Aalto University
School of Business

# A COMBINATION OF MULTI-PERIOD TRAINING DATA AND ENSEMBLE METHODS IN CHURN CLASSIFICATION

## THE CASE OF HOUSING LOAN CHURN

Master's Thesis
Le Thuy
15 May 2017
Information and Service Economy

Approved in the Department of Information and Service Economy

__ / __ / 2017 and awarded the grade

_____

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

**Abstract of master's thesis**

**Author** Le Thuy

**Title of thesis** A COMBINATION OF MULTI-PERIOD TRAINING DATA AND ENSEMBLE METHODS IN CHURN CLASSIFICATION

**Degree** Master of Science in Economics and Business Administration

**Degree programme** Information and Service Economy

**Thesis advisor(s)** Tomi Seppälä

**Year of approval** 2017 **Number of pages** 79 **Language** English

**Abstract**

Customer retention has been the focus of customer relationship management research in the financial sector during the past decade. The first step in customer retention is to classify the customers into binary groups of possible churners, meaning customers that are likely to switch to another service provider, and non-churners, referring to those that are probably staying with the current provider. The second step in customer retention is to take action to retain the most probable churners to either minimize costs or maximize benefits. As a result, churn classification is an important first step in customer retention.

However, the main challenge in churn classification is the extreme rarity of churn events. For example, the churn rate in the banking industry is usually less than 1%. In order to overcome this rarity issue, a great deal of research has been found to improve the two main aspects of a churn classification model: the training data and the algorithm. Regarding the training data, the recently proposed multi-period training data approach is found to outperform the single period training data thanks to the more effective use of longitudinal data of churn behavior. Regarding the churn classification algorithms, the most advanced and widely employed is ensemble method, which combines multiple models to produce a more powerful one. Two popularly used ensemble techniques are random forest and gradient boosting, both of which are found to outperform logistic regression and decision tree in classifying churners from non-churners.

To the best of the author's knowledge, the proposed multi-period training data has not been applied to the ensemble methods in a churn classification model. As a result, the thesis would like to study whether this multi-period training data approach, when employed together with ensemble methods in a churn classification model, produces better churn prediction than with logistic regression and decision tree. The ensemble methods used in this thesis are random forest and gradient boosting.

The study uses empirical data of housing loan customers from a Nordic bank. The churn models are evaluated based on three criteria: misclassification rate, Receiver Operating Characteristics (ROC) index and top decile lift. The key finding of this thesis is models that combine multi-period training data approach with ensemble methods perform the best in the housing loan context based on the aforementioned evaluation criteria.

**Keywords** churn prediction, ensemble methods, random forest, gradient boosting, multiple period training data, housing loan churn

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Background

**Churn classification is the important first step in customer retention**

Customer retention has been the focus of customer relationship management research in the financial sector over the past decades (Zoric, 2016; Gur Ali & Ariturk, 2014; Koh & Chan, 2002, Reichheld & Kenny, 1990). Retaining existing customers is argued to be more economical over the long run for companies than acquiring new ones (Gur Ali & Ariturk, 2014). Van den Poel & Lariviere (2004), in their attempt to translate the benefits of retaining customers over a period of 25 years into monetary terms, concludes that an additional percentage point in customer retention rate contributes to an increase in revenue of approximately 7% (Van den Poel & Lariviere, 2004). Moreover, in their later search, Van den Poel & Lariviere (2005) indicates that acquiring new customers rather than retaining existing ones is not only costlier but also riskier because customers who have earlier switched among different service or product providers are more likely to do so again. Specifically in banking, affluent customers holding different types of assets are usually more skillful in diversifying their portfolios among various financial institutions and hence more difficult to retain than customers with fewer assets (Lariviere & Van den Poel, 2005). As a result, maintaining long term relationships with customers has been the common strategy among leading companies in the financial industry (Nie, et al., 2011), especially in retail banking (A. O. Oyeniyi & A.B. Adeyemo, 2015).

A recent survey conducted by Accenture indicates that retail banking customers worldwide are more knowledgeable to proactively purchase their banking services not only from banks but also from non-traditional service providers, such as fin-tech start-ups (Accenture, 2015). As a result, customer retention has become a strategic priority because the longer the customers stay loyal to the banks, the more likely they are to expand their portfolio with the banks' in-house services and the higher their customer lifetime values become (Reichheld & Kenny, 1990). Due to regulatory requirements, banks must store a vast amount of historical data of customer transactions and interactions, enabling substantial research on customer retention for retail banking customers (A. O. Oyeniyi & A.B. Adeyemo, 2015).

Customer retention generally includes two main steps:

1. The first step involves building a model to identify the so-called ***churn*** events, which refer to the behavior of customers to switch from the current service or product provider to a competitor. The term "churn" is widely adopted in customer retention literature. The terms "churn classification" and "churn prediction" are used to refer to the first step of customer retention, which is to classify customers into binary groups: churners and non-churners. In this context, "churners" refer to the customers that are highly likely to switch to a competitor service or product provider and "non-churners" are the customers who are more loyal to their current services or products. Therefore, the result of the first step in customer retention is the creation of a ***churn classification model***.
2. The second step of customer retention is to determine the customers who are worth retaining the most among the classified churners and take actions to incentivize the continuation of their relationship with the organizations (Ballings & Van den Poel, 2012).

Consequently, in order to achieve a high customer retention rate, being able to predict churn using a good churn classification model plays a vital role (van Wezel & Potharst, 2007). This thesis focuses on the first step of customer retention. A churn classification model trains an algorithm on a specific training data set to classify the observations into binary groups: churners or non-churners. A training data set is a matrix that consists of multiple rows and columns, where observations are presented as rows and the features of each observation, also known as independent variables or predictors, are presented as columns. A churn algorithm is also called a binary classifier in other studies (Gur Ali & Ariturk, 2014; Breiman, 2001) because the dependent variable of the churn classification model can take only binary values: 1 for churners or churn events and 0 or -1 for non-churners or non-churn events depending on the notation of the churn algorithms. Dependent variables in churn classification context can be called churn responses or target variables. Also, churn events and non-churn events are generally called positive and negative events in churn prediction problems. All the mentioned alternative terms are used interchangeably throughout this thesis.

**The rarity issue of churn events in churn classification problems**

The main characteristic, and also challenge, of churn classification studies is that churn is usually a rare event. Churn rate depends on the research domains and how churn is defined. For example, churn rate in telecommunications sector can range from 2.2% to 6% (Lu, et al.,

2014) while churn rate in retail banking context is usually less than 1% (Gur Ali & Ariturk, 2014). This rarity feature makes churn classification difficult for several reasons. First of all, churn is rare both in the number of churners and in proportion to the number of non-churners. These features of absolute rarity and relative rarity (Weiss, 2004) hinder the ability of churn classifiers to predict the churners accurately because the training data is overwhelmed with the majority of non-churners (Lemmens & Croux, 2006). Additionally, in the research conducted by Weiss (2004) on the challenges that frequently manifest in data mining techniques due to the rarity issue, he has criticized that some common metrics for model performance evaluation do not take this rarity characteristic into account; for example churn events have less impact on model accuracy than non-churn events due to their disproportional frequencies, therefore model accuracy should not be considered as an effective criteria for churn model comparison. Moreover, data mining algorithms that partition data into smaller pieces such as decision tree suffers from data fragmentation because the more leaves there are in the tree, the less churn events there are in each leaf. Such partitioning rule decreases the ability of the model to learn about the churn behavior and to generalize on data sets different from the training data (Weiss, 2004). In order to address this rarity issue, Weiss (2004) discusses some of the most widely used solutions in data mining like more appropriate evaluation metrics, various sampling methods like under-sampling of the majority non-churn events or over-sampling of minority events, cost-sensitive learning, or boosting algorithm in ensemble methods (Weiss, 2004).

**Multi-period training data and ensemble methods have emerged in churn classification studies to overcome the rarity challenge**

Over the past years, churn classification literature has evolved to incorporate some of the above mentioned solutions to improve the two main aspects of a churn classification model: the training data and the algorithm (Ballings & Van den Poel, 2012). Regarding the training data, much focus has been drawn to the enhancement of the training data for churn classification models from different angles; most notable are the three followed points:

1. *More diverse data sources for churn predictors, or independent variables* (Baecke & Van den Poel, 2009)

Churn classification models are argued to perform better with churn predictors from more diverse data sources in addition to internally collected data from the organization. Predictors such as macro-economic factors (Gur Ali & Ariturk, 2014; Mavri & Ioannou,

2008) or commercially available data from external vendors (Baecke & Van den Poel, 2009) have been included in churn classification models to capture unseen factors from the macro-environment that might affect churn behavior. Such approach has shown to improve churn model performance compared to the churn predictors merely aggregated from customer transactions (Gur Ali & Ariturk, 2014; Baecke & Van den Poel, 2009; Mavri & Ioannou, 2008).

2. *Inclusion of longitudinal data for churn responses* (Gur Ali & Ariturk, 2014)

The inclusion of longitudinal data of churn response is claimed to improve churn classification performance with statistical techniques such as survival analysis but has not received much attention in churn classification algorithms (Gur Ali & Ariturk, 2014). Gur Ali & Ariturk (2014) stresses the importance of time series data in enabling early detection of churners, providing an opportunity for marketers to act timely in their customer retention campaigns. Churn literature traditionally employs churn responses calculated at the most recent point of time, hence information about customers that have churned prior to the pre-defined churn period is discarded and the variation of churn probability over time due to other aspects such as macro-economic factors is ignored. Moreover, independent time series variables are usually aggregated into static values such as min, max or average over the observation period. Such approach is argued to possibly rid the model of important information from the time series nature of the transactional data (Gur Ali & Ariturk, 2014).

3. *Appropriate time window of analysis* (Ballings & Van den Poel, 2012)

The time window of analysis is also studied to increase model efficiency. Time window of analysis refers mainly to the two periods of time: observational period, used to capture the independent variables, or churn predictors and performance period, used to calculate the dependent variables, or churn responses. Ballings & Van den Poel (2012) focuses on the former time window for the independent variables, which are mainly comprised of transactional data over time in churn classification context, to optimize the duration of such period. In their response to the natural question “How long back in time should the data be included to build the best performing model?”, Ballings & Van den Poel (2012) rejects the common belief that performance improves proportionally with the increasing data volume because it is observed in their research that over the time window of 77 years, after a specific point of time say 15 or 16 years, one additional year of historical data does not solve the lack of churn data issue but only creates computational burden (Ballings & Van den Poel, 2012).

Furthermore, Gur Ali & Ariturk (2014) evaluates models that includes the lagged values for all independent variables and observes that such models perform worse than those without the lagged variables (Gur Ali & Ariturk, 2014).

Taking into consideration all the three aforementioned aspects, Gur Ali & Ariturk (2014) proposes the ***multi-period training data*** approach, in which both churn responses and churn predictors are captured over multiple periods of time. As such, one customer can have many observations over time and historical information about churn responses prior to the defined churn period is retained as much as the operational data allows; therefore, the models are claimed to make more effective use of the historical churn responses rather than throwing them away, mitigating the rarity issue. In their studies, models that employ this approach to construct the training data are compared with models that employ the so-called ***single period training data*** approach. Single period training data refers to the traditional approach, in which both churn responses or dependent variables and churn predictors or independent variables are captured only at a single point of time. Specifically, the churn responses are calculated from the most recent data while the values of the independent variables or churn predictors are aggregated over a period of time prior to the churn period. In a study to predict churn among private customers in a commercial bank, for both models that employ logistic regression and decision tree as churn classification algorithms, the multi-period training data approach is concluded to outperform the single period training data approach (Gur Ali & Ariturk, 2014). However, the proposed multi-period training data approach has been applied to only a few churn classification algorithms such as decision tree or logistic regression. Consequently, there is a need to investigate whether the proposed multi-period training data also performs well in churn classification models using other algorithms.

Regarding churn classification algorithms, researches have recently developed a wide variety of classification algorithms that perform better than logistic regression and decision tree to predict churn. It is worth mentioning that over the history of churn prediction, different techniques have been employed, such as segmentation based on recency, frequency and monetary (usually denoted as RFM) information from customer transactional data; statistical techniques such as logistic regression and survival analysis (Mavri & Ioannou, 2008; Van den Poel & Lariviere, 2004); data mining techniques for large data set such as decision trees (Nie, et al., 2011; van Wezel & Potharst, 2007); and more advanced machine learning techniques like neural networks and support vector machine (Baecke & Van den Poel, 2009). Not until recently has churn prediction literature discovered the superiority of

ensemble method in churn classification compared to logistic regression and decision tree. Ensemble method generally refers to the combination of two or more models into a single and more powerful one (Yaya, et al., 2009; Jinbo, et al., 2007; Lariviere & Van den Poel, 2005). In some studies that employ ensemble methods, the algorithms used in churn classification are also called ***churn classifiers*** (Breiman, 2001). This term will be used throughout the thesis interchangeably with churn algorithms. Among the employed ensemble methods in churn classification, random forest and gradient boosting have been concluded to outperform logistic regression and decision tree in several researches (Van den Poel & Lariviere, 2005 & 2004; Breiman, 2001). Therefore, the thesis considers these two methods from the ensemble family good candidate for the comparative study in the housing loan context.

As a result, this thesis aims to investigate whether the proposed multi-period training data by Gur Ali & Ariturk (2014) improves churn classification accuracy for models that use ensemble methods, namely random forest and gradient boosting, compared to those using logistic regression and decision tree. Moreover, to the best of the author's knowledge, this multi-period training data approach has not been employed in any other researches. In that manner, the thesis is an extension to the research by Gur Ali & Ariturk (2014) by applying their proposed multi-period training data approach to more complexed churn algorithms to examine whether the performance of this approach stays robust in the domain of housing loan churn.

## 1.2 Research Problem and Contribution

In light of the background on churn classification, this thesis focuses on the first step of customer retention, which is to build a churn classification model with the consideration of the proposed ideas to improve the training data and the advanced algorithms: specifically, the multi-period training data proposed by Gur Ali & Ariturk (2014) and the ensemble methods as advanced churn classifiers. The thesis has acknowledged a gap in the application of the proposed multi-period training data approach by in churn classification; the author aims to test the robustness of this approach with ensemble methods, namely random forest and gradient boosting, in the context of housing loan churn prediction. Specifically, the thesis first validates whether the multi-period training data approach performs better than the traditional single period training data approach using logistic regression and decision tree as churn classification algorithms. Secondly, the thesis incorporates this multi-period training

data approach and the advanced ensemble methods in churn classification models and examines their performance in churn classification.

This thesis aims to explore the research problem in churn prediction of housing loan customers in the retail banking segment of a Nordic bank. Despite the large volume of churn prediction studies in retail banking, most of them concentrate on churn in general (Zoric, 2016; Prasad & Madhavi, 2012; Van den Poel & Lariviere, 2004) or in credit card segment (A. O. Oyeniyi & A.B. Adeyemo, 2015; Nie, et al., 2011). Only little research focuses particularly on housing loan customer (Koh & Chan, 2002); therefore, this research also enriches churn literature for this customer segment.

The research problem is detailed into the following research questions.

**Based on the selected evaluation criteria,**

**Question 1: For models that employ logistic regression and decision tree, does multi-period training data approach improve churn classification performance than the single period training data approach?**

**Question 2: For models that employ single period training data, do random forest and gradient boosting improve churn classification performance than logistic regression and decision tree?**

**Question 3: Do models that employ both multi-period training data approach and ensemble methods perform better in churn classification than those in the first question?**

**Question 4: What are the best churn predictors in the housing loan context?**

The first two questions aim to validate whether the multi-period training data approach and the ensemble methods perform better in churn classification for the housing loan customers compared to their counterparts. The third question studies whether their combination improves the churn classification performance even further. If the result in this thesis supports the hypothesis that models, which employ the multi-period training data approach and random forest and gradient boosting as churn classifiers, perform better than the other models created in the thesis, it is sensible to expand the application of this multi-period training data approach proposed by Gur Ali & Ariturk (2014) in churn classification. Finally, the last question is of general managerial interest when it comes to churn prediction to highlight the best churn predictors.

## 1.3 Research Structure

The structure of this thesis is illustrated by Figure 1. Chapter 2 provides an extensive literature review on the selected aspects of churn classification modelling. Chapter 3 reviews the employed research methodologies in this thesis. Chapter 4 describes the data by explaining the procedure to calculate the churn responses for the housing loan customers of the case company, the time windows of analysis, and the preparation of the churn predictors or independent variables. Chapter 5 discusses the model building process and the models' results. Chapter 6 concludes the thesis with its contribution, limitations and suggestions for future research in the churn classification topic.

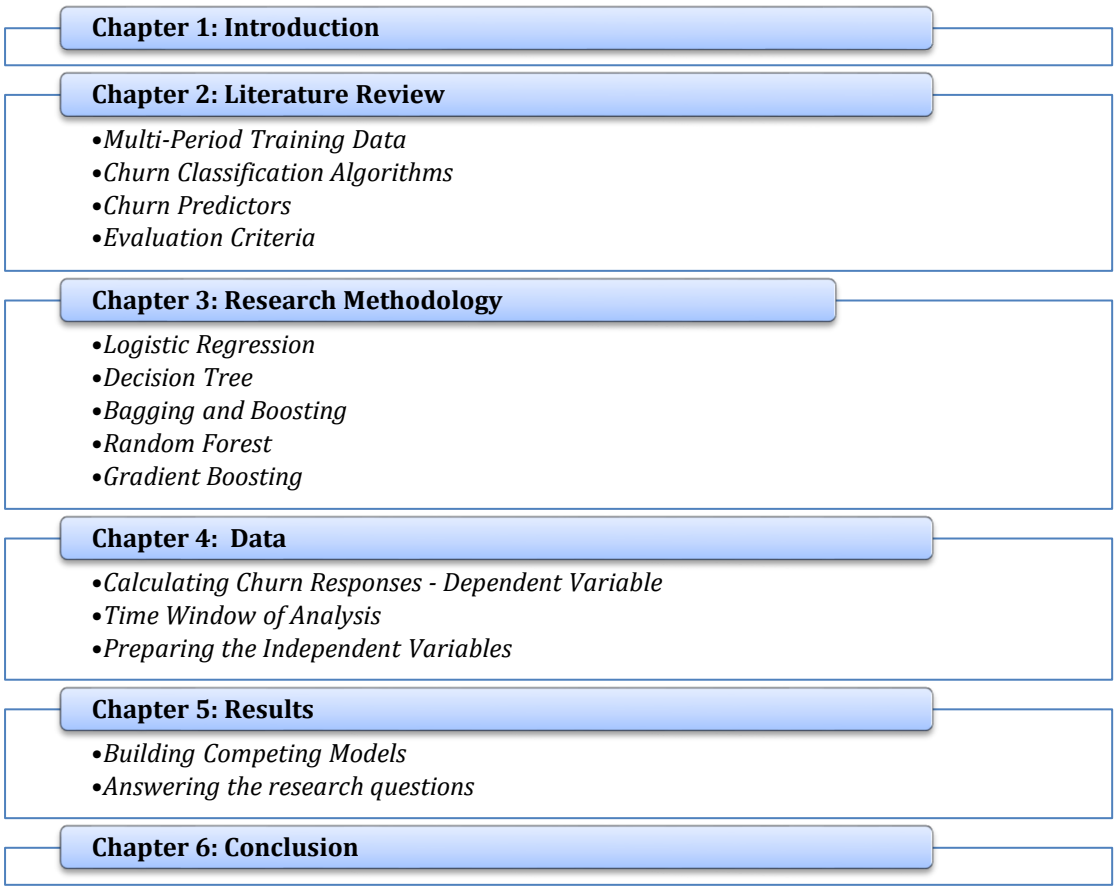


*Figure 1: Thesis structure*

# 2 Literature Review on Churn Prediction

This chapter reviews the selected aspects of churn classification modelling. First, the author describes the proposed multi-period training data approach proposed by Gur Ali & Ariturk (2014). Second, the thesis discusses the most widely used churn algorithms and argues for the selected ensemble methods for this thesis. Then, the thesis reviews the main groups of churn predictors that have been used in churn classification literature. This section concludes with the evaluation criteria that are commonly used to assess churn model performance and are employed in this thesis.

