. 2017**)**

Precision, the fraction of predicted churners that do churn, is calculated as follow:

$$Precision = \frac{TP}{(TP + FP)}$$

Recall, the fraction of real churners which are correctly determined as churners is calculated as follow:

$$Recall = \frac{TP}{(TP + FN)}$$

There is trading off between precision and recall. In order to compare precision and recall numbers, F-Measure, the harmonic average of precision and recall, is used and calculated as follow:

$$F - measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

The closer to one F-measure is the better-combined precision and recalled achieved by the classification model. Similarly, we also have Precision and Recall curve and Area under Precision and Recall to compare classifiers with the same rule, the larger the area under precision and recall is the better.

d) Lift Metrics (Burez and Van den Poel 2009, Counssement et al. 2017)

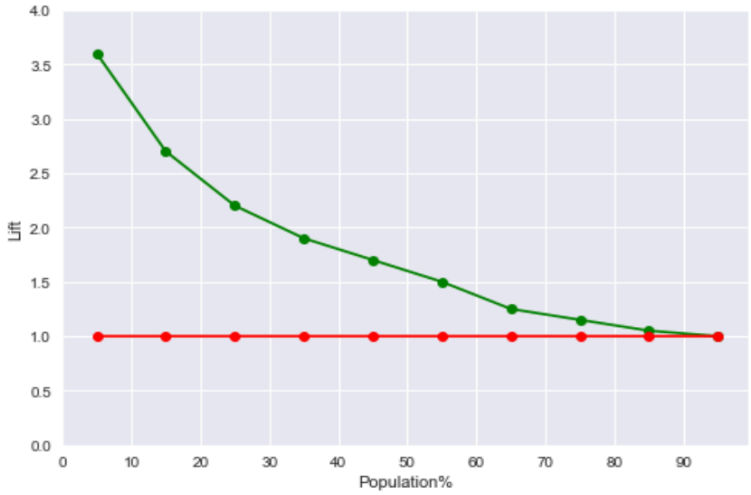


Figure 3: An example of a lift chart of a customer churn model

Lift Metrics is one of the widely used performance evaluation in churn prediction. In the lift metric, the ratio of the correctly classified results using the model is calculated, as well as the one without the model. In a customer churn's Lift chart, customers gather into deciles with their churn probabilities are sorted in descending order in the lift; therefore, it reveals directly customers who are most likely to churn as well as indirectly the profitability of the possible retention campaign. By targeting customers in the top deciles, companies can minimize marketing costs while still maintain a high customer retention rate. As a result, Lift charts are preferred performance evaluation tools in marketing.

Figure 3 presents an example of a lift chart, and the red line presents the positive response of randomly targeting, the blue line represents the positive response by using the classification model. The lift chart shows that by using the model, we gain 3.5 times as many respondents in the top 10 percents than without using the model.

3 METHODOLOGY

Based on the churn prediction process from the previous chapter, given the domain and the research problem from the first chapter, we discuss the method using to solve the research question. The methodology for churn prediction includes four phases. Data understanding and business understanding is the first and foremost phase, where churn is defined. The data preprocessing phase consists of data audit, data cleaning, missing value treatment, variable transformation, and some other data preparing related steps. Then come to the modeling phase, machine learning models used in the study are described in more detail. Finally, the models' performance evaluation is discussed.

3.1 Data and Business Understanding

The first phase of understanding business and the second phase of data understanding is, of course, very crucial for classifying churn. The thesis focuses on the security computer software industry, which uses the data set from F-Secure Corporation: Finnish cybersecurity company. The data set is from actual one of security product name F-Secure SAFE. F-Secure SAFE is an internet security product. SAFE protects users from virus, trojans, and ransomware. SAFE secures users' online banking and shopping and protects users' data if their mobile device is lost or stolen. SAFE make sure that users' internet connection is safe. Users can set limits for internet usage for children so SAFE protection users' entire family on desktop and mobile.

Data set was collected in one year and six months from August 2017 to January 2019. A sample data set of approximately 112 000 subscribers are used for experiment. Customer identity is anonymous in the data set. Four types of attributes can be found in the data set: customer subscription information such as activation dates, created dates, etc., customers device information, customers behavior information, and customer personal information such as location and language.

Churn Definition

Data set is collected from various operators worldwide; each operator can have its subscription type, which is a determinant of churn definition. There are three main types of subscription that are popular to operators: Trial subscription, Predefined subscription, and Continuous subscription.

Trial Churn

F-Secure SAFE offers a free trial for 30 days and on three devices. Users are not obligated to buy the product after the trial period as well as required, providing any credit card or banking details. Therefore, Trial churn is defined as a subscriber who did not purchase F-Secure SAFE after the trial period had ended.

Predefined paid Churn:

Prepaid and paid customers are people who put credit first and then start utilizing F-Secure SAFE. The subscription period is often one year, and the expired date is identified. By graphing the time of the act of extending subscription before and after the expired date, we find that the area of the most extending action is two months around the expired date. Therefore, Predefined paid Churn is defined as a subscriber who did not extend his/her F-Secure SAFE subscription when the subscription had expired.

Continuous paid Churn:

There is no expired date identified; therefore, subscribers can terminate the subscription at any point in time. Continuous paid Churn is defined as a subscriber who terminates his/her subscriptions in the consideration month.

Continuous paid subscription type is the most popular one in the dataset; therefore, the thesis focuses on Continuous paid Churn.

For the following phases: Data preprocessing, Modeling and Evaluation, these phases are needed to carry out iteratively until the expected outcome is met. After understanding the business as well as the data set, we need to preprocess the data into a form that is appropriate for the chosen model. Modeling means feeding our input data into the selected model to get the answer of our question, which is predicting customer churn. By

evaluating the models' performance, several evaluating techniques are applied. The whole processes may need to be repeated. It is most likely that we need to run the data preprocessing or modeling again to get a better modeling performance.

3.2 Data preprocessing

Quality of data plays a vital role in determining the performance of the churn prediction model. Data preprocessing, or cleaning and preparing data for the modeling phase, is a critical step as well as the most time-consuming step. Data sets are from F-Secure data source, a real-world data; therefore, the dataset inclined to be insufficient and noisy. Many activities involved during this preprocessing phase, some of the most important ones are explained bellows.

Data audit

Use statistic techniques to summarize the data such as understanding what the mean, the standard deviations, the maximum, the minimum values of every variable is. Applying descriptive analytics to our data is the best way to have an overview or a complete picture of every data fields and to spot any abnormal point in the data set. For example, in summary, the subscriptions' license info, there should not be any negative value and extreme value or outliers in the data.

Removing noise

During data collections, data goes through a series of processing procedures from collecting a raw data even to saving and transforming the raw data in the form that easy to use by analysts such as in a table format; therefore, the noise is nearly impossible to avoid. For example, One of the noises that occur in the data set is the servers' errors, is removed.

Removing outliers

Many observations serve for testing purpose. Those observations were often very different from typical records. For example, the typical records have the license size from 3 to 5 while the testing records may have license size of 200. Those special observations,

which diverge from normality, were removed. Another type of outliers presented in the data set is customers who open an account but never active in using the product, they never download or have any active devices, those customers, therefore do not give any useful information for the churn behavior, were removed. An identified of top 1% records having outliers or extreme values was removed

Missing value treatments

Missing values occur in many fields in the data sets, especially for fields belong to product improvement data, which may or may not choose to be provided by customers. State another way, there are many data fields in the data set that many customers decided not to report to us. No structure for the missing values in the data set; the values are missing at random. Many variables may have zero values; therefore, avoiding mixing up the actual values with the missing value, missing values were replaced with the number of 9999.

