

**Improved Churn Prediction by More Effective Use of  
Customer Data: The Case of Private Banking Customers**

by

**Umut Arıtürk**

**A Thesis Submitted to the  
Graduate School of Engineering  
in Partial Fulfillment of the Requirements for  
the Degree of**

**Master of Science  
in  
Industrial Engineering**

**Koc University**

**August 2011**

Koc University  
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Umut Arıtürk

and have found that it is complete and satisfactory in all respects,  
and that any and all revisions required by the final  
examining committee have been made.

Committee Members:

---

Asst. Prof. Özden Gür Ali (Advisor)

---

Prof. Süleyman Özekici

---

Assoc. Prof. Serdar Sayman

Date:

---

---

*...to my parents*

*in love and gratitude...*

---

## ABSTRACT

Based on the fact that acquiring new customers in any business is much more expensive than trying to keep the existing ones, customer retention is an increasingly pressing issue in today's ever-competitive commercial arena. This is especially relevant for service related industries. Thus, classification models to detect churning customers have received significant attention in the customer relationship management literature.

In this thesis, we focus on the churn prediction problem in the private banking industry under non-contractual settings. Churn models are typically estimated on cross-sectional data pertaining to a particular time period and used for prediction in subsequent periods. This is appropriate under static settings where the sample size of churners is sufficient; however churn is a process and customer behavior is affected by changes in the environment. We show that modeling next-period churn behavior with multiple observations per customer pertaining to different time periods yields better predictive performance when compared to traditional cross-sectional training data, with or without synthetic oversampling techniques, and provides additional managerial insights. Further, reflecting the idea that customers do not always decide on and carry out the churn action overnight; we propose to model churn with multiple models that predict churn several periods ahead, and to use these predictions in an ensemble for improved next-period churn prediction. This approach provides the company with advance notice on customer's churn propensity, improves out-of-sample next-period churn prediction, and ensures consistency of advance propensity figures with next-period prediction. We evaluate our models with data from the highly dynamic banking industry.

---

## ÖZETÇE

Yeni müşteri edinmenin mevcut müşterileri elde tutmaya kıyasla daha pahalıya mal olduğu ışığında, müşteriyi elde tutma günümüzün rekabetçi ortamında daha baskın bir unsurdur. Bu durum, özellikle hizmet sektöründe faaliyet gösteren firmalar için ön plana çıkmaktadır. Bunun neticesi olarak, bağlılığı azalan ve terk etmeye meyilli müşterileri tayin etmeye yönelik sınıflandırma modelleri müşteri ilişkileri yönetimi yazınında artan bir öneme sahiptirler.

Bu tezde, biz sözleşme dışı koşullardaki müşteri kayıp tahmin problemi üzerinde odaklanacağız. Müşteri kaybı modelleri, genellikle belirli bir zaman dilimine ait kesitsel veride geliştirilip bu zamanı müteakip gelecek zaman dilimlerinde tahmin amacıyla kullanılmaktadırlar. Söz konusu durum, kayıp müşteri sayısının yeterli olduğu statik koşullarda uygundur, ancak müşterinin hizmet aldığı firmayı terk etmesi bir süreçtir ve müşteri davranışı çevredeki değişikliklerden etkilenmektedir. Biz, her bir müşteri için farklı zaman dilimlerine ait birden fazla sayıda gözlemin kullanıldığı boylamsal veri seti üzerinde yapılan ‘gelecek-dönem müşteri kaybı’ modellemesinin, sentetik üst örnekleme uygulandığı ve uygulanmadığı geleneksel kesitsel veri kullanımına kıyasla tahmin performansını artıracak ve yöneticilerin ilave çıkarımlarda bulunmasına imkan sağlayacağını göstereceğiz. Ayrıca, müşterilerin her zaman anlık kararlar doğrultusunda terk etmediği düşüncesi altında, müşteri kaybını muhtelif zaman dilimleri öncesinden tahmin eden birden çok model önereceğiz ve bu modellerden elde edilen tahminleri, tanıtacağımız bir topluluk yönteminde kullanarak ‘gelecek dönem müşteri kaybı’ tahminini geliştireceğiz. Bu yaklaşımlar, firmaya müşterilerin terk etme eğilimlerine ilişkin erken uyarı sistemi sağlamakta, örneklem dışı ‘gelecek dönem müşteri kaybı’ tahmin performansını iyileştirmekte ve elde edilen birden çok sayıdaki kayıp eğilim skorlarının ‘gelecek dönem kayıp’ tahmini ile tutarlı olmasını sağlamaktadır. Modellerimizi, oldukça dinamik olan bankacılık sektöründen elde ettiğimiz veri üzerinde değerlendireceğiz.

---

## ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Asst. Prof. Özden Gür Ali for the continuous support of my graduate study, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me to accomplish this research project. It has been privilege for me to work two years with her.

I would like to thank the members of my thesis committee Prof. Süleyman Özekici and Assoc. Prof. Serdar Sayman for critical reading of this thesis and for their valuable comments.

I would like to thank TUBITAK (The Scientific and Technological Research Council of Turkey) for providing generous scholarship throughout my graduate education, and Mehmet Hamdi Özçelik from Yapı Kredi Bank for providing the data for this study.

I am grateful to Buşra Ürek, Filiz Sayın, and Müge Güçlü not only for being wonderful friends, but also for being a family to me. Without their presence and close friendship, the last two years would not be bearable. I would also like to thank all my friends at Koc University for all the fun and good times we shared together.

Finally, I owe special thanks to my brother Cem, for being my role model, to my sister-in-law Elçin, for the fun she brings to our life, and to my niece Ada who I love so much. I am also grateful to my aunt Emel, for always encouraging me and being supportive, and especially to my cousin Gül who is actually more than a sister to me. Last and most importantly, I am deeply indebted to my parents, Yavuz and Deniz. They have always believed in me and supported each step that I have taken so far. Without their morale support, I would never be the person who I am today. To them I dedicate this thesis.

---

## TABLE OF CONTENTS

<b>List of Tables .....</b>	<b>xi</b>
<b>List of Figures .....</b>	<b>xiii</b>
<b>1. Chapter 1: Introduction .....</b>	<b>1</b>
<b>2. Chapter 2: Literature Review .....</b>	<b>4</b>
2.1. Introduction .....	4
2.2. Churn Definition and Importance of Churn .....	4
2.3. Variables Used in Churn Studies.....	5
2.4. Hypotheses Derived from Churn Studies.....	9
2.4.1. Hypotheses Related to Customer Behavior Predictors.....	9
2.4.2. Hypotheses Related to Customer Demographics Predictors.....	11
2.4.3. Hypotheses Related to Customer-Company Interaction Predictors.....	11
2.4.4. Hypotheses Related to Environmental Indicators .....	12
2.5. Training and Validation Datasets in Churn Studies .....	12
2.5.1. Training Datasets.....	12
2.5.2. Validation Datasets.....	13
2.6. Churn Prediction Techniques .....	14
2.6.1. Logistic Regression .....	14
2.6.2. Classification and Regression Trees.....	17
2.6.2.1. Regression Trees .....	17
2.6.2.2. Classification Trees.....	19

---

2.6.3. Survival Analysis and Proportional Hazard Modeling.....	20
2.6.3.1. Survival Modeling.....	21
2.6.3.2. Proportional Hazard Model.....	22
2.6.3.3. Cox Regression Diagnostics.....	23
2.7. Evaluation Criteria .....	25
<b>3. Chapter 3: Modeling Approach.....</b>	<b>28</b>
3.1. Introduction .....	28
3.2. Problem Definition .....	28
3.3. Churn Definition.....	29
3.4. Modeling Approach.....	30
3.4.1. Training Observations .....	31
3.4.2. Advance Churn Labels .....	32
3.4.3. Benchmark Models to “W1C through WnC Models”.....	35
3.4.3.1. Cox Regression (Proportional Hazard Model).....	36
3.4.3.2. Multinomial Logistic Regression to Obtain Advance Churn Propensity Scores .....	37
3.4.3.3. Ordinal Logistic Regression to Obtain Advance Churn Propensity Scores .....	38
3.4.4. Ensemble Method Generation .....	40
<b>4. Chapter 4: Experiments and Results.....</b>	<b>42</b>
4.1. Introduction .....	42
4.2. Dataset Description .....	43



---

4.3. Predictors and Hypotheses .....	47
4.3.1. Customer Behavior Predictors and Related Hypotheses .....	48
4.3.1.1. Portfolio Information.....	48
4.3.1.2. Account Information .....	49
4.3.1.3. Asset Return Information .....	50
4.3.1.4. Product-specific Ownership .....	50
4.3.1.5. Total Product Ownership .....	51
4.3.1.6. Scorecard.....	52
4.3.2. Customer-Company Interaction Predictors and Related Hypotheses .....	53
4.3.3. Customer Demographics Predictors and Related Hypotheses .....	55
4.3.3.1. Age .....	55
4.3.3.2. Gender .....	56
4.3.3.3. Nationality.....	56
4.3.3.4. Education.....	57
4.3.3.5. Employment Type.....	58
4.3.3.6. Marital Status .....	58
4.3.3.7. Customer Segmentation .....	58
4.3.4. Economic Indicators and Related Hypotheses .....	59
4.4. Experimental Setup .....	59
4.5. Accuracy Results.....	62
4.5.1. MPTD vs. SPTD: Predicting Within-one-period Churn .....	63
4.5.2. MPTD vs. SPTD: Predicting Within-one-period Churn over Time.....	65

---

4.5.3. MPTD vs. SPTD: Predicting Within-one-period Churn with Information Lag .....	69
4.5.4. W2C to W6C Models.....	72
4.5.5. Benchmark Models.....	74
4.5.6. Meta-Combination Ensemble Method .....	78
4.5.7. Contributions.....	83
4.6. Model Outputs and Managerial Insights .....	84
4.6.1. W1C Models.....	84
4.6.2. W2C to W6C Models.....	90
4.6.3. Benchmark Models.....	90
4.6.3.1. Cox Regression .....	91
4.6.3.2. Multinomial Logistic Regression .....	93
4.6.3.3. Ordinal Logistic Regression.....	93
4.6.4. Meta-Combination Ensemble Method .....	94
4.6.5. Contributions.....	95
<b>5. Chapter 5: Conclusion .....</b>	<b>97</b>
<b>Appendices .....</b>	<b>100</b>
<b>Bibliography .....</b>	<b>105</b>
<b>Vita .....</b>	<b>113</b>

---

## LIST OF TABLES

2.1 Literature review of attrition studies .....	7
2.2 Hypotheses generated in churn studies and investigated in this study.....	10
2.3 Confusion matrix.....	26
3.1 Labeling for a customer who churned before m .....	33
3.2 Labeling for a customer who churned in period m .....	34
3.3 Labeling for a non-churner.....	35
4.1 Training and test datasets for W1C to W6C while computing robustness.....	46
4.2 Number of positive cases for each churn label .....	47
4.3 The W1C model's performance: AUC, TDL, and TQL on the first test period .....	64
4.4 MPTD vs. SPTD over time .....	66
4.5 Paired t-test analysis for the difference between MPTD and SPTD: Logistic Regression & without SMOTE .....	68
4.6 Paired t-test analysis for the difference between MPTD-S and SPTD-S: Logistic Regression & with SMOTE .....	69
4.7 How to interpret the performance accuracy when recent data is not available.....	70
4.8 Next-month churn prediction with information lead time: AUC, TDL, and TQL performance on MPTD .....	71
4.9 Several months ahead churn prediction: AUC, TDL, and TQL performance on MPTD.....	73
4.10 Cox regression: AUC and TDL performance .....	75
4.11 Multinomial logistic regression: AUC and TDL performance .....	76
4.12 Ordinal logistic regression: AUC and TDL performance .....	78
4.13 Ensemble models' within-one-period churn prediction: AUC performance .....	79
4.14 Ensemble models' within-one-period churn prediction: TDL performance.....	80
4.15 Paired t-test analysis for the difference between Ensemble (Case5 – 200) and MPTD: Logistic Regression & without SMOTE.....	81

---

4.16 Paired t-test analysis for the difference between Ensemble (Case6 – 1000) and MPTD-S: Logistic Regression & with SMOTE .....	82
4.17 Parameter estimates of the logistic regression models LogReg-SPTD, LogReg- SPTD-S, LogReg-MPTD, and LogReg-MPTD-S for the churn label W1C.....	86
4.18 Hazard ratio estimates of the Cox regression.....	92
A Parameter estimates of the WnC-LogReg models .....	100
B Parameter estimates of the multinomial logistic regression .....	101
C Parameter estimates of the ordinal logistic regression .....	104

---

## LIST OF FIGURES

3.1 MPTD and SPTD .....	32
3.2 Ensemble method overview .....	41
4.1 MPTD, SPTD, and test dataset for the next-period churn label (W1C).....	44
4.2 Longitudinal training data to predict W1C to W6C.....	45
4.3 The regression tree output for Case5-200: Ensemble methodology .....	95

## **Chapter 1**

### **Introduction**

Today's organizational environment is markedly different from that of the past. Increasing global competition, developments in the information technology, and vast and quick changes in global and local economic conditions are forcing management of all types of organizations to entirely rethink and adapt their approach to operational and strategic issues. In this respect, service companies need to closely focus on the customer needs and expectations to sustain their presence in the business environment. Therefore, when the duration of the relationship between the company and the customer is not known, it is of crucial importance to detect the customers decreasing their loyalty to the company and being on the verge to abandon their relationship with the company, also called churners. However, there arise a number of prominent complications that should be considered before developing churn prediction models. These complications are mainly associated with the cross-sectional training data dominating the churn literature, data in which the observations belong to a snapshot in time and each observation corresponds to a unique customer. In this thesis, using the rich data provided by Yapı Kredi Bank, we examine the churn prediction problem for the private banking customers.

First, private banking industry is characterized by its non-contractual relationship with customers, i.e. there is not any binding contract which precisely indicates the ending time of the regular relationship. As a result, the time at which the customer becomes inactive and has already abandoned the company is unobserved. This raises the question when to declare a churn event. Second, the changes in the financial indicators such as the exchange rate and consumer price index are presumed to have impact on the customer behavior, including the churn decision. Therefore, they also

need to be incorporated while developing predictive models. As the extensive churn literature in Table 2.1 indicates, this is skipped in the majority of churn studies. Third, churn is a rare event in service industries (Burez and Van den Poel, 2009), therefore, lacking the sheer number of churn examples, the models can only capture the strongest signal or overfit the peculiarities of the particular time period. Last, after generating the churn model, it is quite possible that the most recent customer information, necessary for obtaining churn scores, may not be available.

In this thesis, to address these key difficulties, we propose to more effectively use the existing customer data and to fully exploit it. Our modeling approach consists of three components: training observation construction, advance churn label generation, and ensemble model generation.

The first component of our modeling approach is to use longitudinal training data which comprises more than one observation per customer. In typical churn studies, cross-sectional training data is employed. In this type of training data, one observation corresponds to a unique customer and mainly, it comprises only the customers who are either active or churn at the measurement time. This implies that neither dynamic environmental information can be mapped into the sample data (the second complication aforesaid) nor the information from previous churners can be incorporated in the model (the third complication aforesaid). To handle these two issues, we propose to split the observation time into customer-period observations and organize the sample data with each observation corresponding to a unique customer time period. For example, if customer  $i$  churned at the end of three (3) months as from we begin to observe the customers, then this is represented by three observations in the data.

The second component of our modeling approach is the advance churn label generation. In private banking industry, there is not a binding contract between the customer and the service provider; therefore it is not for sure when to label a customer as a churning if s/he leaves (the first complication aforesaid). To handle this, we provide a churn definition. In addition, in private banking, there exist a variety of financial instruments offered to customers where some are long-term-based and some short-term.

This, in turn, implies that churn might be a gradual process for some customers, i.e. they do not churn over night. On this ground, it is quite comprehensible that customers' short term churn propensity might differ from the long-term one. To deal with these issues and also to prevent from information loss that occurs due to instant transition of a customer's label from churning to non-churning, we propose to generate multiple advance churn labels, denoting 'churn within  $x$  periods',  $x = 1, 2, 3, \dots, n$ .

In the third component of our modeling approach, we propose an ensemble methodology. It combines the output predictions of independently trained models to predict 'churn within  $x$  periods',  $x = 1, 2, 3, \dots, n$ , in an ensemble to improve the within-one-period churn prediction.

We evaluate our models on private banking customer data from the highly dynamic Turkish banking industry. We use logistic regression and decision tree as the base classifiers. SAS 9.2 is used to prepare the data for the analysis and to conduct the logistic regression. WEKA 3.6 is employed for the decision tree analysis. The computation for the accuracy results is completed with a self-written code in MATLAB 8.

This thesis is organized as follow. In Section 2, we provide relevant literature. Section 3 describes the proposed modeling approach. The experiments involving different modeling techniques and results, which pertain to accuracy and extract private banking managerial insights, are described in Section 4 while Section 5 concludes with interpretation of results and future research opportunities.



## Chapter 2

### LITERATURE REVIEW

#### 2.1. Introduction

This study tackles the customer churn prediction problem. In this thesis, for the variables to be entered into the model with changing values over time, we suggest a modeling approach which is based on the use of panel data for each customer and results in accurate and comprehensible forecasting models by observing one customer over a specific period of time. Further, we present a new ensemble methodology. The relevant streams of literature for this problem are enumerated below:

- ✓ Customer defection and the importance of churn analysis
- ✓ Variables used in churn studies
- ✓ Hypotheses extracted from churn studies
- ✓ Training and validation datasets
- ✓ Classification algorithms and ensemble methods
- ✓ Accuracy measures to evaluate model performances

#### 2.2. Churn Definition and Importance of Churn

Customer defection is defined as the loss of existing customers to a competitor (Chu, Tsai, and Ho, 2007) or the customers whose loyalty towards the service provider decreases are called churners (Glady, Baesens, and Croux, 2009). In this respect, customer retention and acquisition are prominent points in customer relationship management (CRM) literature, points which also hold for the financial services industry. Several studies show the economic value associated with customer retention: The costs arising from acquiring a new customer surpass the expenditure to retain an existing one

(Ennew and Binks, 1996; Dawes and Swailes, 1999); long-term customers buy more and bring in new customers (Reichheld, 1996); for a bank, the longer the customer relationship, the higher the customer's worth (Reichheld and Kenny, 1990); a 5 per cent increase in customer retention generates an increase in customer net present value of between 25 per cent and 95 per cent across a wide range of business environments (Dawkins and Reichheld, 1990); a field experiment conducted by Burez and Van den Poel has shown that preventing attrition of the most fragile 10% of the customers can double the company's profits (2007). On these grounds, we can easily ascertain the increasing importance of the churn analysis. Based on it, companies undertake churn analysis in order to identify the valued customers who are prone to cease their relationships with the company and potentially switch to the competition before they do so. Further, churn analysis provides insights to managers for customizing their offers to retain the targeted customers.

### **2.3. Variables Used in Churn Studies**

One important design parameter while modeling churn is the set of predictors entered into the model. As claimed by Lessmann and Voß (2009), the classification decision is influenced by the input variables, consequently it is important to understand to which extent and in which direction this influence is. This in turn reveals the information based on which customer-centric business processes are acknowledged and/or re-engineered (Lessmann and Voß, 2009).

The extensive literature review and field search made by Van den Poel and Larivière (2004) addresses four distinct predictor categories that directly relate to the attrition: customer behavior, customer demographics, customer perceptions, and macroenvironment. On the other hand, the interaction between a private banking customer and the service provider is assumed to play a significant role while explaining the attrition behavior. So, with a slight modification, we endorse to create one additional category representing it which includes predictors such as the tenure, the

customer representative (branch) id through which the customer gets served, etc. Following this proposed taxonomy to categorize the explanatory variables; Table 2.1 presents a literature overview and covers studies since 2004 (for the previous studies, we kindly ask the reader to refer to Van den Poel and Larivière (2004)) . As pointed out in Table 2.1, churn studies use independent variables from at most three different categories while explaining the churn behavior and the majority of studies do not include time-varying covariates into the models unless dynamic models such as survival analysis are employed.

Table 2.1 Literature review of attrition studies

Reference	Independent Variables										Context			Methodology			Technique
	Customer behavior	Customer perceptions	Customer demographics	Macroenviroment	Customer-Company Interaction	Financial	Retail	Telco	Other	Dependent variables		Independent variables					
										Incidence	Time + incidence	Non-time-varying	Time-varying				
Wei and Chiu (2002)	x				x						x				Multi-classifier class combiner		
Larivière and Van den Poel (2004)	x				x						x				Survival analysis (hazard modeling)		
Van den Poel and Larivière (2004)	x		x	x							x				Survival analysis		
Buckinx and Van den Poel (2005)	x		x						x						Logistic regression, neural network and random forests		
Kim, Shin, and Park (2005)	x		x						x						SVM, back-propagation neural network, artificial intelligence		
Larivière and Van den Poel (2005)	x		x	x	x						x				Random forests, logistic regression		
Zhang (2005)	x		x	x	x						x				Artificial neural network, survival analysis		
Zhao et al (2005)	x		x						x						One-class SVM, artificial neural network, decision tree, Naive Bayes		
Ahn et al (2006)	x		x						x						Binary and multinomial logistic regressions		
Hung, Yen, and Wang (2006)	x		x						x						Decision tree, neural networks and K-means cluster		
Jamal and Bucklin (2006)	x		x						x						Survival analysis (Weibull hazard modeling)		
Lemmens and Croux (2006)	x		x						x						Bagging, stochastic gradient boosting, binary logit model		
Neslin et al (2006)	x		x						x						Logistic regression, neural network, decision tree, discriminant analysis, ...		
Burez and Van den Poel (2007)	x		x						x						Logistic regression (with Markov Chains) and random forests		
Van Wezel and Potharst (2007)	x		x						x						Decision tree, logistic regression, ensemble methods (LogitBoost, MultiBoost)		
Burez and Van den Poel (2008)	x		x						x						Random forests and survival analysis		
Coussemont and Van den Poel (2008a)	x								x						Support vector machine, logistic regression, random forests		
Coussemont and Van den Poel (2008b)	x	x	x						x						Logistic regression		
Guo-en and Wei-dong (2008)	x								x						Support vector machine (SVM), artificial neural network, decision tree, logistic regression, naive bayesian classifier		
Kumar and Ravi (2008)	x		x						x						Multilayer perceptron, logistic regression, decision tree, random forests, radial basis function network, SVM		

Table 2.1 cont'd

Reference	Independent Variables						Context				Methodology		
	Customer behavior	Customer perceptions	Customer demographics	Macroenvironment	Customer-Company Interaction	Financial	Retail	Telco	Other	Dependent variables		Technique	
										Incidence	Time + incidence		Non-time-varying
Burez and Van den Poel (2009)						x	x	x			x	Random forests, logistic regression	
Coussement and Van den Poel (2009)	x	x	x					x			x	Logistic regression, SVM, random forests	
Gladly, Baesens, and Croux (2009)	x					x					x	Logistic regression, decision tree, neural network, AdaCost, cost sensitive decision tree	
Lessmann and Voß (2009)						x		x			x	L1-SVM, RBF-SVM, logistic regression, classification and decision tree	
Lima, Mues, and Baesens (2009)	x							x			x	Logistic regression, decision tree	
Qi et al (2009)	x							x			x	ADTrees Logit model, TreeNet	
Tsai and Lu (2009)								x			x	Hybrid neural network: self organizing maps combined with artificial neural networks (SOM+ANN), ANN+ANN	
Xie et al (2009)	x										x	Improved balanced random forests, artificial neural networks, CWC-SVM, decision tree	
Coussement, Benoit, and Van den Poel (2010)	x										x	Generalized additive models (GAM) and logistic regression	
Tsai and Chen (2010)	x							x			x	Neural network and decision tree	
Karahoča and Karahoča (2011)	x										x	Adaptive neuro fuzzy inference system (ANFIS), fuzzy c-means + ANFIS, decision tree	
Lin, Tzeng, and Chin (2011)	x										x	Rough set theory and flow network graph	
Verbeke et al (2011)	x										x	AntMiner+ and ALBA, C4.5, RIPPER, logistic regression	

## **2.4. Hypotheses Derived from Churn Studies**

The extensive literature review assents that many studies have been conducted to predict churn in various sectors. While one objective of these studies is to correctly distinguish churners among the customer base, researchers are also interested in obtaining managerial insights as to verify and/or create hypotheses that question how the attrition behavior is affected when some predictor is changed in one direction. On this ground, Table 2.2 presents the hypotheses propounded in diverse churn studies. Please note that this table is not an exhaustive list, but contains only those hypotheses that are relevant to our work. Please also remark that + (-) sign indicates that the corresponding predictor has a positive (negative) impact on the churn behavior, i.e. the churn propensity increases (decreases) in the corresponding predictor, whereas zero (0) suggests no relationship.