An overview of these aspects in churn literature over the past years is summarized in Table 1. The details are discussed in the sub-sections below.

*Table 1: An overview of several recent churn studies*

| Reference | Churn problem context | Churn predictor | | | | Training data construction | | Churn algorithms and their performance | Evaluation criteria |
|---|---|---|---|---|---|---|---|---|---|
| | | Demographic | Customer behavior | Relationship characteristic | External data | Time-varying churn predictors | Time-varying churn responses | | |
| (Fitzpatrick & Mues, 2015) | Banking | x | x | | | | | Boosted Regression Trees, Random forests, Generalized Additive Models > Logistic Regression | H measure, Area Under Curve (ROC Index) |
| (Gur Ali & Ariturk, 2014) | Banking | x | x | x | x | x | x | Binary classification > Survival analysis (Multi-period period)<br>Logistic Regression > Decision Tree (Single period prediction) | Area Under Curve (ROC Index), Top Decile Lift |
| (Lemmens & Gupta, 2013) | Telecom. | x | x | x | | | | Stochastic Gradient Boosting with profit loss function > Logistic Regression or Decision Tree | Gain/loss matrix |
| (Ballings & Van den Poel, 2012) | Media | x | x | x | | x | | Bagging of classification trees > Logistic Regression > Decision Tree | Area Under Curve (ROC Index) |
| (Lee, et al., 2012) | Telecom. | x | x | | | x | x | k-nearest-neighbor time series classification | N/A |
| (Nie, et al., 2011) | Banking | x | x | | | x | | Logistic Regression > Decision Tree | Type I Type II error, misclassification cost |
| (Yaya, et al., 2009) | Banking | x | x | | | | | Improved Balanced Random forests > Support Vector Machine > Neural Network > Decision Tree | Lift curve, Top Decile Lift |
| (Burez & Van den Poel, 2009) | Banking, Telecom., Media, Retail | - | - | - | - | - | - | Random forests > Logistic Regression | Area Under Curve (ROC Index), Lift |
| (Qi, et al., 2008) | Telecom. | x | x | x | | x | | ADTreeLogit > ADTree (Alternative Decision Tree) > Logistic regression | Area Under Curve (ROC Index) |
| (Mavri & Ioannou, 2008) | Banking | x | x | x | | x | x | Survival analysis (Cox regression) | N/A |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (van Wezel & Potharst, 2007) | Retail | | x | x | | x | | Multi-boosted CART > Boosted CART > CART | Error rate |
| (Lemmens & Croux, 2006) | Telecom. | x | x | x | | | | Stochastic Gradient Boosting > Bagging of Decision Tree > Binary Logit Model | Error rate, Top Decile Lift, Gini Coefficient |
| (Lariviere & Van den Poel, 2005) | Financial service | x | x | | x | | | Random forests > Logistic Regression | Area Under Curve (ROC Index) |
| *"-" = not mentioned*<br>*"blank" = not used* | | | | | | | | | |

## 2.1 Multi-period Training Data

As can be seen from Table 1, under the header "Training data construction", time-varying churn predictors are employed fairly often among recent research whereas very few studies include time-varying churn responses in their training data. Such an observation confirms the argument by Gur Ali & Ariturk (2014) that even though the inclusion of time series data for both churn predictors and churn responses has been found to increase churn classification accuracy, the approach has not been employed much. The thesis now discusses the proposed multi-period training data approach by Gur Ali & Ariturk (2014), which has been found to outperform the traditional approach, which is called single period training data by Gur Ali & Ariturk (2014) in their research.

First, it is worth starting the discussion with the single period training data approach that captures only one churn response for each customer at a specific point of time. In other words, one customer constitutes only one row in the training data set. Meanwhile, the time-varying predictors are summarized over the most recent period of time into single columns instead of time series using simple aggregation. Examples of such independent variables in some churn prediction studies are average of monthly minutes of use over the past six months, the percentage change compared to the previous six months in a telecommunication churn study (Lemmens & Croux, 2006), or total credit or debit transactions in the last three months in a retail banking churn study (Prasad & Madhavi, 2012). The term "single period training data" is used to refer to such an approach, which is illustrated as below:

$$SPTD_t = [Y_t \quad S \quad D_{t-\delta}]$$

- $SPTD_t$ is the single period training data for the churn period $t$
- $Y_t$ is a vector or an $n$ x 1 matrix of churn responses $y$ at time $t$ of $n$ customers
- $S$ is an $n$ x $s$ matrix of non-time varying variables $s_{ij}$ for $n$ customers. The most popular variables employed in churn prediction that are not changing over time are usually demographic variables like gender or, within the time window of a year, age and educational level.
- $D_{t-\delta}$ is an $n$ x $d$ matrix of time varying variables $d_{ij}$ for $n$ customers at time $t-\delta$, where $\delta$ is the length of either the operational lead time for data to be available in the internal databases or of the lead time reserved for customer

retention action between the observation period and the time when churn behavior is supposed to occur.

Figure 2 illustrates a general time window of analysis in traditional churn prediction studies: the observation period provides the training data for the model, the performance period is the time when churn is defined, and the interval between these two periods is allowed for marketing action to retain worth-while customers (Lu, et al., 2014).

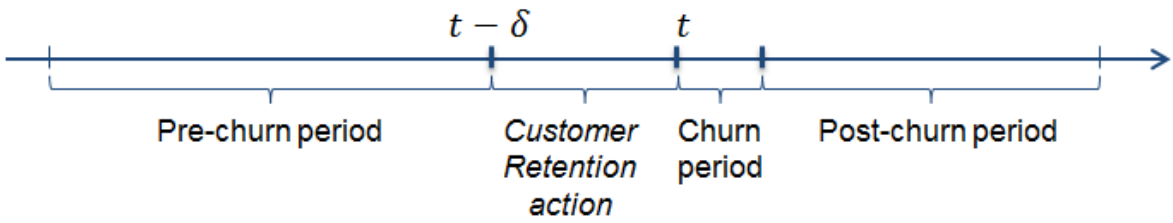


*Figure 2: Time window of analysis illustration for single period training data*

In order to capture the time variation in both churn predictors and churn response, the multi-period training data is composed of multiple single period training data sets ranging from the latest possible point of time $t$ back in time to get the historical churn responses. Such point of time is denoted as $k$ in the below formula for the multi-period training data:

$$MPTD_t = \begin{bmatrix} SPTD_t \\ SPTD_{t-1} \\ \vdots \\ SPTD_{k+1} \\ SPTD_k \end{bmatrix} E \Bigg] = \begin{bmatrix} Y_t & S & D_{t-\delta} & E_{t-\delta} \\ Y_{t-1} & S & D_{t-1-\delta} & E_{t-1-\delta} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{k+1} & S & D_{k+1-\delta} & E_{k+1-\delta} \\ Y_k & S & D_{k-\delta} & E_{k-\delta} \end{bmatrix}$$

- $MPTD_t$ is the multi-period training data for the churn period $t$
- $SPTD_t$ is a matrix of single period training data as shown below, illustrating that the multi-period training data is comprised of multiple single period training data matrices. Although it is more probable that the single period training data sets have various numbers of observations in practice, for the sake of simplicity in mathematical illustration, the single period training data sets are assumed to have equally $n$ observations.
- $E$ is a $n(t-k)$ x $e$ matrix of time-varying macro-environmental factors $e_{ij}$ during the specified period. As can be seen in single period training data and multi-period training data formulas, it is not possible to include environmental predictors in single period training data because only one point of time is included (Gur Ali & Ariturk, 2014).

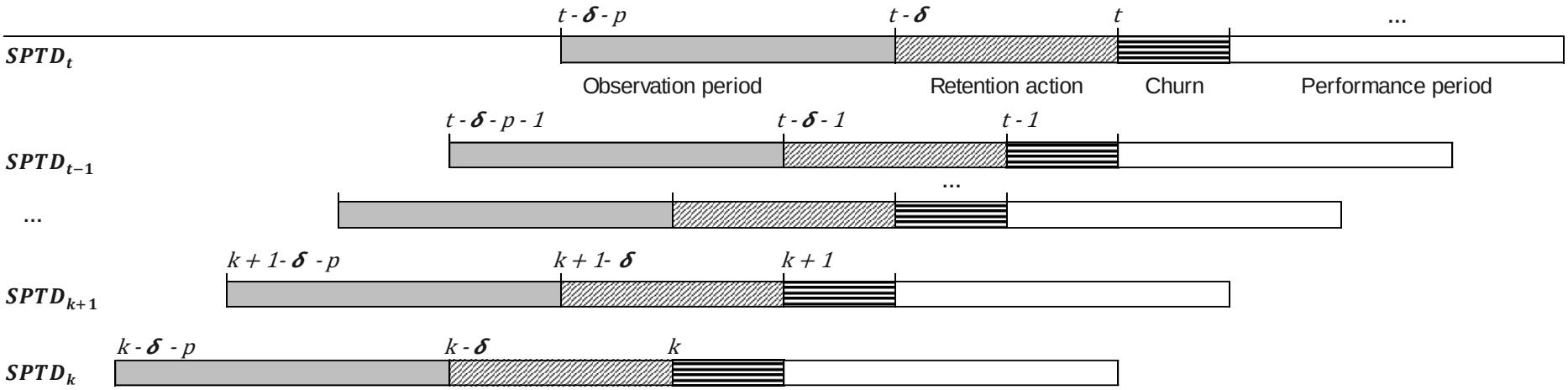


*Figure 3: Time windows of analysis illustration for single period training data and multi-period training data*

- $\boldsymbol{t}$ is the time when the latest churn response is recorded over a pre-defined period of time. The churn period of the housing loan data in this thesis is one month.
- $\boldsymbol{\delta}$ is the length of the period reserved for marketing retention action; hence the starting point of this period is $\boldsymbol{t - \delta}$. For this thesis, $\boldsymbol{\delta}$ **is** assumed to be one month.
- $\boldsymbol{k}$ is the time when the earliest churn response is recorded.
- $\boldsymbol{p}$ is the length of the observation period, when the time series of input variables are aggregated.
- After each churn period, the performance period starts to capture the post-churn criteria for churn response calculation.

Each of the periods is marked and labelled accordingly in each single period training data set. The lengths of these periods can always be adjusted accordingly to fit the business context.

An illustration of the time windows of analysis for single period training data and multi-period training data is provided in Figure 3.

Multi-period training data includes multiple periods of single period training data from time $t$, when the latest churn response is recorded, backwards to time $k$ to obtain the historical churn responses. The factors to determine $k$ include the availability of data, the tenure of the customers and the computational efficiency of the model. If the single period training data sets are assumed to have equally $n$ observations, then the multi-period training data has $n * (t - k)$ observations. This obvious increase in the data size might create computation burden for the model; hence it might not be always a good idea to include all the historical data available.

The additional observations are not generated from the current data set as in different sampling schemes (Burez & Van den Poel, 2009) but are taken from the otherwise discarded data from historical behaviors and churn responses; hence, one customer can have as many observations as the periods of time that they are active. As a result, the models are fed with considerably more new data than in the single period training data approach, a characteristic that is claimed to boost the learning capacity of churn models. Most importantly, even though the churn ratio is the same in both training sets, the multi-period training data has theoretically $(t - k)$ times more churners than the single period training data in absolute number assuming the churn rate remains constant over the period of time, mitigating the absolute rarity problem, which is the main purpose of the proposed approach. Finally, multi-period training data does not suffer from the curse of dimensionality with the large additional data amount because these data are added as rows in the training data set instead of columns. Curse of dimensionality generally refers to the phenomenon that when the sample data size does not increase exponentially in response to the increase in the number of variables, the data becomes sparse in this high dimensional space, causing problems to the models, such as overfitting the data to the model, especially in churn problems when churn events are rare. (Gur Ali & Ariturk, 2014)

The last point is illustrated by constructing also a single period training data + lags training data as shown in the equation below:

$$SPTD + lags_w = [SPTD_t \quad D_{t-\delta-1} \quad D_{t-\delta-2} \quad \dots \quad D_{t-\delta-w-1}]$$
$$= [Y_t \quad S \quad D_{t-\delta} \quad D_{t-\delta-2} \quad \dots \quad D_{t-\delta-w-1}]$$

$SPTD + lags$ includes single period training data plus the predictors' values during the time between the $t$ and the lags $t - w$; consequently, for each independent variable there are $w$ more columns or dimensions in the training data while the total sample size is the same as the single period training data. Among the three approaches, models built from single period training data + lags perform the worst because, among other things, it suffers from the curse of dimensionality explained above.

However, multi-period training data is not without its own limitations. Figure 3 shows that the observation periods for the single period training data sets within multi-period training data are highly overlapping for two consecutive periods. Consequently, the data used to aggregate the time varying variables $d_{ij}$ for the same customers are largely overlapping, and the resulting time-varying predictors are highly correlated. This phenomenon is called multi-collinearity in statistical models, when independent variables in a model are highly correlated with one another, hindering the reliability of the estimated parameters (Van den Poel & Lariviere, 2004). Moreover, time series are prone to be highly correlated with their lagged values, worsen the multi-collinearity among the time-varying variables for the same customers. It can be said that employing longitudinal data from the same customers do not conform to the rule of independent and identical distribution; and such characteristic might bias the estimated parameters in statistic algorithms such as logistic regression (Gur Ali & Ariturk, 2014). In response to these limitations, Gur Ali & Ariturk (2014) emphasizes the goal to improve the performance of binary classification models instead of correcting biased parameters. Their study shows that multi-period training data outperforms the other two approaches single period training data and single period training data + lags in churn prediction for both next period and multiple periods (Gur Ali & Ariturk, 2014). Additionally, as single period training data + lags approach consistently produces the worse performances among the three proposed.

## 2.2 Churn Classification Algorithms

As can be seen from Table 1, under the column "Churn algorithm and their performance", logistic regression and decision tree are employed in nearly all of the reviewed papers, especially when comparison with other more advanced algorithms is needed. This section discusses the churn algorithms that have been used in churn classification and argues for the selected ones to employ in this thesis.

Logistic regression is a statistical method that returns a probability for each observation as the output variable; as a result, it is frequently chosen as the scoring method for the probability to churn among customers. At a specific cut-off (usually 0.5), customers with the estimated probabilities higher than the cut-off point are marked as predicted churners and the rest are predicted non-churners. Logistic regression is frequently chosen in marketing research that deals with individual customers because the model presents the assumed relationship between the independent variables and the dependent variable descriptively, making it possible to interpret how significant and how much an independent variable is able to explain the target variable based on the estimated parameters. Moreover, logistic regression has been found to perform relatively well compared to other more complex models with various data sets from different domains (Fitzpatrick & Mues, 2015). Thanks to its robustness, logistic regression is used a great deal in churn prediction, especially as the baseline in comparison studies (Fitzpatrick & Mues, 2015; Gur Ali & Ariturk, 2014; Lemmens & Gupta, 2013; Ballings & Van den Poel, 2012).

The other popular algorithm in marketing generally and churn classification specifically is decision tree thanks to its clear presentation in the tree structure to show the most significant rules that define the target variable; hence, decision tree has been employed in churn prediction with the ambition to capture the rules from customer behaviors that can detect early signs of churn (Gur Ali & Ariturk, 2014; Lemmens & Gupta, 2013; Nie, et al., 2011). However, decision tree models are usually associated with lack of robustness and sub-optimal performance in churn literature (Lariviere & Van den Poel, 2005). As a result, it can be observed from Table 1 that a wide range of complex models have been developed based on decision trees such as Multi-boosted CART, Boosted CART (Qi, et al., 2008), ADTree (van Wezel & Potharst, 2007), bagging and boosting of classification tree (Ballings & Van den Poel, 2012; Lemmens & Croux, 2006), stochastic gradient boosting (Lemmens & Gupta, 2013), or random forests (Van den Poel & Lariviere, 2005 & 2004). Additionally, more complicated models can be developed not only from decision tree but also from its combination with logistic regression algorithm like ADTreeLogit (Qi, et al., 2008).

Most of the advanced algorithms listed above belong to the ensemble family, revealing the wide application of ensemble methods in churn prediction. In general, ensemble refers to the different methods to combine two or more models into a single yet more powerful one with the aim to improve predictive performance (Yaya, et al., 2009; Jinbo, et al., 2007; Lariviere & Van den Poel, 2005). The most widely used methods in ensemble modelling are

bagging and boosting. Both methods use an algorithm as a base learner to train a model repeatedly on different samples of the original data; finally for each observation, a single predictive score is aggregated from the iterative training. In this case, a base learner is basically an algorithm that is selected to be trained iteratively to produce an ensemble model. The typical base learner in churn classification is decision tree (Chrzanowska, et al., 2009; Qi, et al., 2008; Lariviere & Van den Poel, 2005). A brief review on bagging and boosting is provided in the section 3.3.3 under Research Methodology.

The most powerful and frequently used bagging method in churn prediction is random forest (Yaya, et al., 2009; Lariviere & Van den Poel, 2005; Chen, et al., 2004). Random forest has been found to provide better estimations for binary classification of customer churn than ordinary linear regression and logistic regression (Lariviere & Van den Poel, 2005). In order to improve the capability of churn prediction models in handling imbalanced data, various enhanced versions of random forest models incorporate sampling techniques and cost optimization training. In the research conducted by Chen et al. (2004), the improved versions of random forests are trained on 6 different data sets in various domains from oil exploration to Hypothyroid and Euthyroid diagnoses to produce better churn classification results compared to other existing techniques like logistic regression and decision tree (Chen, et al., 2004). In another research, Yaya, et al. (2009) combines all the improvements suggested by Chen, et al. (2004) into one model called improved balanced random forest, which reveals to provide more accurate churn prediction compared with neural network and decision tree models (Yaya, et al., 2009).

On the other hand, the boosting algorithms popularly employed in churn prediction are stochastic gradient boosting (Lemmens & Gupta, 2013) and AdaBoost (Lu, et al., 2014; Jinbo, et al., 2007). Although stochastic gradient boosting is claimed to be the most sophisticated algorithm in the boosting family, it has been found to outperform binary logit model while performing comparatively with bagging in a churn prediction study of the telecommunications industry (Lemmens & Croux, 2006). In a more recent research, stochastic gradient boosting has been employed in junction with a proposed loss function by Lemmens & Gupta (2013) and has been found to increase the profitability of retention campaigns compared to logistic regression for a telecommunications company (Lemmens & Gupta, 2013).

Due to the superior performance seen in churn literature, this thesis selects random forest and gradient boosting as the advanced algorithm representatives from the ensemble

family. A technical reason for the selection is the possibility to run these models in SAS Enterprise Miner. In conclusion, four algorithms are employed in this thesis to build churn prediction models: logistic regression and decision tree as the standard algorithms, random forest and gradient boosting as the candidates for the advanced methods.

## 2.3 Churn Predictors

The most widely used churn predictors can be categorized into the following groups: demographic data, customer behavior data, customer satisfaction data, and external factors (A. O. Oyeniyi & A.B. Adeyemo, 2015; Nie, et al., 2011; Van den Poel & Lariviere, 2004; Koh & Chan, 2002).

First of all, demographic data such as gender, age, salary, or educational level are static variables that are usually assumed to remain unchanged over the time window of analysis within a year (Gur Ali & Ariturk, 2014; Mavri & Ioannou, 2008; Koh & Chan, 2002). For example, age and gender are found to be significant to predict churn behavior among banking customers (Koh & Chan, 2002). In another study of customer switching behavior for a Greek bank using survival analysis, Mavri & Ioannou (2008) finds that the age group 30 – 40 years old is the most likely to evaluate their cooperation with the existing bank and seek for better terms elsewhere (Mavri & Ioannou, 2008). However, it is often a challenge for banks to obtain high quality for demographic data due to the fact that customers are not mandated to provide all the information in this category. For example, customers do not have to provide their educational level or monthly salary if they do not purchase products that requires such information (Employee, 2016). Therefore, these demographic variables have a high level of missing data and hence, frequently dismissed from the models (Prasad & Madhavi, 2012). Additional to age and salary, lifecycle stage is also employed in churn prediction in the banking context as it is found to significantly impact how customers prioritize their financial differently at different life stages, for example young customers have less money to spend, hence have smaller investment portfolios than their more mature counterparts (Lariviere & Van den Poel, 2005).