Variables transformation

Different transformation techniques are applied depending on the variables' type.

Since our machine learning models can only work with numeric data or boolean data. For example, Logistic Regression cannot work with categorical variables. It can also be seen from data exploration; many variables are object types, string types, or categories. Object or string types can be converted into categories' ones. Categories variables are converted to numerical variables using a one-hot encoding. New features that provide us more meaningful or extra information are created by combining existing features.

For example, the new feature name day-till-activation, which were created from subtracting the date that a subscription was activated to the date that a subscription was created, give us more information on how many delayed days in between the created and activated dates.

Correlation Analysis

Correlation analysis methods are used to identify or study the relationship between numerically or continuous variables in the data set. This statistical method can help to identify features having a strong relationship with the target variable.

Data reduction:

A collection of 800 features are generated. It is possible that not all customer metrics contribute the same quality level in distinguishing churner from non-churners. PCA is applied

Overcome class skews

The last data processing step is data sampling to handle the skewed class problem in churn prediction. By looking at the target feature we can study the actual churn's proportion in the data set.

Figure 4: Percentage of Churn in Dataset

Figure 4 above presents the proportion of churner to non-churner in the data sets. The churn's portion is about five percents, which means the number of churners is so much less than the number of non-churner; therefore our dataset contains the class skew problem. Two techniques, oversampling and undersampling, are used and compared to overcome this class imbalance problem.

Splitting the dataset

Finally, Data set is randomly split into 70%/30% proportion to form the training and test sets, in which training set is used to run algorithm on or to build models while test set is used to evaluate the models' performances and to make decisions regarding what model or parameters to use.

3.3 Modeling

The goal of this study is using state-of-art Machine Learning algorithms in helping F-Secure increase customers by predicting customers who have high propensity to churn so that the company can be proactive and retain these customers. Researchers and practitioners proposed a great diversity of Machine Learning techniques for churn prediction. Among these, Logistic Regression, Decision Tree, and Random Forest are most recommended by their comprehensibility and good predictability. The theory behind these models is discussed in more detail below.

a) **Logistic Regression** (Hastie et al. 2004)

The Logistic Regression model for two classes of a qualitative response Y; Y is either 0 or 1, given a feature vector X of p variables, is defined as follow:

$$\Pr(Y = 1|X) = p(X) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

p(X) is the probabilities that X belong to each class, and p(X) will have values between 0 and 1.

The Logistic Regression model for K classes of a qualitative response Y, given a feature vector X of p variables, is defined as follow:

$$\Pr(Y = k|X) = p(X) = \frac{e^{\beta_{0k}+\beta_{1k} x_1+\cdots+\beta_{pk} x_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l} x_1+\cdots+\beta_{pl} x_p}}$$

Where the class k = 1, …, K-1, p(X) will have values between 0 and 1

b) Decision Tree (Hastie et al. 2004)

Details of the classification tree-building process:

The feature space: X1, X2, .., Xn is partitioned into P separate regions: R1, R2, .., Rp
The same prediction is applied for observations that belong to the same region. The idea is partitioned the dataset into subsets based on descriptive features so that the subset is small enough and all observations belong to that subset fall under one category.
The decision on which feature to split, to build R1, R2, …, Rp, is made based on minimizing the classification error rate E defined as below:

$$E = 1 - max_k(\hat{p}_{mk}).$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

c) **Random Forest** (Hastie et al. 2004)

Random Forest, which is built from Decision Tree, combines the simplicity of decision trees with flexibility; as a result, a significant improvement in accuracy. The idea is the following:

We have a set of n independent observations Z_1, Z_2, …, Z_n; where the observation's variance is σ^2; the variance of the observations' mean $\bar{Z}$ is σ^2/n. Since we can reduce the variance by averaging a set of observations. However, we only have one set; therefore, to create multiple sets, we can bootstrap or take repeatedly sample from the single set. B different bootstrapped sets are generated.

Each of the B decision tree, which is trained on B sets, we get $\hat{f}^{*b}(x)$, the prediction at a point x, and we record the class predicted. The majority vote is implemented, which is the most commonly occurring class among B trees. Moreover, each tree is built based on:

a random selection of m features is selected from the full set of p features, and only those chosen m features are used as split candidates at each split. A fresh selection of m is made at each split, and the number of features m is approximately equal a square-root of p

3.4 Evaluation Metrics

In the evaluation phase, all evaluations' measurement mentioned from the previous chapter such as accuracy, ROC, AUC, Precision and Recall, Area under Precision and Recall, F-Measure, and Lift Metric is used for evaluating the consideration models as well as comparing performances among models. Applying the models and a discussion in which modeling techniques recommended to be deployed is carried out in the next section.

4 EXPERIMENTAL RESULTS

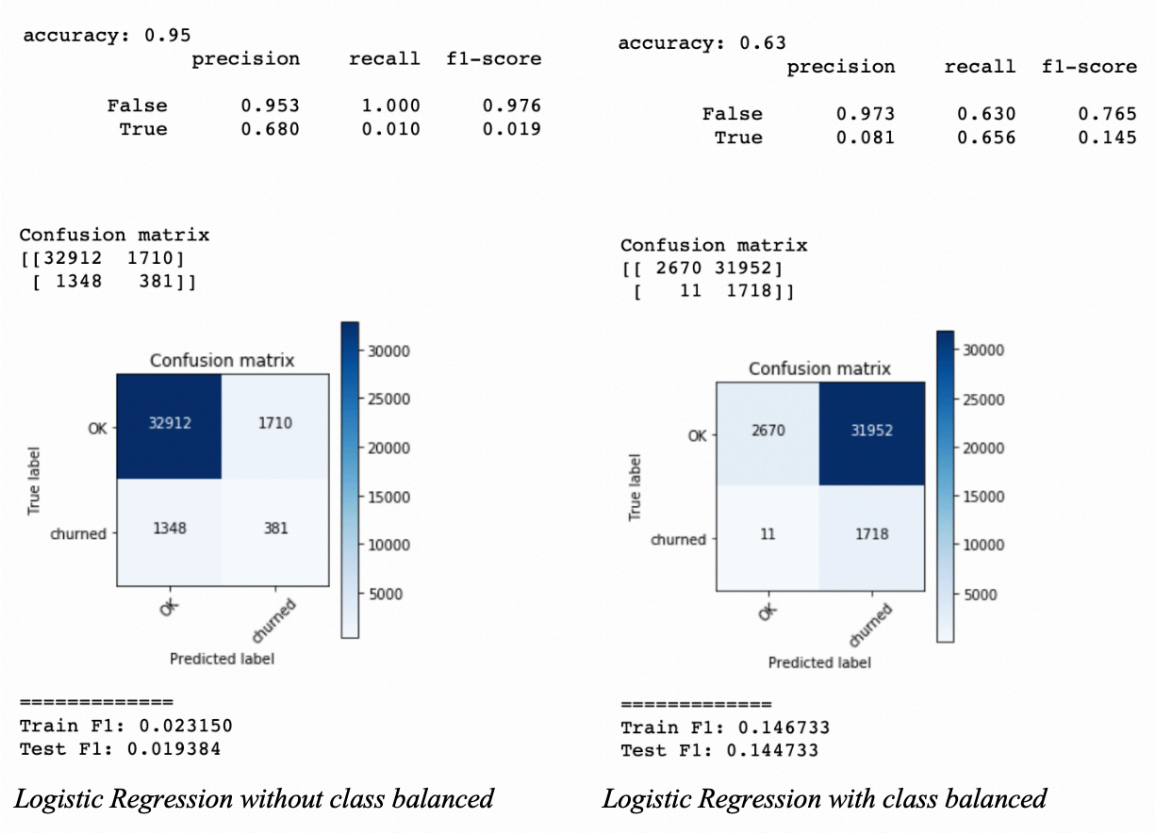
This chapter reports the empirical results by applying the previously described models to the data set based on one year and six months of historical data. The chapter is divided into two sections. The first section presents modeling results from Logistic Regression, Decision tree, and Random Forest, and compare their performances using the previously described evaluation metrics such as accuracy, ROC, AUC, precision and recall, and f-measure. By using a single-number-evaluation metric such as accuracy, AUC, or f-measure, it is easier to compare models. However, the study proved accuracy is not a useful measure since our data set suffer from imbalanced class distribution. F-measure evaluation metric shows the class skews problem and allow us to evaluate the effectiveness of using sampling techniques to overcome class skews issue. The second section discusses limitation as well as proposes future research.