### **2.4.1. Hypotheses Related to Customer Behavior Predictors**

Some researchers have investigated the impact of product ownership on attrition behavior. In their study conducted in the financial services industry, Larivière and Van den Poel (2005) found out that the churn propensity decreases in the total number of products possessed by the customer. In addition, customers owning risky products (including risky saving and investment products) were found to more likely defect compared to customers who do not own these (Larivière and Van den Poel, 2005). In another study also completed by Larivière and Van den Poel (2004), owning ‘high risk products in the long run’ and ‘low risk products on the fixed long run’ were found to decrease the churn probability. On the other hand in the telecommunications industry, Qi et al (2009) reported in their study that customers who buy international call service are less likely to churn and Lemmens and Croux found out a complex trend for the churn propensity: Customers, whose call usage rates increase compared to previous months, are more retention-prone whereas the customers whose corresponding rates remain constant or decrease are more likely to defect.

**Table 2.2** Hypotheses generated in churn studies and investigated in this study

Independent variables used in this study		Supporting reference(s)	Relationship with churn
Predictor category	Explanatory variable		
Customer behavior	Cumulative amount of invoiced to the customer	Jamal and Bucklin (2006)	-
	Total product ownership	Larivière and Van den Poel (2005)	+
	Product specific ownership		
	Savings account	Larivière and Van den Poel (2004)	+
	Savings and investment, high risk	Larivière and Van den Poel (2005)	+
	High-risk products in the long run	Larivière and Van den Poel (2004)	-
	Low-risk products on the fixed long run	Larivière and Van den Poel (2004)	-
	Risky products	Larivière and Van den Poel (2005)	+
	Products with capital risks	Larivière and Van den Poel (2004) <sup>1</sup>	-
	International call service	Qi et al (2009)	-
	Monetary value	Larivière and Van den Poel (2005)	-
	Usage	Lemmens and Croux (2006) <sup>2</sup>	- and +
	Customer demographics	Age	Van den Poel and Larivière (2004)
Coussement, Benoit and Van den Poel (2010) <sup>3</sup>			- and +
Larivière and Van den Poel (2005)			-
Colgate and Danaher (2000) <sup>4</sup>			+
Education level		Jamal and Bucklin (2006)	0
		Mittal and Kamakura	+
		Keaveney and Parthasarathy (2001)	-
Gender <sup>5</sup>		Colgate and Danaher (2000)	0
		Madden, Savage, and Coble-Neal (1999)	+
Marital Status <sup>6</sup>		Ahn, Han, and Lee (2006)	-
		Portela and Menezes (2010)	+
Income level	Lin, Tzebg and Chin (2010)	-	
	Jamal and Bucklin (2006)	0	
Customer-company interaction	Tenure	Madden, Savage, and Coble-Neal (1999)	-
		Coussement, Benoit and Van den Poel (2010)	-
	Cumulative number of contacts to the customer service	Tsai and Chen (2010)	-
	Number of customers served by the salesperson	Jamal and Bucklin (2006)	+
Environmental indicators	GNP per capita	Larivière and Van den Poel (2005)	-
		Van den Poel and Larivière (2004)	+
	Company versus ompetitor performance	Bolton, Kannan, and Bramlett (2000)	- and 0
		Keaveney (1995)	-

<sup>1</sup> For customers who have already acquired a savings account<sup>2</sup> If the call usage increases, churn propensity decreases. Customers with decreasing and constant call usage rates are less likely to defect.<sup>3</sup> For age groups 20-30 and >60: -; for age group 30-60: +<sup>4</sup> Colgate and Danaher investigated the age group 30-49.<sup>5</sup> In terms of male customers: binary variable that takes on 1 if male, else 0<sup>6</sup> In terms of married customers: binary variable that takes on 1 if married, else 0

#### **2.4.2. Hypotheses Related to Customer Demographics Predictors**

As pointed out in Table 2.2, existing empirical evidence is inconclusive with regard to the predictor *age*. While Van den Poel and Larivière (2004) found out a negative relationship between age and the churn attitude, Colgate and Danaher (2000) reported that churn is positively related to it and Jamal and Bucklin (2006) reported no significant relationship.

Several authors have investigated the impact of the education level on the churn behavior. Accordingly, Mittal and Kamakura (2001) argue that better educated people tend to have lower levels of retention. On the contrary, Keaveney and Parthasarathy (2001) claim that customers with higher level of education are less likely to defect. Furthermore, Colgate and Danaher (2000) found no significant relationship between the education level and the attrition behavior.

Similar to *age* and *education level* predictors, the literature is inconclusive in the impact of *gender* and *marital status* on the churn propensity. Madden, Savage, and Coble-Neal (1999) and Portela and Menezes (2010) found out that men are more likely to abandon their relationship whereas Ahn, Han, and Lee (2006) reported the opposite. With respect to marital status, Lin, Tzebg, and Chin (2010) claim that married customers are less likely to churn and Jamal and Bucklin (2006) argue no significant relationship between the marital status and the churn behavior.

As Table 2.2 indicates, authors have also investigated how income level is associated with the attrition attitude. Accordingly, Madden, Savage, and Coble-Neal (1999) found that the churn propensity decreases in the amount of income.

#### **2.4.3. Hypotheses Related to Customer-Company Interaction Predictors**

One of the foremost indicators belonging to this category is the length of the relationship with the company, also called tenure. Several authors have investigated how tenure affects the churn likelihood. As Table 2.2 points out, the empirical evidence is conclusive considering this predictor and tenure has been found out to be negatively



related to the churn propensity (Coussement, Benoit, and Van den Poel, 2010; Tsai and Chen, 2010). In other words, the longer the relationship, the less likely is the customer to churn.

Another predictor which can be listed under this category is *the number of people served by the salesperson*. Larivière and Van den Poel (2005) examined its effect on the churn behavior and found out that churn propensity decreases in it. This may imply that salespeople serving many customers are perceived more reliable, thus increasing the retention rates.

#### **2.4.4. Hypotheses Related to Environmental Indicators**

Other than the predictors which represent the information related solely to customers, environmental predictors have been also examined by some researches. In their study, Van den Poel and Larivière found out that customers display increased churn propensity if GDP index increases. This implies that customers experience higher attrition tendencies in a wealthier macroenvironment. Furthermore, Keaveney (1995) argues that in service industries, companies are more likely to remain their customers if they perform better than the competitors, as perceived by the customer.

### **2.5. Training and Validation Datasets in Churn Studies**

In this section, we will explain training and validation datasets.

#### **2.5.1. Training Datasets**

The majority of churn studies are conducted on cross-sectional training data where subjects are observed only once at the same point of time. In addition, each subject is represented in the data only once, i.e. each observation corresponds to exactly one subject. However, cross-sectional training data does not allow including time-dependency into the classification model. As also pointed out by the churn studies covered in Table 2.1, on the one hand, the historical behavior can be covered in separate

variables (Kim, Shin, and Park, 2005) (for example, credit card balance of the last month, of two months ago and of three months ago can be represented in three variables). The environmental changes (which are measured in discrete unit time) that are the same for all the subjects in the training data (such as the inflation rates or GDP index) can be introduced into models which use time-to-event information such as hazard modeling and survival analysis. On the other hand, however, time-varying variables that take on different values for each subject in the data and the environmental indicators cannot be incorporated.

Contrary to the majority, Jamal and Bucklin (2006) used longitudinal training data to perform hazard modeling, i.e. the training data comprises panel data for each subject included in the analysis. Accordingly, they split the customer duration times into customer-month observations and organized the training data where each observation corresponds to a unique customer time period,  $it$  (i.e., observation  $t$  for customer  $i$ ). This type of data makes the researchers be able to model the impact of the time-varying covariates on the churn behavior. However, it should be noted that, best to our knowledge, this type of data is only used for dynamic models such as the survival analysis, and no study has been conducted to directly claim that the use of longitudinal training data compared to cross-sectional provides improved predictive accuracy.

### 2.5.2. Validation Datasets

After developing a model for prediction, it is necessary to have in place validation processes and in general, all models are evaluated on the basis of a test sample (by definition not included in the training phase). To accomplish this, the trained model is scored on an out-of-sample test dataset, i.e. on the data which include different subjects than the training data either from the same time period (Kim et al, 2005) or from future time period (Lemmens and Croux, 2006). Another way to assess the validation is to score the model on the same set of subjects, but from future time period (Hung, Yen and Wang, 2006). In this study, the latter one will be employed. It should be also noted that in test datasets, one observation corresponds to exactly one subject.

## 2.6. Churn Prediction Techniques

Along with the fact that the customer churn has been studied in different industries (e.g. banking, insurance, telecommunications) and in different contexts (e.g. contractual vs. non-contractual settings, continuous time vs. discrete time), numerous statistical techniques are applied to predict the churn. In Table 2.1, these techniques are listed in the last column.

As illustrated in Table 2.1, the churn classification can be completed either by the use of single models, such as the logistic regression and neural networks, (Zhao et al, 2005; Jamal and Bucklin, 2006; Tsai and Chen, 2010) or it can be done by integrating multiple classifiers and developing variants of the existing algorithms (Lemmens and Croux, 2006; Burez and Van den Poel, 2007; Qi et al, 2009; Tsai and Lu, 2009). Following the taxonomy of Rokach, not only we refer minor variants of the same basic model as ensemble methods, but hybridization of models that are not from the same family are also considered to be ensemble methods. As pointed out, ensemble models mainly improve the predictive performance obtained from the use of single models (2009).

### 2.6.1. Logistic Regression

Logistic regression is an efficient tool for predicting the dependent variable known to have  $K$  different classes. To better understand the logistic regression analysis, it will be helpful to give information about odds and odds ratios in advance. The odds ratio is the ratio of the probability that the outcome is of class  $i$ ,  $i = 1, 2, \dots, K - 1$ , to the probability that the outcome is of the reference class  $K$ . The relationship between probabilities and odds is as follows:

$$O_i = \frac{P_i}{P_K}$$

$$P_i = \frac{O_i}{1 + \sum_i^{K-1} O_i} \quad \& \quad P_K = \frac{1}{1 + \sum_i^{K-1} O_i}, \text{ for } \forall i = 1, 2, \dots, K - 1$$

where  $P_i$  represents the probability of belonging to class  $i$  and  $O_i$  stands for the odds ratio. In linear regression models, the prediction outcome is inherently unbounded, which is a major problem when making predictions via them (Allison, 2003, p. 11), however the probabilities are bounded within the interval of  $[0,1]$ . Transforming the probability to the odds by using the formula given above removes the upper bound. Furthermore, taking the logarithm of the odds removes the lower bound. Setting the result equal to a linear function of the explanatory variables gives the logistic regression model which has the following form for  $i=1, \dots, K$  classes (Hastie, 2003, p. 13)

$$\log \frac{P_i}{P_K} = \beta_{i0} + \beta_i^T x, \text{ for } \forall i = 1, \dots, K - 1$$

The expression on the left hand side of the equation is usually called the log-odds or the logit. As stated by Hastie, Tibshirani, and Friedman, the log-odds equation can be solved using the following equations given below:

$$P_i = \frac{\exp(\beta_{i0} + \beta_i^T x)}{1 + \sum_{n=1}^{K-1} \exp(\beta_{n0} + \beta_n^T x)}$$

$$P_K = \frac{1}{1 + \sum_{n=1}^{K-1} \exp(\beta_{n0} + \beta_n^T x)}$$

This equation has the desired property that no matter what values are substituted for  $\beta$ 's and the  $x$ 's,  $P_i$  will always be a number between 0 and 1. To emphasize the dependence on the entire parameter set  $\theta = \{\beta_{i0}, \beta_i^T\}$ , where  $i=1, 2, \dots, K-1$ , the probabilities are denoted as  $P(y=i|X=x) = p_i(x; \theta) = P_i$  (2009, p. 119).

There are three parameter estimation methods: ordinary least squares, weighted least squares, and maximum likelihood (Allison, 2003, p. 15). Here, the maximum likelihood method will be explained further. In addition, we discuss the two-class case, since the algorithms simplify considerably.

Define a dummy variable such that

$$y_i = \begin{cases} 1, & y = i \\ 0, & \text{otherwise} \end{cases}$$

The log-likelihood for  $N$  observations (Hastie et al., 2009, p. 120) is

$$l(\beta) = \sum_{i=1}^N \{y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})\}$$

To maximize the log-likelihood, we set its derivatives to zero. These score equations are

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^N x_i (y_i - p(x_i; \theta)) = 0,$$

which are  $p+1$  equations nonlinear in  $\beta$ . The algorithm to solve these score equations is the so called Newton-Raphson algorithm which works on an iterative manner; hence the outcome of this algorithm is valid if and only if the convergence criterion is satisfied (Hastie et al., 2009, p. 121).

Following the parameter specification, the global null hypothesis should be checked and it should also be investigated to which extent the model fits the data.

On one hand, the global null hypothesis test evaluates whether the model is significant as a whole or not:

$$H_0: \beta^T = 0$$

$H_a$ : at least one  $\beta$  is different than zero

The likelihood statistics to test the global null hypothesis are likelihood ratio (LR) test, Wald's test, and score test. Under mild assumptions, these three statistics have an asymptotic chi-squared distribution. If “Pr > ChiSq” is less than or equal to the significance level, then  $\beta \neq 0$ .

On the other hand, measures of goodness-of-fit typically summarize the discrepancy between observed values and the values expected under the model in question. One common measure used in the logistic regression analysis is the Hosmer-Lemeshow goodness-of-fit test. Hisarcıklılar (2004) mentions in her lecture notes that the hypothesis pertinent to this set is:

$$H_0: E[Y] = \frac{\exp(\beta_{i0} + \beta_i^T x)}{1 + \exp(\beta_{i0} + \beta_i^T x)}$$

$$H_a: E[Y] \neq \frac{\exp(\beta_{i0} + \beta_i^T x)}{1 + \exp(\beta_{i0} + \beta_i^T x)}$$

In order to calculate this test statistics,

- ✓ sort the estimated event probabilities (also called fitted values) in increasing order
- ✓ group the fitted values into  $c$  classes of roughly equal size ( $c$  is between 6 and 10)
- ✓ compute the observed and expected number of events for each group
- ✓ perform a chi-squared goodness-of-fit test which is as follows:

$$\chi^2 = \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  and  $E_i$  is the total number of observed and estimated event outcomes respectively in the  $i^{\text{th}}$  group (2009). If the test statistics is less than or equal to the significance level, the null hypothesis, that there is no difference between the observed and predicted values of the dependent, is rejected. In other words, if chi-squared goodness-of-fit is NOT significant, then the model has adequate fit (Hisarcıklılar, 2004).

## 2.6.2. Classification and Regression Trees

As mentioned in the lecture notes of George (2010), tree-based methods are classification techniques which partition the feature space into small regions and fit a tree model in each one. They are conceptually simple yet powerful. In this review, two types of trees will be expounded: regression trees and classification trees to be abbreviated by CRT.

### 2.6.2.1. Regression Trees<sup>1</sup>

Assume that the data consists of  $p$  explanatory variables and a dependent variable, for each of  $N$  observations. Consistent with the purpose of the decision tree, namely the

---

<sup>1</sup> This section is prepared applying to Regression Trees of Hastie et al (2009, p. 307).

binary recursive partitioning of the feature space, the corresponding algorithm designates both the splitting variables and split points, together with the shape of the tree.

Starting with all of the data and considering a splitting variable  $j$  and split point  $s$ , the greedy algorithm to undertake the recursive binary partitioning defines the pair of half-planes as

$$R_1(j, s) = \{X|X_j \leq s\} \text{ and } R_2(j, s) = \{X|X_j > s\}.$$

Following this, the objective is to find out the splitting variable  $j$  and split point  $s$  that solve

$$\min\{\min[\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \sum_{x_i \in R_2(j,s)}(y_i - c_2)^2]\}.$$

For any choice of  $j$  and  $s$ , the inner minimization is guaranteed with

$$\hat{c}_1 = \text{ave}(y_i|x_i \in R_1(j, s)) \text{ and } \hat{c}_2 = \text{ave}(y_i|x_i \in R_2(j, s))$$

where  $\hat{c}_1$  and  $\hat{c}_2$  correspond to the responses -defined as constants- in the first and second regions respectively, and  $y_i$  denotes response  $i$ .

The outer minimization is easily obtained by scanning through all the inputs (splitting variables) where the optimal split point  $s$  for each variable is founded very quickly.

After both the splitting variable and split point have been determined, the feature space has been divided into two regions and the above procedure is repeated on all of the resulting regions, independently from each other. However, the problem is to identify how large the tree should be. This decision is important to be made because a large tree may imply overfitting whereas a small tree might not represent the data well. One way to choose the optimal tree size is to stop the partitioning procedure if the decrease in sum of squares does not exceed a predetermined threshold. Another strategy, which is preferred over the first one, is “growing a large tree and tree pruning”. In this second strategy, firstly a large tree  $T_0$  is grown where the splitting procedure terminates when some pre-specified node size has been achieved. Following, this large tree is pruned by applying to the cost-complexity pruning.

Let's define a subtree  $T \subset T_0$ , any tree which can be obtained by pruning  $T_0$ . Assume that  $m$  denotes the terminal nodes, with  $m$  representing region  $R_m$ . Suppose further that  $|T|$  is the number of terminal nodes in subtree  $T$  and  $N_m$  is the total number of observations in region  $R_m$ . As mentioned beforehand, the optimal estimated constant response value for the region  $R_m$  is

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i.$$

Moreover, the impurity function, where the impurity is based on the squared error, is defined as follows:

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2.$$

Using these two equations, the cost complexity criterion is defined as

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|.$$

The idea disguised behind the cost complexity criterion is to find, for each  $\alpha$ ,  $T_\alpha \subset T_0$  to minimize  $C_\alpha(T)$ . Here, the tuning parameter  $\alpha$  plays a balancing role and stands for the trade-off between the complexity and goodness of fit to the data.

The above explanations are valid for general regression tree applications. In our context, we apply to M5P regression tree in WEKA 3.6 which is invented by Quinlan (1992) and improved by Wang (1997).

#### 2.6.2.2. Classification Trees<sup>2</sup>

The classification tree differs from the regression only slightly: the response value  $k$  takes integer values  $1, 2, \dots, K$  where the target is to classify the observations among the  $K$  classes. The classification procedure discussed in the previous section for the binary partitioning of the feature space remains almost the same, where the sole change pertains to the impurity function which is to be defined in three alternative ways:

$$\checkmark \text{ Misclassification error: } \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.$$

---

<sup>2</sup> This section is prepared by applying to Classification Trees of Hastie et al (2009, p. 308).



- ✓ Gini index:  $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk})$ .
- ✓ Cross-entropy or deviance:  $\sum_{k=1}^K \hat{p}_{mk} \log_2 \hat{p}_{mk}$ .

Here,  $N_m$  and  $R_m$  are as defined in Section 2.3.2.1 where  $\hat{p}_{mk}$  denotes the proportion of class  $k$  observations in node (region)  $m$ , formulated as

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

and  $k'$  is the set of classes which does not include the  $k^{\text{th}}$  class.

The working principle of the greedy algorithm and the cost complexity criterion does not change here. The use of Gini index or cross-entropy provides the analyst to be more sensitive to changes in node probabilities, compared to the misclassification error.

The above information holds for general classification tree applications. In our context, we apply J48 in WEKA 3.6 that is introduced by Quinlan (1993).

### 2.6.3. Survival Analysis and Proportional Hazard Modeling<sup>3</sup>

Survival analysis is a statistical method which models the occurrence and timing of an event with the objective of establishing both descriptive and predictive models. As the definition suggests, the survival analysis differs from the previously explained procedures, logistic regression and CRT, in including time concept in model generation.

Before getting into the depth, it would be beneficial to define key concepts pertaining to survival analysis. An **event** is something that terminates an episode, which can be thought as death in a mortality study, birth in a population growth study, churn in a customer relationship management study, etc. **Failure time** for an observation is defined as the time when that specific observation experiences the event. **Censoring is** when an observation is incomplete due to some random cause. The commonest form of censoring is right censoring which is the case if (i) subjects followed until some time, at which the event has not yet occurred, do not further take part in the study, and (ii) the

---

<sup>3</sup> This section is prepared applying to Klein and Moeschberger (2003).

study ends while the subject survives. **Truncation** is the concept used to describe the condition that only those individuals whose lifetimes lie above or below certain value or in a certain interval are to be observed.

### 2.6.3.1. Survival Modeling

Survival analysis measures the probability that an individual will survive given that it lived until the measuring time. The times at which the events occur (also called failure times) are assumed to be realizations of some random process. In other words, the failure time,  $T$ , is a random variable with a probability distribution.

Denote the time elapsed since the starting point and the probability density function (pdf) of the failure time,  $T$ , by  $t$  and  $f(t)$ , respectively. Following, the cumulative density function (cdf) of variable  $T$  is denoted as

$$F(t) = P(T \leq t)$$

with  $F(t)$  referring to the estimated unconditional failure probability from the starting point to the elapsed time  $t$ . In survival analysis, though, it is common to work with the survivor function which is defined as follows:

$$S(t) = 1 - F(t) = P(T > t)$$

As the formula suggests, the survivor function  $S(t)$  describes the estimated unconditional probability of survival up to the elapsed time  $t$ . Knowing the mathematical relationship between pdf and cdf, the latest two equations reveal the following relationships:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

Together with predicting the survival probability, the hazard function also plays a significant role in the survival analysis and is defined as

$$\lambda(t) = \lim_{\Delta t \rightarrow \infty} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{dS(t)}{S(t)dt}$$

denoting the instantaneous risk that the event will occur at time  $t$ . Then, rewriting the above equation as

$$\lambda(t)dt = -\frac{dS(t)}{S(t)}$$

and integrating both sides from 0 to  $t$  yields

$$S(t) = e^{-\int_0^t \lambda(x)dx}$$

Let further  $H_t$  denote the cumulative hazard such that

$$\Lambda(t) \equiv \int_0^t \lambda(x) dx$$

Then, the survival probability until time  $t$  can be rewritten as

$$S(t) = \exp(-\Lambda(t))$$

### 2.6.3.2. Proportional Hazard Model

Proportional hazard model, also known as Cox regression model, dominates the field of the dynamic survival models. The general form of the hazard model is

$$\lambda_Z(t) = \lambda_0(t)C_Z(t)$$

where  $\lambda_0(t)$  denotes a baseline hazard function,  $Z(t) = [Z_1(t), \dots, Z_p(t)]^t$  denotes a set of variables at time  $t$  which may affect the survival distribution, and  $C_Z(t)$  is a multiplier specific to the subjects with the set  $Z(t)$  of variables. Furthermore, for its convenience, the proportion rate  $C_Z(t)$  is considered to be an exponential function with base  $e$ . Based on this observation, the Cox regression model is defined as

$$\lambda_Z(t) = \lambda_0(t) \exp \left[ \sum_{k=1}^p \beta_k Z_k(t) \right]$$

with  $\beta^t = [\beta_1, \dots, \beta_p]^t$  denoting the set of parameter estimates, also known as the Cox regression coefficients. These coefficients are computed by maximizing the log-likelihood function via the Newton-Raphson algorithm and there exist different approaches to constitute the likelihood function: exact likelihood, Breslow likelihood, Efron likelihood, and discrete likelihood. For detailed information, please apply Moeschberger and Klein (2003).

Cox regression is a semi-parametric analysis; where it does not have the objective of estimating the underlying survivor function, but it aims at explaining how and to which extent the predictor variables have an impact on the hazard rate. However, after fitting the regression parameters to the sample data, it would be interesting to find out the survival or failure probability of a new individual who is not a part of the sample data. In accordance with this desire, Breslow suggests an estimator for the cumulative baseline hazard rate, hence for the baseline survival function, as follows:

After the proportional hazards model has been fitted to the sample data and the parameter estimates for the covariates and the estimated covariance matrix have been obtained by maximizing the partial likelihood function, let  $t_1 < t_2 < \dots < t_D$  denote the ordered failure times and  $d_i$  be the number of failures at time  $t_i$ . Regarding this setting, the cumulative baseline hazard rate is estimated as

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{j \in R(t_i)} \exp(\sum_{k=1}^p b_k Z_{jk}(t))}$$

Accordingly, the baseline survival function is given by

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$$

where the baseline values of the covariates' set are set to zero.