Second, customer behavior from transactional databases has been deemed the most important group of predictors for churn literature because of the recency, frequency and monetary (RFM) information that they provide (Ballings & Van den Poel, 2012; Baecke & Van den Poel, 2009). This group of churn predictors also includes information about product

portfolio in terms of volume and their changes over time (Lariviere & Van den Poel, 2005 & 2004).

The third group of churn predictors is the characteristic of customer relationship with the organization, for example, customer interaction with the bank in terms of contacts and tenure (Gur Ali & Ariturk, 2014; Ballings & Van den Poel, 2012; Mavri & Ioannou, 2008). Additionally, customer satisfaction's impact on churn behavior in the banking context has also been studied extensively. Findings suggest that customer satisfaction and customer's perception towards the company's image have a great influence on customer loyalty and the duration of customer relationship. Using proportional hazard method to compare the influence of various predictors on churn behavior, Mavri & Ioannou (2008) discovers that there is a significant difference in churn probability between customers that rank the bank's service as "very important" and "extremely important" (Mavri & Ioannou, 2008). An interesting suggestion from Lariviere & Van den Poel (2005) is the employment of variables related to intermediaries, or particularly sales agents. Examples of such variables are the extent to which the salesperson is prone to cross-sell, the product variety in the offerings from a salesperson, and the number of customers served by a salesperson. In their churn model using random forests, these variables, especially the tendency to sell of the sales person are found to significantly impact customers' decision either in both cross-buying and churn contexts (Lariviere & Van den Poel, 2005).

Last but not least, the fourth group of churn predictors employed in the literature is the external information extracted outside companies' internal data. In dynamic churn literature, churn behavior is argued to be affected by the changing economic conditions that are unseen from the companies' internal databases. Therefore, models, where churn response and independent variables are limited to only a specific time period, cannot employ these time-varying environmental predictors and hence, cannot capture the dynamic variation in churn behavior (Gur Ali & Ariturk, 2014). In their dynamic churn prediction research for private banking customers, Gur Ali & Ariturk (2014) confirms the relevance of macro-economic variables in explaining churn responses. For example, deposit interest rate has a significant effect on private customers' decision to switch banks because most of the private customers' assets are fixed rate deposits (Gur Ali & Ariturk, 2014). Ballings & Van den Poel (2012) also includes a predictor that captures the historical merger and acquisition situation of a financial organization when studying its customers' churn behavior over the period of 77 years. This factor is found to have a statistically significant positive contribution to the retention rate, a

finding that aligns with the market perception about the merger (Ballings & Van den Poel, 2012). Fitzpatrick & Mues (2015) also specifies the hidden non-linear relationship between the churn response and the unseen variables that might be external to the company's context such as unemployment rate in a specific region (Fitzpatrick & Mues, 2015). When comparing with behavior predictors, customer satisfaction and macro-environmental factors are found to have more significant impact on churn behavior (Mavri & Ioannou, 2008).

Baecke & Van den Poel (2009) points out two major limitations of internal data that banks are passively collecting from their customers' information and behavior. First of all, these internal databases represent the limited versions of customers that are observed by the banks. For example, the data only tells what products and by how much a particular customer has purchased but provides no inference about the over-all needs for the total product category. He/she might be purchasing a similar product from a competitor at the same time. However, such data is not available for banks internally. Additionally, focusing on transactional data does not provide banks with an understanding of the motivation and attitudes that drive these purchases. This drawback is argued to possibly hinder the possibility to build long term relationship with customers (Baecke & Van den Poel, 2009). Lariviere & Van den Poel (2005) also advocates the use of geo-demographic data from external data sources to build a more comprehensive picture of the customers (Lariviere & Van den Poel, 2005)

In conclusion, the richer of information that the training data includes, the better the model performs. Nie et al. (2011) compares the performance to predict churn in credit card customers by building multiple models using both single source of information and combination of different sources such as customer demographic data, cards general information, transaction information and risk related information. The model built from a diverse independent variable set outperforms the ones with less diverse data (Nie, et al., 2011). To the author's best knowledge, such approach has not been done for churn problem in the banking context, however, the conclusion is argued to remain that the more diverse the data is, the better the model performs.

## 2.4 Evaluation Criteria

Even though churn literature employs a wide selection of algorithms, as can be seen from Table 1, only a few criteria for performance evaluation are universally selected for model comparison such as misclassification rate, AUC (Area under Curve) or ROC (Receiver

Operating Characteristic) index, and top decile lift (Gur Ali & Ariturk, 2014; Nie, et al., 2011; Yaya, et al., 2009; Burez & Van den Poel, 2009; Lemmens & Croux, 2006). These evaluation metrics and some of their improved forms are argued to take into consideration the imbalanced data set in churn problem (Weiss, 2004).

The next sub-sections provide more detailed reviews on the selected evaluation criteria in this thesis: misclassification rate, ROC index and top decile lift.

### 2.4.1 Misclassification Rate

Binary classification models do not naturally have 100% correct prediction but normally some of the real churners are misclassified as non-churners and vice versa, some of the real non-churners are misclassified as churners. Such statistics are shown in the confusion matrix in Table 2, where churn events are commonly denoted as positive events and non-churn events as negative events. As can be seen from the confusion matrix, among the predicted positive events, there are both true positive events and false positive events. The same applies to the predicted negative events, which include both false negative events and true negative events. True positive events refer to the real churners that are classified as churners by the model. Similarly, true negative events refer to the real non-churners that are classified as non-churners by the model. On the other hand, false positive events refer to the real non-churners that are misclassified as churners by the model while false negative events refer to the real churners that are misclassified as non-churners by the model.

*Table 2: Confusion matrix of churn classification*

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Real Positive** | True Positive | False Negative |
| **Real Negative** | False Positive | True Negative |

From these parameters in the confusion matrix, the accuracy is calculated as followed

$$Accuracy = \frac{number\ of\ correctly\ classified\ events}{number\ of\ all\ events}$$

and the misclassification rate is identified as followed

$$Misclassification\ rate = 1 - Accuracy$$

(Burez & Van den Poel, 2009)

Regarding the confusion matrix in Table 2, the false positive events and false negative events are also referred to as type I and type II errors in the binary classification literature. As a result, type I error refers to the number of the real non-churners that are misclassified as churners while type II error indicates the number of real churners that are misclassified as non-churners. A disadvantage of the misclassification rate is that it views the two types of error equally. Since the would-be churners that are not identified for the retention campaign will churn anyway, the bank will suffer a loss of future profits from the customers not only from housing loan products but also from other possible products that the customers might be or will be holding with the bank. On the other hand, non-churners that are misclassified as churners only cost the bank the unnecessary retention action because they will continue their relationship with the service provider anyway. Consequently, the loss caused by type II error for the company is argued to be much higher than type I error (Nie, et al., 2011). As a result, various loss functions have been designed to capture the actual loss of misclassification that is specific to each organization. Moreover, as it is more costly to falsely identify highly profitable customers than the less profitable ones, customer life time value has also been incorporated in loss functions to capture only the most worth-while customers to retain (Glady, et al., 2009; Lemmens & Croux, 2006). Additionally, Lemmens & Gupta (2013) looks at customer retention from the angle of profit maximization for the marketing campaign and incorporate both customers' value and responsiveness to retention incentives in their loss function using stochastic gradient boosting to design the most profitable target size (Lemmens & Gupta, 2013).

### 2.4.2 ROC Index

ROC stands for receiver operating characteristic (ROC) curve that assumes the form of a graph of $(x, y)$ where

$$x = 1 - \ specificity$$

$$y = sensitivity$$

$$Specificity = \frac{number\ of\ True\ Negative\ events}{number\ of\ Real\ Negative\ events}$$

and

$$Sensitivity = \frac{number\ of\ True\ Positive\ events}{number\ of\ Real\ Positive\ events}$$

For a binary classifier, the horizontal axis of a ROC curve shows the ratio between the number of falsely predicted positive events and the number of real negative events; while the vertical axis presents the ratio of the number of correctly predicted positive events out of the number of real positive events. As a result, the ROC curve plots the true positive rate $(y)$ against the false positive rate $(x)$ (Gur Ali & Ariturk, 2014).

A few examples of ROC curves are demonstrated in Figure 4. The diagonal line connecting the points (0, 0) and (1, 1) represents the random guess that any customer is 50% a churner and a non-churner because on this line, the true positive rate is always equal to the false positive rate. Hence, the ROC curves that closely follow this diagonal line show the inability of the model to identify churn events against non-churn events. An ideal ROC curve is the one that follows closely the vertical axis at first with high true positive rate and low false positive rate and then curves closely towards the horizontal axis. Therefore, between the two lines denoted with ROC 1 and ROC 2 in Figure 4, the model with the ROC 2 curve is more preferable.

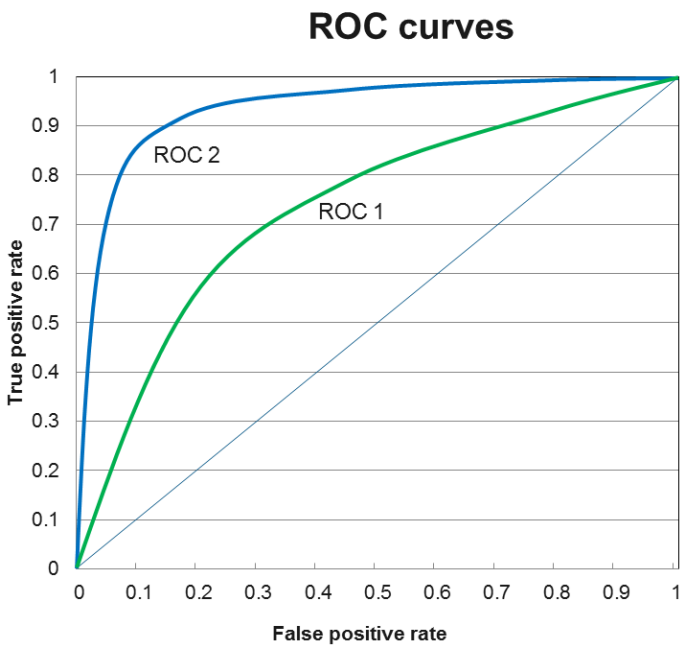


*Figure 4: Example of a ROC chart*

In order to provide the ROC index for model comparison, the area under the ROC curve, also usually referred to as the Area under Curve (AUC) in churn literature, is computed as shown by the formula below.

$$ROC\ index = \int_0^1 \frac{number\ of\ true\ positive\ events}{number\ of\ real\ positive\ events} d \frac{number\ of\ false\ positive\ event}{number\ of\ real\ negative\ events}$$

The ROC index represents the ability of the model to discriminate a positive event from a negative one. Imagine when the sample is divided into two groups of real positive events and negative events, one case is then selected randomly from both groups. A good churn classification model should give a positive event a higher probability to churn than a negative event. In order to get the ROC index, the number of pairs that have the positive event receiving a higher probability from the model than the negative event are divided by the total number of randomly selected pairs. As a result, a ROC index can tell how precisely the model is able to classify true positive events against negative ones. (Burez & Van den Poel, 2009).

Regarding the graph in Figure 4, the larger the area under the curve is, the higher the index becomes and the more precise and preferable the model is. In numerical terms, the diagonal line connecting the points (0, 0) and (1, 1) represents the ROC index of 0.5, which shows no ability to recognize a churn events from a non-churn event. ROC 2 curve is more preferable than ROC 1 curve because the area under the former curve is obviously larger than the latter one. Moreover, an index of less than 0.5 is said to indicate that the model is misleading (Gur Ali & Ariturk, 2014). Consequently, the ROC index has 0.5 as the lower bound and 1.0 as the upper bound.

### 2.4.3 Top Decile Lift

Before discussing the top decile lift, the thesis defines lift to have the followed formula

$$Lift = \frac{\hat{\pi}_{x\%}}{\hat{\pi}_0}$$

- $\hat{\pi}_{x\%}$ is the proportion of churners in the top x% customers with the highest probability to churn given by the model
- $\hat{\pi}_0$ is the proportion of churners in the whole population

When no model is employed, for any $x\%$ of the sample, the expected proportion of churners is $\hat{\pi}_0$. A good churn classification should be able to give higher probability to churn events compared to non-churn events, hence having more churn events among the events receiving the highest probability. As a result, compared to the case without using any model, lift measures the superiority of a churn classification model to identify more churn events among the events that receive the highest probability from the model (Gur Ali & Ariturk, 2014). A lift chart is illustrated in Figure 5 where the horizontal axis is the percentages of the

samples sorted by their probabilities given by the model and the vertical axis is the ratio between the churn ratio in a specific percentile and the churn ratio of the whole population.

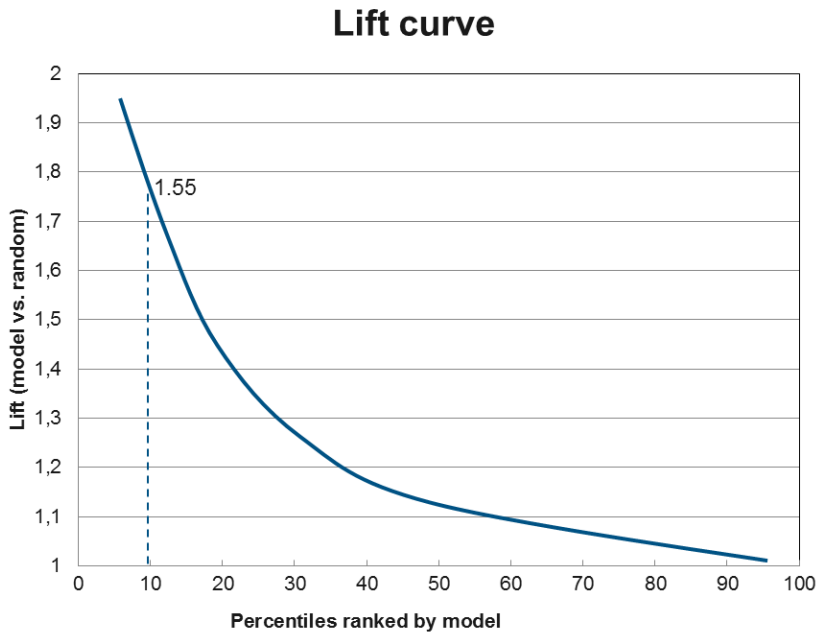


*Figure 5: Illustration of a lift chart*

Lift validates whether the model is able to capture more churners within the top percentiles based on its ranking compared with a random model (Lemmens & Croux, 2006); therefore, a lift of 1 means no lift at all. Figure 5 shows that the lift curve decreases towards 1 when it reaches towards the total population. As a result, 1 is the lower bound of the lift curve.

The popular cut-off used in churn literature when it comes to lift is the top 10%. As the name suggested, top decile lift focuses on the top 10% customers with the highest churn probabilities. Top decile lift is calculated as shown in the formula below:

$$Top\ decile\ lift = \frac{\hat{\pi}_{10\%}}{\hat{\pi}_0}$$

- $\hat{\pi}_{10\%}$ is the proportion of churners in the top 10% riskiest customers based on the model's prediction
- $\hat{\pi}_0$ is the proportion of churners in the whole population

Take the example in Figure 5, the top decile lift is 1.55, meaning that by selecting the top 10% of customers ranked by the model for a retention campaign, it is possible to target at 55% more potential churners than not using the model. The higher the top decile lift is, the

better the model is able to capture churners by its ranking; hence top decile lift is claimed to provide managerial suggestions that are straight-forward and actionable (Gur Ali & Ariturk, 2014).

However, researchers who promote profit maximization for the retention campaign criticize churn models that ignore the different costs between Type I and Type II errors because in such cases, the top decile cut-off is merely an arbitrary target size based on the churn probability ranking that ignores the heterogeneity of customers in terms of the generated profitability and the tendency to response to marketing incentives (Lemmens & Gupta, 2013).

# 3 Research Methodology

In order to answer the research question, four methods are employed in this thesis as churn classification algorithms:

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting

All the four methods will be employed with both multi-period training data and single period training day to create competing models. Specifically, logistic regression and decision tree are used to run the baseline models with single period training data. The other models are then compared with the baseline models in order to answer the research question.

In this section, the four selected algorithms are briefly reviewed. As random forest and gradient boosting belong to the bagging and boosting methods of the ensemble "family", learning the basics of bagging and boosting helps understand the algorithms developed out of these methods: random forest and gradient boosting.

The main goal of a churn classification model is to train an algorithm on a specific training data set with a vector of independent variables or churn predictors $\boldsymbol{X} = (x_1, x_2, \dots, x_i)$ to produce for each observation a binary dependent variable or churn response $Y$, which takes the values of either 1 for churners and 0 or -1 for non-churners depending on the churn algorithms.

## 3.1 Logistic Regression

Given the vector $\boldsymbol{X} = (x_1, x_2, \dots, x_i)$ of independent variables $x_i$ as inputs, $g(\boldsymbol{X})$ is the linear function of $\boldsymbol{X}$.

$$g(\boldsymbol{X}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

As the goal is to predict the probability of churn, logistic regression returns the binary output $Y$ that takes the values of either 1 for churners or 0 for non-churners. Let's denote $F(\boldsymbol{X})$ as the probability of churn:

$$F(\boldsymbol{X}) = P(Y = 1|\boldsymbol{X})$$

The odds ratio is defined as:

$$\frac{F(\boldsymbol{X})}{1-F(\boldsymbol{X})} = \frac{probability\ of\ churn}{probability\ of\ nonchurn}$$

Then the linear regression function $g(\boldsymbol{X})$ is set to equal the log of the odds ratio, producing the model for the logistic regression $F(\boldsymbol{X})$:

$$g(\boldsymbol{X}) = \ln\left(\frac{F(\boldsymbol{X})}{1-F(\boldsymbol{X})}\right)$$

and solving for

$$F(\boldsymbol{X}) = \frac{1}{1+e^{-g(\boldsymbol{X})}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_i x_i)}}$$

Such modelling ensures that the values of $F(\boldsymbol{X})$ are restricted from 0 to 1. In order to illustrate how logistic regression model is interpreted, let's take a simple model with two predictors as an example:

$$\ln\left(\frac{F(\boldsymbol{X})}{1-F(\boldsymbol{X})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 4.51 - 0.38x_1 + 0.05x_2$$

The estimated effect of each predictor is calculated by subtracting the exponential of the estimated parameters from 1. For example, the effect of $x_1$ is $(e^{-0.38} - 1) = -0.26$. It means that, for each unit increase in predictor $x_1$, the odds of the positive event (churn event in this context) decreases by 26% measured in log-scale. Likewise, the effect of $x_2$ is $(e^{0.05} - 1) = 0.05$; meaning for each unit increase in predictor $x_2$, the odds of the positive event or churn event increases by 5%.

## 3.2 Decision Tree

In contrast to logistic regression, decision tree is a non-parametric algorithm. In general, decision tree maps the independent variables to the membership of different classes. In the tree structure, leaves represent classification and branches represent the conditions that lead to the membership of different groups (Chrzanowska, et al., 2009).

Decision trees are grown based on the main ideas of recursive partitioning and pruning. Recursive partitioning continuously splits the predictors into groups and further sub-groups with the aim to create homogeneous groups. Such groups are considered leaves in the tree structure. As the recursive partitioning process applies latter splits based on the previous one, the tree grows into more complexed form. Due to this increasing dependency, complex tree structure is claimed to be sensitive to changes in the training data. As a result, pruning is

employed to prevent the formation of too large trees using different measures for example, misclassification error or Gini coefficient to avoid overfitting the tree to the training data. (Fitzpatrick & Mues, 2015)

Several algorithms are employed in decision tree like ID3, CART, CHAID or C5.0 (Nie, et al., 2011), among which the most frequently used are CART and CHAID. CART stands for Classification And Regression Tree. A classification tree predicts categorical target variables while a regression tree predicts continuous target variables. CHAID stands for Chi Square Automatic Interaction Detector. The preference between these two algorithms is mainly due to empirical performance for a particular data set (Ballings & Van den Poel, 2012).