4.1 RESULT FROM MODELING

Several approaches in modeling for customer churn prediction such as Logistic Regression, Decision Tree, Random Forest are applied. All of these models are highly recommended for customer churn prediction by their comprehensibility and good predictive performance. For every modeling techniques, I firstly applied models using the original dataset with its original ratio of churn to retention; these models are called based model; their performances are reported. The result of these first rounds show high overall accuracy; however, they were biased toward the majority class - retention. The minority class – churn is the one we want to predict. Few techniques are applied to the original dataset to create balanced subsets to overcome the imbalanced class problem. The models have applied again on these balanced subsets to see how they improved the classification rate over the based models. In addition to accuracy, ROC curve, AUC, Precision, and Recall and F-score are used to measure these models. Recall is the percentage of correctly identified churn out of the churn that happens, therefore we want our model to have a high recall for the churn class, while precision is a percentage of identified churn that actually ends up churning, so we also want our model to have decent precision. Accuracy is not the only performance determinant here since or dataset contains a class skew

problem. ROC, AUC, precision and recall and f-measure, and especially lift metric is also used as evaluation criterions for all recommended models. The performances are assessed on the entire dataset, together with the performances' assessment on the top decile. The detail of the results can be found below.

Logistic Regression



Logistic Regression without class balanced *Logistic Regression with class balanced*

Figure 5: Logistic Regression's evaluation measurements

The Logistic Regression based model was very accurate in predicting non-churn customers since it correctly identified 100% of non-churner with 95.3% of precision and made up 97.6% of f1-score. However, it clearly has shown the bias toward the majority class – non-churner. Identifying churn class, on the other hand, experienced poor results of only 1% correctly identified with 68% of precision and made up a modest f-score of 1.9%. The predicting performance was much improved for the minority class: churn after using balanced class techniques on the training set, 65.6% of churner is correctly identified with 8.1% of precision and increased f1-score to 14.5%. Although accuracy was decreased to 63 percents for the later model, accuracy did not show the whole picture here.

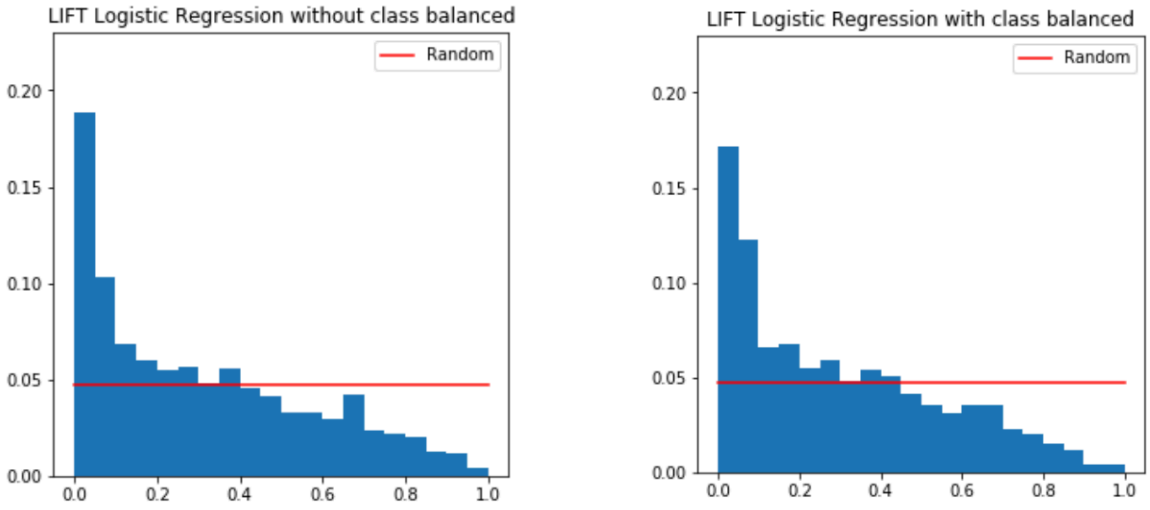


Figure 6: Lift chart for Logistic Regression

The lift charts for Logistic Regression are shown in Figure 6, in which the lift for Logistic Regression based model is shown on the left, the lift for Logistic Regression with class balanced is shown on the right. The red line presents the positive response of randomly targeting, the blue bar chart represents the positive response by using the classification model. The lift on the left performed slightly better than the lift on the right, with the gain of about 3.8 times as many respondents in the top 5 percents than without using the model.

Decision Tree

The Decision Tree based model, similar to the Logistic Regression, shown very impressed predicting power for the majority class - non-churner, while shown modest predicting performance for churn, which is correctly identified 1.6% of churn with 68% of precision. On the other hand, the Decision Tree predicting result after applied class balanced techniques gave better recall rate of 44% with a precision of 8.9%, which made up a total of 14.8% of f1-score.