As stated by Moeschberger and Klein, to estimate the survival function for an individual whose covariate vector corresponds to  $Z^*(t)$ , the following estimator, offered by Breslow, can be used:

$$\hat{S}(t|Z = Z^*(t)) = \hat{S}_0(t) \exp(b'Z^*(t))$$

### 2.6.3.3. Cox Regression Diagnostics

Cox regression should also be assessed in terms of the following criteria: (i) the global null hypothesis, (ii) the proportionality assumption, and (iii) the overall fit of the model.

The statistics in order to perform the so-called global null hypothesis test are likelihood ratio test, Wald's test and score test. As stated in SAS help files (Section

PHREG), under mild assumptions, all these three likelihood statistics have an asymptotic chi-square distribution with  $p$  degrees of freedom where  $p$  corresponds to the dimension of  $\beta$ , the vector of parameter estimates. If the computed score exceeds the corresponding benchmark value to read from the chi-square distribution table, at least one parameter is said to differ from  $\beta_0$ .

One important assumption of the Cox regression model is the proportionality of the hazard rates of individuals with distinct values of a predictor variable. As introduced by Harrel and Lee (1986), checking the proportional hazard assumption is accomplished by finding the correlation between Schoenfeld residuals (Schoenfeld residual is computed for each individual who failed and for each covariate, and is defined as the covariate value for the individual that failed minus its expected value (Klein and Moeschberger, 2003)), for a particular covariate and failure time order. The null hypothesis for this test is  $H_0: \rho = 0$ , with  $\rho$  denoting the correlation coefficient. If the proportional hazard assumption is met, it is expected to have large p-values for the corresponding test, i.e. the correlation coefficient does not significantly differ from zero given the significance level  $\alpha$ .

To assess the overall fit of the Cox regression model, Cox-Snell residuals are used. The Cox-Snell residuals are defined as

$$r_j = \hat{\Lambda}_0(T_j) \exp\left(\sum_{k=1}^p Z_{jk}(t)\beta_k\right), j = 1, \dots, n$$

Here,  $\hat{\Lambda}_0(t)$  is the Breslow's estimator for the baseline hazard at time  $t$ ,  $j$  is the individual with  $n$  denoting the total number of individuals,  $T_j$  is the failure time of the individual  $j$ ,  $\mathbf{Z}$  is the set of variables,  $p$  is the total number of covariates, and  $\beta$  is the set of parameter estimates. The assessment procedure is as follows: If the model is correct, the collection of  $r_j$ 's should be a sample from a unit exponential. Plot of  $r_j$  vs. Nelson-Aalen estimator of the cumulative hazard rate should have a 45-degree slope, where the Nelson-Aalen estimator is defined by

$$\Lambda_{NA}(t) = \begin{cases} 0 & t < t_1, \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & t_1 \leq t. \end{cases}$$

where  $Y_i$  is the total number of individuals at risk at time  $t_i$  (the size of the set  $R(t_i)$ ). The formal test of the hypothesis of proportional hazards is based on the Wald score and the corresponding p-value. A p-value less than the significance level set in advance implies that the model fits the data sample well.

## 2.7. Evaluation Criteria

During the empirical study, several classifiers will be compared. In order to assess the accuracy of each classifier, the loss incurred by wrong predictions or the gain obtained by correct predictions should be quantified. There are numerous ways to evaluate the performance of classification models applied: percentage correctly classified (PCC), misclassification error (ME), precision, recall, false positive rate, receiver operating characteristics curve (ROC curve), area under the receiver operating characteristics curve (AUC), top-decile lift (TDL), hit rate (HR), Gini coefficient (GC) are the prominent evaluation criteria among many others.

Before starting with the AUC, it is necessary to explain some related preliminary metrics. A significant number of evaluation metrics are derived by using the confusion matrix which is illustrated below in Table 2.3 (Hastie et al, 2009, p.301). It should be remarked that the confusion matrix is prepared for one specific cut-off value. Observations whose predicted scores exceed the corresponding cut-off value are predicted to belong to the class positive, otherwise negative. The confusion matrix related accuracy measures described below are explained by applying to Burez and Van den Poel (2009).

**Table 2.3** Confusion matrix

		<b>Predicted</b>	
		Positive	Negative
<b>Actual</b>	Positive	A	B
	Negative	C	D

PCC, also known as the accuracy, and ME give the proportion of the total number of predictions that are correct and incorrect, respectively:

$$PCC = \frac{A + D}{A + B + C + D}$$

$$ME = 1 - PCC$$

The precision (P) corresponds to the proportion of the predicted positive cases that are correct:

$$P = \frac{A}{A + C}$$

The recall (R), also known as the true positive rate (TP), is the proportion of positive cases that are correctly identified.

$$R \text{ or } TP = \frac{A}{A + B}$$

The false positive rate (FP) equals to the proportion of negative cases that are incorrectly classified as positive.

$$FP = \frac{C}{C + D}$$

Although these aforesaid four criteria are applied in comparing different classifiers, Glady et al state that these measures implicitly assume equal misclassification costs. In addition, they are very sensitive to the class distribution and the choice of the cut-off value used to map the classifier output to classes (2009). This problem is tackled by employing ROC curve and AUC. As stated by Burez and Van den Poel, ROC curve is a two dimensional drawing of true positive rate versus false positive rate as its cut-off value (the discrimination threshold) is varied. The closer the ROC plot is to the upper left corner, the higher the overall accuracy of the model. Here, AUC gives the area

under this plot and it considers the individual performance of classes for all possible threshold values (2009). Theoretically, AUC is a real number between 0 and 1. If this value is smaller than or equal to 0.5, then the model generated is said not to distinguish between two classes and to be indifferent from a random model. On the other hand, values greater than 0.5 imply that the model in question can be applied for predictive purposes and the predictive performance increases in the AUC scores, i.e. when the value increases up to 1. Furthermore, it is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

Another measurement used to assess the performance in binary classification problems is top-decile lift (TDL). To employ this measure, observations need to be sorted in decreasing order based on the computed posterior scores. In churn context, it corresponds to the percentage of the 10% of customers predicted to be most likely to churn who actually churned, divided by the baseline churn rate (Ghorbani, Taghiyareh, and Lucas, 2009). Intuitively, it is to claim that the accuracy of classification models, including churn prediction models, increases in the TDL score. In addition, this evaluation criterion is particularly useful when applied for the imbalanced sample datasets where the class, for which the corresponding score is computed, constitutes only a small part of the entire data.

The last criterion used in the empirical study is the top-quartile lift (TQL). This measure is a criterion which we suggest to apply to assess the predictive performance of the churn models generated during the empirical study. The data on which churn models are trained are of the private banking customers. Considering the nature of this dataset, first, the dataset incorporates a relatively low number of customers and second, these customers are all important for the service provider based on their high portfolio sizes. When tailoring retention campaigns, not only the top 10% are targeted, but more. On this ground, TDL is expanded and TQL is created. Accordingly in the churn context, TQL measures to which extent the model built outperforms a random model in identifying churners if 25% of customers, who are ranked most likely to churn, are contacted.



## Chapter 3

### MODELING APPROACH

#### 3.1. Introduction

In this chapter, we introduce our problem, define customer churn in our context, and propose our modeling approach that is particularly based on more efficient use of customer data while training classification models. This proposed approach includes three important elements: generation of longitudinal training dataset for all customers and generation of advance churn label with multiple lag times. The first element puts forward a new way to generate the training dataset as to include time series for each customer included in the analysis. The second element suggests that the customers may not churn abruptly, thus requiring the use of advance churn labels to train churn prediction models. The third element proposes an ensemble methodology that uses the output predictions of independently trained models as input attributes.

#### 3.2. Problem Definition

In this study, we deal with the churn prediction problem in a non-contractual setting under dynamic environments and the primary objective is to detect the customers, who are likely to cease their relationship, before they do so.

This problem is difficult to handle because of the following reasons: First, under non-contractual settings, the time at which the customer becomes inactive is unobserved. The customer can churn anytime without explicitly declaring it to the company; also the customer can abruptly or gradually cease the relationship with the company. This raises the question of when to declare a churn event.

Second, not only the customer characteristics, but also environmental conditions can trigger or prevent the churn event. However traditional prediction models (using static data) cannot cover time dependency while modeling customer churn and the corresponding cross-sectional dataset generally comprises fragmentary information depending on the measurement time. This, in turn, does not allow the detection of the environmental effects.

Third, as stated by Burez and Van den Poel (2009), churn is often a rare event in service industries, but of great interest and great value; therefore constructing good models requires many positive examples. Unless the training dataset incorporates sufficient number of churn event, the models will only capture the strongest signal.

Last, it is a common fact that it takes a certain amount of time to gather the customer data from different departments of the company and prepare all the independent variables using the entire customer base, to be called as “information lead time”. On this ground, it is quite possible that the most recent information may not be available at the time of scoring.

### **3.3. Churn Definition**

In the absence of a binding contract with the company, the churn event cannot simply be defined as ‘not renewing the subscription after a specific time period’. On the other hand, a customer who does not do business with the company has churned even if the account is still open. Some customers may churn abruptly; some may gradually evolve to churn.

As Glady et al (2009) claim, most definitions of churn use the product activity of a customer and a threshold value fixed by a business rule. If the activity of the customer has fallen below the threshold, this customer is considered as a churner. Consistent with this approach and with respect to the business rule of the company that provides the data for the empirical study, we state the churn definition as follows: If a customer’s

level of activity drops below a specified threshold value (TH) and stays that way for  $n$  consecutive periods, then this customer is assumed to have churned.

### 3.4. Modeling Approach

In Section 3.2, we enumerated the challenging points that we are confronted with while developing churn prediction models in the non-contractual settings. To handle those issues, we propose the following modeling approaches:

Two of the prominent challenges to appear in the churn analysis are that churn is rare event and that time-dependency cannot be captured in the typical churn models. Typical models do only comprise the customers who are either active or churn at the measurement time, thus filtering out the customers having churned previously. In addition, since the customers are only incorporated as one observation in the dataset, it is not possible to map the dynamic variables and environmental changes into the data. To address these issues, we recommend using the customer data more effectively by generating customer-month observations instead of using only one observation per customer, as explained in detail in Section 3.4.1.

Two further difficulties while dealing with the churn prediction problem under non-contractual settings are the information lead time and the exact time to declare a customer as a churner. Since the data needs to be pre-processed before entered into the analysis and since sometimes there is delay in collecting the information from different departments of the company, the recent information may not be available. Also, some customers may abruptly churn whereas for some others this may be a gradual process. To address these issues, we propose to develop advance churn labels denoting ‘churn within  $x$  periods’,  $x = 1, 2, 3, \dots, n$ . The way the advance churn labels are generated is discussed in Section 3.4.2.

The  $n$  models each run separately for the advance churn label ‘churn within one period’ through ‘churn within  $n$  periods’ give  $n$  churn probability scores. This is similar to the output of duration models, for example Cox regression. In addition, using

*months-to-churn* label as the dependent variable, being nominal or ordinal, and running multinomial logistic regression and ordinal logistic regression on the train data also provide models from which propensity scores for within-one-period churn to within-*n*-periods churn can be obtained similarly. On these grounds, we devote Section 3.4.3 to give details about these three algorithms which will be used as benchmark models.

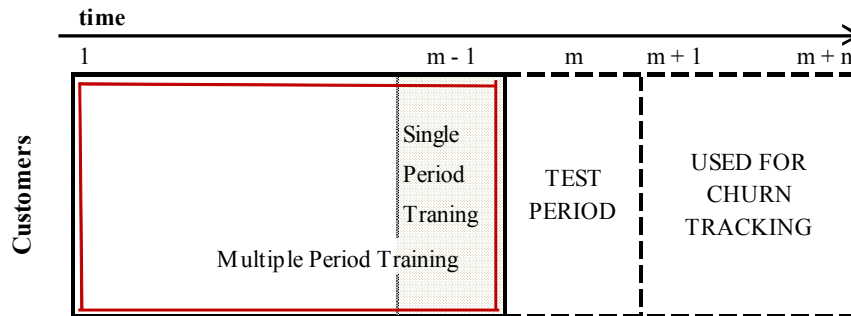
We employ classification algorithms for each advance churn label we generate. This, in turn, reveals multiple output predictions for each customer. Using these multiple propensity scores as input attributes, we propose an ensemble approach that is explained in Section 3.4.4.

### 3.4.1. Training Observations

The majority of the studies in the churn literature develop prediction models on the datasets where the observations belong to a snap shot in time. However, cross-sectional training data by its nature cannot capture the process of churn. In cross-sectional datasets, a customer is labeled only once, disregarding the fact that a customer who appears to be active this period may be getting ready to churn the next time period. In addition, cross-sectional training data also does not allow variation in the dynamic environmental variables; therefore, it cannot associate the environmental factors with change in churn propensity. Lacking the sheer numbers of churn examples, the models can only capture the strongest signal or overfit the peculiarities of the particular time period.

To address these three issues, we propose to train the churn model on a longitudinal training dataset with multiple observations per customer. Lately, this approach has been used in some churn studies (Jamal and Bucklin, 2006; Burez and Van den Poel, 2007). However, in this study, our main objective is to show that the use of longitudinal training dataset significantly improves the performance accuracy compared to the use of cross-sectional customer data, which has not been investigated yet. In our study, we will use customer-month observations with each observation containing the current month's

information about the customers' transactions. Differently from aforementioned studies, we add economic indicators and include a summary of historical customer behavior.



**Figure 3.1** MPTD and SPTD

Figure 3.1 illustrates the cross-sectional training data, which we call single period training dataset (SPTD), and the proposed multiple period training dataset (MPTD). As seen in the figure, MPTD uses each customer period as an observation containing a summary of previous periods' customer behaviors, current environmental conditions in addition to current customer behavior and characteristics. As also illustrated in Figure 3.1, the cross-sectional training dataset, SPTD, is formed to incorporate customer observations belonging to period  $m$ , the period which is nearest to the test period.

### 3.4.2. Advance Churn Labels

To tackle the information lead time, capture the process of churning and prevent possible information loss that occurs due to instant transition of a customer's label from non-churner to churner we generate  $n$  advance churn labels: within-one-period-churn label ( $W1C$ ) through within- $n$ -periods-churn label ( $WnC$ ). Tables 3.1, 3.2, and 3.3 illustrate the churn label generation. In the following examples,  $m$  refers to the length of the analysis time interval.

**Table 3.1** Labeling for a customer who churned before  $m$ 

Period	Level of Activity	W1C label	W2C label	W3C label	..	..	WnC label
1	> TH	0	0	0	0	0	0
:	> TH	0	0	0	0	0	0
k-n+1	> TH	0	0	0	:	:	1
:	> TH	0	0	0	:	:	1
:	> TH	0	0	0	:	:	1
:	> TH	0	0	1	1	1	1
k-1	> TH	0	1	1	1	1	1
k	> TH	1	1	1	1	1	1
k+1	< TH	.	.	.	.	.	.
:	< TH	.	.	.	.	.	.
:	< TH	.	.	.	.	.	.
m-1	< TH	.	.	.	.	.	.
m	< TH	.	.	.	.	.	.
:	< TH	.	.	.	.	.	.
k+n	< TH	:	:	:	:	:	:
:	:	:	:	:	:	:	:
m+n	:	.	.	.	.	.	.

Table 3.1 exhibits how the churn labels are generated for a customer that churns in some period  $k$ ,  $k < m$ . As shown in the table, the customer is active at the beginning and the level of activity drops below the threshold value in period  $k+1$  and stays that way until period  $k+n$ . Consistent with the objective of detecting churners before they do so, the *W1C* label is assigned 1 at period  $k$  (not at the period  $k+1$ ) and 0 beforehand. Also notice that once the customer churns, no observation is generated in the subsequent periods. With respect to *W2C*, the customer is recorded as a churner in periods  $k-1$  and  $k$ , and a non-churner beforehand. This is repeated with the same logic until all the labels *W3C* through *WnC* are obtained. Missing values are generated for periods following churn label.

**Table 3.2** Labeling for a customer who churned in period  $m$ 

Period	Level of Activity	W1C label	W2C label	W3C label	..	..	WnC label
1	> TH	0	0	0	:	:	0
2	> TH	0	0	0	:	:	0
3	> TH	0	0	0	:	:	0
4	> TH	0	0	0	:	:	0
5	> TH	0	0	0	:	:	0
:	> TH	0	0	0	:	:	0
$m-n$	> TH	0	0	0	:	:	0
$m-n+1$	> TH	0	0	0	:	:	1
:	> TH	0	0	0	:	:	.
:	> TH	0	0	0	:	:	.
:	> TH	0	0	1	.	.	.
$m-1$	> TH	0	1	.	.	.	.
$m$	> TH	1	.	.	.	.	.
$m+1$	< TH	.	.	.	.	.	.
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
$m+n$	< TH	.	.	.	.	.	.

Table 3.2 displays an example customer who is active at the beginning of the timeframe. As the table indicates, the level of activity for the corresponding customer falls below the threshold value in period  $m+1$  and stays that way until period  $m+n$ . Here, *W1C* is assigned 1 in period  $m$  and 0 beforehand.

The non-churner behavior is displayed in Table 3.3. Accordingly, the customer is labeled as a non-churner in the *W1C* label in all periods up to and including  $m$ . The label is missing for the subsequent time periods ( $m+1$  to  $m+n$ ), since “next  $n$  time periods” are not observable and the churn label cannot be determined. Similarly, *W2C* label is set to missing from period  $m$  onwards, *W3C* label from period  $m-1$  onwards etc. For consistency purposes the same convention is followed in Table 3.2.

**Table 3.3** Labeling for a non-churner

Period	Level of Activity	W1C label	W2C label	W3C label	..	..	WnC label
1	> TH	0	0	0	:	:	0
2	> TH	0	0	0	:	:	0
3	> TH	0	0	0	:	:	0
4	> TH	0	0	0	:	:	0
5	> TH	0	0	0	:	:	0
:	> TH	0	0	0	:	:	0
m-n	> TH	0	0	0	:	:	0
:	> TH	0	0	0	:	:	0
:	> TH	0	0	0	:	:	.
:	> TH	0	0	0	:	:	.
:	> TH	0	0	0	.	.	.
m-1	> TH	0	0	.	.	.	.
m	> TH	0	.	.	.	.	.
m+1	> TH	.	.	.	.	.	.
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
m+n	> TH	.	.	.	.	.	.

Following the generation of the binary churn label, a continuous dependent variable, periods to churn, is also proposed. This label indicates the remaining periods until the churn event. For a customer who is labeled as churning in period  $m$ , this label is assigned 1, 2, 3, etc. in periods  $m$ ,  $m-1$ ,  $m-2$ , etc., respectively. For a customer who is not churning within  $n$  periods, this label is set to ‘non-churner’.

### 3.4.3. Benchmark Models to “W1C through WnC Models”

Generating the training dataset as to obtain customer-period information at rows, panel data for each customer, allows us to create advance churn labels  $W1C$  through  $WnC$  and to acquire various advance churn notice models separately for each of them. Although these labels are dichotomous and can be modeled independently, when cogitated together, they resemble the duration models and/or the multinomial classification models. On this ground, we benchmark the accuracy of the classification models using  $W1C$  to  $WnC$  with Cox regression, multinomial logistic regression, and



ordinal logistic regression. For these methods, it is assumed that the information lead time equals zero, i.e. the most recent information is available at the time of scoring.

#### 3.4.3.1. Cox Regression (Proportional Hazard Model)

The first benchmark model used is the proportional hazard model. We construct it for the tenure variable which denotes the length of the relationship between the customer and the service provider.

In the dataset, we confront the common problem of having the duration time (i.e. number of months from the time the customer has become a customer of the bank) right censored. In addition, we can observe the customer for only a specific period of time and most of the customers come under observation only some known time after the natural time of origin. In other words, the dataset we deal with is subject to left truncation. These in turn make a hazard modeling approach a natural solution. Since we have customer-month observations in the MPTD, we are able to model the impact of the time-varying covariates on the hazard function (Here, we should remark that the majority of the exploratory variables used in the previous analyses is time-dependent, hence the generation of the dataset as to split customer observations into customer-month observations is particularly appropriate for hazard modeling).

While training the proportional hazard model, we organize the training data such that each observation corresponds to a unique customer time period (i.e. we use the longitudinal training dataset) and employ the entry and exit options of the procedure “proc phreg” in SAS 9.2.

Using the survival probability estimates of consequent tenure times, it is quite possible to compute the within-n-periods churn probability at test periods, corresponding to the churn propensity scores as obtained separately from *WIC* through *WnC* models. This is accomplished following the below steps:

- ✓ Denote  $T$  as the tenure time at the corresponding test period where  $T \in 0, 1, \dots, t_u$  with  $t_u$  denoting the greatest tenure time for which survival probability can be computed.

- ✓ Denote  $S_i(t)$  as the survival probability of  $i^{\text{th}}$  customer for tenure  $t$ .
- ✓ The probability of within- $n$ -periods churn for  $i^{\text{th}}$  customer, given that s/he has survived up to  $T$ , at a test period is computed by:

$$\frac{S_i(T) - S_i(T + n)}{S_i(T)}$$

It should be noted that the Cox regression does estimate the survival probabilities only for those tenure times when at least one event is observed. For the remaining intermediate tenure times, missing values are assigned. To tackle this problem, we decide to impute the missing survival probability scores by linearly joining the adjacent non-missing ones.

#### 3.4.3.2. *Multinomial Logistic Regression to Obtain Advance Churn Propensity Scores*

The second benchmark model to compare the performance of within- $n$ -periods advance churn models is the multinomial logistic regression (MLR). It is similar to the binary logistic regression, with the sole change of dealing with a polychotomous response value, i.e. taking more than two categories.

In MLR, the event of interest is observing a particular category out of  $K$  categories. As described in Section 2.6.1, one category serves as the reference point. For the remaining  $K-1$  categories, the MLR outputs  $K-1$  logit functions and it is important to remark that the estimated coefficients are not the same, i.e.  $K-1$  different logit functions with different coefficient estimates are generated.

The dependent variable “periods to churn” can be considered as multinomial variable because we have a total of  $n+1$  categories, where  $n$  is the same number used in the churn definition. In this approach, the below steps are followed:

- ✓ Generate the “months-to-churn” dependent variable as described in Section 3.3.
- ✓ Assign ‘non-churner’ to this variable for customers who do not churn within  $n$  periods and for non-churners.
- ✓ Use ‘non-churner’ as the reference category and run the MLR on MPTD.

This will output  $n$  different logit functions. Using these, the probability for belonging to category 1 to  $n$  is computed. In addition, the following should be noticed: “Months-to-churn =  $m$ ” denotes that the corresponding customer is going to churn between  $(m-1)^{\text{st}}$  and  $m^{\text{th}}$  periods as of the current period. In other words, the probability score,  $p_m$ , computed for the category  $m$ ,  $m \in 1, 2, \dots, n$ , equals to the probability of that the particular observation is going to churn within  $(m-1)^{\text{st}}$  and  $m^{\text{th}}$  periods. Taking this into consideration and considering that belonging to categories are mutually exclusive; the within-x-periods churn probability (WxC-P) is calculated as below:

$$\text{WxC} - \text{P} = \sum_{i=1}^x p_i$$

- ✓ After having obtained the churn probability scores, evaluate the model performance in terms of the binary dependent variables  $W1C$  through  $WnC$ .

#### 3.4.3.3. Ordinal Logistic Regression to Obtain Advance Churn Propensity Scores

The third benchmark model we employ is the ordinal logistic regression (OLR). It is similar to the multinomial logistic regression, but with slight modifications. Instead of considering the probability of an individual event, you consider the probability of that event and all events that are ordered before it.

In OLR, the event of interest is observing a particular class or less. Accordingly, the odds are modeled as follows (Adeleke and Adepoju, 2010):

$$0_j = \frac{\text{probability of belonging to class } j \text{ or less}}{\text{probability of belonging to classes greater than } j}$$

$$0_j = \frac{P(\text{class} \leq j)}{P(\text{class} > j)} = \frac{P(\text{class} \leq j)}{1 - P(\text{class} \leq j)}$$

The ordinal logistic regression model for the independent variable set  $\mathbf{X}$  is then

$$\ln(0_j) = \beta_{j0} + \beta^T \mathbf{X}$$

where  $j$  goes from 1 to the number of categories minus 1 (one category serves as the reference point).