## 3.3 Bagging and Boosting

### 3.3.1 Bagging

Bagging is shortened from bootstrap aggregating. Bootstrap refers to the sampling technique of random selection with replacement to create multiple training data sets for the associated base learner (Efron, 1979). Base learner is basically a chosen algorithm for the ensemble method, for example logistic regression or decision tree, to be trained iteratively on the bootstrap samples. Base learner can be defined by the function below:

$$h(\boldsymbol{X}; A)$$

- $h$ is usually a simple function of the independent variables $\boldsymbol{X}$ such as decision tree or logistic regression.
- $\boldsymbol{A}$ is the vector of the estimated parameters of $\boldsymbol{X}$ in the function $h$

On the other hand, “aggregating” refers to the technique used to generate the final score from the results of the iterations, in which each iterative model gets an equal weight. In a nutshell, bagging generates multiple independent sub-models simultaneously and aggregate the final result by taking the average of the sub-models’ results for numerical outcome or the class with the maximum votes for a class outcome. (Lemmens & Croux, 2006)

As illustrated in Figure 6, bagging creates multiple bootstrap samples and grows simultaneously a tree from each sample and at the end, aggregates the results to form the final score (Maldonado, et al., 2014).

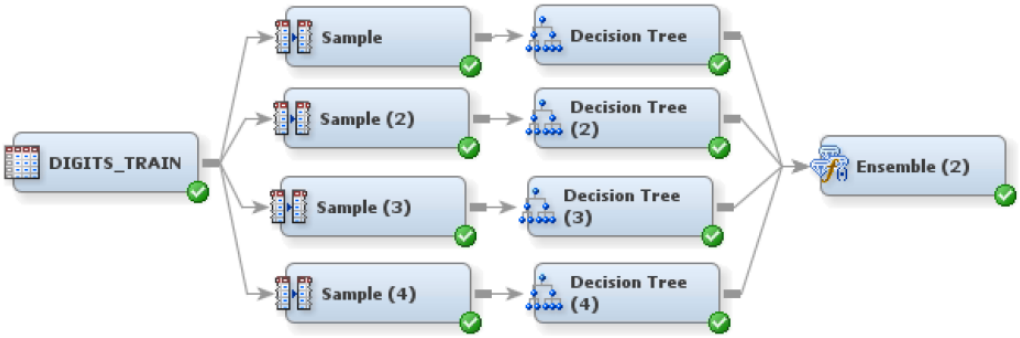


*Figure 6: Illustration of bagging procedure in SAS Enterpsie Miner*

### 3.3.2 Boosting

Boosting algorithm also generates multiple models but as a sequence, which means that after each of the iterations, the previous model's result impacts the way how the training data for the next model is sampled. Instead of generating the training data sets for the associated models randomly as in bagging, boosting algorithm emphasizes more on the examples that are misclassified in previous iterations and selects them more frequently in the subsequent iterations. The final model combines the results from the iterations, each of which receives different weights based on their performances. Boosting is found to achieve better performance than single model through the iterative learning process, which makes more effective use of the rare churn events (Chrzanowska, et al., 2009).

Figure 7 shows that boosting generates the multiple models in sequence. In this figure, the base learner is a decision tree. As can be observed, the first tree's performance is assessed and given a weight. The sample for the second tree is then created by selecting more frequently the misclassified cases from the first tree. The procedure goes on until a pre-defined number of trees are grown. The final score is a weighted sum of all the trees' results. (Maldonado, et al., 2014)

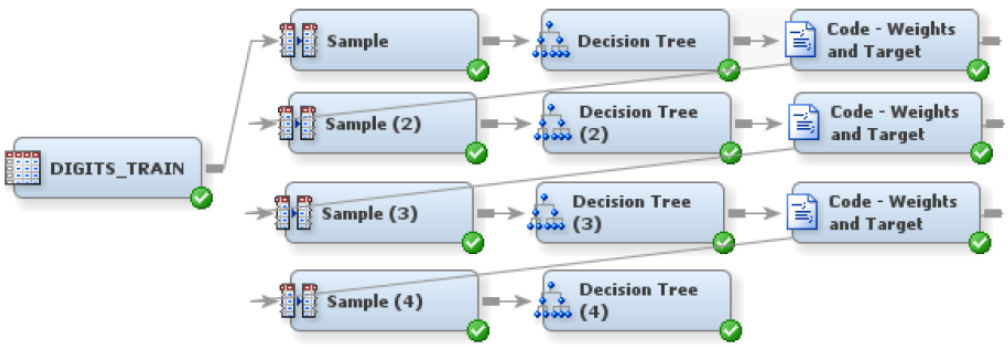


*Figure 7: Illustration of boosting procedure in SAS Enterprise Miner*

## 3.4 Random Forest

The random forest method proposed by Breiman (2001) has been widely used in the last decades in different areas thanks to its robust performance and efficiency over large data sets (Van den Poel & Lariviere, 2005 & 2004; Breiman, 2001).

There are many variations of random forests in the literature (Breiman, 2001) such as bagging of decision tree by Breiman (1996), random split selection (Dietterich, 1998), random feature selection (Ho, 1998), or most widely used random forest proposed by Breiman that incorporates random feature selection (2001). A general definition of all random forests is quoted below from Breiman (2001):

*"A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(x, \Theta_k), k = 1, \dots\}$ *where the* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $x$*."* (Breiman, 2001)

The definition summarizes the common characteristic of all kinds of random forests, which all train independent trees in parallel on random subsets of the original data set. Due to their simultaneous creation, the trees within a random forest are independent from one another and the formation of one tree does not depend on that of another tree. Random forests then aggregate equally the results from all the trees to produce the final ensemble result: the class that receives the most votes from the trees for a classification problem or the average of the tree's estimates for an estimation problem (Burez & Van den Poel, 2009). The described procedure is similar to that shown in Figure 7.

The aggregated results are proved mathematically to reduce the generalization error by Breiman (2001). The generalization error is negatively proportional to the number of trees grown within the forest, meaning the higher the number of trees grown, the lower the generalization error. Moreover, such generalization error is also claimed to prevent random forests from over-fitting when more trees are adding while curbing the generalization error. (Breiman, 2001)

Once the general definition of random forest is presented, the thesis now dissects the random forest model proposed by Breiman in 2001, which specifically combines bagging with random feature selection.

Let's assume a training data set $T$ with $n$ observations $(\boldsymbol{X}, Y)$, where $\boldsymbol{X}$ is a vector of independent variables or churn predictors $\boldsymbol{X} = (x_1, x_2, \dots, x_i)$ and $Y$ is a binary dependent variable or churn response.

Random forest by Breiman (2001) creates bootstrap training data sets $\boldsymbol{T_k}$ by random sampling with replacement, where $k = 1, 2, \dots K$. Each training set $T_k$ has $n'$ observations with $n' < n$. From these $T_k$ bootstrap samples, $k$ decision trees are run to get their predicted value of $Y$ denoted by $Y'_k$.

Since churn classification is a binary classification model, random forest then aggregates the votes of the predicted churn responses $Y'_k$ to find the class with the most votes as the final result for each of the observations in the model. The observations that are not included in the creation of the $k$ decision trees form out-of-bag data set. Random forests also runs the trees with these out-of-bag examples, aggregates the votes to get the out-of-bag result and estimates the generalization error (Breiman, 2001). Consequently, there are always three figures in the results of random forests for the three data sets:

- Training
- Validation
- Out-of-bag

Furthermore, a typical characteristic of the random forest by Breiman (2001) is the random feature selection, which means that only a sub-set of the independent variables are used to grow the trees. Breiman (2001) points out the trade-off between correlation and strength associated with the number of independent variables: a high number of input variables creates correlation but provides better strength. In his experiment to test the impact of the number of inputs in growing the forest, Breiman (2001) tests model performance with two values: one and $int(\log_2 M + 1)$ where M is the number of input variables. The formula returns the maximum integer that is less than $(\log_2 M + 1)$. Breiman (2001) observes that for small data sets of less than 3300 observations, there is no significant difference in the generalization error between the two values; therefore, only one randomly chosen input variable is surprisingly sufficient. However, different results are obtained from larger data sets of more than 10,000 observations when the generalization error decreases with the higher number of input variables (Breiman, 2001). Since some data sets for this thesis include at least 100,000 observations, the author employs the formula $int(\log_2 M + 1)$ for the number of input variables. Another parameter required in the random forests proposed by Breiman

(2001) is the number of trees to be grown. Since the generalization error decreases when the number of trees increases, it is recommended to employ a high number of trees (100 in Breiman's (2001) experiment). (Breiman, 2001)

## 3.5 Gradient Boosting

Like bagging, there are also various boosting techniques that have been employed in churn literature while this thesis selects gradient boosting proposed by Friedman (2002). This part of the thesis reviews briefly the method.

First of all, given the observations $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}$ is a vector of independent variables or churn predictors $\boldsymbol{X} = (x_1, x_2, \dots, x_i)$ and $Y$ is the binary dependent variable or churn response, the goal of gradient boosting is to find a function $F(\boldsymbol{X})$ that map the independent variables $\boldsymbol{X}$ to the churn responses $Y$ while minimizing the expected value of a specific loss function.

Gradient boosting approximates the function $F(\boldsymbol{X})$ by aggregating the estimates from $m$ iterations of a base learner, which is already defined in the section 3.3.3.1 Bagging:

$$h(\boldsymbol{X}; A)$$

- $h$ is a function of the independent variables $\boldsymbol{X}$. For gradient boosting, $h$ is a decision tree
- $\boldsymbol{A}$ is the vector of the estimated parameters of $\boldsymbol{X}$ in the function $h$

As shown in Figure 7, boosting develops the decision tree in sequence so that the next tree is formed based on the performance of the previous tree. Therefore, gradient boosting also follows the same logic. The base learner $h(\boldsymbol{X}; A)$, or decision tree, is trained sequentially, each of which obtain a weight $\beta_m, m = 0, 1, \dots, M$ in the final aggregation procedure of gradient boosting as shown below:

$$F(\boldsymbol{X}) = \sum_{m=0}^{M} \beta_m h(\boldsymbol{X}; A_m)$$

The gradient boosting model's output is a weighted average of the base learners' results. In contrast to bagging, boosting models do not execute these iterations of the base learner simultaneously but starts with a guess $F_0(x)$ and produces the next estimates by adding the estimated decision tree and its weight:

$$F_1(\boldsymbol{X}) = F_0(\boldsymbol{X}) + \beta_1 h(\boldsymbol{X}; A_1)$$

And the procedure continues with the subsequent iterations:

$$F_m(\boldsymbol{X}) = F_{m-1}(\boldsymbol{X}) + \beta_m h(\boldsymbol{X}; A_m)$$

The procedure stops when the pre-defined number of iterations is reached. As a result, the weights $\beta_m$ and the vector of parameters $\boldsymbol{A}_m$ for the decision trees should be solved in order to minimize a loss function. According to Friedman (2002), gradient boosting handles this possibly challenging optimization problem by fitting the base learner decision tree $h(\boldsymbol{X}; A_m)$ to the residual of the estimate from the previous iteration $F_{m-1}$ and solving for $A_m$ by differentiating the loss function. Then the two parameter optimization problem above becomes one parameter optimization problem for $\beta_m$. More detailed mathematical explanation is presented in Friedman (2002).

# 4 Data

This thesis now describes the data of housing loan customers from a large Nordic bank to be used for the churn classification models. Before building the models, it is important to prepare the training data, specifically collecting and processing the dependent variable or the churn responses and the independent variables or churn predictors within the selected time window of the study is explained. Consequently, this chapter first describes the housing loan churn problem to define the parameters for the churn responses, then identifies the time window of analysis, and finally explains the preparation of the churn predictors. There are several aspects to be considered when calculating the independent variables such as data pre-processing, and variable selection. These steps are also briefly discussed in this thesis.

## 4.1 Calculating Churn Responses – Dependent Variable

In a typical churn classification study, a set of criteria must be defined to classify housing loan customers as churners. Although churn in general refers to the customers that switch from one service provider to another, the detailed criteria is important to calculate the churn response or the dependent variable of the model. Churn definition depends on the business problem that the churn model is targeted to solve, however there are several common aspects that need to be considered when defining churn. In retail banking context, one customer can possess a wide variety of products in their portfolio such as accounts, loans and investment products. When the customer switches all of her investment portfolio to a competitor's

investment funds but still keep her loans and accounts active in the current bank, whether such customer is identified as churn depends on the business problem (Glady, et al., 2009). Moreover, churn definition is mainly based on the status of customer relationship such as active customer, involuntary churn and voluntary churn. It is difficult for banks to capture such status because customers are free to switch among the competitors at a minimal to no cost and the treatment of customer data is heavily regulated. For example, banks cannot close customers' accounts even though the customers are no longer transacting with the banks without customers' consent (Bank's employee 1, 2016). Therefore banks with large customer base are unable to capture their customers' real statuses (Prasad & Madhavi, 2012).

An important question in churn definition is whether to apply a fixed business rule on all customers or to track the variation of a particular customer behavior. It is more common in churn literature to apply a threshold dictated by a business rule to a particular behavior for churn definition. For example, customers holding assets below a specific amount or purchasing at a lower frequency than a specified rate are considered churners (Glady, et al., 2009). However, such business rules are argued to be too robust to capture the variation in customer behavior over time, a point well illustrated by the following extreme example. A business rule assumes that a customer carries out six or less transactions per year is considered churn. Therefore, a customer who has made six transactions per year previously but only five this year is considered a churn event. However, another customer who has made 100 transactions annually in the previous year but only six this year is not considered a churn case (Glady, et al., 2009). This simple example demonstrates the failure to detect churn with fixed business rules. Moreover, the interest of churn study is to be able to identify customers to retain, therefore churn definition should help to identify customers with early sign of churning, rather than churned (Gur Ali & Ariturk, 2014; Glady, et al., 2009). Qi et al. (2008) confirms that churners do behave unusually before their time of churning; hence their behaviors will signal some early warnings (Qi, et al., 2008). As a result, tracking the evolution of customer behavior is recommended (Nie, et al., 2011). For instance, it is suggested that a threshold should be defined on the slope of product usage rather than its face value to define churn (Glady, et al., 2009).

This thesis focuses on churn housing loan agreement, a contractual product with an end date. Naturally, customers who have paid off their loans when their mortgage contracts are matured are not considered churn. Involuntary churn might refer to customers who sell their houses during the loan period, hence stop the loan (Koh & Chan, 2002); therefore those

customers are not considered churn either because they do not switch their housing loans to other competitors. As a result, voluntary churn is the case of interest. Voluntary churn refers to customers that pro-actively terminate their active housing loan agreements with this bank long before the maturity date while there is still a high remaining principle of the loan and take the loan from elsewhere. However, such definition might be too broad because it might also include cases of early pay-off of the remaining loan due to a sudden increase in the customer's wealth from, for example, winning a lottery or inheritance. Unfortunately, the bank does not have a clear indicator for customers that have switched their housing loans to a competitor nor any recorded explanation when a housing loan agreement is stopped long before the maturity date. Consequently, the author has to apply the business rule for housing loan applications, which particularly requires customers to direct their salaries to the bank when they open the housing loan agreements (Bank's employees, 2016). Therefore, customers who have decided to continue their housing loan contracts with a competitor will experience a significant decrease in their salary inflows within this bank. This rule helps eliminate the cases of possible early housing loan repayment because in such cases, if everything else is assumed to remain the same, the customers are still with the bank and there should be no considerable drop in salaries. Another reason to resort to this business rule is that the early signs of churn in customer behavior cannot be captured in the housing loan volume, because loan repayment is mostly automatic nowadays and customers who are going to churn do not necessarily change their loan repayment behavior until the point of churn (Bank's Employee, 2016).

Figure 8 illustrates the above outlined definition for churn responses employed in this thesis for the Finnish customers that have active housing loan agreements for at least 10 months with the bank. The green line in the upper chart represents the balance for the housing loan volume with the regular monthly decrease due to loan repayment until the drop in month t, when the churn actually happens. The red line in the lower chart represents the balance of the salary account. In the pre-churn period, the salary account sees the normal ebbs and flows within a monthly cycle. On every pay-day of the months, the incoming salaries increase the account balance, creating the peaks of the red line as marked in the chart. After the pay-days, loan repayments and consumption decrease the balance. However, in the post-churn period, the salary account is expected to become flat because salaries have been transferred to another bank where customers have switched their housing loan agreements to.

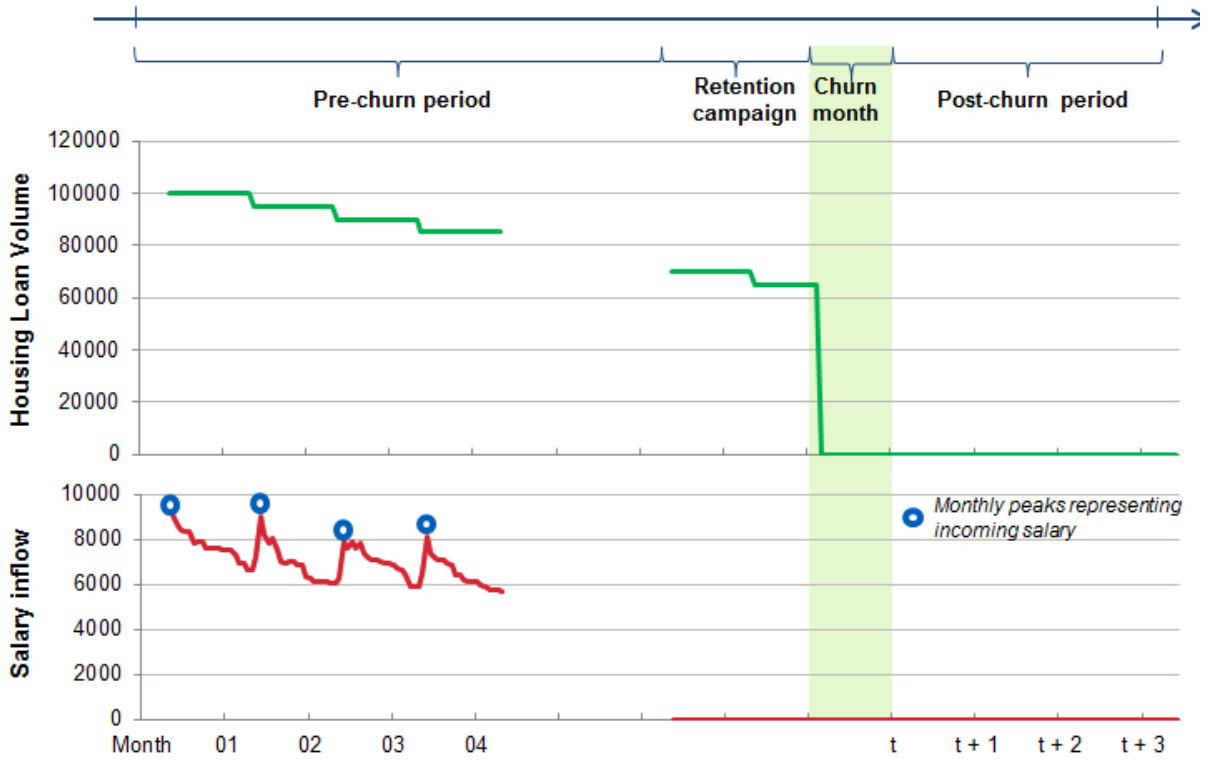


*Figure 8: Illustration of the churn definition employed in this thesis*

It is also observed from Figure 8 that the salary account has become flat even before the churn period, indicating the possibility that salaries can be directed to the other bank even before the official churn month when the housing loan agreement is terminated. The monthly loan repayment pattern remains until the churn month, indicating the automatic repayment scheme as discussed above.

It is worth noting that the monetary scale in Figure 6 only illustrates one case of the above outlined definition for a churned customer. The euro terms are not applicable to all the churned customers due to the difference in customers' salaries and the needed volumes for a housing loan.

In conclusion, two main thresholds are defined in this thesis to define churn.

1. *Remaining housing loan principle*

A threshold of the high remaining housing loan principle is defined to avoid possible pay-off within the next period. In this thesis, this threshold is set to 30000 euro, meaning the churners should have housing loan volumes of more than 30000 euro in the last month of the observation period but have no housing loan agreement in the churn month.