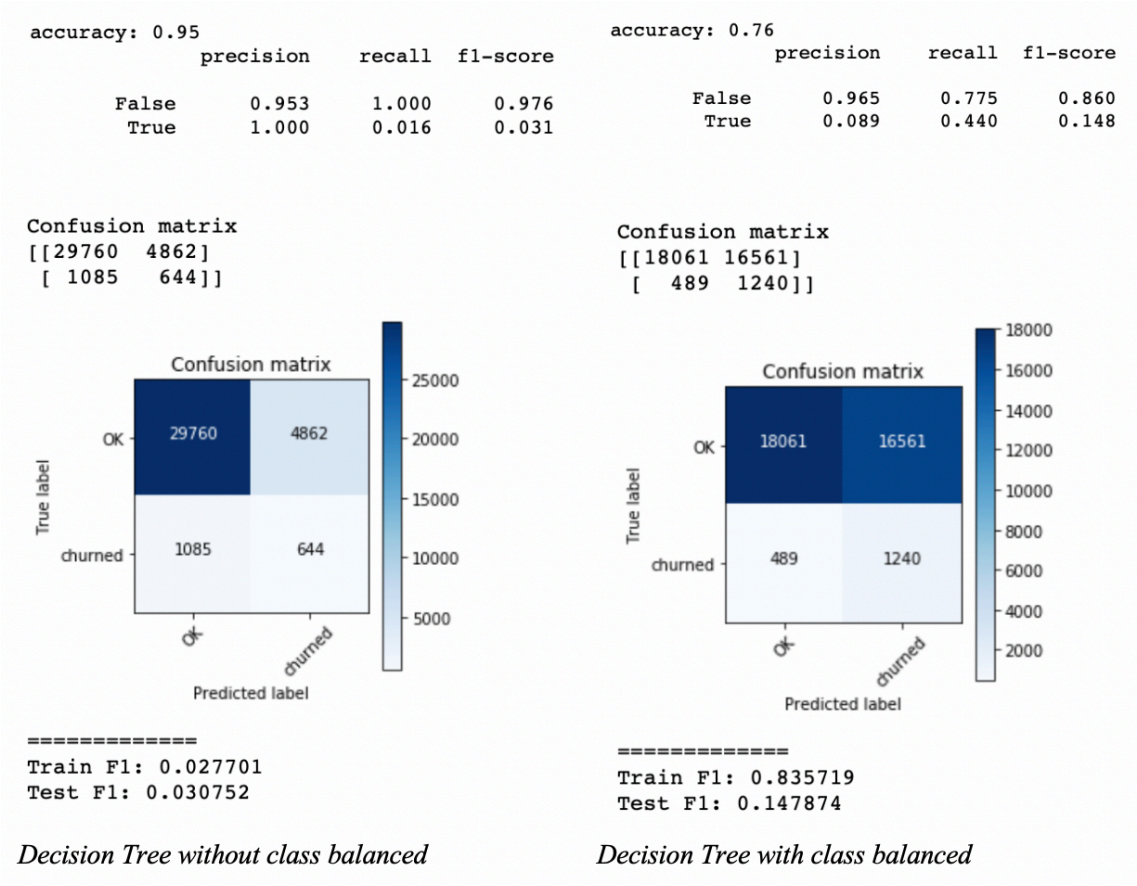


Figure 7: Decision Tree's evaluation measurements

Figure 8 presented the lift chart for Decision Tree. The lift of Decision Tree based model is shown on the left. The lift for Decision Tree applied the balanced class technique is shown on the right. The red line presents the positive response of randomly targeting, the blue bar chart represents the positive response by using the classification model. The lift chart on the left shows that by using the model, we gain about 4.5 times as many respondents in the top 5% than without using the model while the lift on the right presents the gain of 2.9 times.

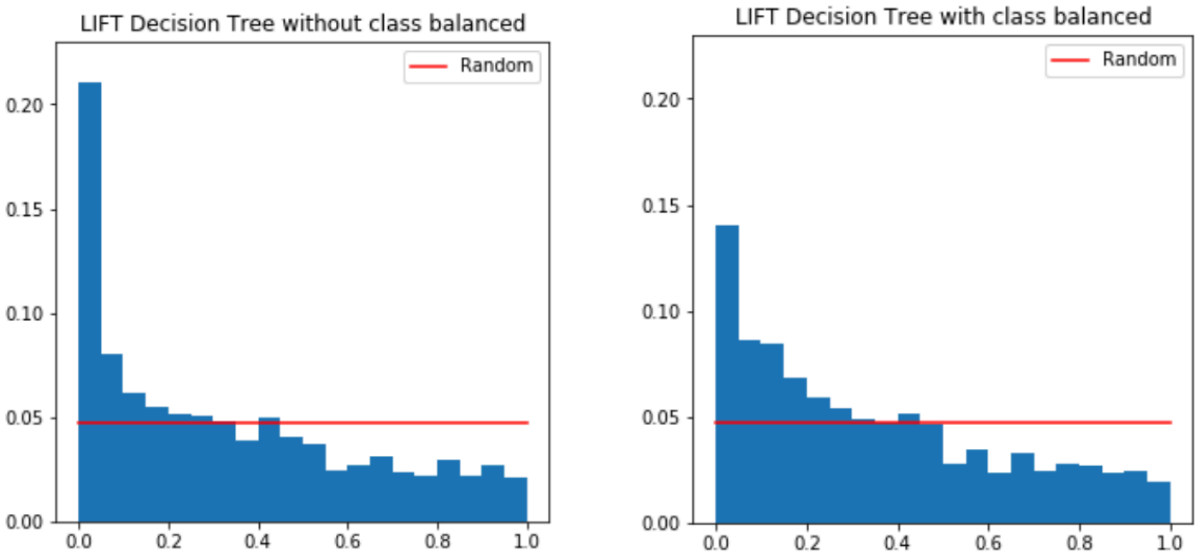


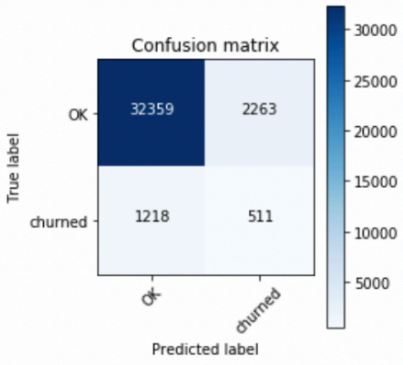
Figure 8: Lift chart for Decision Tree

Random Forest

```
accuracy: 0.95
              precision    recall  f1-score

       False      0.953     1.000     0.976
        True      0.711     0.019     0.036
```

```
Confusion matrix
[[32359  2263]
 [ 1218   511]]
```



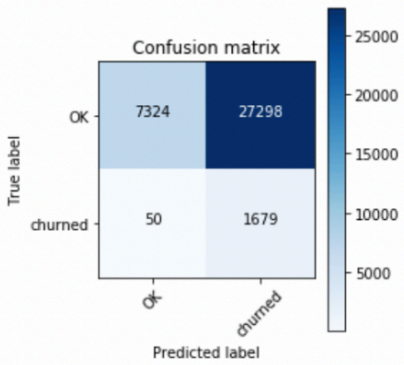

```
=============
Train F1: 0.081333
Test F1: 0.036077
```

Random Forest without class balanced

```
accuracy: 0.76
              precision    recall  f1-score

       False      0.969     0.774     0.861
        True      0.101     0.510     0.169
```

```
Confusion matrix
[[ 7324 27298]
 [   50  1679]]
```




```
=============
Train F1: 0.799949
Test F1: 0.168836
```

Random Forest with class balanced

Figure 9: Random Forest's evaluation measurements

Although the accuracy of Random Forest dropped from 95% to 76% across the two models of based line and with class balanced, we observed a significant increase in f1-score that implied increasing in churn predicting power within the data. The recall rate increased from 1.9% for Random Forest based model to 51% for Random Forest with class balanced, while the precision rates are dropped from 71% to 10.1%, which, however, still made up f1-score increased from 3.6% to up to 16.9%.