As the above last equation indicates, each logit has the same coefficient  $\beta^T$ , but different intercept estimate. This implies that the impact of the independent variable is the same for different logit functions. In other words, as stated by Norusis, OLR assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. This is called the proportional odds assumption or the parallel regression assumption (2010). This is one of the main differences of the OLR from the multinomial logistic regression where the logit functions possess different coefficient estimates. In addition, this also can be considered as a drawback, because observations belonging to different categories are expected to behave differently, hence need to be differentiated in terms of the parameter estimates. On the other hand, fitting multinomial logistic regression model is computationally intensive, hence using it with stepwise selection procedure is not recommended for big datasets (Cherrie, 2007). The relatively quick parameter estimation for OLR is an advantage when variable selection procedures are used.

The dependent variable “periods to churn” can be considered as ordinal variable because (i) we observe  $n+1$  different response values (where  $n$  denotes the value used in the churn definition) and (ii) there is an inherit order in the response values. The OLR yields  $n+1$  class probabilities for each. In addition, it is ensured that the predicted class probability increases in the response value. In this approach, the below steps are followed:

- ✓ Run OLR on the MPTD, as illustrated in Figure 3.1, to obtain cumulative class probability scores.
- ✓ Use the probability of belonging to class 1 – namely the class which denotes that the customer is going to churn within one period of time – as the within-one-period churn probability.

- ✓ Use the cumulative probability of belonging to classes 1 or 2 as the within-two-periods churn probability, 1 or 2 or 3 as the within-three-periods churn probability, etc.:

$$WxC - P = p_x$$

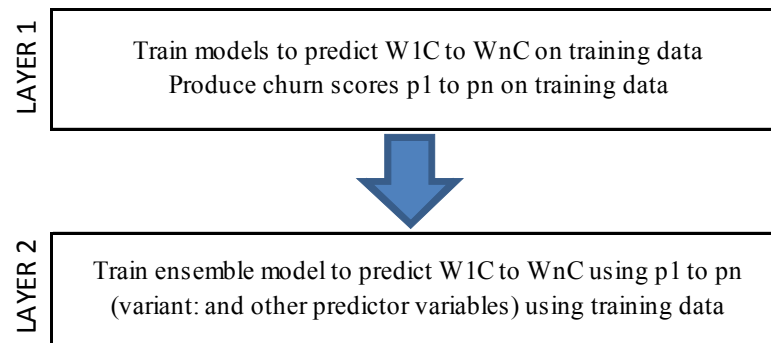
To remember, “months-to-churn =  $m$ ” denotes that the corresponding observation is going to churn between  $(m-1)^{\text{st}}$  and  $m^{\text{th}}$  periods as of the current period. In addition, the events for this dependent variable are mutually exclusive. Furthermore, it should be remembered that the OLR outputs cumulative probability scores for belonging to a category and its antecedents. On these grounds, the cumulative probability score estimated for “months-to-churn” by the OLR entirely corresponds to the within- $n$ -periods churn score, hence to be compared with the binary logistic regression models trained separately for  $WIC$  through  $WnC$ .

- ✓ After having obtained the churn scores, evaluate the model performance in terms of the binary dependent variables  $WIC$  through  $WnC$ .

#### 3.4.4. Ensemble Method Generation

Many studies have shown that ensemble methods provide significant improvement in the predictive accuracy when compared to single model use (Leigh et al., 2002; Lin et al. 2006; Lemmens and Croux, 2006). We think that the  $n$  independently trained churn models that have a different advance notice window will capture different churn patterns, and combining them in an ensemble should improve the prediction power while maintaining interpretability.

Churn behavior is the outcome of a dynamic process. The churn decision is generally not made instantaneously, but the customers gradually evolve to churn and this transition might be hidden in the customers’ transactions, behavior or environmental changes. To capture this process, we propose the following ensemble method as illustrated in Figure 3.2.



**Figure 3.2** Ensemble method overview

Figure 3.2 provides an overview of the ensemble method, which consists of two layers. In the first layer, we train a classifier for each churn label  $W1C$  through  $WnC$ . This outputs  $n$  churn scores in total, which represent the churn propensity for different multiple-periods-ahead. In the second layer, we use the  $n$  advance churn scores,  $p1$  through  $pn$ , to predict within-one-period-churn. The prediction in the second layer of the ensemble method has three variants according to the explanatory variables. The first variant includes only the advance churn scores, the second one includes these scores along with the all explanatory variables that were available to the  $WnC$  models whereas the third variant uses some selected explanatory variables along with these six churn scores. The rationale here is that the ensemble may be able to identify under which conditions (e.g. customer or environmental characteristics) the advance signals or their combinations are useful in predicting imminent churn.

In terms of the classification taxonomy proposed by Rokach (2009), this procedure can be considered as ensemble diversity such bagging or and meta-combination such as stacking, because our ensemble uses output predictions of independent classifiers trained to model churn multiple periods ahead.

## Chapter 4

### EXPERIMENTS AND RESULTS

#### 4.1. Introduction

As one of the more commonly studied areas in services research, the banking industry provides an appropriate setting for comparing churn models. In addition, by its nature, both retail and private banking constitute non-contractual settings (disregarding credit card contracts), which complicates the churn prediction. Hence, we select private banking industry in particular as the empirical population for this study. As claimed by Walfried, Manolis, and Winsor (2000) private banking customers are unique in that they have large deposits and high accounts. On this ground, it can be asserted that private banking distinguishes itself by extremely high quality service in terms of advice, execution, and responsiveness.

In this chapter, we first describe the dataset and explain the variables used and hypotheses generated. Then, we briefly explain the experimental setup. Following this, we give the empirical results under two major categories: accuracy measures and managerial insights.

In the category “accuracy measures” – where the accuracy is calculated as the AUC, top-decile lift (TDL) and top-quartile lift (TQL), we

- ✓ evaluate how the performance accuracy improves when MPTD is employed instead of SPTD while predicting next period churn and how this performance remains robust over time
- ✓ assess how the performance accuracy deteriorates when the most recent data is not available at the time of scoring

- ✓ evaluate how the additional advance churn notice models (to predict churn multiple periods ahead) perform
- ✓ benchmark the advance churn notice models to Cox regression, multinomial logistic regression, and ordinal logistic regression
- ✓ investigate whether the ensemble method provides improvement

In the category of “managerial insights”, we give the parameter estimates obtained from advance churn notice models, following the order as given above for the category “accuracy measures”, and provide interesting private banking managerial insights resulting mainly from MPTD models.

In this study, we use three different softwares. The statistical software SAS 9.2 is used for the descriptive analysis and preparation of datasets and as well as for the logistic regression analysis. The CRT analyses are performed in the open source data mining software WEKA 3.6. The accuracy computation is completed under MATLAB R2008a software with self-written code.

## 4.2. Dataset Description

For this study, private banking customers’ transaction, and demographics information is collected from one of the leading Turkish banking companies. The total dataset consists of 13,468 private banking customers and includes customer transactions from July 2008 to October 2010, corresponding to 28 months in total. The bank operates in an emerging market and the market conditions are very dynamic. The time frame of the dataset covers the financial crisis period, which took place in early 2009, and post-crisis period.

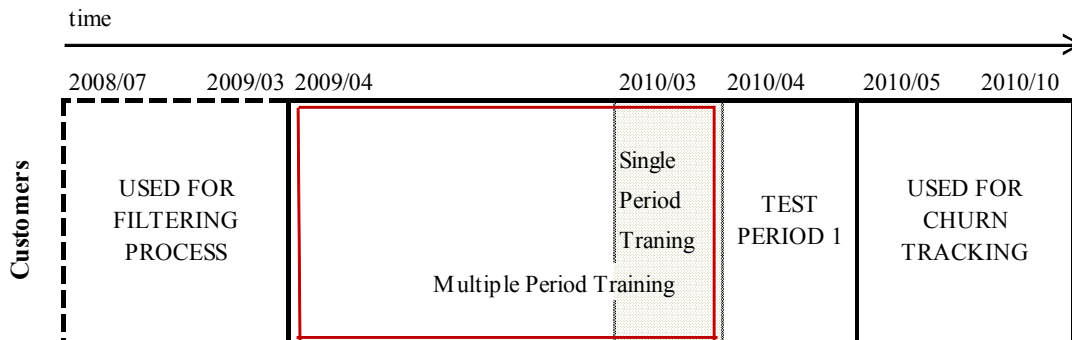
Figure 4.1 visualizes the time window of the analysis. The first ten periods of this time frame are used to filter out the inactive customers while the last six periods are used for churn tracking. Accordingly, only customers with a total portfolio size of plus 250,000 TL between November 2008 and April 2009 and with an average total portfolio size of plus 250,000 TL between July 2008 and April 2009 are decided to be



included in the modeling process, revealing only active customers. As a result, April 2009 is the first period for which we can assign the churn label.

The bank defines a customer as having churned if his total portfolio size falls below 50,000 TL and stays that way for six consecutive months. From Section 3.3, this yields a total of six dichotomous churn labels  $W1C$  through  $W6C$ . As depicted in Figure 4.1, the last six periods are only used to track the churn behavior and customers can not be labeled in these periods. Therefore, these six periods are excluded from the analysis after the customers have been labeled as churners and non-churners for the preceding periods as explained in Section 3.3.

As a result, we have 7,203 customers and 90,963 customer-month observations in total and the maximum number of periods to observe for a customer is 13. After we have generated the variables including the churn labels, we standardize the explanatory variables by setting the mean to zero and standard deviation to one.



**Figure 4.1** MPTD, SPTD, and test dataset for the next-period churn label ( $W1C$ )

As illustrated in Figure 4.1, we have two training datasets, MPTD and SPTD, and one common test dataset of future time period. The MPTD is built such that it comprises the customer observations of periods April 2009 through March 2010. On the other hand, the second training dataset, the SPTD, includes the customer observations for period March 2010 only. Using the MPTD and the SPTD, we train a model with each classifier to predict the within-one-period-churn and evaluate their accuracy on the same test dataset of future time period (TD). The training data for the classifier  $W2C$

contains only those labels that can be created with data up to but not including the period April 2010. For example, a customer who is labeled as within-one-period churner in period April 2010 is labeled as within-two-months churner in period March 2010 (namely in the last period of the training data). This implies that the information from further periods (November 2010) is used to generate  $W2C$  label in period March 2010; hence the corresponding observation is not included in the training data while modeling  $W2C$ . The same logic holds for  $W3C$  to  $W6C$ . Accordingly, Figure 4.2 summarizes which periods are used while training  $W1C$  through  $W6C$ .

The test dataset comprises the last period with churn label, namely period April 2010. Notice that both the SPTD and the TD include one observation per customer and the validation dataset is from future time period. Also notice that the SPTD is adjacent to the test time period, and hence the environmental conditions should be similar in both time periods. This, in turn, should affect the accuracy performance of SPTD positively and should provide an optimistic measure.

Period	W1C	W2C	W3C	W4C	W5C	W6C
200904	MPTD	MPTD	MPTD	MPTD	MPTD	MPTD
200905						
200906						
200907						
200908						
200909						
200910						
200911						
200912						
201001		MPTD	MPTD	MPTD	MPTD	MPTD
201002						
201003						
201003						

**Figure 4.2** Longitudinal training data to predict  $W1C$  to  $W6C$

After the two training datasets and the test dataset of future time period had been generated, the bank has provided us with further customer data belonging to periods November 2010 to February 2011. These new four periods are used to generate four

additional test datasets of subsequent future time periods. These additional test periods provide us with the opportunity for evaluating to which extent the models remain robust. Accordingly, we compute the accuracy scores not only for the classification model trained for the *W1C* label, but also for the remaining advance churn labels *W2C* through *W6C*. However, please notice that not all the six advance churn labels can be generated for the test periods, hence impeding computation of the accuracy performance for the corresponding labels. For example, in order to assign a value to *W2C* label at time  $t$ , the portfolio size information for periods  $t+1$  to  $t+7$  should be known. Nevertheless, the last period with customer information is February 2011 and we therefore are not capable of determining the within-two-periods churn state of customers at August 2010, i.e. the test period of August 2010 cannot be used to score the performance of models with *W2C* being the dependent variable. Table 4.1 illustrates which test period can be used to evaluate which churn label.

**Table 4.1** Training and test datasets for *W1C* to *W6C* while computing robustness

Churn Label	Test Periods				
	2010/04	2010/05	2010/06	2010/07	2010/08
<i>W1C</i>	✓	✓	✓	✓	✓
<i>W2C</i>	✓	✓	✓	✓	
<i>W3C</i>	✓	✓	✓		
<i>W4C</i>	✓	✓			
<i>W5C</i>	✓				
<i>W6C</i>					

The MPTD, the SPTD, and the TD1 of *W1C* consist of 84180, 6821, and 6783 observations respectively. In the SPTD and the TD1, we have only 39 and 50 churners, respectively. These correspond to a churn rate of 0.57% in the SPTD and 0.73% in the TD1. Considering the MPTD, the number of positive cases (churn events) and the corresponding churn rates for each dependent variable (*W1C* through *W6C*) are summarized in Table 4.2.

**Table 4.2** Number of positive cases for each churn label

Label	# of positive cases	churn rate
W1C	420	0.499%
W2C	773	0.999%
W3C	1049	1.489%
W4C	1252	1.969%
W5C	1394	2.460%
W6C	1455	2.928%

As the churn rates in Table 4.2 indicate, our training datasets are highly imbalanced. To equally represent the classification categories, we oversampled the training datasets by applying the ‘‘Synthetic Minority Over-sampling Technique’’ (SMOTE), which has been successfully used in churn prediction problems in the literature, e.g.; Kumar and Ravi (2008). As stated by Chawla et al. (2002), ‘‘the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the  $k$  minority class nearest neighbors’’. Our implementation uses the five nearest neighbors and the amount of oversampling needed is determined so as to have a 1:1 ratio between non-churn and churn events. We call the oversampled MPTD and SPTD as MPTD-S and SPTD-S, respectively. This procedure is repeated separately for each dependent variable  $W1C$  through  $W6C$ . Please notice that no oversampling is employed in the test datasets. This is because we will comparatively evaluate the performance of each model generated, either on oversampled or non-oversampled training data.

### 4.3. Predictors and Hypotheses

In this section, we are going to present the variables used in this study and the related hypotheses which presume some relationship and its direction of the impact of the corresponding variables on the churn attitude. Based on the general variable categorization and the nature of our dataset, we decide to split up our variables into four

main groups: customer behavior predictors, customer-company interaction predictors, customer demographics predictors, and environmental indicators.

### 4.3.1. Customer Behavior Predictors and Related Hypotheses

The following subsections discuss the customer behavior covariates considered in this study and the hypotheses which are meant to be significantly related to the churn behavior of private banking customers.

#### 4.3.1.1. Portfolio Information

In this sub-category, we include four continuous variables -*the total portfolio size at the month end, the percentage change in the portfolio size, ratio of the shared portfolio size to the total portfolio size and ratio of the shared and owned portfolio size to the total portfolio size*- and the following four dummy variables:

- Portfolio Bucket 1 = 1 if the customer's month-end portfolio size is less than 500.000 TL, 0 otherwise
- Portfolio Bucket 2 = 1 if the customer's month-end portfolio size is in [500.000, 1.000.000) TL, 0 otherwise
- Portfolio Bucket 3 = 1 if the customer's month-end portfolio size is in [1.000.000,5.000.000) TL, 0 otherwise
- Portfolio Bucket 4 = 1 if the customer's month-end portfolio size is greater than or equal to 5.000.000 TL, 0 otherwise

Parallel with the finding of Lemmens and Croux (2006), the decline in the portfolio size and belonging to a lower-leveled portfolio bucket may be indicative for the churn behavior. The following hypotheses are generated for this sub-category:

Hypothesis 1: The decline in the total portfolio size increases the churn propensity.

Hypothesis 2: The churn propensity is positively related to an increase in the percentage change of the portfolio size.

Hypothesis 3: If a customer's portfolio size falls below 500.000 TL (if the dummy variable *Portfolio Bucket 1* is assigned 1), then the customer is more prone to churn.

Hypothesis 4: If a customer's portfolio size exceeds 5 million TL (if the dummy variable *Portfolio Bucket 4* is assigned 1), then the customer is less prone to churn.

#### 4.3.1.2. Account Information

In the banking industry, it is allowed that the customers open accounts which they share with other people. On this ground, we ponder that the account information related variables might explain the churn behavior. This sub-category contains the following predictors: number of all accounts a customer possesses, number of people who a customer shares account with, ratio of the number of shared-accounts to the number of all accounts, ratio of the number of non-shared accounts to the number of all accounts, and ratio of the number of shared-and-owned accounts to the number of all accounts. The literature lacks conclusive evidence for the impact of these predictors on the churn behavior; however we consider the following hypotheses to influence the churn decision.

Hypothesis 5: The more the number of shared accounts, the more loyal is the customer.

Hypothesis 6: The churn behavior is negatively related with the percentage of shared accounts and positively related with the percentage of shared and owned accounts.

Hypothesis 7: The more the number of all accounts, the more loyal is the customer.

Hypothesis 8: The churn propensity decreases in the number of customers whom a customer shares account(s) with.

Further, on the one hand we hypothesize that the more accounts a customer possesses accounts, the longer she or he is likely to remain a customer. On the other hand, we suppose that the more a customer has single accounts, i.e. she or he does not share accounts, the more power he or she possesses on the churn decision. Based on this contradiction, the next hypothesis for this sub-category is generated as follows:

Hypothesis 9: We only hypothesize an impact of the number of single accounts and percentage of single accounts on the churn attrition without a priori expectation of the direction of the impact.

#### 4.3.1.3. Asset Return Information

In the banking industry, especially for the private banking customers, the degree to which extent a customer gains profit from his or her investments plays significant role in the churn and switching behavior. On this ground, we decided to include the following five variables in this study: last one month return (in percentage), last three months return (in percentage), last six months return (in percentage), last twelve months return (in percentage), and twelve-monthly relative return which is computed as follows:

$$\text{Relative Return}_{ij} = \frac{(1 + \text{Last Twelve Months Return}_{ij})}{(1 + \text{Average Twelve Months Return}_j)}$$

where

$$\text{Average Twelve Months Return}_j = \frac{\sum_i \text{Last Twelve Months Return}_{ij}}{\# \text{ of active customer at period } j}$$

with  $i$  and  $j$  denoting the customer and the period (month), respectively.

Best to our knowledge, the relationship between asset return and the likelihood to churn were not investigated yet. Nevertheless, we presume that the churn decision is shaped by a win-win relationship:

Hypothesis 10: The churn attrition is negatively related with an increase in the asset return.

#### 4.3.1.4. Product-specific Ownership

The bank identified 42 different financial instruments held by the customers (we call them products) and we grouped them into 14 asset groups. The instruments and their

group classifications will not be shared due to confidentiality reasons. For each specific product, we created two dummy variables: whether the customer possesses it in the corresponding month and whether the customer has ever obtained it over his or her lifetime until the corresponding month. Similarly, we generated an indicator variable expressing ownership of the asset groups at the end of the corresponding month. Given the different characteristics of products and the inconclusive evidence in the literature, we only hypothesize that product-specific ownership and asset group ownership might influence the customer's decision to leave the company.

Hypothesis 11: Owning product X influences the churn decision.

Hypothesis 12: The churn behavior is affected by ever-use of product X.

Hypothesis 13: Obtaining product group Y influences the churn decision.

#### 4.3.1.5. Total Product Ownership

For this sub-category, two dummy variables were generated: *the total number of instruments currently used by the customer in the corresponding month* and *the total number of instruments ever used by the customer until the corresponding month*. Huber, Lane, and Pofcher (1998) revealed in their study that the customer retention in banking industry increases in the number of products he or she owns. In addition, Van den Poel and Larivière (2004) showed in their study that in financial services industry, an increase of one additional product lowers the switching likelihood with 99.9%. On these grounds, the hypotheses we generate with respect to this sub-category are as below:

Hypothesis 14: An increase in the total number of products currently used is negatively related to the attrition decision.

Hypothesis 15: An increase in the total number of products ever used is negatively related to the attrition decision.



#### 4.3.1.6. Scorecard

For each customer-month observation, we prepared a scorecard based on the customer's monthly portfolio holdings where these scorecard variables were recomputed every month. After having specified products and product groups, we first calculate the overall portfolio weight of the asset group in the customer's portfolio at the end of the month as

$$w_i = \frac{(\text{Value of investment})_i}{\text{Value of overall portfolio}} \text{ where } i = 1, 2, \dots, 14$$

Next, we calculate "love" scores that are the total portfolio weights of instruments with a certain shared common feature (Sayman and Demiroğlu, 2011). For example, customers that allocate a large share of their investments into interest paying instruments such as deposits and bonds are considered to be interest lovers. On this ground, we create five distinct love scores: foreign exchange, gold, cash, easy-to-sell, and interest. Due to confidentiality reasons, it will not be explained how the financial instruments (products) are mapped into five love groups. The love scores are computed similarly as given in the above formula.

Last, we examine the monthly risk attitude of each customer. We start by assigning a risk score to each of the 42 financial instruments where it ranges between 1 and 6. Next, we create one discrete variable, "risk score", and one continuous variable, "average risk score". Each month, the risk score predictor is assigned the risk score of the riskiest financial instrument that a customer possesses in his or her portfolio at the corresponding month. The average risk score predictor is computed as the weighted average of the instruments' risk scores where the weights correspond to the weight of the financial instruments in the customer's total portfolio size.

The broad literature review we conducted does not reveal any study which uses this type of variables as potential churn predictors. Nevertheless, we summarize our expectations through the following hypotheses:

Hypothesis 16: The churn propensity decreases in the risk score (*variant*: average risk score).

Hypothesis 17: The churn propensity increases in the easy-to-sell love score (*ets*score).

Hypothesis 18: The churn propensity increases in the cash love score (*cash*score).

#### 4.3.2. Customer-Company Interaction Predictors and Related Hypotheses

Given that the private banking customers are assigned a customer representative and that they are differently treated compared with other clientele, the interaction between them and the service provider is expected to play a significant role in the churn decision. On this ground, the following variables were decided to be included in the classification analysis: the number of customers served by each customer representative, the number of customers served in each branch, two continuous variables denoting the tenure of the client both as a banking customer and private banking customer, and four dummy impact variables which are explained as follows:

The first two impact variables generated are customer representative historically significant positive impact and negative impact on his or her clients' attrition. To form the variables, the below steps are followed:

- ✓ Given only the active customers who survived the filtering process described in Section 4.2, we computed the average churn rate and the corresponding binomial standard deviation for each customer representative at each period. It should be kept in mind that it is calculated based on the historical information at each time period.

$n_{ij}$  ... the cumulative number of customers served by  $i^{\text{th}}$  customer representative from period 1 to  $j$

$r_{ij}$  ... the average churn rate for the  $i^{\text{th}}$  customer representative from period 1 to  $j$

$stdev_{ij}$  ... the binomial standard deviation corresponding to the computed churn rate  $r_{ij}$

$$\text{stdev}_{ij} = \sqrt{(r_{ij} * (1 - r_{ij})/n_{ij})}$$

- ✓ Following the computation of the binomial standard deviation, the upper and lower bounds of the customer representative average churn rate were computed at significance level 0.05.

$$\text{upper bound } r_{ij} = r_{ij} + z_{0.025} * \text{stdev}_{ij}$$

$$\text{lower bound } r_{ij} = r_{ij} - z_{0.025} * \text{stdev}_{ij}$$

- ✓ After the upper and lower bounds of average churn rates had been obtained, the corresponding dummy variables were generated as follows:

$$\text{RM\_POS\_IMPACT} = \begin{cases} 1, & \text{if } \hat{r}_j > \text{upper bound } r_{ij} \text{ and } n_{ij} \geq 5 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{RM\_NEG\_IMPACT} = \begin{cases} 1, & \text{if } \hat{r}_j < \text{lower bound } r_i \text{ and } n_{ij} \geq 5 \\ 0, & \text{otherwise} \end{cases}$$

with  $\hat{r}_j$  and RM denoting the average churn rate of period  $j$  and customer representative, respectively.

The remaining two dummy variables, branch positive impact and branch negative impact, were prepared similarly as explained above for the customer representative impact predictors.