2. *Salary change*

A threshold is defined for the salary change between the pre-churn and post-churn periods. This second threshold is set to 95% for this thesis, meaning the churners should experience a salary drop in the post-churn period of at least 95% of the pre-churn period's average salary. However, in some cases, considerable drops in salaries might not be caused by churn but by unemployment or unused salary accounts with balances of even a few euros, creating misleading cases. As a result, the target group selected for this thesis includes housing loan customers with a monthly salary of at least 2000 euro.

Consequently, customers who satisfy the two criteria for housing loan balance and salary inflow are classified as churners. However, these thresholds are coded to be flexible for changes in this thesis so that they can always be adjusted to fit the changing business contexts.

## 4.2 Time Window of Analysis

It is important to identify the time window of analysis, from which the data is collected and analyzed. Figure 9 illustrates the time window of analysis for the single period training data. The observation period to calculate the independent variables lasts for three months from August 2015 to October 2015. The performance period to calculate the churn responses takes three months from January to March 2016 in December 2015. The month November 2015 marked with "Retention" is indeed the time set aside for customer retention activities.

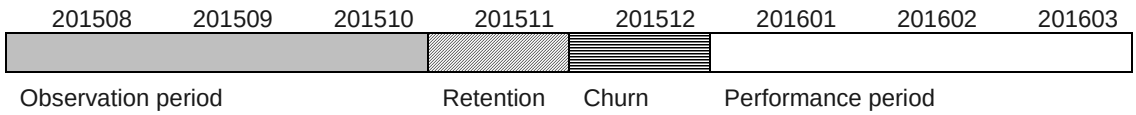


*Figure 9: Time window of analysis for single period training data of churn in December 2015*

The timeline provides the empirical context for the research questions above. The models are set to predict the housing loan customers that are going to churn in two months from the time running the model.

As such, the timeline to construct the multi period training data is illustrated in Figure 10. By definition, multi-period training data includes multiple single period training data sets, in this case the churn responses of the multi-period training data are comprised of five sets of churn responses from August to December 2015. The observation periods, retention periods, churn measurement and performance periods are moved backward correspondingly for each single period training data set and marked differently in the figure.

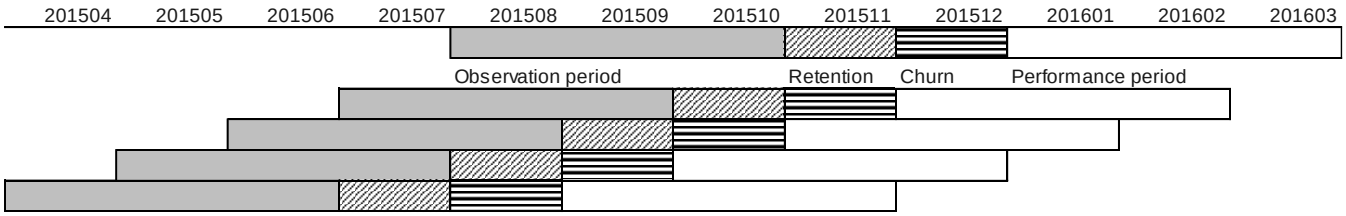


*Figure 10: Time window of analysis for multi-period training data*

The churn rates for the churn periods from August to December 2015 are illustrated in Figure 11. This aligns with the significant rarity issue of churn prediction in the banking context.

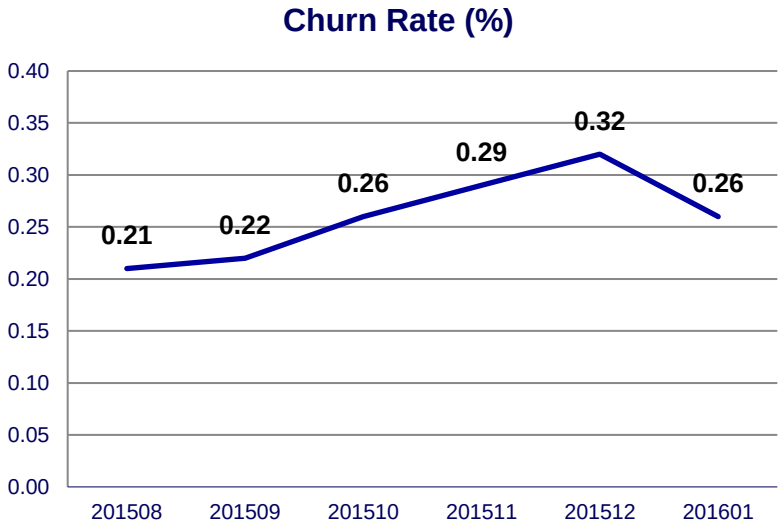


*Figure 11: Housing loan churn rates within the time window of anaylsis*

As a result, two training data sets are created for model construction: the latest single period training data includes active housing loan customers in October 2015 while the multi-period training data consists of five single period training data sets with active housing loan customers from June to October 2015. Single period training data has 111887 observations with a churn rate of 0.32%, corresponding to the churn period of December 2015. On the other hand, multi-period training data has 622032 observations with the churn rate of 0.26%, which is the average churn rate during the period from August to December 2015. More details in the construction of these two data sets are explained in section 5.1.

## 4.3 Preparing the Independent Variables

### 4.3.1 Employed Churn Predictors

The predictors employed in this thesis include information from all the four main groups as recommended in the literature: demography, customer behavior, characteristics of customer

relationship and macro-environmental factors. The churn predictors are summarized in Table 3.

The first group includes the widely employed demographic variables in churn literature such as age, gender and salary. As mentioned before, most of the bank's products do not require the customers to reveal the number of their family members; however the bank can capture this information through the family members that are customers of the bank (Bank's Employee, 2016). The population of customers' residential areas is also obtained as it is argued to correlate with housing activities. For example, when the housing loan market gets busy, the effect is seen most obviously in big cities such as Helsinki and Turku in Finland while less so in smaller cities.

*Table 3: List of churn predictors employed in this thesis*

| Groups | Information provided |
|---|---|
| **Demographic**<br>(5 variables) | - Age<br>- Gender<br>- Salary<br>- Number of family members that transact with the bank<br>- Residential population |
| **Customer behavior**<br>(24 variables) | - Housing loan related variables: count, volume, ownership<br>- Credit and debit transactions related variables: count, volume<br>- Other loans related variables: count, volume<br>- Deposits related information: volume |
| **Characteristics of customer relationship with the bank**<br>(7 variables) | - Tenure with the bank<br>- Tenure as a housing loan customer<br>- Total contacts and last contact days with the bank<br>- Customer satisfaction indicators with the bank's contact center and first time resolution services |
| **Macro-environmental factors**<br>(5 variables) | - OMXH25<br>- Inflation rate<br>- Consumer confidence index<br>- EUR USD exchange rate<br>- Unemployment rate |

The second group of customer behavior variables consists of the most variables: 24 variables, aligned with the traditional approach in churn prediction to focus on customer behavior. These variables contain information like number of products, volumes and ownership for customers' portfolio with the banks, including deposits, housing loans and other loans, and credit and debit transactions. These variables are coded to get the most recent values and their changes compared to the previous month for each single period training data sets.

The third group is comprised of 7 variables describing the characteristics of customer relationship with the bank. Seniority of customers is an important indicator in studying churn; as explained previously, it is worth retaining more tenure customers because the longer the

customers stay, the larger portfolio they would grow with the bank (Reichheld & Kenny, 1990). As a result, two types of seniority are included in this study: tenure with the bank since the customers initiate their first transaction and tenure as a housing loan customer since the first housing loans are started with the bank. Moreover, information about customers' contacts with the bank is also captured by the total number contacts during the customer relationship and number of days since last contact. These contacts are compiled from all channels such as letter, telephone, internet communication channel, and physical or online meetings. Contacts can mean either possibilities of selling new products or negative feedback and on-going issues that need actions from the bank. Frequent contacts to the bank before the churn period might indicate the latter case when customers are dissatisfied about the service in the bank; and hence churn away (Bank's Employee, 2016). Last but not least, the satisfaction indicators are collected through survey data on a monthly basis in 2015 about the bank's contact center and first time resolution for the retail banking services. Changes in these indicators might reflect the churn behavior in housing loan service. Naturally, the monthly customer satisfaction indicators are included only in the multi-period training data sets.

Finally, macro-environment predictors are also captured monthly for the multi-period training data set. These external variables are collected based on the recommendation from the literature, including:

- The monthly open rate of OMXH25, the Helsinki stock index compiling the 25 most traded stocks (Nordnet, 2017)
- Monthly nominal inflation rate in Finland (Triami Media BV, 2017)
- Monthly consumer confidence index in Finland (OECD, 2017)
- Monthly exchange rate between euro and US dollar (X-rates, 2017)
- Monthly unemployment rate in Finland (Statistics Finland, 2017)

Although all of them reflect the economic health, unemployment rate and consumer confidence index relate directly to the possibility of buying new houses or changing houses, the opportunity for customers to research for better housing loan agreements elsewhere if they are not satisfied with the current agreements at this bank (Bank's Employee, 2016).

### 4.3.2 Data Pre-processing

The raw data from the Analytic CRM team of the bank for household customers is categorized into multiple levels such as customer level, agreement level, or service level in

various Tables stored in Hadoop. SQL codes are employed in SAS Enterprise Guide 7.1 to extract and combine the needed data on the customer level. The final data sets are then sent to SAS Enterprise Miner 14.1 for modelling.

In order to prepare the data for modelling, data pre-processing steps such as handling missing value or variable transformation are crucial (Gur Ali & Ariturk, 2014). First of all, most research sets a threshold to exclude variables with too many missing values, which might hinder model building (Lemmens & Croux, 2006). However, for those inputs with some degrees of missing values, it is important to handle them. Two main methods to handle missing values: synthetic distribution methods and estimation. In synthetic distribution methods, a fixed number is used to replace the missing value so that it impacts the least the assumed distribution of the input. This method is also referred to as one-size-fits-all approach because it handles all missing values in the same manner. The most popular fixed value used in the literature is the mean for continuous value and the most frequent category for categorical value. On the other hand, the estimation methods consider the input with missing values as the target of a predictive model, which is trained on other inputs to predict the missing value. However, missing values are not only caused by non-disclosure or non-matching of available data. In some cases, missing values mean non-applicable. For example, a customer with a missing value for housing loan volume indeed does not have a housing loan with the bank. Moreover, missing values might even contribute to the explanation of the target variable. For example, in a response prediction problem, customers with a missing value for the variable "number of responded marketing emails" might indicate that they have not answered to marketing emails and that they will continue to behave in the same manner. Therefore, binary dummy variables are usually created to indicate whether values of an input are missing or not to study the contribution of missing values to the target. (Jinbo, et al., 2007; Lemmens & Croux, 2006)

In this thesis, there are four variables with missing values: two categorical variables for residential areas and their population and two interval variables for deposit amount and change in deposit amount compared to the previous month. For the categorical variables, an "Unknown" value is imputed to the missing ones. For the continuous variables, since missing values regarding deposit amount usually indicate that the customers no longer hold deposit accounts with the bank. As a result, missing value imputation is not employed in this thesis.

Besides handling missing data, variable transformation for continuous variables is also performed to reduce skewness in the distribution of the raw data. This thesis employs log transformation and standardization.

### 4.3.3 Variable Selection

Given the abundant data that banks constantly collect, it is most convenient to include all the data available during data collection into modelling. However, such behavior might cause input redundancy, which in turn causes multicollinearity, producing biased parameters for regression models (Burez & Van den Poel, 2009). Therefore, the author performs variable selection before running logistic regression models.

The most frequently used variable selection methods are stepwise with both forward and backward selection (Ballings & Van den Poel, 2012; Nie, et al., 2011; Burez & Van den Poel, 2009), CHAID - Chi Square Automatic Interaction Detector (Lu, et al., 2014), principal component analysis (Lemmens & Croux, 2006) and AUC or ROC (Qi, et al., 2008).

In this thesis, the following methods are employed in SAS Enterprise Miner before running the logistic regression models: a decision tree with CHAID method, step wise selection, and a variable selection node using R square procedure. In the decision tree variable selection method, as the name suggested, a decision tree model is run with all the input variables and the variables that are chosen for the splits in the tree are selected (Sarma, 2007). Step wise variable selection starts by running a regression model with all the variables. Then, the variable with the highest p-value that exceeds a pre-defined significance level $\alpha$ is eliminated. A new regression model is produced with the remaining variables; and the elimination happens similarly based on the p-value. The procedure continues until all the remaining variables have a lower p-value than the chosen significance level in the revised regression model (Burez & Van den Poel, 2009). Finally, R square variable selection is a two-step procedure. First, a simple regression model is run for each input variable to obtain the R-square. The input variables that produce higher R square values than a predefined R square minimum are selected for the second step. Then, a forward selection procedure starts with the input variable that has the highest correlation coefficient by running a regression model with that variable, and continues with the input variable with the second highest correlation coefficient by running a regression model with all the selected inputs, and so on. The procedure stops until the incremental increase in the R square value reaches the pre-defined minimum value. (Sarma, 2007)

# 5 Results

This chapter first discusses the factors that are employed to build the competing models: training data construction approach, sampling and churn classification algorithms. Then, the performances of the constructed models are compared using the selected evaluation criteria, namely misclassification rate, ROC index and top decile lift with the purpose to answer the research questions.

## 5.1 Building Competing Models

In order to build the competing models, three main factors are taken into consideration: training data, sampling and churn prediction algorithms as detailed in Table 4. As discussed previously in the literature review, two training data construction approaches and four churn classification algorithms are employed as shown in the table.

*Table 4: Factors to build competing models*

| Factors | Details |
|---|---|
| **Training data construction approach** | • Multi-period training data (with macro-economic data)<br>• Single period training data |
| **Sampling** | • 10 balanced samples with equal share of churners and non-churners<br>• $n_{multi-period\ training\ data} = 2564$<br>• $n_{single\ period\ training\ data} = 568$ |
| **Algorithm** | • Decision Tree<br>• Logistic Regression<br>• Random Forest<br>• Gradient Boosting |

As the thesis selected only one option for sampling as presented in Table 4, sampling methods do not constitute a section in the literature review. However, it is worth discussing sampling here to justify for the sampling method selected for this thesis. First of all, sampling is an important step that has been widely exploited in churn literature to tackle the rarity issue. The most basic sampling techniques are under-sampling and over-sampling. According to Weiss (2004), under-sampling removes observations from the majority class, or non-churners in this context while over-sampling repeatedly select the minority class, or the churners. However, over-sampling is criticized to possibly cause overfitting while introducing no new data and burdening the computation of large data set (Gur Ali & Ariturk, 2014). In the most basic form of under-sampling, the proportion between churners and non-churners is also worth a discussion. Multiple researches opt for balance sampling to include an equal number of churners and non-churners in the training data rather than the real

imbalanced proportion to improve the model's learning (Nie, et al., 2011; Burez & Van den Poel, 2009; Lemmens & Croux, 2006). Such approach is claimed to reduce the misclassification rate because the model is not as overwhelmed with non-churners' characteristics as in the case when the data set uses the real proportion. For example, Nie et al. (2011) compares models built from samples comprised of different proportion between credit card churners and non-churners with the respective ratio of 1:1, 1:2 and 1:4 and observes that the model with balanced sampling performs the best in most of the iterations while the models with 1:4 ratio performs the worst (Nie, et al., 2011). Although there exists more advanced under-sampling methods such as SMOTE, SHRINK, SMOTEboost (Gur Ali & Ariturk, 2014; Chen, et al., 2004), Gur Ali & Ariturk (2014) do not find significant improvement when using these complex sampling methods compared to the basic under-sampling with multi-period training data (Gur Ali & Ariturk, 2014).

In conclusion, this thesis employs balance sampling with equal proportion of churners and non-churners for all the training data set. Sampling is done in SAS Enterprise Guide to build the training data sets.

*Table 5: Sampling the training data sets*

| | Training data construction | Churn month | Original size | Churn rate | Balanced sample size |
|---|---|---|---|---|---|
| 1 | Single period training data | August | 126 038 | 0.21 % | |
| 2 | Single period training data | September | 132 406 | 0.22 % | |
| 3 | Single period training data | October | 132 052 | 0.26 % | |
| 4 | Single period training data | November | 119 649 | 0.29 % | |
| **5** | **Single period training data** | **December** | **111 887** | **0.32 %** | **568** |
| **6** | **Mutl-period training data** | **December** | **622 032** | **0.26 %** | 2 564 |

Table 5 summarizes the training data sets and their information. The last two bolded training data sets numbered 5 and 6 are selected as the samples for single period training data and multi-period training data approaches correspondingly. Specifically, five single period training data sets at the churn months from August to December 2015 are combined to form the multi-period training data set with 622,032 observations as the total of observations from the five single period training data sets, while the latest single period training data set of December 2015 as the churn month with 111,887 observations is used to run the models for the single period training data approach. As a result, the single period training data set is part of the multi-period training data set. The churn rate of the single period training data set is 0.32% while the churn rate for multi-period training data set is calculated as the average

churn rate of the five single period training data sets at 0.26%. From the multi-period training data and latest single period training data sets, ten samples are created with balanced sampling of equal proportions for churners and non-churners. For each sample, 80% of churn data from the original data set is recruited, giving the sample sizes of 568 observations for the single period training data sample and 2,564 observations for the multi-period training data sample. Moreover, this thesis only considers simple random sampling of the original data set without experimenting more advanced sampling methods in order to study the effect of bootstrap sampling and weighted sampling in bagging and boosting.

The samples are run with the model construction diagram as shown in Figure 12 in SAS Enterprise Miner. After the metadata node, each sample goes through data partition nodes to split the data into 70% for training and 30% for validation. Since the random forests node is available only in high performance nodes (starting with HP) in SAS Enterprise Miner, a high performance data partition node is called prior to the random forests model. The partitioned data is used to immediately generate a decision tree while for logistic regression models, variable transformation and variable selection steps are performed as recommended in the literature prior to running the models. As a result, three logistic regression models are built: model (1) right after partition with step wise selection method, model (2) after R square variable selection and model (3) after a decision tree node as marked in Figure 10. Since hybrid variable selection methods are also employed in literature such that Chi-square variable selection is performed to choose the most important independent variables for boosting models (Lu, et al., 2014), this thesis also experiments such approach for the two ensemble methods. Two gradient boosting models are run, model (1) after partitioning and model (2) after a variable selection node employing linear regression with forward selection. Similarly, two random forests models are run, model (1) after partitioning and model (2) after R square variable selection.

In total, 160 models are built in SAS Enterprise Miner.