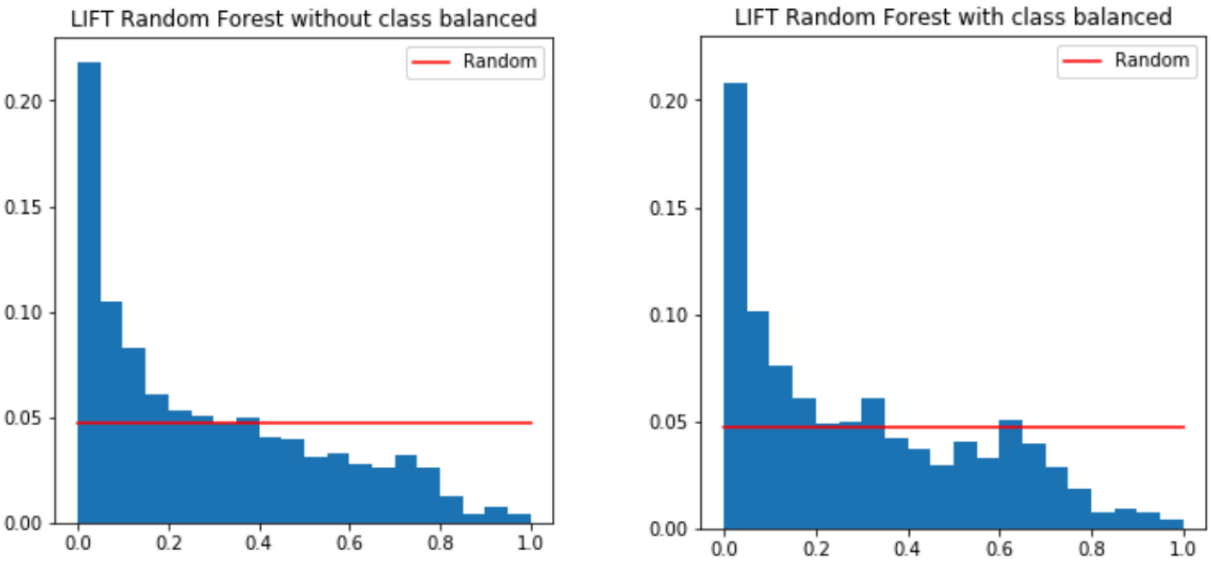


Figure 10: Lift chart for Random Forest

Figure 10 presents the lift chart for Random Forest. The lift for Random Forest based model is shown on the left. The lift for Random Forest used a sampling technique to balance classes in the training set is shown on the right. The red line presents the positive response of randomly targeting, the blue bar chart represents the positive response by using the classification model. The lift chart of Random Forest shows that by targeting the top three deciles, we arrest about 50% of churners. Both lift charts shown a similar pattern with the slightly better gain is from Random Forest based model, by using the model we gain almost 4.6 times as many respondents in the top 5 percents than without using the model.

The comparison

The confusion matrixes above were created with a threshold of 0.1. The thresholds can be set from 0 to 1. How the models' evaluation performances vary with the thresholds and How we can determine which threshold is the best for each model. In order to evaluate and compare the effectiveness of all models at the whole range of thresholds, ROC graphs and Precision and Recall curves of all models are depicted to provide a simple way to summarize all of the information and AUC of these curves are calculated to compare the ROC curve of one model to the one of another.

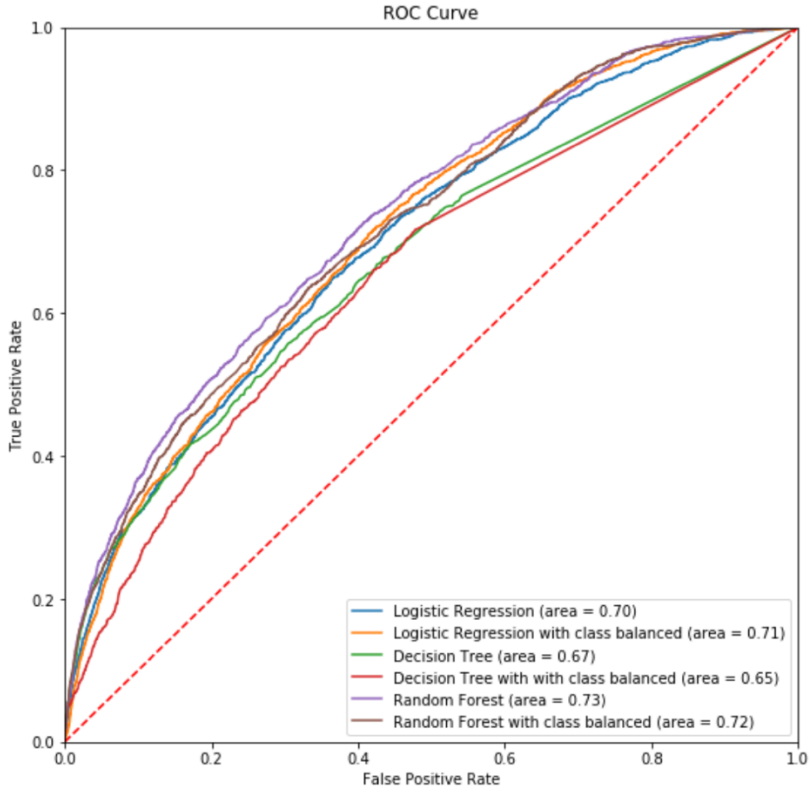


Figure 11: ROC curves of different models for the whole dataset

The ROC curve for each classification models is plotted in figure 11. The horizontal axis presents the False Positive Rate, while the vertical axis presents the True Positive Rate. It is easy to be implied from the ROC graph that Random Forest based model curve outperformed all other models' one with almost the whole range of thresholds. The AUC of Random Forest based model is 73%, the largest compared to the ones of other models, followed by the AUC of 72% for Random Forest with class balanced. Decision Trees are the most underperformed models with the AUCs of 67% for the based model and 65% for the one with class balanced.

Other metrics used to summarize and compare all models is Precision and Recall graphs. In the Precision and Recall curve, the horizontal axis presented the recall or the proportion of actual churners were correctly classified and the vertical axis showed the precision or the portion of predicted churners that were correctly classified. Since our dataset contain much more non-churners, precision might be more useful than the False Positive Rate because Precision does not include the number of True Negatives in its calculation and is not affected by the imbalance