Some researchers investigated the impact of firm-customer interaction on the attrition behavior. Larivière and Van den Poel (2005) revealed in their study that the number of financial services customers served by the salesperson is inversely proportional to the churn attitude and they assumed that this impact is based on the impression that intermediaries who serve fewer customers are less reliable when compared with other salesperson with a wide clientele base. In addition, in various studies, tenure was ascertained to be negatively related with the churn and switching behavior (Burez and Van den Poel, 2007; Buckinx and Van den Poel, 2005; Coussement, Benoit and Van den Poel, 2010; Hung et al, 2006). Parallel with these

findings, the following hypotheses were presumed to be possibly validated in the classification analysis:

Hypothesis 19: An increase in the number of customers served by a customer representative (variant: in a branch) decreases the churn attrition.

Hypothesis 20: The churn propensity increases in the historical customer representative negative impact (variant: branch negative impact).

Hypothesis 21: The churn propensity decreases in the historical customer representative positive impact (variant: branch positive impact).

Hypothesis 22: The churn probability decreases in the length of the relationship that a customer sustains with the service provider.

### **4.3.3. Customer Demographics Predictors and Related Hypotheses**

The extensive literature search revealed that many studies in the field of churn prediction use demographic variables as potential predictors of the churn event. The following paragraphs discuss the demographic variables considered in this study.

#### *4.3.3.1. Age*

This variable is among the predictors which are included in the classification problems when demographics variables are covered. Consistently, we also decided to include age and created one continuous variable denoting the age of the customer in terms of years. As shown in Table 2.2, the existing churn literature is inconclusive in the direction of this variable on the churn attitude. We therefore consider age to possibly affect the churn decision, without a priori expectation of the direction of the relationship.

#### 4.3.3.2. Gender

Along with the age variable, gender information is also commonly used in the churn classification problems. On this ground, we entered two dummy variables into our classification models:

$$\text{GENDER\_MALE} = \begin{cases} 1, & \text{if the customer is male} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{GENDER\_MISSING\_FLAG} = \begin{cases} 1, & \text{if the gender information is missing} \\ 0, & \text{otherwise} \end{cases}$$

However, as Table 2.2 reveals, the literature lacks unanimity for the impact of this predictor on the churn behavior. While Portela and Menezes (2010) found out in the telecommunications industry in Portugal that women have a higher probability to stay with the company, Ahn, Han, and Lee again in the telecommunications industry in Korea (2006) on the contrary demonstrated that women have a higher probability to quit. As such, we are unable to formulate a relationship between gender and the customer's churn attitude.

#### 4.3.3.3. Nationality

The customer of the bank that provided us with the necessary data includes foreign customers; we therefore included the following two variables in our analysis:

$$\text{NAT\_TC} = \begin{cases} 1, & \text{if the customer is Turkish} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{NAT\_MISSING\_FLAG} = \begin{cases} 1, & \text{if the nationality information is missing} \\ 0, & \text{otherwise} \end{cases}$$

Best to our knowledge, the nationality information was not considered as a possible predictor in the previous churn studies for the financial services industry. In addition, considering the vast number of variables created for this study, we believe that the first covariate aforementioned will not have an impact on the churn behavior. On the other

hand, along with the gender missing flag, the second predictor given above might influence the churn decision because the descriptive statistics performed before classification analysis showed that the average churn rate for customers whose gender and nationality information are missing is higher than the of customers with full information.

#### 4.3.3.4. Education

The level of education is another predictor type used in the classification analyses by many researchers. Nevertheless, as Table 2.2 points out, the literature is inconclusive in the direction of the relationship between the education level and the retention. The data provided by the bank includes six different categories: not known, elementary school, high school, associate degree, bachelor degree, and master or doctorate degree. Instead of creating five dummy variables per observation in order to know to which categorical group he or she belongs in the corresponding month, we performed descriptive analysis and computed the average churn rate and its confidence interval for each education group. Based on this analysis, we decided to create only one dummy variable, *edu\_college*, expressing whether the specific customer has an associate or higher degree in the corresponding period (the value equals 1) or not (the value equals 0).

The decision rules generated by Lin, Tzeng, and Chin (2010) to explain the non-churner behavior in credit card industry indicate that education groups other than graduate school, university/college, high school, and under junior high school constitutes one node in the path. The same education groups are also included in the decision rules path created for the involuntary churn. Moreover, the set of decision rules to explain the voluntary churn cover the high school and university school graduates. On this ground, we are unsure about the existence and, if there is any, direction of the education's impact on the churn propensity.

#### 4.3.3.5. *Employment Type*

The employment variable includes 11 categories. As explained above for the education, we performed descriptive analysis to compute the average churn rates and their confidence interval for each employment category and, based on the results, we decided to create two dummy variables as below: *emp\_own\_business* expressing whether the customer has his or her own business in the corresponding month or not (if yes, the value equals 1, 0 otherwise) and *emp\_employed* expressing whether the customer is employed or not (if yes, the value equals 1, 0 otherwise). The literature lacks any relationship between these and the churn propensity. We therefore could not develop a precise hypothesis on the relationship between employment and the attrition.

#### 4.3.3.6. *Marital Status*

One dummy variable was generated expressing whether the customer is married in the corresponding month (the value equals 1) or not (the value equals 0). In their study, Lemmens and Croux (2006) investigated the relationship between the marital status and the churn behavior in the wireless telecommunications industry in U.S. and found none. On this ground, we expect the customer's being married or not to have no impact on the churn decision

Hypothesis 23: The marital status does not affect the churn behavior of the private wealth customers.

#### 4.3.3.7. *Customer Segmentation*

The customer segments generated by the bank for the private banking customers will not be shared in order not to expose the company into unfair competition. For this variable, we generated one dummy variable, *sbu\_type\_nm\_uh*, expressing whether the customer belongs to the corresponding segment in the corresponding period (the value equals 1) or not (the value equals 0). We hypothesize that belonging to this segment does not influence the churn propensity.

Hypothesis 24: Being grouped into segment x does not have an impact on the churn attitude.

#### 4.3.4. Economic Indicators and Related Hypotheses

Churn is a process and the customers do not make abrupt decisions when terminating their relationship with the companies. Especially for the financial services company, both micro- and macroenvironmental factors constitute an important part of the customer churn behavior. Considering the majority of churn studies, we observe that the economic indicators are not included, mostly due to the cross-sectional training datasets. Our modeling approach on the other hand, the use of panel data for each customer, allows the dynamic variables to be directly included. On these grounds, we generated the following variables: *consumer confidence index*, *twelve-monthly inflation for the producer price index*, *EUR/TL buying exchange rate*, *USD/TL buying exchange rate*, *IMKB100 index*, *TRYOND\_line*, *TRIMT\_rr\_line*, *MSCIEF*, *DJI\_SMA*, *TRIYT\_rr\_line\_minus\_TRIMT\_rr\_line*, *atipi\_ykb\_vs\_market*, *btipi\_ykb\_vs\_market*. The latter two predictors denote how the bank product returns of Type A and B compare to the market indices in the corresponding month. These two variables are positive continuous and values greater (less) than 1 imply that the bank product is more (less) lucrative for the customers. On this ground we expect the churn propensity to decrease in these two predictors. For the remaining variables, we could not formulate specific relationships between them and the attrition behavior.

Hypothesis 25: The more the bank makes their customers earn compared to the market, the less prone are they to churn.

#### 4.4. Experimental Setup

The experiments performed for this study follow the below steps:

- ✓ The primary objective is to assess whether the way the dataset is generated is an important design parameter, regardless of the base classifier and regardless



of the use of oversampling. Accordingly, we train logistic regression and classification tree models on SPTD, MPTD, SPTD-S and MPTD-S for the churn label  $W1C$  and test it on the most adjacent test period.

- ✓ Next, it is also aimed to show that the models generated on longitudinal training data continues to outperform the models obtained from the cross-sectional training data. Accordingly, we test the models trained in the first step for the successive test periods.
- ✓ In addition, we investigate how the  $W1C$  models perform when the customer data is not available at the time of scoring.
- ✓ Further, we also have the purpose to verify the use of advance churn models, particularly based on the fact that churn may be a gradual process and the  $W1C$  models do not cover it. By employing the prevailing base classifier of the first two parts, we train logistic classification models on MPTD and MPTD-S for the advance churn labels  $W2C$  to  $W6C$ . The predictive performance of these models will be compared to the  $W1C$  classification models, to measure the loss in accuracy. Other alternatives for providing churn prediction multiple periods ahead are survival model, ordinal logistic regression and multinomial logistic regression that predict periods to churn. These methods yield within- $n$ -periods churn probabilities, hence being used as benchmark models for the advance churn labels that are trained via the prevailing base classifier on the longitudinal training data.
- ✓ The advance churn notice models trained on  $W1C$  to  $W6C$  are generated independent from each other. Yet, the outputs of these models can be used in an ensemble which in turn should create added value. Following this, in the last step, we train meta-combination ensemble model for the next-period-churn prediction ( $W1C$ ). The predictions will be compared to the  $W1C$  models, assuming zero information lead time.

Before starting with the accuracy results, it is also helpful to inform the reader in detail how the ensemble method is constructed. As formerly depicted in Figure 3.2, the ensemble model consists of two layers. In the first layer, we compute six advance churn scores, which we label as  $p1$  through  $p6$ . We use the output predictions  $p1-p6$  of the prevailing classifier (logistic regression or classification tree) in the second layer of the ensemble method. To remember, the first score  $p1$  is obtained by modeling  $W1C$ . The remaining scores  $p2$  through  $p6$  are similarly generated by training churn models for each advance churn label  $W2C$  through  $W6C$ . Here, we experiment with two variants of the training data, MPTD with and without synthetic oversampling.

In the second layer of the ensemble method, in order to predict within-one-period churn, we perform eight regression tree analyses. We experiment with three variants in terms of input variables: (i) only the advance churn scores  $p1-p6$ , or (ii)  $p1-p6$  plus all independent variables of the first layer, or (iii)  $p1-p6$  plus selected variables of the first layer where the selection of the variables is completed as to include at least one predictor from each predictor category, as to reflect our hypotheses through them and as to incorporate the managerial inductions such as the sophistication impact. Accordingly, the selected variables are: *portfolio size, percentage of the number of shared accounts, twelve monthly relative asset return, total number of products ever used; customer age, customer tenure, nationality missing flag, college education dummy variable; monthly deposit interest rate, USD-TL exchange rate, "A tipi" percentage gain from YKB investments relative to the market index, "B tipi" percentage gain from YKB investments relative to the market index*. Based on the first part results, we also generate two further variables -percentage change in the USD-TL exchange rate between consecutive months and monthly relative asset return- to enter them into the regression tree.

For the first and second variants, the minimum number of observations per leaf is set to be 1% of the total number of observations included in the training dataset, namely 85. For the third variant, though, we experiment two further variants in terms of the minimum leaf size: (i) 85 and 200 for the case where  $p1-p6$  are obtained from MPTD,

and (ii) 842 and 1000 for the case where  $p1-p6$  are obtained from MPTD-S. Here, we should remark that these values for the minimum leaf size are not set randomly. For each regression tree of the third variant in terms of the predictors, we repeat the analysis for various values of the minimum leaf size and present only those which succeed in improving the accuracy performance observed in the first layer of the ensemble for the within-one-period churn label.

#### 4.5. Accuracy Results

We use the *proc logistic* procedure of SAS 9.2 to perform logistic regression where we apply stepwise selection procedure for determining the final set of covariates with a significance threshold of 0.1 for introducing and removing variables.

To conduct the classification tree and regression analyses, the *J48* and *M5P* procedures, respectively, of the open source software WEKA 3.6 are chosen. While developing the trees, we specify the minimum number of observations per leaf as 1% of the total number of observations and allow pruning on a confidence level of 0.25, unless said otherwise.

To perform Cox regression, the *proc phreg* procedure of SAS 9.2 implemented. As in the logistic regression, we introduce stepwise selection procedure with the same significance level 0.1. To impute the missing output predictions (the survival probabilities), we apply the *proc expand* procedure, again on SAS 9.2.

The ordinal logistic regression is implemented on SAS 9.2, using the *proc logistic* procedure. To be consistent with the previous analyses, we again apply stepwise selection for the variables on a significance threshold of 0.1.

The multinomial logistic regression is performed using the *proc logistic* procedure of SAS 9.2, with setting the function linking the response probabilities to the linear predictors as the generalized logit function ( $link = glogit$ ). Due to computational intensity, instead of performing stepwise variable selection procedure in the multinomial logistic regression, we introduce only the variables, which have been

chosen by at least one logistic regression procedure to model  $WIC$  to  $W6C$ , and thus avoid the selection of variables.

We use AUC, TDL, and TQL to compare the accuracy of models' performances that are described above in five steps. Further, to better distinguish amongst the models while expressing them, we name each model in the following format: *churn label-classification method-training data*. For example, the model generated to train  $WIC$  on MPTD via the logistic regression is named W1C-LogReg-MPTD model.

#### 4.5.1. MPTD vs. SPTD: Predicting Within-one-period Churn

Here, we evaluate the predictive accuracy of the models developed on SPTD and MPTD, with and without oversampling, via the logistic regression and classification tree for the dependent variable  $WIC$  for the first test period. We investigate whether the results are classification technique and oversampling dependent.


The classification tree analysis outputs a tree with a specific number of leaves. Since the churn probability of customers in the same leaf is the same, we obtain multiple observations with exactly same churn propensity score and being labeled differently. This in turn raises the question of how to accomplish the sorting procedure before computing the performance scores which require the observations to be sorted in decreasing order in terms of the predicted probabilities. To deal with this problem, we implement the following steps:

- ✓ Sort the data first by descending churn score and then by descending actual class. This corresponds to the optimistic case. All the observations with equal churn scores are sorted such that the churners are ranked before the non-churners.
- ✓ Sort the data first by descending churn score and then by ascending actual class. This corresponds to the pessimistic case. All the observations with equal churn scores are sorted such that the non-churners are ranked before the churners.

- ✓ Compute the AUC, TDL, and TQL for (1) and (2). Take the simple average and record them as the final AUC, TDL, and TQL.

**Table 4.3** The W1C model's performance: AUC, TDL, and TQL on the first test period

Churn Label	Classifier	Training Data	AUC	TDL	TQL
W1C	J48 <sup>1</sup>	SPTD	-	-	-
W1C	J481	SPTD-S	0.584	2.298	1.680
W1C	J48 <sup>1</sup>	MPTD	-	-	-
W1C	J48 <sup>1</sup>	MPTD-S	0.602	3.396	1.680
W1C	LogReg	SPTD	0.670	2.601	1.680
W1C	LogReg	SPTD-S	0.652	2.601	1.840
W1C	LogReg	MPTD	0.740	3.996	2.320
W1C	LogReg	MPTD-S	0.758	4.396	2.560



<sup>1</sup> No tree was built.

Table 4.3 reports the AUC, TDL, and TQL results for the test data set that is most adjacent to the training periods. The first finding is that models trained on MPTD outperform models trained on SPTD, independently from which classification algorithm is used and whether oversampling is applied. Accordingly, the W1C-Logreg-SPTD model yields an AUC score of 0.670 whereas this corresponds to 0.740 in the W1C-LogReg-MPTD model. This, in turn, implies a 10%-improvement in the AUC score. When oversampling is used, a 16% improvement in the AUC score is accomplished when the longitudinal training medium is favored instead of the cross-sectional data (from 0.652 to 0.758). As to deduce from Table 4.3, the findings for the evaluation metrics TDL and TQL are similar to those of AUC. Comparing logistic regression models trained without SMOTE, the W1C-LogReg-MPTD model yields a TDL score of 3.996 whereas the W1C-LogReg-SPTD model can perform 2.601 times better than a random model. This, in turn, corresponds to an improvement of 54% in the TDL score.

The second finding is that no tree is built without oversampling (W1C-J48-SPTD and W1C-J48-MPTD), i.e. all the observations are predicted to be non-churner. With oversampled training datasets, the trees developed on the longitudinal training dataset, namely W1C-J48-MPTD-S, outperforms the W1C-J48-SPTD-S model.

The third finding is that logistic regression models outperform the classification trees, when comparing on the same training medium. Further, even in the imbalanced training datasets, logistic regression is capable to develop models while classification tree cannot.

In brief, accuracy results illustrated in Table 4.3 point out that regardless of the base classifier used and with or without application of the oversampling procedure, models obtained from the longitudinal training data yield always better performance than models trained on the cross-sectional training data.

#### **4.5.2. MPTD vs. SPTD: Predicting Within-one-period Churn over Time**

Along with the superior performance of the models, which are trained on the longitudinal training medium (MPTD), for the first test period, it is also important to verify that these models are robust and keep outperforming the rival classification models obtained from cross-sectional data (SPTD) in the subsequent time periods.

Our objective in this section is to evaluate the robustness of the predictive accuracy of the models trained on MPTD over time. Accordingly, this sub-section reports the evaluation results of the within-one-period churn models developed in the first part for five consecutive test periods and makes pairwise comparisons between MPTD and SPTD to assess both the performance in each test period and the robustness. In addition, to investigate whether the MPTD-models significantly differ from the SPTD-models, the evaluation process is repeated on 10 bootstrap samples generated for each training medium with or without SMOTE (namely SPTD, SPTD-S, MPTD, MPTD-S).

Table 4.4 presents the AUC and TDL scores of the classification models of Section 4.5.1 for the test periods April 2010 through August 2010. Before interpreting the results, we need to remark that the TQL is not included in the table because we observed that the results corresponding to it are entirely parallel with the of TDL.

**Table 4.4** MPTD vs. SPTD over time

Model Name	Eval. Crit.	2010/04	2010/05	2010/06	2010/07	2010/08
<b>W1C-LogReg-MPTD</b>	<b>AUC</b>	0.740	0.734	0.817	0.723	0.691
	<b>TDL</b>	3.996	3.330	4.000	3.198	2.499
<b>W1C-LogReg-SPTD</b>	<b>AUC</b>	0.670	0.673	0.690	0.598	0.614
	<b>TDL</b>	2.601	1.776	3.000	2.398	1.943
<b>W1C-LogReg-MPTD-S</b>	<b>AUC</b>	0.758	0.742	0.821	0.727	0.721
	<b>TDL</b>	4.396	3.330	4.667	2.398	3.050
<b>W1C-LogReg-SPTD-S</b>	<b>AUC</b>	0.652	0.680	0.796	0.663	0.662
	<b>TDL</b>	2.601	2.664	4.333	1.199	1.388
<b>W1C-J48-MPTD-S</b>	<b>AUC</b>	0.602	0.679	0.634	0.660	0.591
	<b>TDL</b>	3.396	3.328	3.333	2.796	2.025
<b>W1C-J48-SPTD-S</b>	<b>AUC</b>	0.584	0.425	0.554	0.594	0.559
	<b>TDL</b>	2.298	1.332	1.333	2.798	1.943

As the results in Table 4.4 indicate, no matter which base classifier is employed and no matter whether oversampling is applied or not, the longitudinal training dataset reveals better performance accuracy in terms of both AUC and TDL in the immediate test period (2010/04) and continues to do so for the next four periods. Considering the case of “logistic regression and no SMOTE”, the TDL scores corresponding to the MPTD (namely W1C-LogReg-MPTD model) are 3.996, 3.33, 4, 3.198, and 2.499, whereas these equal 2.601, 1.776, 3, 2398, and 1.943, respectively for SPTD (namely W1C-LogReg-SPTD model) in test periods April to August 2010 in the SPTD. Accordingly, in terms of the TDL, we provide an improvement that ranges between 28.5% and 87.5%. We also see that the W1C-LogReg-MPTD-S is the best performing model in terms of AUC and TDL (with only one exception for the TDL score in July, 2010).

We showed that the classification models trained on the longitudinal training media provide superior accuracy results compared to the models obtained from the cross-sectional training media. However, all the evaluation scores correspond to the points estimates. On this ground, yet we cannot claim that these point estimates are significantly different from each other. To tackle this problem, we follow the below steps:

- ✓ Generate ten bootstrap samples for each training medium SPTD and MPTD while ensuring the original ratio of churners against non-churners. This corresponds to MPTD and SPTD datasets for the bootstrap samples.  
As explained by Hastie et al (2009, p. 249), in the bootstrap method, the basic idea is to randomly draw datasets with replacement from the original training data, each sample the same size as the original training dataset. This is done  $B$  times, where  $B$  equals to 10 in our case.
- ✓ Apply SMOTE to the bootstrap samples as to obtain one-to-one ratio between event and no-event. This corresponds to MPTD-S and SPTD-S datasets for the bootstrap samples.

Following the generation of bootstrap samples, we repeat all the procedures applied for the original data for the prevailing base classifier logistic regression. Accordingly, for each training medium (SPTD, MPTD, SPTD-S, MPTD-S), we obtain ten AUC, TDL, and TQL scores at each test period April 2010 through August 2010, with respect to Table 4.1.

To evaluate whether MPTD significantly outperforms SPTD, we perform Student t-test. This test is particularly developed to evaluate the corresponding hypothesis for each of the three evaluation measures, to be abbreviated as EM, (AUC, TDL, and TQL) is built as:

$$H_0: EM_{MPTD} - EM_{SPTD} = 0$$

$$H_1: EM_{MPTD} - EM_{SPTD} > 0$$

Based on the degrees of freedom (df) which equals to “*number of bootstrap samples* –  $I$ ”, the paired t-test outputs a p-value using the mean difference and its standard error. If this value falls below the predetermined significance level  $\alpha$ , then  $H_0$  is rejected. In other words,  $EM_{MPTD}$  is significantly greater than  $EM_{SPTD}$  in case of “*p-value* <  $\alpha$ ”. In this study, we set the significance level  $\alpha$  to 0.05.



**Table 4.5** Paired t-test analysis for the difference between MPTD and SPTD: Logistic Regression & without SMOTE

AUC	2010/04	2010/05	2010/06	2010/07	2010/08
<b>mean difference</b>	0.07557	0.08828	0.13489	0.04544	0.02034
<b>st. deviation</b>	0.03609	0.03784	0.06286	0.08098	0.04917
<b>st. error</b>	0.01141	0.01196	0.01988	0.02561	0.01555
<b>t-score</b>	6.6212	7.37822	6.78561	1.77444	1.30806
<b>p-value</b>	4.8E-05	2.1E-05	4E-05	0.05487	0.11164

TDL	2010/04	2010/05	2010/06	2010/07	2010/08
<b>mean difference</b>	0.85894	1.13199	0.86666	0.15973	0.16089
<b>st. deviation</b>	0.5733	0.702	1.20495	1.31152	1.03084
<b>st. error</b>	0.18129	0.22199	0.38104	0.41474	0.32598
<b>t-score</b>	4.73783	5.09923	2.27448	0.38513	0.49356
<b>p-value</b>	0.00053	0.00032	0.0245	0.35454	0.31672

TQL	2010/04	2010/05	2010/06	2010/07	2010/08
<b>mean difference</b>	0.16797	0.38205	0.89311	0.43194	0.15552
<b>st. deviation</b>	0.32754	0.40619	0.45335	0.50617	0.3747
<b>st. error</b>	0.10358	0.12845	0.14336	0.16007	0.11849
<b>t-score</b>	1.62168	2.97436	6.22975	2.69852	1.31251
<b>p-value</b>	0.06966	0.0078	7.7E-05	0.01223	0.11092

Table 4.5 reports paired t-test statistics for the difference between the logistic regression models trained for the churn label *WIC* on the non-oversampled longitudinal and cross-sectional training media (MPTD and SPTD, respectively). As the p-values indicate, we can reject the null hypothesis at the 0.05 significance level, for each evaluation criterion in the first three time periods (except for TQL for April 2010).

**Table 4.6** Paired t-test analysis for the difference between MPTD-S and SPTD-S: Logistic Regression & with SMOTE

AUC	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	0.07407	0.08698	0.15039	0.05995	0.06472
st. deviation	0.05135	0.05014	0.05541	0.05872	0.04146
st. error	0.01624	0.01585	0.01752	0.01857	0.01311
t-score	4.56128	5.48598	8.58276	3.22862	4.93609
p-value	0.00068	0.00019	6.3E-06	0.00517	0.0004
TDL	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	1.2787	1.37639	1.30001	0.27979	0.97162
st. deviation	1.1152	0.65186	0.86709	0.75486	0.5889
st. error	0.35266	0.20614	0.2742	0.23871	0.18623
t-score	3.62591	6.67707	4.74111	1.17211	5.21737
p-value	0.00276	4.5E-05	0.00053	0.13562	0.00028
TQL	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	0.37591	0.53311	0.94642	0.28797	0.52207
st. deviation	0.27722	0.36031	0.37937	0.39759	0.33138
st. error	0.08766	0.11394	0.11997	0.12573	0.10479
t-score	4.28811	4.67893	7.88896	2.2904	4.98193
p-value	0.00101	0.00058	1.2E-05	0.02387	0.00038

Table 4.6 reports paired t-test statistics for the difference between the logistic regression models trained for the churn label *WIC* on the oversampled longitudinal and cross-sectional training media (MPTD-S and SPTD-S, respectively). As the p-values indicate, we can reject the null hypothesis at the 0.1 significance level, for each evaluation criterion at each time period, with only exception for TDL which yields a p-value of 0.13562 in the fourth time period.