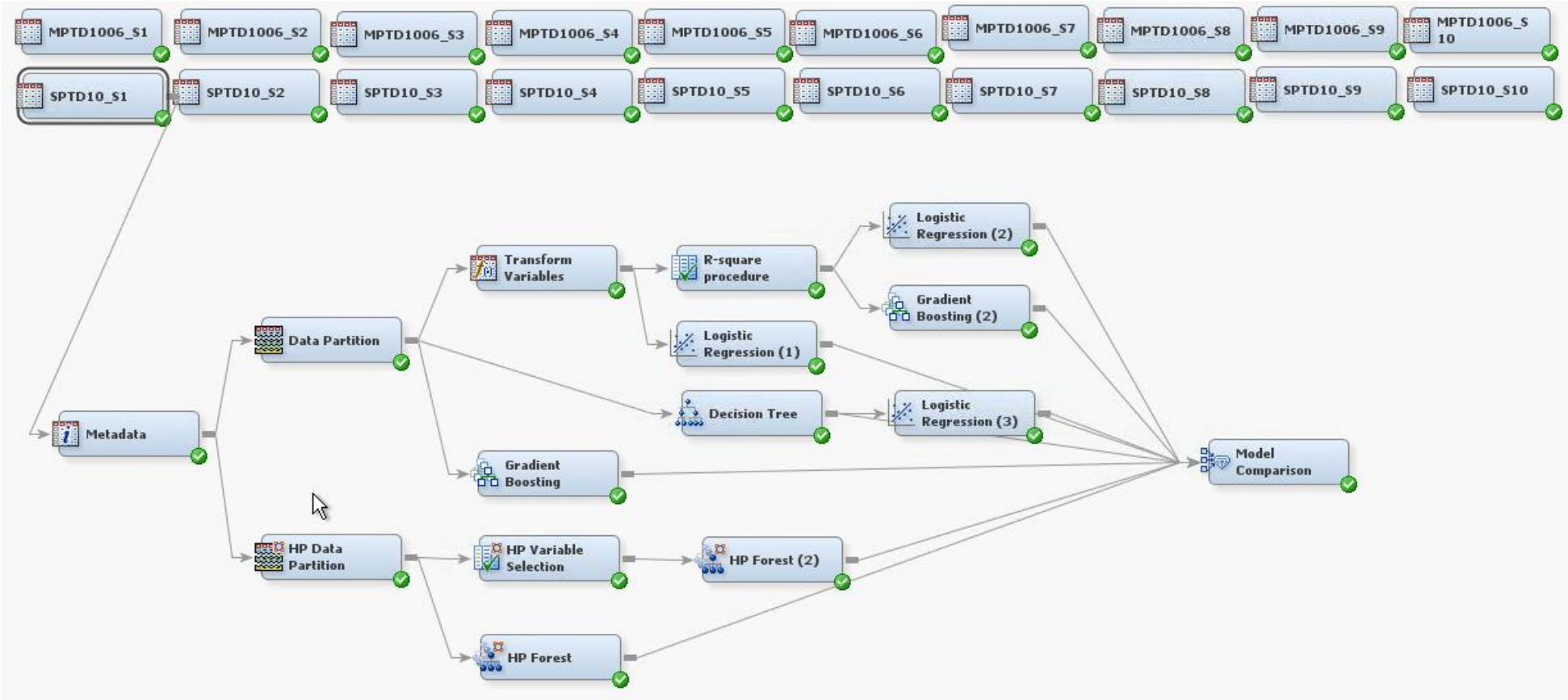


*Figure 12: Modelling diagram in SAS Enterprise Miner 7.1*

## 5.2 Answering the Research Questions

The thesis now answers the research questions that have been outlined earlier in section 3.2 using the selected evaluation criteria: misclassification rate, ROC index and top decile lift. For each type of models, the estimates of these criteria on the validation data sets are aggregated over the 10 samples for each type of models created from the combination of the factors: training data construction approaches and churn classification algorithms. The aggregated estimates of the criteria are then compared to answer the research question. Those aggregated estimates are presented in the form of line graph of the means with the error bars showing the standard deviation for both sides.

### 5.2.1 Question 1: Multi-period Training Data versus Single Period Training Data

In order to answer this question, logistic regression and decision tree are examined with both multi-period training data and single period training data. Figures 13, 14 and 15 present the misclassification rate, ROC index and top decile lift for these models. It can be observed that, multi-period training data improves the performance of these churn classifiers consistently and considerably. The error bars are reasonably short, showing the consistent results over the ten iterations. This result aligns with the expectation that models using the multi-period training data approach perform better than those with single period training data.

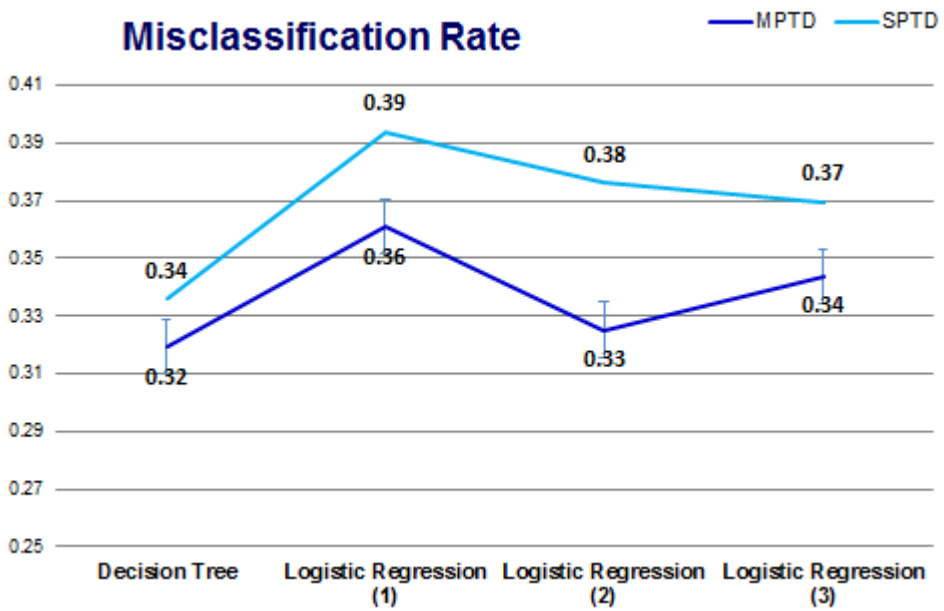


*Figure 13: Misclassification rate comparison between multi-period training data and single period training data for models with logistic regression and decision tree*

It can also be observed that for both approaches to construct the training data, all the average misclassification rates are much lower than 0.5; all ROC indices are higher than 0.5

and all top decile lifts are higher than 1.0. This result shows that balanced sampling of equal numbers of churners and non-churners in the training data set helps mitigate the rarity issue to produce decent models. On the contrary, the first preliminary runs with the training data using the real churn ratio, no models could be achieved due to the little presence of churn events.

Furthermore the MPTD lines in all three figures have better values than the SPTD lines, representing the potential that the models could further improve by employing multi-period training data proposed by Gur Ali & Ariturk (2014). Models with logistic regression see more significant increase in performance than those with decision tree. Misclassification rate and top decile lift are improved the most for logistic regression model 2 while ROC indices increase the most for logistic regression models 1 and 2.

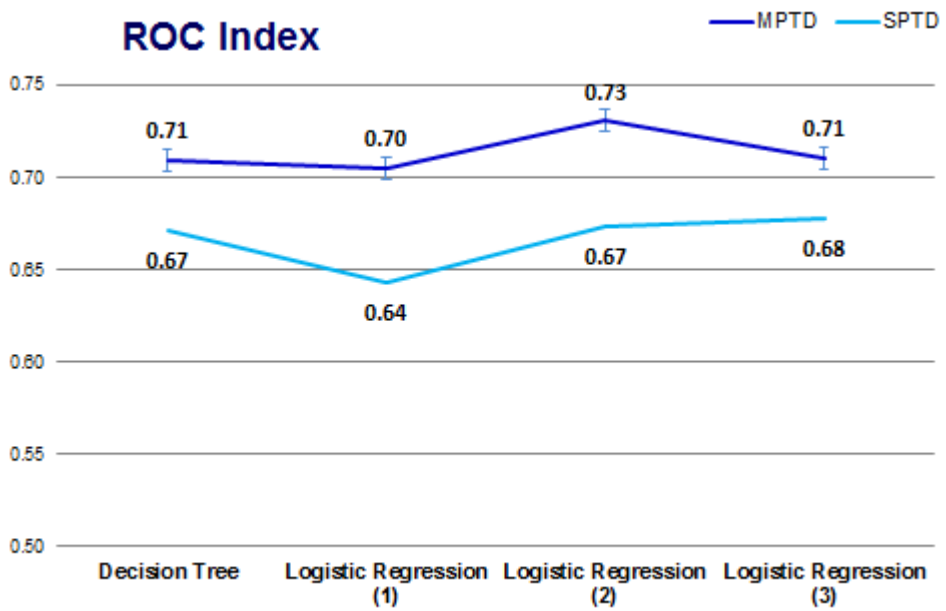


*Figure 14: ROC index comparison between multi-period training data and single period training data with models with logistic regression and decision tree*

It is worth noting the different performances among the logistic regression models with single period training data due to the ways they are created. As shown in the SAS Enterprise Miner diagram in Figure 12, logistic regression model 1 is run with step-wise selection immediately after data partition while logistic regression model 2 is run after variable transformation and R-square variable selection. It is expected that data pre-processing steps like variable transformation and variable selection are necessary to improve the performance of parameterized algorithms like logistic regression. As a result, among the logistic

regression models with single period training data, model 2 performs best regarding the three evaluation criteria.

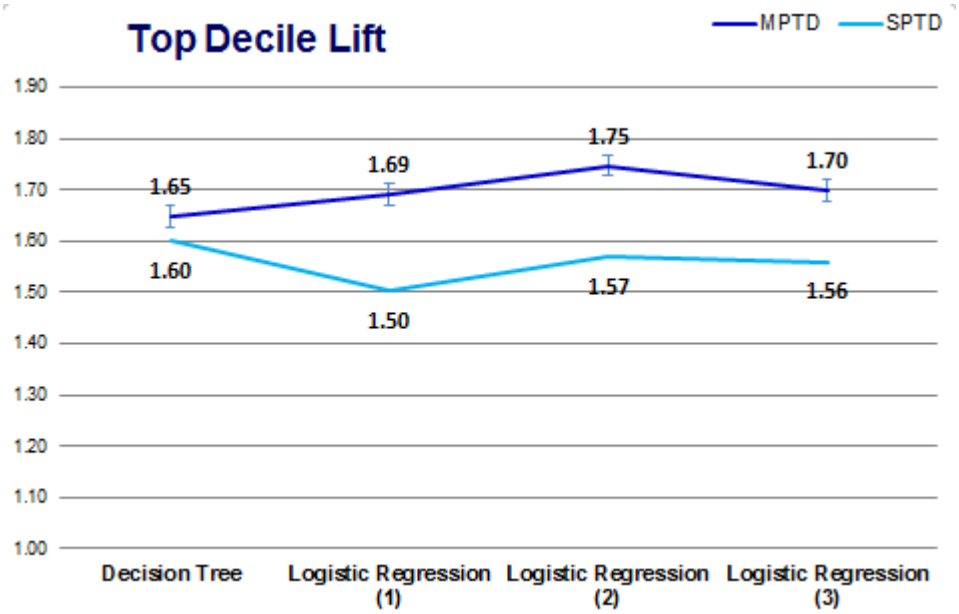


*Figure 15: Top decile lift comparison between multi-period training data and single period training data for models with logistic regression and decision tree*

An interesting result here is that decision tree is among the best models as seen from the figures compared to the logistic regression models.

### 5.2.2 Question 2: Random Forest and Gradient Boosting versus Logistic Regression and Decision Tree

The second comparison focuses on the employment of random forest and gradient boosting while using only the single period training data approach. The misclassification rates, ROC indices and top decile lifts in figures 16, 17 and 18 show that the models with random forest and gradient boosting consistently perform better than logistic regression and decision tree.

The first observation from all the figures is that the models on right hand side with random forest and gradient boosting indeed improve all the estimated evaluation criteria, i.e. decreasing the misclassification rate and increasing the ROC index and the top decile lift.

Since both random forest and gradient boosting are the ensemble methods to improve the base learner decision tree, a direct comparison between decision tree and its more advanced algorithms is necessary. As seen from all the figures, the selected ensemble methods improve significantly the performance of a single decision tree while the single decision tree even performs better than the three logistic regression models in terms of misclassification rate and top decile lift.

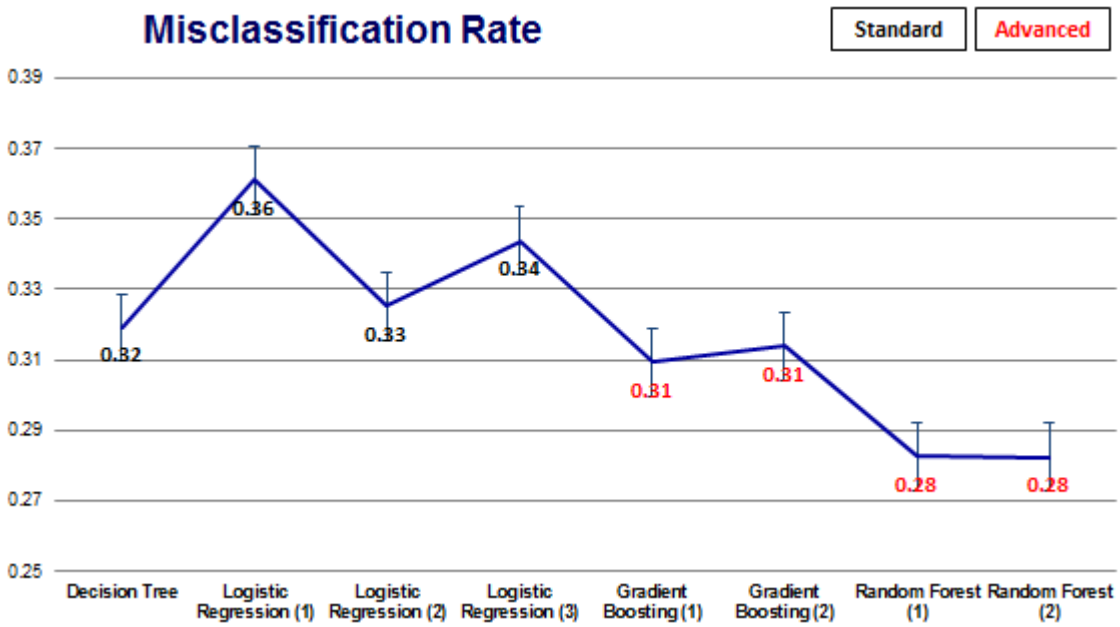


*Figure 16: Misclassification rate comparison between random forest & gradient boosting and logistic regression & decision tree for models with single period training data*

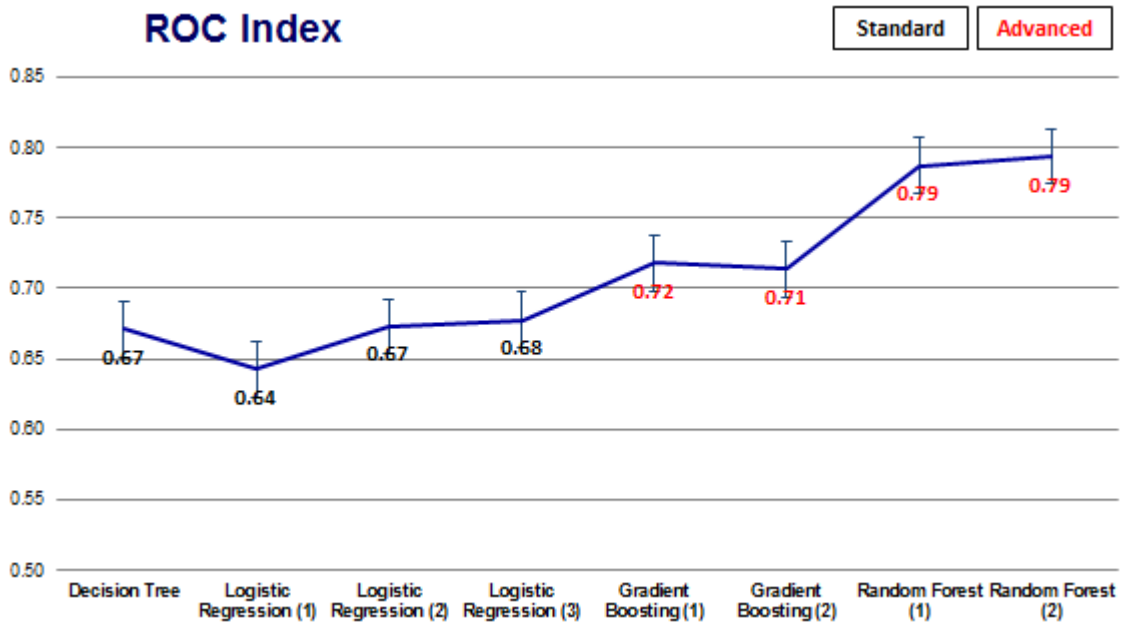


*Figure 17: ROC index comparison between random forest & gradient boosting and logistic regression & decision tree for models with single period training data*

Moreover, random forest model 2 after an R-square variable selection node consistently outperforms the other models with ensemble methods.

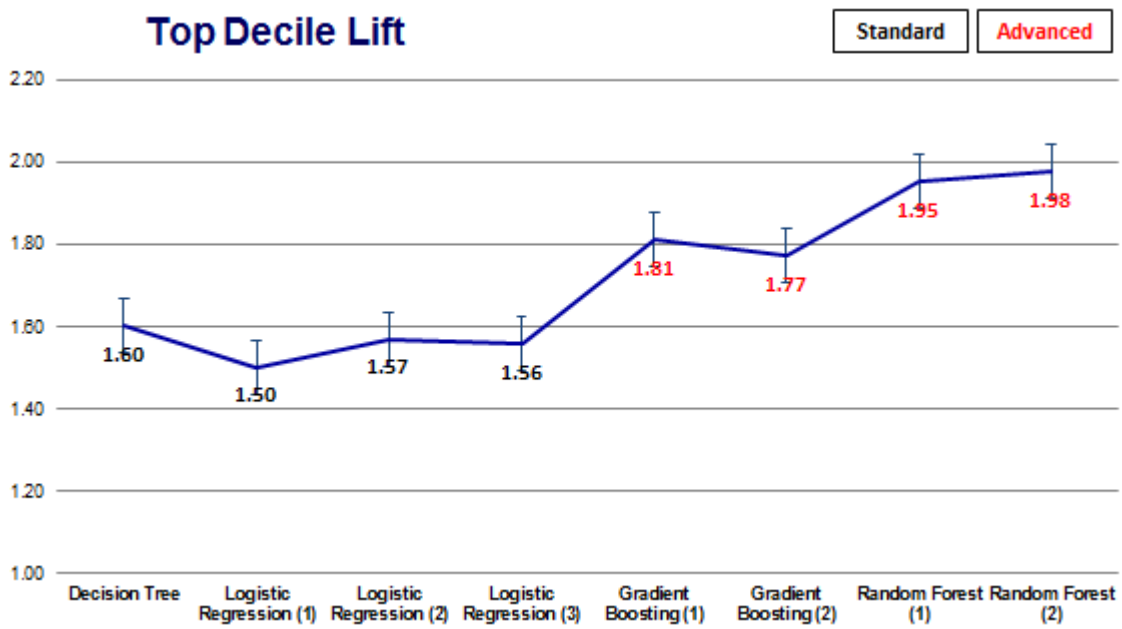


*Figure 18: Top decile lift comparison between random forest & gradient boosting and logistic regression & decision tree for models with single period training data*

In order to understand how the selected ensemble methods outperform logistic regression and decision tree, the thesis examines the results from some models with random forest and gradient boosting.

The result of a random forest model is analysed first. Random forest model is produced by the HP (high performance) random forests node in SAS Enterprise Miner. As mentioned above, the random forest proposed by Breiman (2001) recommends 100 for the number of trees and $int(\log_2 M + 1)$ for the number of input variables or 6 variables in this case (42 input variables in the original training data set).

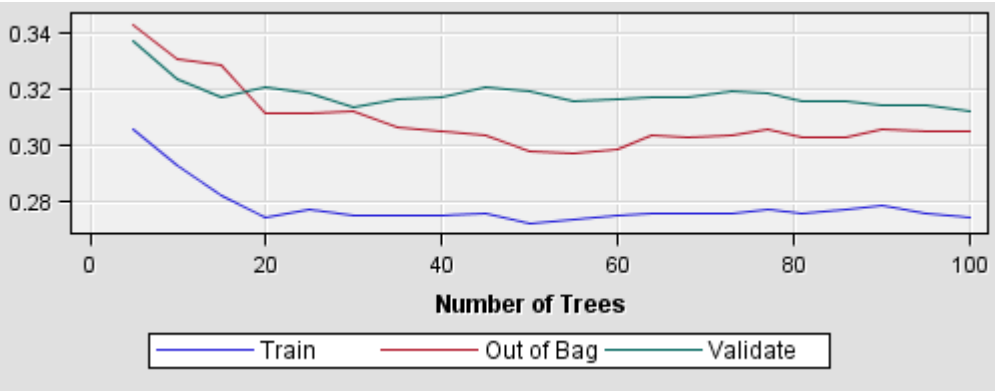


*Figure 19: Misclassification rate versus the number of trees for a random forest model*

Figure 19 shows the misclassification rate against the number of trees for the training, validation and the out-of-bag data sets (see section 3.3.4 Random forest for the explanation of "out-of-bag" data sets). It can be observed that all the misclassification rates significantly

decrease after 20 trees and the drop continues for the out-of-bag data set until about 50 trees. Moreover, it is expected that the misclassification rate in the training data set is considerably lower than those in the validation and out-of-bag rates because the training data is used to grow the trees and overfitting is possible. That also confirms the need to have at least validation and out-of-bag data sets to objectively assess the model performance.