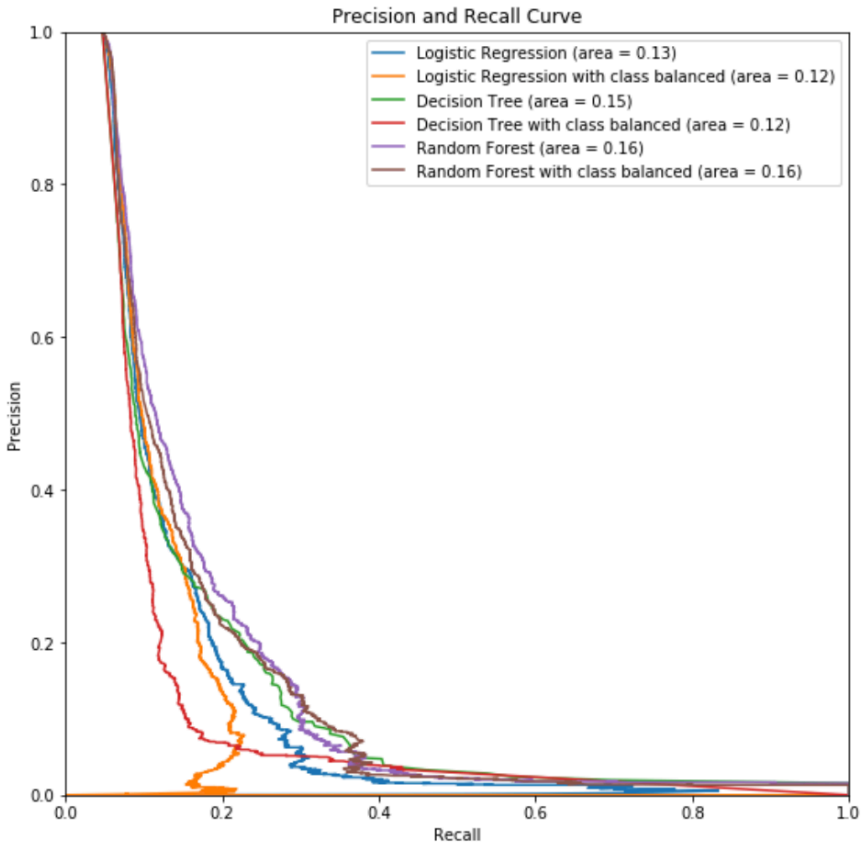


Figure 12: Precision and Recall curves of different models for the whole dataset

The Precision and Recall curves of all the experimented classification techniques are depicted in the figure 12. The outperformed precision and recall curve is of Random Frest based model. The best AUCs are of the Random Forest with and without class balanced of 13%. The most underperformed curves with AUC of 12% are of Logistic Regression with class balanced and Decision Tree with class balanced. The precision and recall curve is for the whole test set.

The performances on the top decile

An important area of applied customer churn prediction results is marketing, with the goal of supporting marketing strategies by understanding which customer to target. We assessed the models' performance on the whole population, which does not necessarily tell the entire picture. In marketing, there is always a cost associated with targeting a customer. It is not efficient or infeasible to target the whole dataset because of many reasons; a limited budget is worth to mention here. Instead of targeting the entire population, only the top decile customers with the highest probabilities of churn are reached

For the above reason, the evaluation of models performance in a portion of the population, or the top 10% ranked cases are described and compared below.

Table 3: Compare F-measure values of different models for the top decile

```
=============
1. Logistic Regression                      Test F1: 0.069570
2. Logistic Regression with class balanced  Test F1: 0.266158
3. Decision Tree                            Test F1: 0.097297
4. Decision Tree with class balanced        Test F1: 0.203164
5. Random Forest                            Test F1: 0.097946
6. Random Forest with class balanced        Test F1: 0.267810
```

1. Logistic Regression

	precision	recall	f1-score
False	0.000	0.000	0.000
True	0.039	1.000	0.074

2. Logistic Regression with class balanced

	precision	recall	f1-score
False	0.000	0.000	0.000
True	0.154	1.000	0.266

3. Decision Tree

	precision	recall	f1-score
False	0.861	1.000	0.925
True	1.000	0.051	0.097

4. Decision Tree with class balanced

	precision	recall	f1-score
False	0.000	0.000	0.000
True	0.113	1.000	0.203

5. Random Forest

	precision	recall	f1-score
False	0.845	0.995	0.914
True	0.689	0.053	0.098

6. Random Forest with class balanced

	precision	recall	f1-score
False	0.000	0.000	0.000
True	0.155	1.000	0.268

For the top decile, precision defines the classifiers exactness in predicting customer churn obtained 3.9% for Logistic Regression, 100% for Decision Tree and 68.9% for Random Forest. To understand the completeness, recall is used whereby 100%, 5.1%, and 5.3%. For those models with class balanced have similar performance results, with a precision of 15.4%, 11.3% and 15.5% respectively, recall is 100% for all.

Base on the table, among six models, Random Forest with class balanced provides the highest f1-score of 26.8%

The results from the table are threshold dependence. For an independent threshold measure, the precision and recall curves and AUC under precision and recall considered all the possible threshold values of all models for the top decile are depicted below.

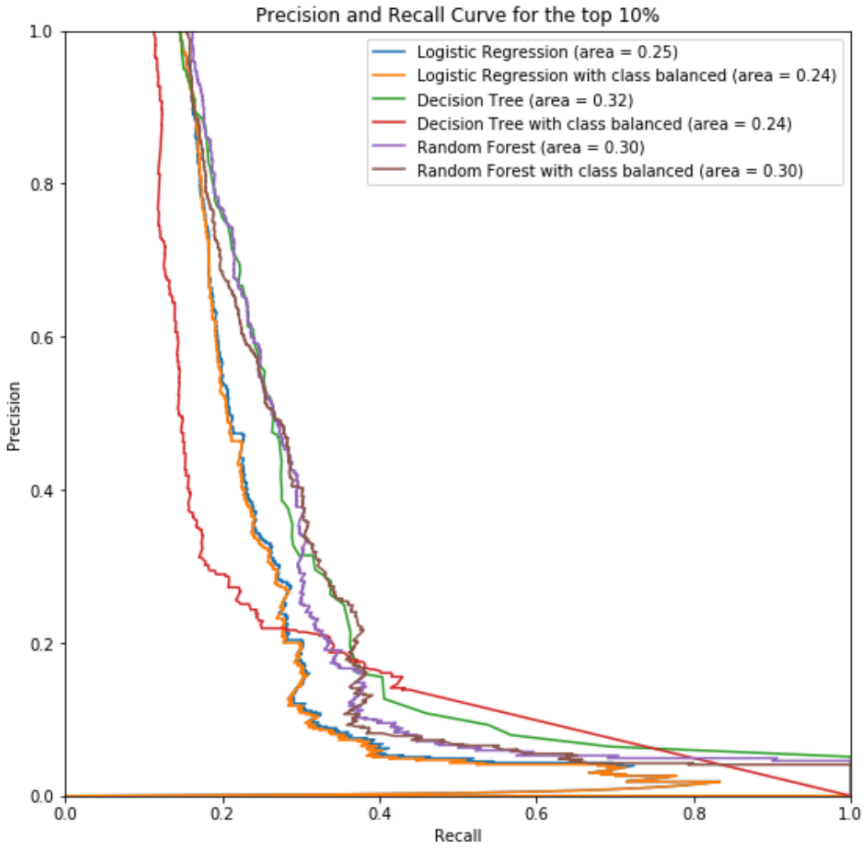


Figure 13: Precision and Recall curves of different models for the top decile

For the top 10% ranked customer group, the Decision Tree based model curve has the best precision and recall AUC of 32%, followed by the one of Random Forest based model and the Random Forest with class balanced, both of 30%. The smallest AUC of 24% is from both Logistic Regression with class balanced and Decision Tree with class balanced. Depending on the trade-off between precision and recall, we can choose the optimal point based on this graph.