#### 4.5.3. MPTD vs. SPTD: Predicting Within-one-period Churn with Information Lag

Sometimes, the managers may lack the recent information at the time of scoring after they have generated classification models to predict next-period churn. On this ground, in this sub-section, we investigate the performance accuracy when there exists a lead time for the information. We use the *WIC*-LogReg-MPTD model and the *WIC*-LogReg-SPTD model to obtain the scores. Please follow the below instructions,

together with Table 4.7, to better understand how we score in the absence of the recent information.

**Table 4.7** How to interpret the performance accuracy when recent data is not available

Information Lead Time	Period X	Period X+1	Period X+2
0	score1_0	score2_0	score3_0
1	score1_1	score2_1	score3_1
2	score1_2	score2_2	score3_2
.	.	.	.
.	.	.	.
k-1	score1_k	score2_k	score3_k
k	-	score2_k+1	score3_k+1
k+1	-	-	score3_k+2

Assume that we are currently at period  $X$  and we have already generated a classification model to predict next-period churn. Further assume that the last period included in the training data is  $X-k$ . While computing the performance scores  $score1_0$  through  $score1_k$  (any accuracy measure which we represent in Table 4.7 as *score*), we always use the same set of customers, but the variables take on different values. For example, to calculate  $score1_0$ , we use the customer information belonging to the current period  $X$ , i.e. the information has zero lead time. If the current period's customer information is not available at the time of scoring, namely at period  $X$ , we employ the customer data of one month ago, namely of period  $X-1$ , hence compute  $score1_1$ .

With respect to Table 4.7, it should be noticed that the set of customers used to score differ across columns (namely time periods), but remain the same across rows (namely information lead time). Accordingly, customers who churn at period  $X$  are not included in the set of customers when we compute the corresponding scores at successive period  $X+1$ . Table 4.8 presents the AUC, TDL, and TQL scores for the W1C-LogReg-MPTD and W1C-LogReg-SPTD models considering various lead times for the information.

**Table 4.8** Next-month churn prediction with information lead time: AUC, TDL, and TQL performance on MPTD

<b>AUC</b>	<b>MPTD</b>					<b>SPTD</b>				
<b>Information Lead Time</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>0</b>	0.740	0.734	0.817	0.723	0.691	0.670	0.673	0.690	0.598	0.614
<b>1</b>		0.693	0.780	0.765	0.687		0.681	0.667	0.647	0.588
<b>2</b>			0.714	0.737	0.676			0.673	0.685	0.582
<b>3</b>				0.731	0.660				0.650	0.583
<b>4</b>					0.659					0.572

<b>TDL</b>	<b>MPTD</b>					<b>SPTD</b>				
<b>Information Lead Time</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>0</b>	3.996	3.330	4.000	3.198	2.499	2.601	1.776	3.000	2.398	1.943
<b>1</b>		3.552	3.666	4.395	2.774		2.442	2.333	3.198	1.665
<b>2</b>			2.999	4.000	2.775			2.333	3.600	1.387
<b>3</b>				3.600	1.942				3.200	1.942
<b>4</b>					2.219					1.387

<b>TQL</b>	<b>MPTD</b>					<b>SPTD</b>				
<b>Information Lead Time</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>0</b>	2.320	2.221	2.799	2.400	2.222	1.680	2.132	1.866	1.280	1.333
<b>1</b>		2.132	2.667	2.400	2.111		2.132	1.333	1.760	1.333
<b>2</b>			2.133	1.760	2.000			1.733	1.760	1.444
<b>3</b>				2.079	1.999				1.600	1.222
<b>4</b>					1.999					1.222

As illustrated in Table 4.8 for the longitudinal training dataset, for the second, third, and fifth test periods, the AUC scores become worse as the information lead time increases and this is entirely in line with the expectations. The customer decision, in our context the decision to leave, is mostly covered by the covariate values of the last periods both before and when churn occurs, and the more the information is missing, the less churn tracks are included while scoring. With respect to period July, we observe some fluctuation, but this can be explained with the extra-ordinary performance of the regular W1C-LogReg-MPTD model at June, 2010 (To compute the AUC score at July 2010 with the current information not available, we employ the customer information of June, 2010). With respect to TDL, similar conclusions can be made as done for AUC. In terms of TQL, the claim “the performance decreases in the quantity of missing

information” holds without any exception for all the test periods included in this analysis. Same interpretation can also be made for the model trained on the cross-sectional training data. When to compare both models, we can easily assert that the MPTD model also performs better when the recent information is not available.

#### 4.5.4. W2C to W6C Models

Sometimes, the companies are interested not only in predicting the next-period churn, but they also want to obtain insights about the long-term churn probability. On this ground, in this sub-section, we investigate the performance of various models to predict the attrition several months ahead. Since the logistic regression has outperformed the classification tree analysis in terms of modeling within-one-period-churn, here, we only apply the former classification technique when modeling churn several months ahead, both on the oversampled and non-oversampled longitudinal training dataset<sup>4</sup>. For ease of comparison, we repeat the results for the within-one-period churn prediction (where churn label corresponds to *W1C*) above the models generated for the remaining churn labels, in Table 4.9. To remember: By generating *WnC* churn labels, we increase the number of positive examples as to see in Table 4.2, but on the other hand, we use less customer information in model training as to see in Figure 4.2, i.e. the predictive models are trained on customer information with the last time period becoming farther from the validation period. While scoring, we assume that the information lead time is zero, i.e. the current period’s information is available at the time of scoring. Please also consider that *W2C* through *W6C* models cannot be scored for each test period, as explained in Section4.2.

---

<sup>4</sup> It is important to note that it is not possible to create W2C through W6C for the cross-sectional training dataset.

**Table 4.9** Several months ahead churn prediction: AUC, TDL, and TQL performance on MPTD

<b>AUC</b>		<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	
<b>W1C</b>	0.740	0.734	0.817	0.723	0.691	
<b>W2C</b>	0.728	0.759	0.796	0.708		
<b>W3C</b>	0.723	0.721	0.711			
<b>W4C</b>	0.709	0.719				
<b>W5C</b>	0.693					
<b>SPTD</b>	0.670	-	-	-	-	

<b>TDL</b>		<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	
<b>W1C</b>	3.996	3.330	4.000	3.198	2.499	
<b>W2C</b>	3.575	3.463	4.182	2.785		
<b>W3C</b>	3.516	3.097	2.528			
<b>W4C</b>	3.263	3.232				
<b>W5C</b>	3.223					
<b>SPTD</b>	2.601	-	-	-	-	

<b>TQL</b>		<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>	
<b>W1C</b>	2.320	2.221	2.799	2.400	2.222	
<b>W2C</b>	2.273	2.452	2.763	2.033		
<b>W3C</b>	2.240	2.039	2.021			
<b>W4C</b>	2.080	1.999				
<b>W5C</b>	1.899					
<b>SPTD</b>	1.680	-	-	-	-	

Table 4.9 reports the accuracy results for the five test data sets of consecutive future time periods, in line with Table 4.1. For example, the AUC of the W2C model in predicting churn for the 2010/05 or 2010/06 time periods using information available as of 2010/04 is 0.728. As illustrated in Table 4.9, as expected, the AUC performance deteriorates as we predict the attrition more months ahead, i.e. when less customer information is used for training the model, with only one exception at period May 2010. The *W2C*'s AUC exceeds the *W1C*'s AUC. In addition, it is remarkable to note that the W5C-LogReg-MPTD model, which predicts within-five-periods churn and hence doesn't use the customer information belonging to periods December 2009 through

March 2010, outperforms the W1C-LogReg-SPTD model, namely the model that is trained on the cross-sectional training data of time period March 2010. As to deduce from Table 4.9, the same findings are also valid for the remaining two evaluation criteria TDL and TQL.

#### 4.5.5. Benchmark Models

As explained in Section 3.4.3, the logistic regression models separately generated for the dichotomous labels W1C to W6C resemble the duration and time-to-event models, when considered as a unique entire model. So, in this part, we report the accuracy performance for three benchmark models - Cox regression, multinomial logistic regression, and ordinal logistic regression - each trained on the MPTD for the dependent variable *months-to-churn* and we only use the AUC and TDL scores while comparing among the models.

The first benchmark procedure applied is the Cox regression modeling. After the model has been obtained, we score the test periods' data and obtain survival probability estimates for each customer and for each time point when at least one churn event has been observed. Following this, we impute the missing values of survival probability estimates for each time period by interpolation and compute the within-one-period to within-five-periods churn probability, with respect to the steps explained in Section 3.4.3.1. Table 4.10 reports the accuracy results for the five test data sets of consecutive future time periods, in line with Table 4.1. Comparing these scores to the accuracy results obtained for the *W1C* through *W5C* logistic regression model (as summarized in Table 4.9), we see that the Cox regression cannot outperform the logistic regression models separately run for each binary dependent variable *W1C* to *W5C* on MPTD. This is important to note due to the following reason: While generating the Cox regression model, we use the longitudinal training dataset and observations belong to time periods from April 2009 to March 2010. So, while computing the within-n-periods churn probability, we always use the full information (including March 2010) no matter what  $n$  takes on. On the other hand, as we described beforehand, we could not make utilize

the customer information of the last  $n$  periods of time interval April 2009 and March 2010 while training within- $(n+1)$ -periods logistic regression churn model. For example, while modeling within-two-periods (within-three-periods) churn on MPTD, the customer information of March 2010 (February and March 2010) is not included in the training data. In other words, even though the Cox regression model exploits the full information available, it is beaten by the logistic regression trained on MPTD with missing information.

**Table 4.10** Cox regression: AUC and TDL performance

<b>AUC</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	0.708	0.637	0.759	0.665	0.676
<b>W2C</b>	0.704	0.670	0.745	0.683	
<b>W3C</b>	0.684	0.689	0.732		
<b>W4C</b>	0.692	0.696			
<b>W5C</b>	0.687				

<b>TDL</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	3.796	1.332	2.333	1.999	1.944
<b>W2C</b>	3.365	1.332	2.273	2.621	
<b>W3C</b>	2.197	1.698	2.747		
<b>W4C</b>	2.130	2.204			
<b>W5C</b>	2.900				

As illustrated in Table 4.10, the Cox regression model identifies 3.796 times more within-one-period churners in the top-decile in April 2010 compared to a random model and the corresponding AUC score is computed to be 0.708. But, these scores are lower than the scores of W1C logistic regression model on MPTD, 3.996 and 0.740 respectively. In addition, differently from expectations and the previous results, both the AUC and the TDL performance of within- $n$ -periods churn label fluctuate over time.

The second benchmark model developed is the multinomial logistic regression (MLR). Following the instructions given in Section 3.4.3.2, we use the MPTD and months-to-churn as the training dataset and the dependent variable, respectively. It should be noted that the training dataset used in the MLR analysis is the same as the



training medium employed for the W6C logistic regression models. Please remember, that the month-to-churn variable comprises seven categories, 1 to 6 and non-churner within 6 periods (reference category). No selection procedure for variables is employed due to the computational complexity; instead, we introduce all the variables, which have been selected by at least one WnC logistic regression model, into the model. Accordingly, MLR outputs six different logit functions with different coefficients. Using the parameter estimates, we compute the probabilities for belonging to each category and convert these into within-n-periods churn probability scores summarized in Section 3.4.3.2. Table 4.11 reports the corresponding accuracy scores.

**Table 4.11** Multinomial logistic regression: AUC and TDL performance

<b>AUC</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	0.758	0.694	0.725	0.689	0.696
<b>W2C</b>	0.718	0.726	0.763	0.686	
<b>W3C</b>	0.696	0.702	0.700		
<b>W4C</b>	0.671	0.698			
<b>W5C</b>	0.695				
<b>TDL</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	3.397	3.552	3.000	2.798	1.943
<b>W2C</b>	3.365	3.596	3.818	2.457	
<b>W3C</b>	3.357	3.197	3.077		
<b>W4C</b>	3.263	2.938			
<b>W5C</b>	3.169				

When comparing the MLR's accuracy scores to the performance results obtained for the W1C through W5C logistic regression models (as summarized in Table 4.9), we see that WnC models outperform the MLR with only two and five exceptions for the AUC and TDL criteria, respectively. In the first test period, April 2010, the MLR model identifies 3.397 times more within-one-period churners in the top-decile lift than a random model, and this is less the TDL score for the W1C model on MPTD (3.996); but the corresponding AUC score computed via the MLR is greater than the W1C-

LogReg model (0.758 vs. 0.740). In terms of the within-two-periods churn probability in the first test period, MLR is beaten by the W2C-LogReg-MPTD model considering both AUC and TDL (TDL and AUC scores are 3.365 and 3.575, and 0.718 and 0.728, respectively) whereas it beats the W5C-LogReg-MPTD model (TDL scores are 3.169 and 3.115, respectively). In general, the comparative analysis shows that MLR is surpassed by the WnC logistic regression models.

The last benchmark model employed is the ordinal logistic regression (OLR). Following the instructions given in Section 3.4.3.3, we use the MPTD and months-to-churn as the training dataset and the dependent variable, respectively. Please note that the training dataset used in the OLR analysis is the same as the training medium employed for the W6C logistic regression models. Accordingly, OLR outputs six probability score estimates  $p1$  through  $p6$  and these correspond to within-one-period to within-six-periods churn probabilities, respectively. Along with the binary churn labels  $W1C$  to  $W6C$ , these probabilities are used to compute the AUC and TDL scores in test periods April to August 2010. Table 4.12 reports the corresponding accuracy measures.

When comparing the OLR's accuracy scores to the performance results obtained for the W1C through W5C logistic regression model (as summarized in Table 4.9), we see that OLR is beaten by the WnC models. In the first test period, April 2010, the OLR model identifies 3.397 times more within-one-period churners in the TDL than a random model, but this is less than 3.996 obtained for the W1C-LogReg-MPTD model. The same is valid for the AUC (0.719 vs. 0.740). In addition, as the scores indicate, the accuracy performance of OLR model hugely deteriorates in the third test period and this implies that the model generated is also not as robust as the WnC models.

**Table 4.12** Ordinal logistic regression: AUC and TDL performance

<b>AUC</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	0.719	0.732	0.680	0.632	0.709
<b>W2C</b>	0.709	0.710	0.676	0.651	
<b>W3C</b>	0.698	0.708	0.674		
<b>W4C</b>	0.694	0.699			
<b>W5C</b>	0.690				

<b>TDL</b>	<b>Test Periods</b>				
<b>Churn Label</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>W1C</b>	3.397	3.552	2.080	1.599	2.553
<b>W2C</b>	3.260	2.930	1.636	1.966	
<b>W3C</b>	2.877	3.097	1.978		
<b>W4C</b>	2.797	2.865			
<b>W5C</b>	2.793				

#### 4.5.6. Meta-Combination Ensemble Method

Table 4.13 reports the AUC results for the five test data sets of consecutive future time periods (Please note that we use Case ID while interpreting the performance results and these IDs are to be found in Table 4.13). For ease of comparison, we repeat the results of the W1C models trained on oversampled and non-oversampled longitudinal training datasets - the W1C-LogReg-MPTD and the W1C-LogReg-MPTD-S models - below the ensemble models. Here, it should be noticed that we make comparisons only among the models which have been trained on the same training dataset. In other words, models trained on MPTD are not compared to models trained on MPTD-S.

For the performance results of the first test dataset, April 2010, the table shows that the ensemble models which use MPTD-S as the training dataset cannot improve the AUC score obtained from the W1C-LogReg-MPTD-S model, namely 0.758. On the other hand, in the case of MPTD (longitudinal dataset without use of SMOTE), a slight improvement can be accomplished as to increase the AUC score from 0.740 up to 0.750 for Case5-85. However, this point estimate differences do seem to be insignificant. In terms of the second test period, an improvement in the AUC score is achieved for none

of the ensemble models. For test period June 2010, the last four ensemble models - Case5-85 to Case6-1000 -, we observe slight advancement in the AUC scores, however the significance of this improvement is suspect. For test periods July and August 2010, Case6-842 and Case6-1000 reveal improved AUC scores compared to the W1C-LogReg-MPTD-S model and Case1 slightly improves the AUC score obtained from the W1C-LogReg-MPTD model.

**Table 4.13** Ensemble models' within-one-period churn prediction: AUC performance

Case ID	Training Data	Variables	2010/04	2010/05	2010/06	2010/07	2010/08
Case1	MPTD	p1-p6	0.749	0.728	0.759	0.796	0.708
Case2	MPTD-S	p1-p6	0.737	0.727	0.755	0.785	0.716
Case3	MPTD	p1-p6, all x	0.730	0.734	0.723	0.721	0.711
Case4	MPTD-S	p1-p6, all x	0.714	0.712	0.720	0.752	0.695
Case5-85	MPTD	p1-p6, selected x	0.750	0.704	0.842	0.681	0.690
Case5-200	MPTD	p1-p6, selected x	0.743	0.713	0.848	0.680	0.689
Case6-842	MPTD-S	p1-p6, selected x	0.747	0.737	0.827	0.726	0.740
Case6-1000	MPTD-S	p1-p6, selected x	0.744	0.737	0.824	0.724	0.748
	<b>MPTD</b>	<b>W1C</b>	<b>0.740</b>	<b>0.734</b>	<b>0.817</b>	<b>0.723</b>	<b>0.691</b>
	<b>MPTD-S</b>	<b>W1C</b>	<b>0.758</b>	<b>0.742</b>	<b>0.821</b>	<b>0.727</b>	<b>0.721</b>

Another aspect when evaluating the models is to assess their robustness. As Table 4.13 shows, the ensemble models corresponding to the Case-6-842 and Case6-1000 seem to be the most robust ones, with no degrading for five periods at all. On the other hand, although the Case-5-85 ensemble model provides enhanced AUC score in the first test period, it performs worse as of the second test period and starts to deteriorate in the fourth test period. On these grounds, it is difficult to put forward one model from others in terms of the AUC score.

Table 4.14 reports the top-decile lift point estimates for the eight ensemble model. Accordingly, the Case5-85 and Case5-200 ensemble models seem to outperform the LogReg-MPTD model for all the test periods. For the Case5-85 model, the percentage improvement in the TDL score is 12.5%, 13.3%, 37.5%, 18.8%, and 16.7% for the first to fifth test periods, respectively. In terms of the Case5-200 model, this improvement corresponds to 15%, 20%, 37.5%, 18.8%, and 16.7%, respectively. On the other hand,

the last two ensemble models generated (Case6-842 and Case6-1000 which are the ensemble models where  $p1-p6$  derive from the LogReg-MPTD-S models) cannot beat the W1C-LogReg-MPTD-S model on 2010/04 however they yield better TDL scores in the test data of the second through fifth test period. In other words, the ensemble built on MPTD-S yields more robust models whose predictive power remains high as the recency of the training data versus test data deteriorates. On these grounds, we claim that if the managerial objective is to target top-decile churners in the short term, then the Case5-200 ensemble model is recommended whereas Case6-1000 models should be preferred if robustness is desired. Similar findings have been obtained for the TQL accuracy criterion, we therefore won't report it.

**Table 4.14** Ensemble models' within-one-period churn prediction: TDL performance

Case ID	Training Data	Variables	2010/04	2010/05	2010/06	2010/07	2010/08
Case1	MPTD	p1-p6	4.196	3.663	3.667	3.397	2.221
Case2	MPTD-S	p1-p6	3.796	2.886	3.167	2.198	2.360
Case3	MPTD	p1-p6, all x	4.395	2.664	3.167	2.198	3.054
Case4	MPTD-S	p1-p6, all x	4.196	2.109	4.000	3.397	3.887
Case5-85	MPTD	p1-p6, selected x	4.495	3.774	5.500	3.797	2.915
Case5-200	MPTD	p1-p6, selected x	4.595	3.996	5.500	3.797	2.915
Case6-842	MPTD-S	p1-p6, selected x	4.196	3.552	4.833	2.598	3.331
Case6-1000	MPTD-S	p1-p6, selected x	4.196	3.885	4.833	3.198	3.609
	<b>MPTD</b>	<b>W1C</b>	<b>3.996</b>	<b>3.330</b>	<b>4.000</b>	<b>3.198</b>	<b>2.499</b>
	<b>MPTD-S</b>	<b>W1C</b>	<b>4.396</b>	<b>3.330</b>	<b>4.667</b>	<b>2.398</b>	<b>3.050</b>

We showed that the ensemble models Case5-200 and Case6-1000 surpassed the accuracy performance of the W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models, respectively. However, all the evaluation scores given correspond to the points estimates. On this ground, yet we cannot claim that these point estimates are significantly different from each other. To tackle this problem, we follow the same procedure as we applied in Section 4.5.2: bootstrap sample generation and repeating everything from scratch for these samples. Please remark, that we employ the same bootstrap samples which we utilized in the previous variability analysis.

To evaluate whether the ensemble provides a significant improvement against the use of W1C-LogReg-MPTD model in terms of the performance criteria (AUC, TDL, and TQL), we perform Student t-test. The corresponding hypothesis for each of the three evaluation measures, to be abbreviated as EM, is built as:

$$H_0: EM_{Ensemble} - EM_{MPTD} = 0$$

$$H_1: EM_{Ensemble} - EM_{MPTD} > 0$$

If the mean difference is negative, then the alternative hypothesis  $H_1$  is specified as below:

$$H_1: EM_{Ensemble} - EM_{MPTD} < 0$$

Based on the degrees of freedom (df) which equals to “*number of bootstrap samples* –  $I$ ”, the paired t-test outputs a p-value using the mean difference and its standard error. If this value falls below the predetermined significance level  $\alpha$ , then  $H_0$  is rejected. In other words,  $EM_{Ensemble}$  is significantly greater than  $EM_{MPTD}$  in case of “*p-value* <  $\alpha$ ”. In this study, we set the significance level  $\alpha$  to 0.05.

**Table 4.15** Paired t-test analysis for the difference between Ensemble (Case5 – 200) and MPTD: Logistic Regression & without SMOTE

AUC	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	0.00817	-0.0114	-0.075	0.00644	0.04403
st. deviation	0.05523	0.04036	0.07639	0.06947	0.0312
st. error	0.01747	0.01276	0.02416	0.02197	0.00986
t-score	0.46777	-0.8947	-3.1034	0.29315	4.46334
p-value	0.32553	0.19711	0.00633	0.38803	0.00078
TDL	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	-0.4194	-0.3772	-0.6667	-0.2796	-0.3274
st. deviation	0.72621	0.70358	0.70622	1.01117	0.60608
st. error	0.22965	0.22249	0.22333	0.31976	0.19166
t-score	-1.8262	-1.6953	-2.9852	-0.8745	-1.7085
p-value	0.05055	0.06213	0.00766	0.20228	0.06086
TQL	2010/04	2010/05	2010/06	2010/07	2010/08
mean difference	-0.024	-0.0355	-0.4932	-0.256	-0.1999
st. deviation	0.44623	0.6029	0.67108	0.53431	0.27608
st. error	0.14111	0.19065	0.21221	0.16896	0.08731
t-score	-0.1702	-0.1864	-2.3241	-1.5149	-2.2901
p-value	0.43433	0.42815	0.02259	0.08205	0.02388

Table 4.15 reports paired t-test statistics for the difference between the ensemble method Case5-200 and the logistic regression model trained for *WIC* on the non-oversampled longitudinal training medium (MPTD). Here, the mean difference is computed by subtracting the mean accuracy score of the WIC-LogReg-MPTD model from the of the ensemble model Case5-200. As the p-values indicate, we cannot reject the null hypothesis at the 0.05 significance level, with only four exceptions in 15 criterion-period combination (AUC-2010/08, TDL-2010/06, TQL-2010/06, and TQL-2010/08).