Similarly, the results of gradient boosting model also generate the graphs shown in Figure 20 to plot the misclassification rate against the iterations. The decrease in the error rate can be seen most considerably within the first 10 iterations. An interesting fact in this gradient boosting's result is that the misclassification rate in the validation data set is much lower than that of the training set after the first 10 iterations. The expectation is usually the other way around as the training data is used to create the model. However, the result of average squared error in all 10 samples show the normal pattern that the error of the models run with training data is much lower than those run with validation data.

The blue vertical bar at iteration 49 in Figure 20 shows the optimal model where the minimum misclassification rate is found in the training data. However, for validation data, we can see that the optimal error rate is already discovered after iteration 34 at the rate of 0.29.

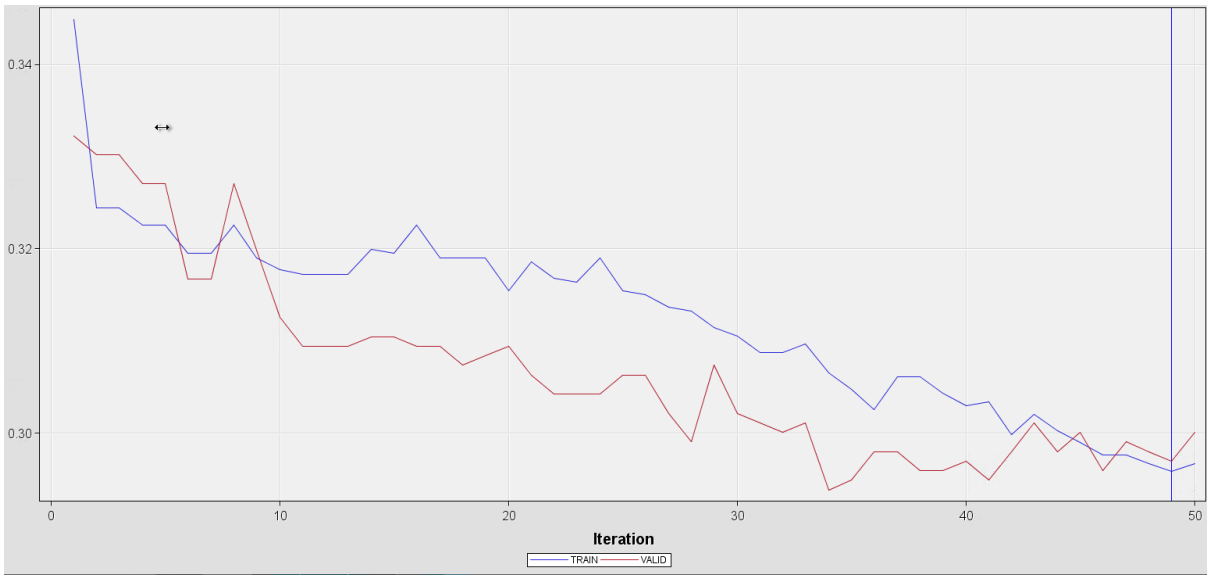


*Figure 20: Misclassification rate against the iterations for a gradient boosting model*

Consequently, iterative training of decision trees in random forest and gradient boosting shows compelling improvement in misclassification rate of the model rather than building only one tree. It also provides a convincing reason to grow multiple trees with better generalization of the model rather than fine tuning for the single best decision tree.

In conclusion, it can be confirmed that both multi-period training data and ensemble methods actually improve churn classification performance compared to their counterparts in this housing loan context.

### 5.2.3 Question 3: Combining Random Forest and Gradient Boosting with Multi-period Training Data

In order to answer the third research question, all models are considered in the comparison. The results are shown in Figures 21, 22 and 23.

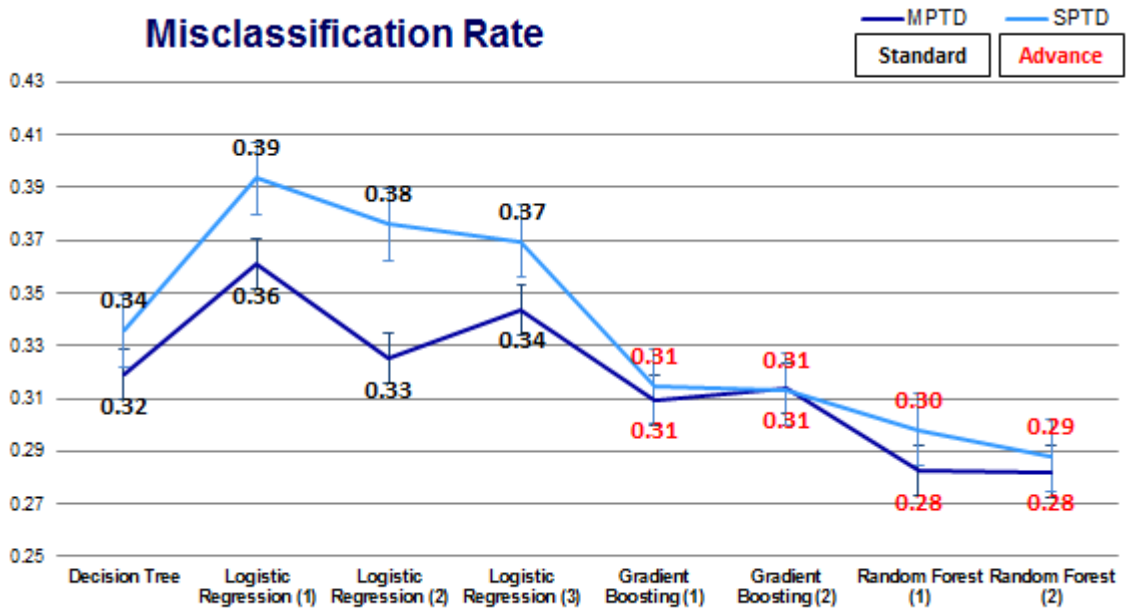


*Figure 21: Misclassification rate comparison of the models with random forest & gradient boosting and multi-period training data against the other models*

Regarding the models with random forest and gradient boosting, it can be seen from the figures that multi-period training data improves the misclassification rates and ROC indices for most of the models compared to single period training data.

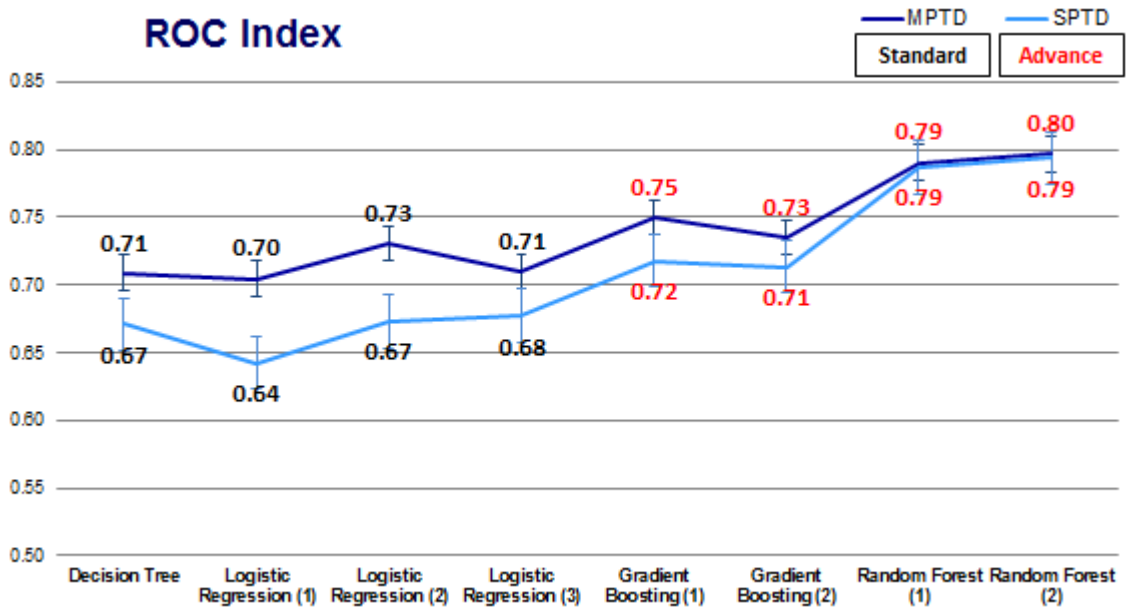


*Figure 22: ROC index comparison comparison of the models with random forest & gradient boosting and multi-period training data against the other models*

However, the opposite conclusion is seen from the top decile lifts in Figure 23, where the lifts for single period training data models are higher than those with multi-period training data regarding the advanced algorithms. Models with random forest algorithm return extremely high cumulative lifts of 2 for the validation top decile for most of the single period training data samples, showing the possibility of over-fitting in the model with the small validation data sets.

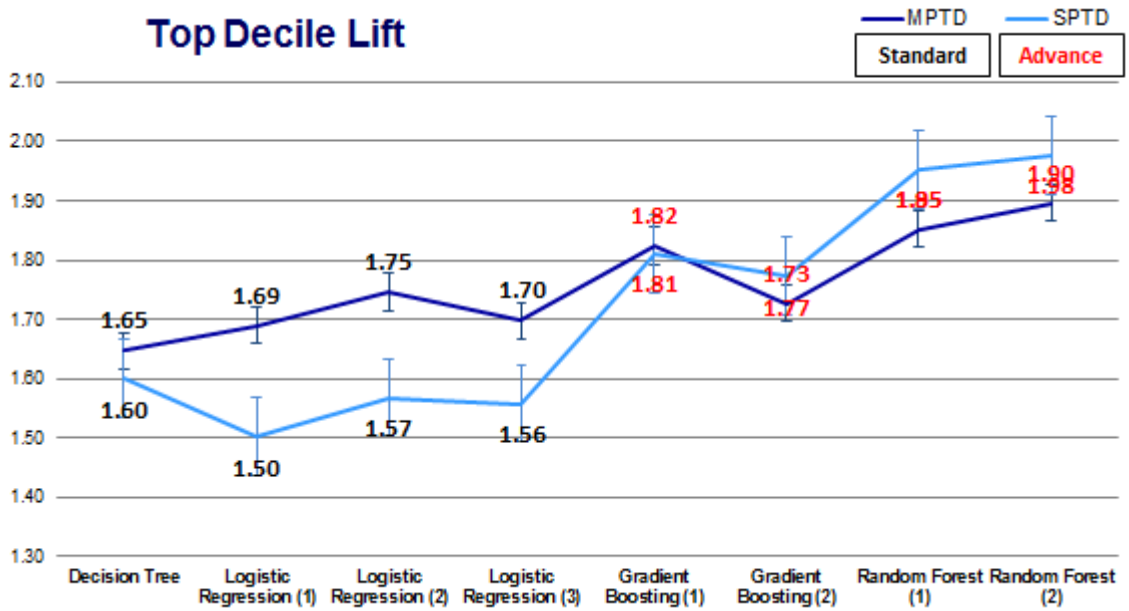


*Figure 23: Top decile lift comparison comparison of the models with random forest & gradient boosting and multi-period training data against the other models*

Regarding the models with both multi-period training data and the selected ensemble methods, it can be concluded from the figures that they perform the best compared with all the other models built in this thesis for most cases.

### 5.2.4 Question 4: The Best Churn Predictors

The last research question focuses on the most significant of churn predictors selected by the ensemble methods and the variable selection node in SAS Enterprise Miner. Ensemble models like random forests and gradient boosting provide statistics for the importance of variables in explaining the target variable after the iterations. Table 6 summarizes the top 10 churn predictors from the three methods arranged in the order of their importance. The variables are ranked in the order of number of splitting rules in for a random forest, variable importance index in boosting and R square variable selection that are aggregated from the 20 iterations of single period training data and multi-period training data for each method. The detailed statistics of the variable ranking for the three methods are provided in Appendix A.

*Table 6: Top 10 churn predictors from ensemble methods and R square variable selection*

| Order | Random Forest | Gradient Boosting | R-square variable selection |
|---|---|---|---|
| **1** | Count of family members | Count of family members | Count of family members |
| **2** | Total contacts | Days since last contact | Days since last contact |
| **3** | Days since last contact | Age | % housing loan of total agreements |
| **4** | Age | Count of shared agreements | Tenure before 1st housing loan |
| **5** | Tenure as housing loan customer | Minimum balance | Age |
| **6** | % housing loan of total agreements | Count of all agreements | Change in number of debit transactions since last month |
| **7** | Volume of housing loans | Tenure before 1st housing loan | Tenure as housing loan customer |
| **8** | Count of shared agreements | Volume of housing loans | **Unemployment rate** |
| **9** | Tenure before 1st housing loan | Total contacts | Salary |
| **10** | % housing loan of total exposure | Tenure as housing loan customer | Count of housing loans |

When reviewing the variables based on their rankings from top down, it can be seen that the demographic predictor “Count of family members” is the most important in all three methods. The second most significant demographic predictor is “Age”. The results from decision trees also show that the count of family members is most of the times selected for the first split as shown in Figure 24. Customers with less than 1.5 family members, or in other words, living alone have higher probability to churn than those living with their partners or families. However, the variable “Age” is consistently selected among the first splits.

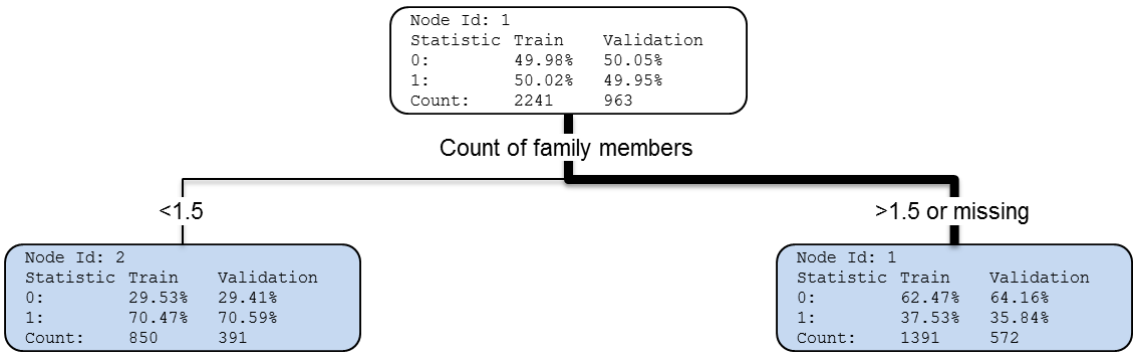


*Figure 24: First split from a random decision tree model with multi-period training data*

Similarly, the results from logistic regression models also show the significance for these two variables with p-value < 0.0001. The result from a random logistic regression model with stepwise selection using multi-period training data is provided in Table 6 to illustrate the impact of these two variables on the churn behaviour. The impact can be interpreted from the exponentials of the estimates. With this value of 0.532, one more additional family member decreases significantly the probability to churn on active housing loans by 46.8%. On the other hand, according to the result, an additional year to customers' age makes the customers 3% less likely to churn. It means that customers that have been loyal to the bank for 10 years would be 30% less likely to churn in their housing loan agreements, an impact deemed considerable in practice. It can be seen that logistic regression model could considerable facilitate managerial decision. Given the impact of these demographic churn predictors, the bank could invest on updating the information about the number family members, which plays a significant role in customers' churn behaviour.

*Table 7: Maximum likelihood analysis result of a logistic regression model on multi-period training data*

| Parameter | DF | Estimate | Standard Error | Pr > ChiSq | Exp(Est) |
|---|---|---|---|---|---|
| **Intercept** | 1 | 3.6675 | 0.4057 | <0.0001 | 39.155 |
| **% housing loan of total agreements** | 1 | -1.0691 | 0.2825 | 0.0002 | 0.343 |
| **Age** | 1 | -0.0314 | 0.0059 | <0.0001 | 0.969 |
| **Count of family members** | 1 | -0.631 | 0.0568 | <0.0001 | 0.532 |
| **Days since last contact** | 1 | -0.0012 | 0.0002 | <0.0001 | 0.999 |
| **Total contacts** | 1 | 0.2661 | 0.0623 | <0.0001 | 1.305 |

N = 1795, for training data

The variables in the next highest ranks in Table 7 describe the characteristic of the relationship with the bank such as "Total contacts", "Days since last contact" and "Tenure as

housing loan customer" or "Tenure before 1st housing loan". Also, Table 7 shows that "total contacts" and "number of days since last contact" are significant with p-value < 0.0001. The measures for their impact are 0.999 and 1.305 correspondingly; meaning one additional day since the last contact reduces the probability to churn by 0.1% while an increase in the total contact by one increases the probability to churn by 30.5%. Thus, the impact of the total contacts is severely considerable on the churn probability.

Most of the remaining variables belong to the customer behaviour group: housing loan related information such as housing loan volume, the percentage of the whole housing loan agreement or total exposure per customer; total portfolio related variables such as total number of agreements made with the bank; or transactional information such as minimum balance or the change in the number of debit transactions compared to last month. These variables have quite lower ranks compared to demographic and relationship related variables, a result that aligns with other churn papers (Mavri & Ioannou, 2008).

However, among the macro-economic factors modelled in this thesis, only "unemployment" is ranked rather low in R-square variable selection (rank 8) and none of them are present in the top ten most important variables in the ensemble methods. This might be explained by the short time window of analysis within only a year for these external factors such as inflation rate and consumer confidence index to show their effect because these indices are usually examined on a yearly basis (Ballings & Van den Poel, 2012).

# 6 Discussion and Conclusion

## 6.1 Main Findings

This thesis studies the extended application of the proposed multi-period training data approach by Gur Ali & Ariturk (2014) to a churn classification model that employs ensemble methods like random forest and gradient boosting using the empirical data of the housing loan customers from a large Nordic bank. From the results presented in Chapter 5, based on the selected evaluation criteria such as misclassification rate, ROC index and top decile lift, the main findings can be summarized as follow:

1. **Answer to research question 1:** For churn classification models that run with logistic regression and decision tree, the thesis has validated that multi-period training data produces better churn prediction than single period training data.

The results show that the multi-period training data approach improves the churn classification accuracy by decreasing the misclassification rate on average by 0.02 for the models with decision tree and on average by 0.04 for the models with logistic regression. The improvement in misclassification rate can be translated into the reduction of type II error, referring to the error of false negative, or misclassifying real churn events into non-churn events. For models with decision tree and logistic regression respectively, the multi-period training data approach reduces the false negative error rate by 0.11 and 0.08 on average. To put this into perspective, regarding the multi-period training data set with 770 observations in the validation data set (30% of the original data set), there are 350 real churn events, for which a reduction of 0.1 in false negative rate means 35 real churners would be classified correctly by models that employs multi-period training data. As mentioned previously in the literature review, type II error is usually of more interest than type I error because misclassifying a would-be churner as a non-churner and hence, taking no action to retain that customer will leave the bank with a loss of future profits from that customer not only from housing loan products but also from other possible products that the customer might be or will be holding with the bank. On the other hand, misclassifying a non-churner as a would-be churner only costs the bank the unnecessary retention action. In other words, such improvement in type II error is truly considerable for a bank with a large customer base.

Top decile lift evaluation criteria is said to provide the best managerial guidance for churn retention among the top ten most probable churners (Gur Ali & Ariturk, 2014). When comparing two top decile lifts by taking their ratio as followed

$$\frac{Top\ decile\ lift_1}{Top\ decile\ lift_2} = \frac{\hat{\pi}_{10\%,1}}{\hat{\pi}_0} : \frac{\hat{\pi}_{10\%,2}}{\hat{\pi}_0} = \frac{\hat{\pi}_{10\%,1}}{\hat{\pi}_{10\%,2}}$$

the result is a ratio of the proportion of churners that model 1 can identify from the top 10% customers with the highest probability compared to that of model 2.

Making the same comparison for the top decile lifts of models that run with decision tree and logistic regression between the two approaches of constructing the training data, the ratio between multi-period training data improves the top decile lift on average by 0.03 for the models with decision tree and by 0.11 for the models with logistic regression. The result means that among the top 10% customers that receive the highest probability to churn given by the models, the multi-period training data approach has classified 3% and 11%, respectively, more churners compared with the single period training data approach.