Diagnosing Bias and Variance

The inability of a machine learning method to capture the true relationship between the features and the target outputs: churn is called bias. The different in fits between data sets are called variances. High bias classifiers indicate underfitting, which may be failed to capture important regularities. High variance classification models indicate overfitting, which may be well performed in the training set but performed poorly on the test set. The ideal model has low bias and can accurately model the true relationship and has low variability by producing consistent predictions across different datasets. In order to diagnose bias and variance of our learning models, learning curves are plotted. The learning curve of Random Forest with class balanced is depicted in figure 14.

A learning curve plots the test set performance measure against the different number of training samples. Since the training set contains about 80000 subscribers, Random Forest with class balanced was trained separately on an increasing number of the training's example such as 100, 300, .., 10000, 20000, .., 80000, then the test set performance measure is plotted, together with the plot of the training performance measure, varies with the training set size. The learning curve gives an estimation of the impact of adding more data.

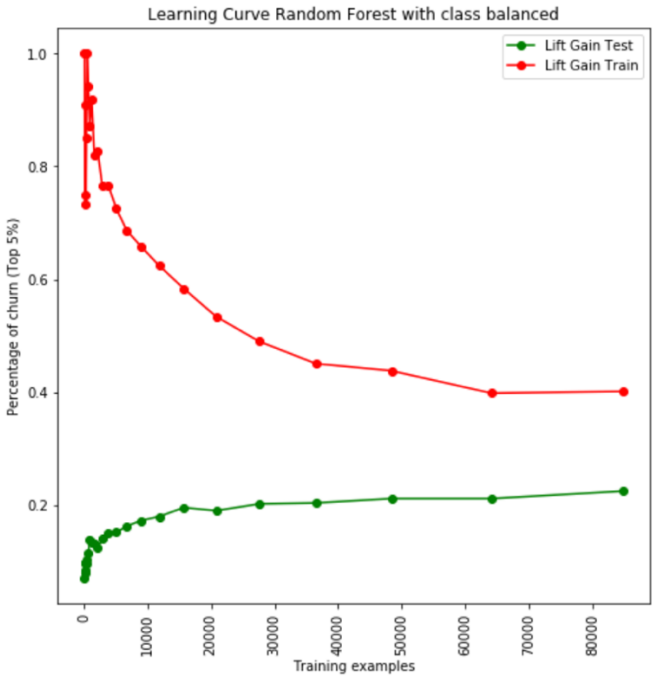


Figure 14 (a): Learning curve of Random Forest with class balanced

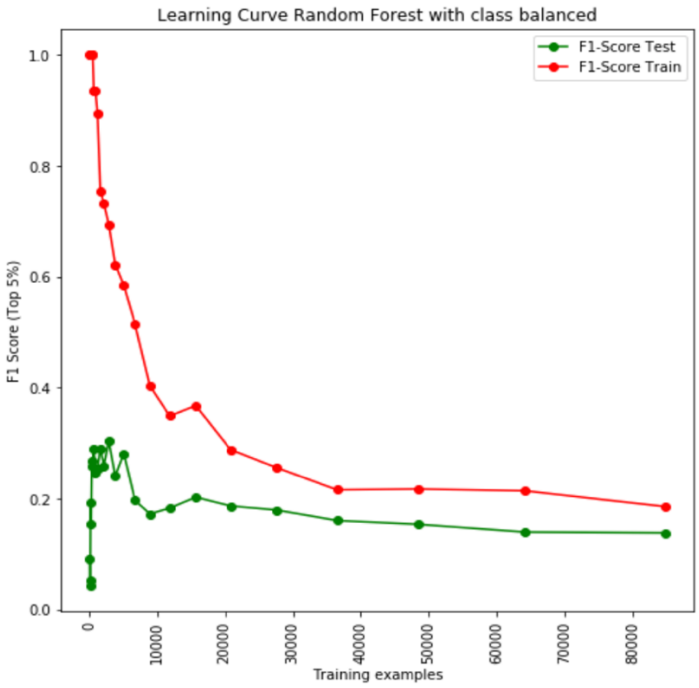


Figure 14 (b): Learning curve of Random Forest with class balanced

Figure 14 (a) is the learning curve using the lift gain on the top 5% ranked customer as the performance measurement for Random Forest after applied class balanced techniques. In this graph, the training gain should decrease while the test gain should increase as the training set size grows, which is depicted from the figure. The trend of the learning curve is relatively clearly shown here. The training percentage of churn captured seems to continue to slight decrease if we add more data, and the test percentage also tends to increase as more data adding. The gap between training and test curves is quite large, indicating high variance. Therefore, we can still improve the percentage of churn captured by adding more data.

Figure 14 (b) used f-score on the top 5% ranked customer as a performance measurement to describe the learning curve for Random Forest with class balanced. For a small training set between zero and 10000 training observations, the green test curve dramatically increase, followed by randomly fluctuate. The test curve then slightly increased when the training set grows from 10000 to 17000, then gradually decreased after training size grows to 20000 and more. The red training curve dropped abruptly in the beginning then gradually decreased after the training size reaches 20000 and more. Both the training and

test curves had flattened out in the end. The gap between the training and test curve is quite small, indicating small variance. By adding more data does not help in increasing f-score.

4.2 DISCUSSION

a) Managing expectation

The purpose of the thesis is to analyze and predict F-Secure SAFE's churn customers to help increase customers by retaining current ones. Predicting which customers are likely to churn is more important than predicting customers who are not likely to churn. Therefore, it is crucial for the models to classify the positive observations – the minority class – churn correctly. State another way, minimizing the percentage of misclassifications on the label 1 – the false positive is more important than minimizing the percentage of misclassification on the label 0 – the false negatives. Since our dataset experiences an imbalanced class problem, our based models have no problem identifying non-churner customers, but they have problem classifying churn ones. By using sampling techniques to alter the distribution of training's example can help to overcome this class skew problem. Although the overall accuracy of models with class balanced dropped slightly, we observed a significant increase in the ability to classify churn within the data. Also, by using more appropriate evaluation metrics such as ROC, AUC, precision, and recall and lift can give us an adequate picture of our models performance.

Moreover, An essential area of applied customer churn prediction results is marketing, with the goal of supporting marketing strategies by understanding which customer to target. When we assessed the models' performance on the whole population does not necessarily tell the whole picture. Instead of targeting the entire population, only the top decile customers with the highest probabilities of churn are reached. Therefore, how the models' performance in classifying customer churn in the top decile is very useful and relevant information.

b) Solution limitation

F-Secure has two kinds of data to collect; service data is the data that requires for running F-Secure product; product improvement is the data that we need consent from customers to collect. The product improvement data gives information about how our customers

using our products. It is normal that not all of our customers permit us to collect the data; therefore, missing values may affect the overall solution

Another limitation of data is data quality. Companies regularly drops or add new features in data collection. Some data fields are not collecting across all customers throughout a long period. There are quite many derived features that only added recently, therefore, the historical data before the feature added is missing.

There are various reasons for churning that we cannot capture. For example, the company gets acquired by another company where they go out of business, or subscriber churn effected by the influential peers. These are things that we cannot predict by unpredicting events or by lacking relevant information.

c) Future research

The thesis focuses on Continuous paid churn but Predefined paid churn and Trial churn. In particular, for Trial churn, customers are given 30 days of the free trial period that is using the product free of charge; after this free period, many customers terminate their subscription. F-Secure SAFE has a large of subscribers in this category, if we can build the model to try to understand and predict customer who terminate or not recharge in this trunk, so the company can approach them and give them some offers that can encourage them to recharge or stay, will bring a great benefit for the company. Moreover, subscriber churn may be affected by the influential peers, therefore, Social Network analysis used for modeling relationship between subscribers is also a useful topic for future research. Last but not least, utilize customers supported data or customer feedback by doing sentiment analysis for customer churn prediction is a fruitful topic for future research.