**Table 4.16** Paired t-test analysis for the difference between Ensemble (Case6-1000) and MPTD-S: Logistic Regression & with SMOTE

<b>AUC</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>mean difference</b>	0.00221	0.0069	-0.0058	0.00432	0.02376
<b>st. deviation</b>	0.03254	0.03157	0.03622	0.05161	0.04517
<b>st. error</b>	0.01029	0.00998	0.01145	0.01632	0.01428
<b>t-score</b>	0.21477	0.69122	-0.5029	0.26467	1.66351
<b>p-value</b>	0.41737	0.25343	0.31356	0.39861	0.06529
<b>TDL</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>mean difference</b>	-0.2198	-0.2442	-0.4333	-0.5996	0.11105
<b>st. deviation</b>	0.49347	0.43716	1.19724	0.60324	0.57342
<b>st. error</b>	0.15605	0.13824	0.3786	0.19076	0.18133
<b>t-score</b>	-1.4084	-1.7664	-1.1445	-3.1429	0.61241
<b>p-value</b>	0.0963	0.05557	0.14096	0.00594	0.27771
<b>TQL</b>	<b>2010/04</b>	<b>2010/05</b>	<b>2010/06</b>	<b>2010/07</b>	<b>2010/08</b>
<b>mean difference</b>	-0.2319	0.07108	-0.1999	0.048	-0.1777
<b>st. deviation</b>	0.34238	0.25752	0.47129	0.52275	0.33611
<b>st. error</b>	0.10827	0.08143	0.14903	0.16531	0.10629
<b>t-score</b>	-2.1422	0.87286	-1.3416	0.29037	-1.6722
<b>p-value</b>	0.0304	0.20271	0.1063	0.38906	0.0644

Table 4.16 reports paired t-test statistics for the difference between the ensemble method Case6-1000 and logistic regression model trained on the oversampled longitudinal training medium (MPTD-S). As the p-values indicate, we cannot reject the null hypothesis at the 0.05 significance level, for each evaluation criterion at each time period, with only exception for TQL at April 2010.

On these grounds, we deduce that we cannot create added value in the identification of churners just by training a regression tree in which we use the output predictions of independently trained binary logistic regressions as input attributes. However, it should be also noted that the point estimates and variability analysis from bootstrap samples do not correspond to the optimal regression tree. We did not test for each “confidence level” and “minimum number of observations per leaf” combination.

#### 4.5.7. Contributions

On the basis of the three performance scores and the paired t-test statistics, we are able to claim that the way the training dataset is generated improves the predictive accuracy, i.e. the models developed on longitudinal training datasets outperform the models trained on cross-sectional data and this improvement is significant. The MPTD models continue to outperform the SPTD models even several periods after the training period. Further, it is also remarkable that the W1C-LogReg-SPTD-S model performs worse than the W1C-LogReg-MPTD model which implies that creating artificial observations of positive cases cannot compete with entering the historical information of customers as separate observations to obtain more accurate predictions.

On the ground of findings shared in Section 4.5.4, we can assert that when most recent information cannot be utilized in the model training phase, the advance churn notice models (WnC) which predict the churn  $n$  periods ahead give satisfactory results. As expected, their predictive performance deteriorates with information lead time, but the predictive performance of even the model that uses 5 periods old information is better than the model generated on the cross-sectional training data with most recent customer information.

The benchmark Cox regression used the full customer training data whereas the WnC logistic regression models cannot, for  $n > 1$ . Nevertheless, this benchmark model does not perform as good as the WnC models while identifying within- $n$ -periods churners in test periods April to August 2010. In other words, the proposed way to generate the training dataset makes the binary logistic regression models perform better



than the duration model, even with missing information. In addition, with respect to MLRR and OLR which use months-to-churn as the dependent variable, WnC binary logistic regression models display better predictive performance.

As explained in Section 4.5.6, the ensemble method that we proposed does not provide any significant improvement in the accuracy performance. The accuracy improvement in point estimates is only valid for the data which we used.

## **4.6. Model Outputs and Managerial Insights**

### **4.6.1. W1C Models**

For the managerial purposes, it is important first to acknowledge that the churn model developed is valid and then to understand the churn drivers. For the former one, the convergence results and the goodness-of-fit statistics are presented whereas we introduce the parameter estimates for the latter one. Since the logistic regression models yield significantly better performance results and since the decision trees are known to be unstable, i.e. one major problem with trees is their high variance, this sub-section is devoted only to the logistic regression models.

For all the four logistic regression models trained on SPTD, SPTD-S, MPTD, and MPTD-S the convergence criterion is satisfied. When we evaluate these models whether they are significant as a whole or not, the global null hypothesis test shows that they all are significant, i.e. the parameter estimates are different than 0. The next goodness-of-fit measure used to evaluate the models is the Hosmer-Lemeshow test. The test outputs reveal that the models without the use of the oversampling technique has the adequate fit for the training data (for the SPTD, this statement holds at the significance level of 0.01) whereas the models trained on the oversampled datasets reveal lack in the fit in the training data, i.e. a difference exists between the predicted and observed values of the dependent variable. The inadequate fit for the oversampled cases might be caused by the creation of artificial observations.

Table 4.17 reports the parameter estimates for the corresponding four models. The first finding is that the models trained on MPTD and MPTD-S choose more variables as to explain churn triggers. Accordingly, W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models have selected 25 and 33 variables, respectively, to explain churn attitude whereas W1C-LogReg-SPTD and W1C-LogReg-SPTD-S models have identified 13 and 16 covariates, respectively. Next, percentage change in the portfolio size (*CR\_PERC\_CHANGE\_ENDBAL*), total number of products currently owned (*TOT\_NUM\_PROD\_CURRENT\_USE*), and the binary covariate to represent whether the customer has ever owned government bonds (*EU\_FI5*) are covered by these four models in common. The former two variables are negatively related to the churn behavior whereas the latter one positively.

W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models include “last one month return” and “last three month return” variables which are not covered by the W1C-LogReg-SPTD and W1C-LogReg-SPTD-S models. This is important to note because it implies that the short term asset return also plays a role in the churn attitude of customers which cannot be detected when cross-sectional data is used. Similarly, the company-customer interaction via the branch and customer representative negative impact variables (*br\_neg\_impact* and *rm\_neg\_impact*) cannot be covered by the models trained on the cross-sectional data and the longitudinal data format allows more product-ownership variables to enter into the model. Further, W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models also identify the interest rate for the monthly deposit rate (*TRIMT\_RR\_Line*) and the difference between the interest rates for monthly and yearly deposit rates (*TRIYT\_RR\_Line\_minus\_TRIMT\_RR\_Line*) to relate to the customer churn behavior whereas the W1C-LogReg-SPTD and W1C-LogReg-SPTD-S models cannot appoint economic indicators as churn drivers by their nature. In general, a number of interesting findings emerge from our analysis, as to be deduced from the parameter estimates in Table 4.17.

**Table 4.17** Parameter estimates of the logistic regression models LogReg-SPTD, LogReg-SPTD-S, LogReg-MPTD, and LogReg-MPTD-S for the churn label W1C

Parameter Estimates		<i>SPTD</i>	<i>SPTD-S</i>	<i>MPTD</i>	<i>MPTD-S</i>
Predictor Category	Predictor				
	Intercept	-6.225	-1.9612	-5.9249	-1.006
Cust. Behavior	CR_PERC_CHANGE_ENDBAL	-0.3506*	-0.4911	-0.226	-0.289
Cust. Behavior	PORTFOLIO_BUCKET1		0.3818	0.1947*	0.2759
Cust. Behavior	NUM_OF_ALL_ACC	-0.6608**	-1.215		
Cust. Behavior	CR_LAST_1_MONTH_RETURN_PERCENT			0.3118	0.3224
Cust. Behavior	CR_LAST_3_MONTH_RETURN_PERCENT			-0.0994***	
Cust. Behavior	CR_LAST_12_MONTH_RETURN_PERCENT	0.1981***			
Cust. Behavior	RELATIVE_RETURN		0.5139	0.1435*	
Cust. Behavior	TOT_NUM_PROD_CURRENT_USE	-0.9593*	-1.27	-0.9426	-0.5746
Cust. Behavior	CU_BTFF2			0.1763*	
Cust. Behavior	CU_BTFF3			-0.2853*	-0.5409
Cust. Behavior	CU_BTFF6		0.345		
Cust. Behavior	CU_DVZ1_EUR			0.2974	0.1416
Cust. Behavior	CU_DVZ1_USD			0.1657*	
Cust. Behavior	CU_FI1_other			0.2127*	0.1519
Cust. Behavior	CU_FI1_EUR			-0.25*	-0.3807
Cust. Behavior	CU_FI1_USD			-0.1134***	-0.3163
Cust. Behavior	CU_FI1_YTL				-0.2055
Cust. Behavior	CU_FI5	-0.5116**			
Cust. Behavior	CU_FI7_YTL			0.1189*	
Cust. Behavior	CU_HSF2			0.0665*	
Cust. Behavior	CU_HSF5	0.2685*	0.3232		
Cust. Behavior	CU_HSF6			0.1681*	0.2254
Cust. Behavior	CU_PPF2	0.3607**	0.3788		-0.142
Cust. Behavior	TOT_NUM_PROD_EVER_USED		0.7519		
Cust. Behavior	EU_BTFF1		-0.7254		
Cust. Behavior	EU_BTFF2				0.1588
Cust. Behavior	EU_BTFF6				-0.212
Cust. Behavior	EU_DVZ1_EUR				0.1503
Cust. Behavior	EU_DVZ1_USD	0.3749**			
Cust. Behavior	EU_FI1_other	0.3456**			
Cust. Behavior	EU_FI4			0.1298**	
Cust. Behavior	EU_FI5	0.7036	0.7642	0.2434	0.2233
Cust. Behavior	EU_HSF1			0.1177*	0.1988
Cust. Behavior	EU_HSF5			0.1461*	
Cust. Behavior	EU_HSF7				
Cust. Behavior	EU_PPF1			0.1706*	0.1526
Cust. Behavior	aver_riskscore_cr		-0.496	-0.4565	
Cust. Behavior	etsscore_cr				0.2026
Cust. Behavior	dBond				-0.3416
Cust. Behavior	dEurobond		-1.0014		
Cust. Behavior	dFundA				0.1901
Cust. Behavior	dFundB			-0.252*	-0.3781
Cust. Behavior	dGold				-0.1514
Cust. Behavior	wEurobond				-0.1605
Cust. Behavior	wFX				-0.169
Cust. Behavior	wFundA			-0.5865***	-3.3026
Cust. Behavior	wFundB			0.1907*	
Interaction	br_neg_impact				0.2688
Interaction	rm_neg_impact				0.113
Interaction	CUST_MONTH_OF_YKB	-0.4555*			-0.2474
Environment	TR1MT_RR_Line				-0.2464
Environment	TR1YT_RR_Line_minus_TR1MT_RR_Line				0.1399
Demographics	GENDER_MISSING_FLAG	0.1901	0.7428		0.3042
Demographics	EMP_OWN_BUSINESS		0.4691		
Demographics	MAR_STAT_EVLI				-0.1257
Demographics	SBU_TYPE_NM_UH	0.265**	0.4282		

\* Significant at 0.01; \*\* Significant at 0.05; \*\*\* Significant at 0.10; if no asterix: Significant at &lt;0.0001

In terms of demographic characteristics, all but the W1C-LogReg-MPTD model predict the customers with missing demographics information (indicated by the gender missing flag) to be more likely to churn. The unwillingness to provide information may be interpreted as lack of commitment to the relationship with the bank by the customer.

The W1C-LogReg-MPTD-S model asserts that individuals experience lower attrition tendency when the monthly interest for deposit accounts are high, nevertheless the churn propensity increases when the difference between the interest rates of monthly and yearly deposit accounts grows.

For all the models, an increase in the total number of financial instruments currently used results in advanced retention whereas the total number of products ever used is positively related with the churn propensity in the W1C-LogReg-SPTD-S model. The models built on MPTD and MPTD-S do not include the total number of products ever used as to explain the churn. Instead, they relate the churn attitude to whether the customer has ever possessed specific financial instruments.

Models trained on the longitudinal data include more interaction predictors to explain the churn attitude.

In the following paragraphs, we discuss the findings in detail, and consider their managerial implications while relating them to the hypotheses we generate. We have chosen to mainly focus on the output resulting from the W1C-LogReg-MPTD-S and W1C-LogReg-MPTD models which exhibit the best performance accuracy as discussed in Section 4.5.1 and Section 4.5.2.

#### *Customer Behavior Predictors*

For each model generated, an increase in the total portfolio size indicates a decreasing churn tendency, which is in line with the expectations and satisfies the Hypothesis 1. Having a portfolio size less than 500.000 TL is also specified as an indicator to increase the churn propensity that satisfies our Hypothesis 3. Next, as in line with the expectations and summarized via the Hypothesis 16, the customers who possess more risky products are more likely to remain (W1C-LogReg-MPTD model).

The rationale may be that risky products may be more difficult to liquidate without incurring losses. In addition, as specified through the Hypotheses 17 and 18, the W1C-LogReg-MPTD-S model indicates that increasing easy-to-sell scores are positively related with the churn attitude. One interesting and surprising outcome in terms of the behavioral predictors is the parameter estimates for the asset return predictors. As explained by the Hypothesis 10, we expect all the return variables negatively related to the churn tendency, but as the parameter estimates of the W1C-LogReg-MPTD model indicate, the attrition propensity increases when the short term return increases (variable: *cr\_last\_1\_month\_return\_percent*) and when the customers obtain more-than-average long-term return (variable: *relative\_return*). This shows that customers who are literate in the financial services industry are more prone to churn which can be summarized as the “sophistication impact”.

With respect to product ownership information, total number of products currently in the portfolio decreases churn propensity, as captured by all models. On the other hand, number of products ever-used is found to have no impact on the churn decision. Instead, the W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models pick specific products that if the customer has ever used them, the churn propensity increases. Paper by Kamakura et al (1991) postulates that as customer sophistication in financial services increases the customer proceeds to more complicated and risky products. Hence, having tried different products can be seen as a measure of customer sophistication. Taken together with the finding about the returns increasing churn propensity, a likely conclusion is that more sophisticated customers are more likely to churn. Furthermore, while foreign exchange demand deposit accounts are associated with higher churn propensity, time deposits are associated with lower churn.

#### Customer-Company Interaction Predictors

A total of three interaction variables are determined to be indicative in the churn tendency: branch negative impact, customer representative negative impact and tenure. Parallel with expectations given in Hypothesis 20, customers whose branch and

customer representative they work with display historically significantly higher-than-average churn rates, they are more likely to churn (corresponding parameters are estimated to be 0.2688 and 0.113, respectively in the W1C-LogReg-MPTD-S model). Further, satisfying the Hypothesis 22, customers with higher tenure are more likely to continue their relationship (corresponding parameter is estimated as -0.2474 in the W1C-LogReg-MPTD-S model).

#### Environmental Predictors

The advantage of the longitudinal training dataset is that it allows the dynamic environmental changes to be entered into the logistic regression model as separate variables which cannot be accomplished in the cross-sectional training media. As illustrated in Table X, the W1C-LogReg-MPTD-S model identifies two environmental predictors to affect the churn behavior: the yearly deposit interest rate and the difference between the monthly and yearly deposit interest rates. The results show that increasing monthly deposit rates result in lower churn tendency (following the parameter estimate of -0.2464 for *TRIMT\_RR\_Line*), but if the difference between monthly and yearly deposit rates increase, then customers are more likely to terminate their relationship, as indicated by the parameter estimate 0.1399 for the corresponding predictor. Further, as the parameter estimates pertaining to *atipi\_ykb\_vs\_market* and *btipi\_ykb\_vs\_market* indicate, the Hypothesis 25 - the more the bank makes their customers gain in comparison to the market, the more likely are they to stay – is not verified in any of the models generated.

#### Demographic Predictors

All the models but the W1C-LogReg-MPTD model identify the customers with missing demographics information to be more likely to churn which should be closely investigated by the bank. In addition, as identified by the W1C-LogReg-MPTD-S model, married customers are less likely to. Accordingly, we can assert that customers who are freer in decision making are more prone to churn.

#### 4.6.2. W2C to W6C Models

For all the models applied, the convergence criterion is satisfied and as the global null hypothesis shows, all the parameter estimates are significantly different than 0. In addition, as the Hosmer-Lemeshow test results indicate, the churn prediction models generated on the MPTD fit the data whereas the models trained on MPTD-S lack it. This is to explain by the creation of artificial observations.

Considering the predictors entered into the models for each specific case and the corresponding parameter estimates, the prediction results are summarized in Table A in Appendix A. These prediction outputs reveal that all the models in question choose the total number products currently in use, the current state of holding Euro, and the percentage change in the portfolio size, no matter which label is modeled and no matter which longitudinal training medium is selected. In addition, except the W1C-LogReg-MPTD model, the customer representative negative impact and customer tenure predictors are specified to be indicative in the churn attitude for all the models. This in turn helps managers to deduce some important insights for the future, especially if the classification analysis cannot be performed. Accordingly, carefully monitored should be the customers (i) who are served by customer representatives with higher-than-average churn rates, (ii) whose portfolio size decreases, (iii) who hold Euro currency in his or her portfolio, (iv) who are relatively new customers, and (v) who currently possess less number of instruments relative to other customers.

#### 4.6.3. Benchmark Models

In this sub-section, we will report the convergence, model fit, parameter estimates, and managerial insights for the Cox regression, MLR, and OLR.

#### 4.6.3.1. Cox Regression

First, we examine whether the proportional hazard assumption is satisfied and evaluate the overall fit of the Cox regression model. Next, we provide the parameter insights and state the related managerial insights.

To test whether the Cox regression model violates the proportionality assumption, we use the method proposed by Karrel and Lee which is described in Section 2.6.3.3. To remember, in this technique, Schoenfeld residuals are computed for each covariate included in the final model. If the correlation coefficient between the failure time order and Schoenfeld residuals is not significantly different from zero for each single covariate, then the proportional hazard assumption is met (i.e. the corresponding p-value should be greater than the significance level). Among 29 variables, the minimum p-value computed for correlation coefficients is 0.0165, hence satisfying the assumption at 0.01 significance level.

To test the overall fit of the Cox regression model, we apply the procedure explained in Section 2.6.3.3 and plot the Cox-Snell residuals against the Nelson-Aalen estimator of the cumulative hazard rate and this yields approximately 45-degree slope. This implies that the model fits the data.

Even though the proportional hazard model does not yield performance as good as the binary logistic models run on MPTD as explained in Section 4.5.5, it is still interesting to investigate the parameter estimates of the Cox regression, because semi-parametric models as the Cox regression analysis are useful for comparative analysis (for example: how does the hazard change if a subject ages by one year?) as claimed by Aytug (2008). Table 4.18 presents the variables and the corresponding parameter estimates. To compute the hazard ratios for one unit increase, we follow the below steps:

- ✓ In the original dataset (the dataset before the variables have been standardized) find two observations for which the values of one of the corresponding variables take on values  $x+1$  and  $x$ .



- ✓ In the dataset trained on the standardized variables, find the corresponding values for those particular observations and compute the difference.
- ✓ Setting this difference into the following equation yields the hazard ratio.

$$\text{hazard ratio} = \exp(\text{difference} * \text{parameter estimate})$$

The following findings among many can be deduced from the hazard ratios, as illustrated in Table 4.18:

**Table 4.18** Hazard ratio estimates of the Cox regression

Predictor	Parameter Estimates
CR_PERC_CHANGE_ENDBAL	-0.22248
PORTFOLIO_BUCKET1	0.19063
CR_LAST_1_MONTH_RETURN_PERCENT	0.2538
TOT_NUM_PROD_CURRENT_USE	-1.178
CU_BTF2	0.15245
CU_BTF3	-0.33476
CU_DVZ1_EUR	0.35049
CU_DVZ1_USD	0.18026
CU_DVZ1_YTL	0.2681
CU_FI1_other	0.22102
CU_FI1_EUR	-0.19906
CU_FI7_YTL	0.14572
CU_HSF2	0.07643
CU_HSF6	0.18625
EU_FI4	0.11829
EU_FI5	0.25454
EU_HSF1	0.1312
EU_HSF5	0.15731
EU_PPF1	0.13744
dFundB	-0.33357
wBond	-0.17271
wFundA	-0.68312
wFundB	0.13693
br_neg_impact	0.18011
rm_neg_impact	0.10641
TR1YT_RR_Line_minus_TR1MT_RR_Lin	-1.7168
GENDER_MISSING_FLAG	0.13374
CU_PPF1	0.20533
dTL	-0.21411

A one unit increase in the total number of products currently used decreases the churn probability by 40.1%, which is in line with the findings of the WnC binary logistic regression models and which satisfies Hypothesis 14.

The short time return covariate (*CR\_LAST\_1\_MONTH\_RETURN\_PERCENT*) has been found to be positively related to the attrition behavior for the W1C logistic regression model on the MPTD, as explained in Section 4.6.1. We observe the same for the proportional hazard modeling. Accordingly, 10% increase in the one-monthly asset return increases the churn probability by 24.4%. This, in turn, violates Hypothesis 10.

As we have discovered in the W1C-LogReg-MPTD model in Section 4.6.1, the historical customer representative and branch negative impact are positively related to the attrition behavior and this is also discovered by the Cox regression. Accordingly, if these variables take on the value 1, the corresponding churn probabilities increase by 120.9% and 227.1%, respectively. This finding is parallel with Hypothesis 20.

#### *4.6.3.2. Multinomial Logistic Regression*

A total of 71 exploratory variables are used in the MLR. This in turn corresponds to 432 parameter estimates (we therefore will not share the coefficients here. For the list of variables, please apply to Table A in Appendix A). Since our objective is not to gain managerial insights from MLR, but to benchmark it to advance churn notice models in terms of accuracy performance, we won't discuss the parameter estimates here. To see the parameter estimates corresponding to MLR, please apply to Table B in Appendix B.

#### *4.6.3.3. Ordinal Logistic Regression*

A total of 16 variables have been included in the final model. The global null hypothesis test yields a p-value less than 0.001 for the likelihood ratio, i.e. the model built is significant. Similar to MLR, we do not aim to extract managerial insights from OLR. For the parameter estimates, please apply to Table C in Appendix C.

#### 4.6.4. Meta-Combination Ensemble Method

In terms of operational use, the advance churn scores obtained in the first layer can play an important role in planning and adjusting retention activities over time. Here, we will present the output tree of Case5-200 which was one of the recommended ensembles based on accuracy evaluation. It uses the selected variables as the independent variables in the second layer of the ensemble method, along with the probability scores  $p1$ - $p6$  obtained from the LogReg-MPTD models in the first layer.

As illustrated in Figure 4.3, a total of 23 rules have been generated. We will not report the regression functions for each of the rule generated; but report the important findings to be deduced from this tree. As the tree indicates, the following variables are used while building the regression tree: within-one-period churn probability ( $p1$ ), within-two-periods churn probability ( $p2$ ), within-three-periods churn probability ( $p3$ ), portfolio size (*END\_MONTH\_BALANCE*), customer age (*age*), customer tenure (*CUST\_MONTH\_OF\_YKB*), “A tipi” percentage gain from YKB investments relative to the market index (*atipi\_ykb\_vs\_market*), one monthly relative asset return (*RELATIVE\_RETURN\_1*), and twelve monthly relative asset return (*RELATIVE\_RETURN\_12*).

The within-one-period churn probability score is the exploratory variable used to split the observations into two segments in the first level and the observations with very low  $p1$  scores ( $p1 \leq 0.001$ ) are further divided into two parts based on the *age* variable. The regression function coefficients corresponding to these two segments (decision rules LM1 and LM2) differ from each other in the intercept estimate: The intercept belonging to the second rule is greater than the intercept of the first rule. Accordingly, the following finding can be deduced: For customers who are very unlikely to churn within one period, those who are older, are more likely to churn. The same interpretation is valid for the decision rules LM14 and LM15.