2. **Answer to research question 2:** For churn classification models that run with single period training data, it is observed in the thesis that ensemble methods such as random forest and gradient boosting improve churn classification performance of the models compared to logistic regression and decision tree.

It can be seen from the results that among the models using single period training data, models with random forest and gradient boosting increase the ROC index by 0.12 and 0.05 while decrease the misclassification rate by 0.04 and 0.01 on average compared to the models with decision tree. In terms of type II error rate, models with the selected ensemble methods have reduced the rate by roughly 0.15 compared to models with decision tree, showing the strong improvement in churn classification performance. Regarding top decile lift, models with random forest and gradient boosting considerable improve the top decile lift by 0.2 on average compared to models with decision tree. In practice, such an increase means that by employing the selected ensemble methods, the bank can reach 20% more of potential churners by targeting the retention campaign to the top 10% customers based on the model.

However, the improvement in top decile lift between models with the selected ensemble methods and those with logistic regression is only minimal using the single period training data. Once again, this result is similarly observed previously due to the robust performance of logistic regression.

3. **Answer to research question 3:** Churn classification models that employ both multi-period training data and ensemble methods like random forest and gradient boosting have the best performance in churn prediction compared to the other models only based on misclassification rates and ROC index but not for top decile lift.

Let's focus on the performance of churn classifiers or churn algorithms. In comparison of all the models built with both training data construction approaches and four algorithms, models with random forest and gradient boosting have the best aggregated measures of the evaluation criteria. Particularly between these two selected ensemble methods, random forest consistently produces models with better results. Therefore, the so-called best models based on all criteria in this thesis are those that run with random forest. The best models with random forest based on misclassification rates and ROC index are those with multi-period training data; however, based on top decile lift, the best models are those with random forest and single period training data. Also, when comparing the models with ensemble methods to those with logistic regression and decision tree, it can be observed that the gaps between the multi-period training data lines and the single period training data lines for models with ensemble methods are considerably smaller for misclassification rate and ROC index while the single period training data line is higher than its counterpart for top decile lift.

However, in most cases, the so-called best models are those with random forest and multi-period training data. On the other hand, the models that have the worst results in all evaluation criteria are those with single period training data and logistic regression without variable selection. Consequently, variable selection methods using R-square or decision tree have indeed improved the performance of logistic regression in churn classification. To illustrate the extension of the improvement in churn classification using the proposed multi-period training data and ensemble methods, the so-called best and worst models are compared. The average misclassification rate is reduced from 39% to 28%. Such an improvement is translated into 11% reduction in the type II error rate. The average ROC index is increased from 0.64 to 0.8, attributing to 25% improvement. Finally, the average top decile lift is improved from 1.5 to 1.98, meaning 32% more churners might be targeted in a retention campaign to the top 10% most probably churners based on the best models.

4. **Answer to research question 4:** It is observed that the best churn predictors in this thesis are those from the demographic and customer relationship variable groups while the behavioral variables and macro-economic factors do not prove to be significant in explaining the churn behavior.

Given the results presented in section 5.2.4, the most important churn predictor is "the number of family members that transact with the bank". It is understandable that changes, either upward or downward, in the number of family members considerably impact customers' behaviors regarding their housing loan agreements. For example, families with new members usually need more spacious apartments or own houses; hence new housing loan agreements. On the other hand, divorced couples generally move away from their existing houses or apartments and look for new ones; as a result, they might terminate their existing shared housing loan agreements and open new ones. In those cases, customers might not only look for offers from the bank, where they are having their existing housing loan agreements, but also look for offers from the other banks. If better offers are found from other loan providers, the customers are more inclined to terminate the existing agreements with the current bank and open new ones elsewhere. In general, changes in the number of family members are meaningful milestones in customers' lives that will impact the decision for their housing loans. As a result, it would be beneficial for the bank to be able to capture or predict these changes in advance.

The second most important churn predictor is customers' age from the demographic group. It also makes sense that the older a customer gets, the more probable that she/he will experience the housing related milestone events in their lives such as getting married, having children, getting divorced and so on. However, the impact of one year increase in customers' age on churn probability is likely to depend on the life stage that the person is in. For example, young people who are living in rented apartments are less probable to churn than mature people who have accumulated wealth over time and are capable of purchasing their own houses.

Another highly important churn predictor among the top 10 belongs to the group that describes customer relationship with the bank: "total contacts with the bank". Most noticeably, one of the models suggests that a unit increase in the total contacts with the bank increases the probability to churn by more than 30%. This severely considerable impact might be explained by the possibly negative nature of those contacts to the bank. In general, customers with more contacts with the service providers might indicate a more interactive relationship with the bank, which usually implies more positive and profitable relationship. However, in the banking context, given the focus on omni-channel and seamless customer experience in the current digital age, recent and frequent contacts might also mean issues that

need support from the bank or even negative feedback. Such experience might drive the customers away from the bank's services.

Regarding the models with multi-period training data, although it is advisable to include external macro-economic factors to explain the time-varying churn behavior, none of the selected environmental variables in this study are convincingly present among the top ten most important churn predictors. This might be due to the short time window of this study of less than a year, a period that might be too short for these variables as they usually exhibits their impact on a yearly time scale.

Finally, it is interesting to see that no churn predictor from the customer behavior group is consistently suggested to be important by the employed variable selection methods. This is aligned with other researches' conclusion that customer transactional data might not be a good source for churn predictors compared to other groups.

## 6.2 Practical Contribution

The contribution of the thesis is two-fold in terms of academic contribution to churn classification literature and practical contribution to the commissioned bank.

**Academic contribution to churn classification in housing loan**

Regarding the training data construction approach in churn classification models, the thesis confirms the application of the proposed multi-period training data approach by Gur Ali & Ariturk (2014) in churn classification to improve considerably the performance of the models that run with logistic regression and decision tree. Improvement can also be observed, though not highly considerable, for models that run with ensemble methods like random forest and gradient boosting in comparison with the single period training data approach. The improvement is mainly thanks to the more effective use of churn events that are usually rarely available in real life data. Specifically, in contrast to the single period training data approach, which captures churn only at a specific period of time and discards the churn events that has happened prior to that period, the multi-period training data approach allows the employment of historical churn events, providing more real data regarding churn events to the model and mitigating the rarity issue in churn prediction. Consequently, the author highly recommends other studies in churn classification to employ the multi-period training data approach together with ensemble methods to achieve the best possible classification models.

Furthermore, the thesis has enriched the churn prediction literature in the housing loan domain, which has not received sufficient attention from churn researchers even though much churn studies have been done within the retail banking industry. First of all, as far as the author's knowledge, there has not been an operationalized definition for a churn response for housing loan customers. Therefore, this thesis has provided the housing loan churn literature with an operational definition of a housing loan churn event. Second, the thesis has not only conducted thorough literature review on the most popular churn predictors for banking customers in general and for housing loan customers in specific, but also employed those churn predictors in the thesis to show their importance. Consequently, subsequent studies in housing loan churn prediction can use the list of churn predictors and the thesis's answer to the research question 4 as the benchmark for further comparative studies.

**Practical solution to the bank's housing loan churn classification problem**

The most important contribution to the commissioned bank is that the author has created the first churn classification models for the bank's Finnish housing loan customers. As mentioned above, regarding the methodologies applied in churn classification model, the thesis has employed a recently proposed training data construction approach and ensemble methods to create churn classification models for the housing loan customers in Finland. The results in the thesis support that such approach considerably improves churn prediction performance. Regarding the variables, the thesis has invented the first operational definition of a churn event for housing loan customers to calculate the dependent variable and has identified the most important churn predictors based on the employed methodologies.

Moreover, the data collection process is coded in highly reusable programs in SAS Enterprise Guide in SQL code. Changes can be made easily to reflect the changing business environment in different aspects. For example, the operational definition of churn events is coded with adjustable thresholds for the parameters. The bank can create different versions of churn operational definition for customers from other countries based on the calculation logic created by the author. Besides, the collection and processing procedure of the churn predictors or independent variables are also coded to guide other analysts where to collect the data for which time period, how to pre-process the data and how to calculate the variables for the models. This thesis would provide the benchmark for future churn models in terms of data collection. The author has also created a program that produces single period training data sets and combines them into the multi-period training data set in SAS Enterprise within a pre-defined time window of analysis.

## 6.3 Limitations and Future Research

As the focus of this thesis is on the first step in customer retention, which is to classify the customers into potential churners and non-churners, the thesis only provides ranking of the customers based on their probability to churn. As it was impossible to collect profitability related data for the housing loan customers within the thesis's time frame, the thesis cannot provide any analysis on the monetary impact on the retention campaign conducted based on the suggestion of the models created. For example, the thesis cannot answer such questions as how much more profit the retention campaign can make based on the best models in the thesis, or how costly it is for the bank if it is not able to retain the probable churners. Such questions are in general of managerial interest. As a result, the logical next step with the created models is to employ the relevant data to produce profitability analyses of the retention campaigns based on the created models. For example, it is highly advisable in churn literature to incorporate customers' heterogeneous profitability or customer life time value into the churn prediction model in order to maximize the profitability of the customer retention campaign based on the model. Besides customer life-time values, the costs to carry out the retention action should also be included in order to estimate the possible gain (or loss) to retain (or lose) a possible churner. For the banks that perform targeted marketing to their customers individually, the retention cost might also be heterogeneous for the customers. For example, it is more profitable for the banks to invest more in the retention actions for probable churners that are worth more to the bank with more loans or assets managed by the bank.

Moreover, the limited availability of data during the thesis's timeline also narrows the time window of this analysis to less than a year, a period of time that is too short for any model built from the employed data in this thesis to be able to generalize for the bank's all housing loan customers. Besides, this short time window also makes it hard for the impact of the macro-economic factors to manifest. The effect of changing interest rates or customer confidence index is usually observed over the years; therefore, little to none of the selected macro-economic factors in the thesis have proved to be important in churn classification within the time window of less than a year.

# Reference

A., O. & A., A., 2015. Customer Churn Analysis In Banking Sector Using Data Mining Techniques. *African Journal of Computing & ICT,* pp. 165-174.

Accenture, 2015. *Banking Customer 2020 - Rising Expections Point to the Everyday Bank,* s.l.: Accenture Strategy.

Baecke, P. & Van den Poel, D., 2009. Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligience Information Systems,* p. 367–383.

Ballings, M. & Van den Poel, D., 2012. Customer Event History for Churn Prediction - How Long Is Long Enough?. *Expert Systems with Applications,* pp. 13517-13522.

Breiman, L., 2001. Random Forests. *Machine Learning,* pp. 5-32.

Burez, J. & Van den Poel, D., 2009. Handling Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications,* pp. 4626-4636.

Chen, C., Liaw, A. & Breiman, L., 2004. *Using Random Forest to Learn Imbalanced Data,* Berkeley: University of California BerKeley Library.

Chrzanowska, M., Alfaro, E. & Witkowska, D., 2009. The Individual Borrowers Recognition: Single and Ensemble Trees. *Expert Systems with Applications,* pp. 6409-6414.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics,* 7(1), pp. 1-26.

Employee, 2016. *Credit Process Analyst* [Interview] (5 September 2016).

Employee, 2016. *Senior Business Developer* [Interview] (October 2016).

Fitzpatrick, T. & Mues, C., 2015. An Empirical Comparison of Classification Algorithms for Mortgage Default Prediction: Evidence from a Distress Mortgage Market. *European Journal of Operational Research,* p. 427–439.

Freund, Y. & Schapire, R. E., 1999. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence,* pp. 771-780.

Friedman, J. H., 2002. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis,* p. 367–378.

Glady, N., Baesens, B. & Croux, C., 2009. Modeling Churn Using Customer Lifetime Value. *European Journal of Operational Research,* p. 402–411.

Gur Ali, Ö. & Ariturk, U., 2014. Dynamic Churn Prediction Framework with More Effective Use of Rare Event Data: The Case of Private Banking. *Expert Systems with Applications,* pp. 7889-7903.

Jinbo, S., Xiu, L. & Wenhuang, L., 2007. *The Application ofAdaBoost in Customer Churn Prediction.* s.l., s.n., pp. 513-518.

Koh, H. C. & Chan, K. L. G., 2002. Data Mining and Customer Relationship Marketing in the Banking Industry. *Singapore Management Review,* pp. 1-27.

Lariviere, B. & Van den Poel, D., 2005. Predicting Customer Retention and Profitability by Using Random Forests and Regression Forest Techniques. *Expert Systems with Applications,* pp. 472-484.

Lee, Y.-H., Wei, C.-P., Cheng, T.-H. & Yang, C.-T., 2012. Nearest-Neighbor-Based Approach to Time-Series Classification. *Decision Support Systems,* pp. 207-217.

Lemmens, A. & Croux, C., 2006. Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research,* pp. 276-286.

Lemmens, A. & Gupta, S., 2013. *Managing Churn to Maximize Profits,* Massachusetts: Harvard Business Review.

Lu, N., Lin, H., Lu, J. & Zhang, G., 2014. A Customer Churn Prediction Model in Telecom Industry Using Boosting. *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS,* pp. 1659-1665.

Maldonado, M., Dean, J., Czika, W. & Haller, S., 2014. *Leveraging Ensemble Models in SAS® Enterprise Miner™,* s.l.: SAS.

Mavri, M. & Ioannou, G., 2008. Customer Switching Behaviour in Greek Banking Services Using Survival Analysis. *Managerial Finance,* pp. 186-197.

Nie, G. et al., 2011. Credit Card Churn Forecasting by Logistic Regression and Decision Tree. *Expert Systems with Applications,* pp. 15273-15285.

Nordnet, 2017. *Nordet.* [Online]
Available at: https://www.nordnet.fi/mux/web/nordnet/index.html?gclid=CK60-

aeSytICFU5kGQodN4oAcw&ef_id=V8BhgAAABXGWViwS:20170309192057:s
[Accessed 17 December 2016].

OECD, 2017. *OECD Data.* [Online]
Available at: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm#indicator-chart
[Accessed 17 December 2016].

Prasad, U. D. & Madhavi, S., 2012. Prediction of Churn Behavior of Bank Customers Using Data Mining Tools. *Indian Jounral of Marketing.*

Qi, J. et al., 2008. ADTreesLogit Model for Customer Churn Prediction. *Annals of Operations Research.*

Reichheld, F. F. & Kenny, D. W., 1990. The Hidden Advantages of Customer Retention. *Journal of Retail Banking*, p. 19.

Sarma, K. S., 2007. *Variable Selection and Transformation of Variables in SAS® Enterprise Miner™ 5.2,* White Plains Newyork: Ecostat Research Corp..

Statistics Finland, 2017. *Findicator.* [Online]
Available at: http://www.findikaattori.fi/en/34?_ga=1.80184471.1604733451.1482679702
[Accessed 17 December 2016].

Triami Media BV, 2017. *Worldwide Inflation Data.* [Online]
Available at: http://www.inflation.eu/inflation-rates/finland/historic-inflation/cpi-inflation-finland-2015.aspx
[Accessed 17 December 2016].

Van den Poel, D. & Lariviere, B., 2004. Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. *European Journal of Operational Research*, pp. 196-217.

van Wezel, M. & Potharst, R., 2007. Improved Customer Choice Predictions Using Ensemble Methods. *European Journal of Operational Research*, pp. 436-452.

Weiss, G. M., 2004. Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets*, 1 June, pp. 7-19.

X-rates, 2017. *X-rates.* [Online]
Available at: http://www.x-rates.com/average/?from=USD&to=EUR&amount=1&year=2015
[Accessed 17 December 2016].

Yaya, X., Xiu, L., E.W.T., N. & Weiyun, Y., 2009. Customer Churn Prediction Using Improved Balanced Random Forests. *Expert Systems with Applications*, p. 5445–5449.

Zoric, A. B., 2016. Predicting Customer Churn in Banking Industry Using Neural Networks. *Interdisciplinary Description of Complex Systems*, pp. 116-124.

# Appendix A: Variable Importance

Detailed variable importance statistics are provided from random models run with random forest, gradient boosting and R-square variable selection.

*Table A1 Variable importance statistics of a random forest model for the top ten most important variables*

| Variable Name | Number of Splitting Rules | Train: Gini Reduction | Train: Margin Reduction | OOB: Gini Reduction | OOB: Margin Reduction | Valid: Gini Reduction | Valid: Margin Reduction |
|---|---|---|---|---|---|---|---|
| Count of family members | 51 | 0.01 | 0.02 | 0.00 | 0.02 | 0.02 | 0.03 |
| Total contacts | 50 | 0.01 | 0.03 | 0.01 | 0.02 | 0.01 | 0.02 |
| Days since last contact | 34 | 0.01 | 0.02 | 0.00 | 0.01 | 0.00 | 0.01 |
| Age | 25 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Tenure as housing loan customer | 22 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| % housing loan of total agreement | 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Volume of housing loan | 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Number of shared agreements | 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Tenure before housing loan | 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| % housing loan of total exposure | 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Table A2 Variable importance statistics of a gradient boosting model for the top ten most important variables*

| Variable Name | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance | Interaction Importance |
|---|---|---|---|---|---|
| Count of family member | 3 | 1.00 | 1.00 | 1.00 | 0.11 |
| Days since last contact | 3 | 0.86 | 0.47 | 0.54 | 0.07 |
| Age | 1 | 0.56 | 0.00 | 0.00 | 0.03 |
| Count of shared agreements | 1 | 0.50 | 0.30 | 0.60 | 0.02 |
| Minimum balance | 1 | 0.47 | 0.38 | 0.82 | 0.01 |
| Count of all agreements | 1 | 0.44 | 0.77 | 1.75 | 0.02 |
| Bank tenure before mortgage | 1 | 0.31 | 0.28 | 0.90 | 0.01 |
| Volume of housing loan | 1 | 0.28 | 0.12 | 0.42 | 0.01 |
| Total contacts | 1 | 0.28 | 0.12 | 0.42 | 0.01 |
| Tenure as housing loan customer | 1 | 0.28 | 0.12 | 0.42 | 0.01 |

Exceptionally, the variable importance of the R-square variable selection method is presented as a model in SAS

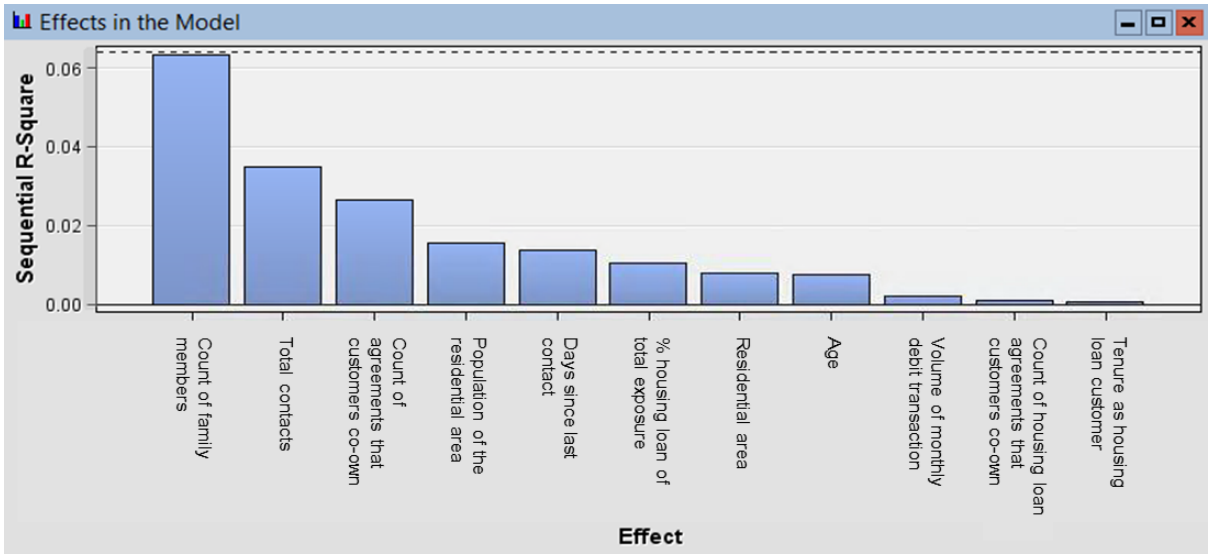


*Figure A1 Variable importance statistics of a R-square variable selection node for the top ten most important variables*