5 CONCLUSION

This paper presents a solution for customer churn prediction in the computer software industry, which is the first study on customer churn prediction in this industry according to our best knowledge, in particular, in the context of the combination of telecommunication industry and security product. Three modeling techniques: Logistic Regression, Decision Tree, and Random Forest were employed. The experiments were implemented on a real dataset from a cybersecurity company - F-Secure Corporation.

The first chapter presented the background information about the business and the practical problem at hand. The second chapter went through what had already consolidated and recommended in the literature in term of data processing, data mining tools, evaluation metrics, and a suggested framework for the study to carry on. In the third chapter the methodology was described by firstly understanding data and business, explaining the structure of dataset available to study, the characteristics of data as well as the specific problem we faced during handling with the data, secondly describing the classifiers used to predict customer churn, and finally giving a brief discussion on evaluation metrics used to asset models performance. And the fourth chapter has shown the research results and the comparison among the studied models. A result discussion, solution limitation, as well as potential future research, are proposed in the last chapter.

The study also attempted to apply different sampling techniques as mentioned in the second chapter such as ROS, RUS, SMOTE on training sample before modeling to overcome the class skew problem, however, not much different in term of classification performance among these techniques. In the end, it can be seen that customer churn is not an easy problem to model since churn is a rare event. The study further demonstrated that accuracy is not the best performance determination in this class imbalanced problem case.

The result suggested that Random Forest outperformed all other models. The AUC of Random Forest ROC curve is 73%, and the f1-score of Random Forest with class balanced reaches to 26.8% for the top decile population. The lift chart of Random Forest shows that by targeting the top three deciles, we arrest about 50% of churners. By applying Random Forest models in targeting marketing, we gain 4.6 times as many respondents

than without using any model or by randomly picking target customers in the top 5% of the population.

Since the study concentrates on continuous churn, for the future direction, we recommend research in other types of churn as Predefined paid churn and Trial churn. Utilize customers supported data or customer feedback by doing sentiment analysis for churn prediction is a fruitful topic for future research.

REFERENCES

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining Association Rules Between Sets of Items in Large Databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93. ACM, New York, NY, USA, pp. 207–216.

Burez, J., Van den Poel, D., 2009. Handling class imbalance in customer churn prediction. Expert Systems with Applications 36, 4626–4636.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357.

Coussement, K., De Bock, K.W., 2013. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, Advancing Research Methods in Marketing 66, 1629–1636.

Coussement, K., Lessmann, S., Verstraeten, G., 2017. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems 95, 27–36.

Eichinger, F., Nauck, D.D., Klawonn, F., n.d. Sequence Mining for Customer Behaviour Predictions in Telecommunications.

Fayyad, U.M., Irani, K.B., 1992. On the handling of continuous-valued attributes in decision tree generation. Mach Learn 8, 87–102.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, 463–484.

Ghoting, A., Parthasarathy, S., Otey, M.E., n.d. Fast Mining of Distance-Based Outliers in High-Dimensional Datasets 5.

Gür Ali, Ö., Arıtürk, U., 2014. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systems with Applications 41, 7889–7903.

Hall, M.A., 2000. Correlation-based feature selection of discrete and numeric class machine learning (Working Paper). University of Waikato, Department of Computer Science.

Hassouna, M., Tarhini, A., Elyas, T., AbouTrab, M.S., 2015. Customer Churn in Mobile Markets A Comparison of Techniques [WWW Document]. undefined. URL /paper/Customer-Churn-in-Mobile-Markets-A-Comparison-of-Hassouna-Tarhini/dfcb012b63e3549888a784c1ce13275df7666afa (accessed 4.7.19).

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2004. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Math. Intell. 27, 83–85.

He, B., Shi, Y., Wan, Q., Zhao, X., 2014. Prediction of Customer Attrition of Commercial Banks based on SVM Model. Procedia Computer Science, 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014 31, 423–430.

Jadhav, R.J., Pawar, U.T., 2011. Churn Prediction in Telecommunication Using Data Mining Technology. International Journal of Advanced Computer Science and Applications (IJACSA) 2.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics. Springer-Verlag, New York.

Kai Ming Ting, 2002. An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14, 659–665.

Kawale, J., Pal, A., Srivastava, J., n.d. Churn Prediction in MMORPGs: A Social Influence Based Approach.

Ke, S.-W., Lin, W.-C., Tsai, C.-F., 2014. Dimensionality and data reduction in telecom churn prediction. Kybernetes 43, 737–749.

Khodabandehlou, S., Rahman, M.Z., 2017. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. J. Systems and IT 19, 65–93.

Kim, K., Jun, C.-H., Lee, J., 2014. Improved churn prediction in telecommunication industry by analyzing a large network. Expert Systems with Applications 41, 6575–6584.

Kirui, C.K., Hong, L., Cheruiyot, W.K., Kirui, H., 2013. Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining.

Longo, S., 2016. The Cost of Customer Acquisition vs Customer Retention. Kapost Blog. URL https://marketeer.kapost.com/customer-acquisition-versus-customer-retention/ (accessed 3.16.19).

Milošević, M., Živić, N., Andjelković, I., 2017. Early churn prediction with personalized targeting in mobile social games. Expert Systems with Applications 83, 326–332.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., Mason, C.H., 2006. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research 43, 204–211.

Prasad, U.D., Madhavi, S., 2012. Prediction of Churn Behaviour of Bank Customers Using Data Mining Tools. Indian Journal of Marketing 42, 25-30–30.

Reinartz, T., 2002. A Unifying View on Instance Selection. Data Mining and Knowledge Discovery 6, 191–210.

Tibshirani, S., Friedman, H., n.d. Valerie and Patrick Hastie 764.

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

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

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218, 211–229.

Verbeke, W., Martens, D., Baesens, B., 2014a. Social network analysis for customer churn prediction. Applied Soft Computing 14, 431–446.

Verbeke, W., Martens, D., Baesens, B., 2014b. Social network analysis for customer churn prediction. Applied Soft Computing 14, 431–446.

Weiss, G.M., 2004. Mining with Rarity: A Unifying Framework. SIGKDD Explor. Newsl. 6, 7–19.

Wilson, D.R., Martinez, T.R., 2000. Reduction Techniques for Instance-Based Learning Algorithms. Machine Learning 38, 257–286.

Yang, J., Olafsson, S., 2006. Optimization-based feature selection with adaptive instance sampling. Computers & Operations Research, Part Special Issue: Operations Research and Data Mining 33, 3088–3106.

Zhang, S., Zhang, C., Yang, Q., 2003. Data preparation for data mining. Applied Artificial Intelligence 17, 375–381

Zhu, B., Baesens, B., vanden Broucke, S.K.L.M., 2017. An empirical comparison of techniques for the class imbalance problem in churn prediction. Information Sciences 408, 84–99.