For customers whose within-one-period churn probability score is computed to be greater than 0.024, the main splitting criteria are the portfolio size and the customer

tenure. Following the regression coefficient estimates, we infer that the customers with greater within-one-period churn probability and lower customer tenure (LM21) are more likely to cease their relationship than those with greater p1 scores and greater tenure. This is parallel with expectations of that churn is negatively related customer tenure and satisfies Hypothesis 22.

```

p1 <= 0.001 :
  AGE <= 0.868 : LM1
  AGE > 0.868 : LM2
p1 > 0.001 :
  p1 <= 0.007 : LM3
  p1 > 0.007 :
    p1 <= 0.013 :
      END_MONTH_BALANCE <= -0.214 :
        END_MONTH_BALANCE <= -0.233 : LM4
        END_MONTH_BALANCE > -0.233 :
          AGE <= -0.346 :
            atipi_ykb_vs_market <= 0.972 :
              RELATIVE_RETURN_1 <= 0.993 : LM5
              RELATIVE_RETURN_1 > 0.993 :
                RELATIVE_RETURN_12 <= -0.246 : LM6
                RELATIVE_RETURN_12 > -0.246 :
                  CUST_MONTH_OF_YKB <= -0.845 : LM7
                  CUST_MONTH_OF_YKB > -0.845 : LM8
            atipi_ykb_vs_market > 0.972 : LM9
          AGE > -0.346 : LM10
      END_MONTH_BALANCE > -0.214 :
        END_MONTH_BALANCE <= -0.16 : LM11
        END_MONTH_BALANCE > -0.16 :
          p2 <= 0.013 : LM12
          p2 > 0.013 :
            p3 <= 0.02 : LM13
            p3 > 0.02 :
              AGE <= -1.02 : LM14
              AGE > -1.02 : LM15
    p1 > 0.013 :
      p1 <= 0.024 :
        END_MONTH_BALANCE <= -0.22 :
          END_MONTH_BALANCE <= -0.245 : LM16
          END_MONTH_BALANCE > -0.245 : LM17
        END_MONTH_BALANCE > -0.22 : LM18
      p1 > 0.024 :
        END_MONTH_BALANCE <= -0.227 :
          END_MONTH_BALANCE <= -0.268 : LM19
          END_MONTH_BALANCE > -0.268 : LM20
        END_MONTH_BALANCE > -0.227 :
          END_MONTH_BALANCE <= -0.14 :
            CUST_MONTH_OF_YKB <= -1.429 : LM21
            CUST_MONTH_OF_YKB > -1.429 : LM22
          END_MONTH_BALANCE > -0.14 : LM23

```

**Figure 4.3** The regression tree output for Case5-200: Ensemble methodology

#### 4.6.5. Contributions

Along with the increase in the predictive performance, the models trained on the longitudinal training dataset also provide managerial insights which cannot be extracted from the models on cross-sectional training media. For example, in terms of the product

types owned currently, the W1C-LogReg-SPTD model can associate the churn behavior with only three different instruments whereas the W1C-LogReg-MPTD model identifies ten financial instruments whose current ownership has an impact on the churn attitude. This implies that the MPTD models are capable of generating more profound insights.

One surprising finding is that the customers' churn propensity increases as their short-time asset return increases and as their asset return is getting relatively higher than the average customer. This implies that customers who are literate in financial services are more prone to churn, and we call this as the sophistication impact. In addition, it has been found that number of products ever-used increases churn propensity, as captured by SPTD-S model and other models pick specific products that if the customer has ever used them, the churn propensity increases. Kamakura et al (1991) indicate that the more the customer sophistication grows in the financial services, the more risky and complicated instruments customers tend to own. On this ground, having tried different financial instruments can be seen as a measure of customer sophistication. When cogitated together with the parameter estimates of short-time asset return and long-term relative asset return covariates, it can be deduced that more sophisticated customers are more prone to cease their relationship.

With respect to the ensemble method, even though no significant improvement can be provided in terms of accuracy performance, the second layer of the ensemble method outputs a regression, hence dividing customers into segments from which further helpful managerial insights can be extracted, as we did in Section 4.6.4.

## Chapter 5

### Conclusion

In this study, we tackled the churn prediction problem in non-contractual settings and under dynamic environments. There are four complications that make it difficult: First, under non-contractual settings, the customer can switch to competition anytime he or she wants to and her intention is not observable until she churns. Second, beyond the customer characteristics, time varying environmental conditions also play an important role in a churn decision which cannot be detected in typical, cross-sectional churn analyses. Third, churn is a rare event; hence generally, few positive examples are included in the churn data. Last, at the time of scoring, the current information might not be available.

This study developed a method to address the above difficulties of customer churn prediction. It uses a longitudinal training dataset with multiple observations per customer from different time periods and thus increases the number of positive examples without introducing artificial noise like the synthetic methods. By virtue of multiple time periods the model is exposed to different values of the environmental variables and learns how they affect customer behavior. Finally, the heterogeneity in customer behavior is captured in multiple advance churn notice score models that are combined with a regression tree into a churn prediction method which didn't succeed. In terms of the ensemble taxonomy proposed by Rokach (2009), our ensemble method can be classified as a meta-combination method where the classifiers are induced independently. The mechanism for generating diversity in the predictions is prediction of churn multiple months ahead, which should correspond to the heterogeneity in customer behavior.

We are not the first researchers in using the longitudinal training data to develop churn prediction models. It has been applied previously by Jamal and Bucklin (2006). However, we are the first in explicitly comparing the accuracy performances of models generated on longitudinal (MPTD in our context) and cross-sectional (SPTD in our context) customer data and in demonstrating that the use of MPTD provides significant improvement, both in terms of predictive accuracy and managerial insights.

Our results demonstrate that training a churn classification model to predict the within-one-period churn on the MPTD via logistic regression improves the AUC, TDL, and TQL scores by 10.4%, 53.6%, and 38.1%, respectively for the first test period April 2010, when compared to the logistic regression model trained on SPTD. Even the prediction model obtained from the oversampled SPTD, namely SPTD-S, could not approach the performance of the MPTD. This can be explained as follows: The MPTD possesses more observations and is capable of projecting the environmental changes into the model because it uses longitudinal data from different time periods. In addition, all the observations are real. On the other hand, the SPTD includes only cross-sectional data, thus having much fewer churn examples and not incorporating time variation. This in turn lets the trained model not to generalize, but to memorize. Other than the improvement in the accuracy performance, we have also provided one important managerial insight: customer sophistication impact. Accordingly, thanks to the use of longitudinal training data, we have been able to identify the customers, who are more literate in financial services, as more prone to churn.

Next, we have found out that the MPDT accuracy continues to be better than the SPTD for at least five future time periods (until August 2010). In addition, the significance of this improvement has been verified by conducting Student's t-test on ten bootstrap samples. Further, we have shown that the MPTD model yields accurate results if the recent customer information is missing at the time of scoring, even better than the SPTD model which is assumed not to confront any 'information lead time' problem.

Further, we developed additional advance churn notice models with the purpose of predicting within-two-periods through within-six-periods churn. Even though these models use less customer information (by dropping out the recent customer information) in the model generation phase, as described in Section 4.2, the performance results imply that they outperform the model trained on SPTD with the most recent customer information. The findings show that the accuracy deteriorates as we predict churn for greater time periods, but they are still useful. In addition, these models help to distinguish among the churn propensities from different time perspective.

Last, we have further improved the point estimate predictive performance by applying a two-layer ensemble method that uses within-x-periods churn scores,  $p1$  to  $p6$ , that are meant to reflect the customer's churn intention, in conjunction with other explanatory variables regarding demographics, customer behavior and environmental conditions. A further 15%-improvement in the TDL has been achieved for the ensemble method compared with the MPTD. However, as the Student's t-test on ten bootstrap samples points out, this improvement is not found to be statistically significant. In addition, the overall TDL improvement from SPTD to the ensemble method corresponds to 76.6%.

This research can be extended into at least three directions. The first research direction is to repeat the analysis in this paper for some other research data from private banking or another industry to determine under which conditions this modeling approach to use longitudinal training data provide more favorable results. The second research direction includes using  $L_1$ -norm regularization for the variable selection and repeating the analysis from scratch for the same research data. The third research direction involves developing another ensemble methodology to significantly increase the predictive performance compared to the accuracy results obtained from the W1C-LogReg-MPTD and W1C-LogReg-MPTD-S models.



APPENDICES

Appendix A: Parameter Estimations of the WnC-LogReg Models

Table A. Parameter estimates of the WnC-LogReg models

Predictor Category	Predictor	WnC		WSC		WnC		WSC		WnC		WSC	
		SPTD	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS	MPDTS
Cust. Behavior	Intercept	-6.225	-1.9612	-5.9249	-1.006	-5.0972	-0.6993	-4.5305	-0.5054	-4.3051	-0.5012	-3.8261	-0.4836
Cust. Behavior	CR_FEE_CHANGE_ENDBAL	-0.3506*	-0.4911	-0.236	-0.289	-0.1615	-0.1868	-0.2123	-0.1959	-0.233	-0.2368	-0.2318	-0.2449
Cust. Behavior	PORTFOLIO_BUCKET1	0.3818	0.1947*	0.7759	0.1905	0.2335	0.1852	0.1852	0.2236	0.1661	0.1896	0.1534	0.1451
Cust. Behavior	PORTFOLIO_BUCKET4	-0.6608**	-1.215	0.3118	0.3224	-0.1795	-0.1795	-0.1498	-0.1498	-0.1498	-0.1498	-0.164	-0.1887
Cust. Behavior	NUM_OF_AIL_ACC	0.1939	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598	0.1598
Cust. Behavior	CR_LAST_1_MONTH_RETURN_PERCENT	0.1981***	-0.0994***	0.1544	-0.0994***	0.1544	-0.0994***	0.1544	-0.0994***	0.1544	-0.0994***	0.1544	-0.0994***
Cust. Behavior	CR_LAST_3_MONTH_RETURN_PERCENT	0.5139	0.1435*	0.1544	0.1435*	0.1544	0.1435*	0.1544	0.1435*	0.1544	0.1435*	0.1544	0.1435*
Cust. Behavior	CR_LAST_12_MONTH_RETURN_PERCENT	-0.9593*	-1.27	-0.9426	-0.9745	0.1544	-0.9426	0.1544	-0.9426	0.1544	-0.9426	0.1544	-0.9426
Cust. Behavior	RELATIVE_RETURN	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	TOT_NUM_PROD_CURRENT_USE	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_BTF5	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_BTF6	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_DVZ1_EUR	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_DVZ1_USD	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_other	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_EUR	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_USD	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_YTL	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_EUR	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_FH_YTL	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_HRF2	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_HRF5	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_HRF6	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	CU_PPF2	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	TOT_NUM_PROD_EVER_USED	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_BTF2	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_BTF6	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_DVZ1_EUR	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_DVZ1_USD	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_FH_other	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_FH	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_FH5	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_HRF1	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_HRF5	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_HRF7	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	EU_PPF1	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	aver_riskscore_cr	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456	-0.456
Cust. Behavior	estscore_cr	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026	0.2026
Cust. Behavior	dBond	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014	-1.0014
Cust. Behavior	dEurobond	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901
Cust. Behavior	dFundA	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*	-0.322*
Cust. Behavior	dFundB	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781	-0.3781
Cust. Behavior	dGold	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514	-0.1514
Cust. Behavior	dImdeposit	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605	-0.1605
Cust. Behavior	wBond	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165	-0.165
Cust. Behavior	wEurobond	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025	-0.3025
Cust. Behavior	wFundA	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*	0.1907*
Cust. Behavior	wFundB	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395	0.3395
Cust. Behavior	wPioneer	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	wStock	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Cust. Behavior	wTimeDeposit	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Interaction	br_neg_amaact	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*	-0.4555*
Interaction	br_neg_impact	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901	0.1901
Interaction	CRUST_MONTH_OF_YKB	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Environment	TROND_Line	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464	-0.2464
Environment	TRIMT_PP_Line	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999	0.1999
Environment	TRIMT_PP_Line_minus_TRIMT_PP_Line	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Environment	USD_buying	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Environment	GENDER_MISSING_FLAG	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	NAT_MISSING_FLAG	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	EDU_COLLEGE	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	EVP_OWN_BUSINESS	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	EVP_EMPLOYED	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	NAT_STAT_EU1	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Demographics	SBC_TYPE_ENM_UH	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345

\* Significant at 0.01; \*\* Significant at 0.05; \*\*\* Significant at 0.10; if no label: Significant at <0.0001

**Appendix B: Parameter Estimations of the Multinomial Logistic Regression**

**Table B.** Parameter estimates of the multinomial logistic regression

Parameters	Months to churn					
	1	2	3	4	5	6
Intercept	-	-	-	-	-	-
aver_riskscore_cr	.	.	.	.	.	.
br_neg_impact	+	.	.	.	.	.
cashscore_cr	.	.	.	.	.	.
CR_LAST_1_MONTH_RETURN_PERCENT	+	.	+	.	.	.
CR_LAST_12_MONTH_RETURN_PERCENT	.	.	.	.	.	.
CR_LAST_3_MONTH_RETURN_PERCENT	.	.	.	.	.	.
CR_PERC_CHANGE_ENDBAL	-	-	-	-	.	-
CU_BTF2	.	.	.	.	.	.
CU_BTF3	-	.	.	.	.	.
CU_BTF6	.	.	.	.	.	.
CU_DVZ1_EUR	.	.	.	+	+	+
CU_DVZ1_USD	.	.	.	.	.	+
CU_DVZ1_YTL	.	.	.	.	.	+
CU_FI1_EUR	-	-	-	-	-	.
CU_FI1_other	.	.	.	.	.	.
CU_FI1_USD	.	.	.	.	.	.
CU_FI1_YTL	.	.	.	.	.	.
CU_FI5	.	.	.	.	.	.
CU_FI6_EUR	.	.	.	.	.	.
CU_FI7_YTL	.	.	.	.	.	.
CU_HSF2	.	.	.	.	.	.
CU_HSF5	.	.	.	.	.	.
CU_HSF6	.	.	.	.	.	.
CU_PPF2	.	.	.	.	.	.
CUST_MONTH_OF_YKB	-	-	-	-	-	-
dBond	.	.	.	.	.	.
dEurobond	.	.	.	.	.	.
dFundA	.	.	.	.	.	.
dFundB	-	-	-	-	.	.
dGold	.	.	.	.	.	.
dTimedeposit	.	-	.	.	.	.
EDU_COLLEGE	.	.	.	.	+	+
EMP_EMPLOYED	.	.	.	.	.	.
EMP_OWN_BUSINESS	.	.	.	.	.	.

**Table B.** cont'd

Parameters	Months to churn					
	1	2	3	4	5	6
EU_BTF1	.	.	.	.	.	.
EU_BTF2	.	.	.	.	.	.
EU_BTF6	.	.	.	.	.	.
EU_DVZ1_EUR	+	.	.	.	.	.
EU_DVZ1_USD	.	.	.	.	.	.
EU_FI1_other	.	.	.	.	.	.
EU_F14	+	+	+	.	.	.
EU_F15	.	+	+	+	+	+
EU_HSF1	+	.	.	.	.	.
EU_HSF5	.	.	.	.	.	.
EU_HSF7	.	.	.	.	.	.
EU_PPF1	.	.	.	+	+	.
GENDER_MISSING_FLAG	+	.	.	+	.	.
IMKB100	.	.	.	.	.	.
MAR_STAT_EVLI	.	.	.	.	.	.
NUM_OF_ALL_ACC	.	.	.	.	.	.
PORTFOLIO_BUCKET1	.	+	+	+	.	.
PORTFOLIO_BUCKET4	.	.	.	.	.	.
RELATIVE_RETURN	.	.	.	.	.	.
rm_neg_impact	.	.	.	+	.	.
SBU_TYPE_NM_UH	.	.	.	.	.	.
TOT_NUM_PROD_CURRENT_USE	.	.	.	.	-	-
TOT_NUM_PROD_EVER_USED	.	.	.	.	.	.
TR1MT_RR_Line	.	.	.	.	.	.
TR1YT_RR_Line_minus_TR1MT_RR_Line	.	.	.	.	.	.
TRYOND_Line	.	.	.	.	.	.
USD_buying	.	.	+	.	.	.
wBond	.	.	.	.	.	.
wEurobond	.	.	.	.	.	.
wFundA	.	.	.	.	.	.
wFundB	.	.	.	.	.	.
wFX	.	.	.	.	-	.
wPioneer	.	.	.	+	+	+
wStock	.	.	.	.	.	.
wTimedeposit	.	.	.	.	.	.

Please note that here in Table B, we only report the sign of the parameter estimates. The plus sign (+) denotes that the corresponding variables have a positive impact on the churn behavior, i.e. an increase in the value of those variables increases the churn probability. Please also remark “.” denotes that the corresponding variable is found not to be statistically significant in explaining the churn attitude, at a significance level of 0.1.

**Appendix C: Parameter Estimations of the Ordinal Logistic Regression****Table C.** Parameter estimates of the multinomial logistic regression

Predictor	months-to- churn	Parameter Estimate
Intercept	1	-6.0307
Intercept	2	-5.363
Intercept	3	-4.9914
Intercept	4	-4.736
Intercept	5	-4.5464
Intercept	6	-4.403
CR_PERC_CHANGE_ENDBAL		-0.2061
PORTFOLIO_BUCKET1		0.1426
TOT_NUM_PROD_CURRENT_USE		-0.6481
CU_BTF2		0.157
CU_BTF3		-0.1888
CU_DVZ1_EUR		0.2611
CU_FI1_other		0.1374
CU_FI1_EUR		-0.2685
EU_FI5		0.2718
dBond		-0.1889
dFundA		0.2084
wPioneer		0.0502
rm_neg_impact		0.1256
TUFE_12month_inflation		-0.4468
DJI_SMA		-0.2899
CUST_MONTH_OF_YKB		-0.2501

## BIBLIOGRAPHY

Adeleke, K. A., & Adepoju, A. A. (2010). Ordinal logistic regression model: An application to pregnancy outcomes. *Journal of Mathematics and Statistics* 6 (3) , 6 (3), 279-285.

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy* , 30, 552-568.

Allison, P. D. (2003). *Logistic Regression Using the SAS System: Theory and Application*. Cary, NC: SAS Institute Inc.

Aytuğ, H. (2008, June 17). Conference Notes: Survival Analysis. Koc University, Istanbul, Turkey.

Bolton, R. N., Kannan, P. K., & Bramlett, M. D. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academic of Marketing Science* , 28 (1), 95-108.

Breiman, L. (1996). Bagging predictors. *Machine Learning* , 24 (2), 123-140.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research* , 164, 252-268.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications* , 32, 277-288.

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

Burez, J., & Van den Poel, D. (2008). Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. *Expert Systems with Applications* , 35, 497-514.

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

Cherrie, J. A. (2007). Variable screening for multinomial logistic regression on very large datasets as applied to direct response modeling. Brewster, NY. Retrieved June 23, 2011, from SAS Global Forum: <http://www2.sas.com/proceedings/forum2007/081-2007.pdf>

Chu, B. H., Tsai, M. S., & Ho, C. S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems* , 20, 703-718.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of marketing* , 28 (3), 375-387.

Coussement, K., & Van den Poel, D. (2008a). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications* , 34, 313-327.

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications* , 36, 6127-6134.

Coussement, K., & Van den Poel, D. (2008b). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management*, 45, 164-174.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, 37, 2132-2143.

Dawes, J., & Swailes, S. (1999). Retention sans frontières: Issues for financial service retailers. *International Journal of Bank Marketing*, 17 (1), 36-43.

Dawkins, P., & Reichheld, F. F. (1990). Customer retention as a competitive weapon. *Directors and Boards*, 14, 41-47.

Dekimpe, M. G., & Degraeve, Z. (1997). The attrition of volunteers. *European Journal of Operational Research*, 98 (1), 37-51.

Ennew, C. T., & Binks, M. R. (1996). The impact of service quality and service characteristics on customer retention: Small businesses and their banks in UK. *British Journal of Management*, 7 (3), 219-230.

George, C. P. (2010). Retrieved July 14, 2011, from University of Florida - Elements of Statistical Learning Web site:  
[http://www.cise.ufl.edu/class/cis6930sp10esl/downloads/Clint\\_9\\_2.pdf](http://www.cise.ufl.edu/class/cis6930sp10esl/downloads/Clint_9_2.pdf)

Ghorbani, A., Taghiyareh, F., & Lucas, C. (2009). The application of the locally linear model tree on customer churn prediction. *International Conference of Soft Computing and Pattern Recognition*, (pp. 472-477).

Gladly, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197, 402-411.



Guo-en, X., & Wei-dong, J. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering - Theory & Practice* , 28 (1), 71-77.

Harrel, F., & Lee, K. (1986). Verifying assumptions of the proportional hazards model. *Proceedings of the Eleventh Annual SAS User's Group International Conference*, (pp. 823-828).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). New York, NY: Springer.

Hisarcıklılar, M. (2004). Lecture Notes: İkili Seçim Modelleri (Binary Choice Models). Istanbul, Turkey: Istanbul Technical University.

Huber, C. P., Lane, K. R., & Pofcher, S. (1998). Format renewal in banks - it's not easy. *McKinsey Quarterly* , 1998 May (2), 148-156.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications* , 31, 525-524.

Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. *Journal of Interactive Marketing* , 20, 16-29.

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing* , 8, 329-349.

Karahoca, A., & Karahoca, D. (2010). GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system. *Expert Systems with Applications* , 38 (3), 1814-1822.

Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing* , 59 (2), 71-82.

Keaveney, S. M., & Parthasarathy, F. D. (2001). Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. *Journal of the Academy of Marketing Science* , 29 (4), 374-390.

Kim, S., Jung, T., Suh, E., & Hwang, H. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications* , 31, 101-107.

Kim, S., Shin, K. S., & Park, K. (2005). An application of support vector machines for customer churn analysis: Credit card case. *Lecture Notes in Computer Science* , 3611, 636-647.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2 ed.). New York, NY: Springer.

Kumar, D. A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies* , 1 (1), 4-28.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications* , 27, 277-285.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forest techniques. *Expert Systems with Applications* , 29, 472-484.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research* , 43, 276-286.

- Lessmann, S., & Voß, S. (2009). A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research* , 199, 520-530.
- Lima, E., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of the Operational research Society* , 60, 1096-1106.
- Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications* , 38 (1), 8-15.
- Mittal, V., & Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research* , 38 (1), 131-142.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* , 43, 204-211.
- Norusis, M. J. (2010). *PASW Statistics 18 Advanced Statistical Procedures: Sample Chapter*. Retrieved July 28, 2010, from [http://www.norusis.com/pdf/ASPC\\_v13.pdf](http://www.norusis.com/pdf/ASPC_v13.pdf)
- Pendharkar, P. C. (2009). Genetic algorithm based neural network approaches for predicting churn in cellular wireless network. *Expert Systems with Applications* , 36, 6714-6720.
- Portela, S., & Menezes, R. (2010). An empirical investigation of the factors that influence the customer churn in the Portuguese fixed telecommunications industry: A survival analysis application. *The Business Review* , 14 (2), 98-104.

- Prinzie, A., & Van den Poel, D. (2006). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. *Decision Support Systems* , 42, 508-526.
- Qi, J., Zhang, L., Liu, Y., Li, L., Zhou, Y., Shen, Y., et al. (2009). ADTreesLogit model for customer churn prediction. *Annals of Operations research* , 168 (1), 247-265.
- Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.
- Quinlan, R. J. (1992). Learning with continuous classes. *5th Australian Joint Conference on Artificial Intelligence*, (pp. 343-348). Singapore.
- Reichheld, F. F. (1996). Learning from customer defections. *Harvard Business Review* , 74, 56-69.
- Reichheld, F. F., & Kenny, D. W. (1990). The hidden advantages of customer retention. *Journal of retail Banking* , 12 (4), 19-23.
- Sayman, S., & Demiroğlu, C. (2011). *Golden Offer project: Customer scorecard report*. Koc University, College of Administrative Sciences and Economics, Istanbul.
- Tsai, C. F., & Chen, M. Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications* , 37, 2006-2015.
- Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications* , 36, 12547-12553.
- Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazards models. *European Journal of Operational Research* , 157, 196-217.

Van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research* , 181, 436-452.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications* , 38 (3), 2354-2364.

Walfried, M. L., Manolis, C., & Winsor, R. D. (2000). Service quality perspectives and satisfaction in private banking. *Journal of Services Marketing* , 14 (3), 244-271.

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunication call details to churn prediction: A data mining approach. *Expert Systems with Applications* , 23, 103-112.

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balance random forests. *Expert Systems with Applications* , 36, 5445-5449.

Zhang, G. (2007). Customer retention based on BP ANN and survival analysis. *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, (pp. 3406-3411). Shanghai.

Zhao, Y., Li, B., Li, X., & Ren, S. (2005). Customer Churn Prediction Using Improved One-Class Support Vector Machine. In *Advanced Data Mining and Applications* (Vol. 3584/2005, pp. 300-306). Berlin: Springer.

**VITA**

Umut Arıtürk was born in Istanbul, Turkey, on April 27, 1985. He graduated from Sankt Georg Austrian College in 2004. He received his B.Sc degree in Management Engineering from Istanbul Technical University, Istanbul 2009. Same year, he joined the M.Sc program in Industrial Engineering at Koc University as a research and teaching assistant.