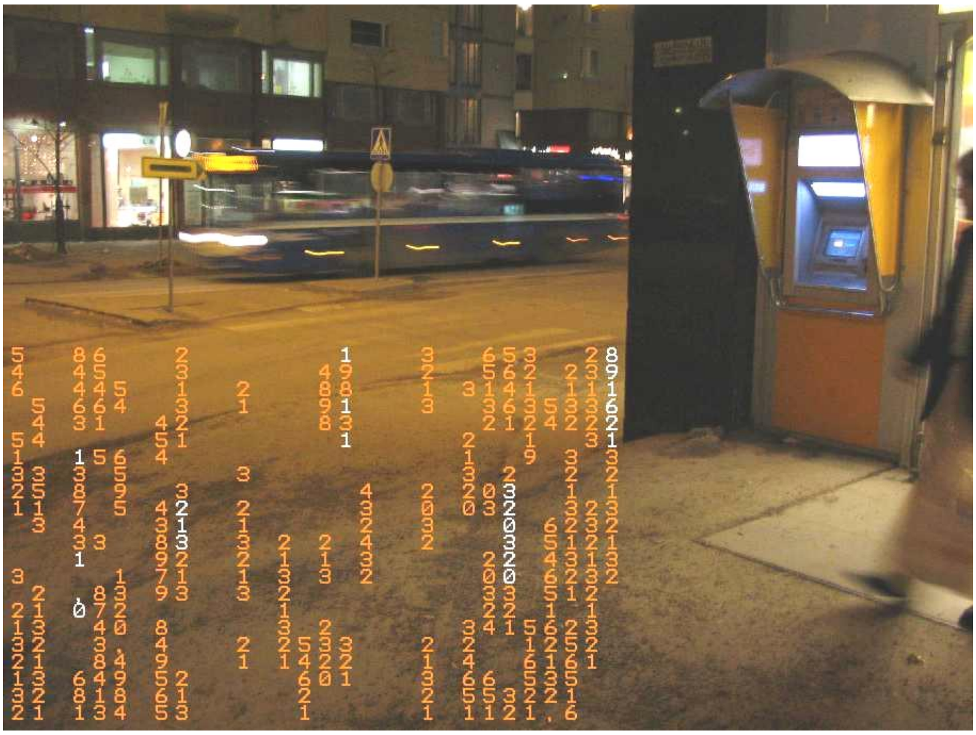
RESEARCH REPORT No VTT-R-01184-06 15.3.2006



Customer churn analysis – a case study

Author Teemu Mutanen

TABLE OF CONTENTS

ABSTRACT

Customer value analysis is critical for a good marketing and a customer relationship management strategy. An important component of this strategy is the customer retention rate. Customer retention rate has a strong impact on the customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Conventional statistical methods are very successful in predicting a customer churn. The goal of this study is to apply logistic regression techniques to predict a customer churn and analyze the churning and no-churning customers by using data from a personal retail banking company.

1. Introduction

The subject of customer retention, loyalty, and churn is receiving attention in many industries. This is important in the customer lifetime value context. A company will have a sense of how much is really being lost because of the customer churn and the scale of the efforts that would be appropriate for retention campaign. The mass marketing approach cannot succeed in the diversity of consumer business today. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers.

Personal retail banking sector is characterized by customers who stays with a company very long time. Customers usually give their financial business to one company and they won't switch the provider of their financial help very often. In the company's perspective this produces a stabile environment for the customer relationship management. Although the continuous relationships with the customers the potential loss of revenue because of customer churn in this case can be huge.

This paper will present a customer churn analysis in personal retail banking sector. The goal of this paper is twofold. First the churning customers are analyzed in banking context. The second objective is a forecast of churning customers based on a logistic regression model.

After the introduction this paper has 6 sections. The background for customer lifetime value concept is presented in the chapter 2. There is also a literature review about the customer churn included in the chapter 2. The methods used in this study are presented in the chapter 3 while a closer look at the case data is taken in the chapter 4. The focus of this case study is described in the chapter 5. All the results of the churn prediction are presented in the chapter 6. And conclusions of this study are left for the chapter 7. The analysis part of this paper was conducted by using both Matlab [20] and SPSS [21]software.

2. The need for customer churn prediction

Our case data in this paper was provided by a company operating in a retail banking sector. In personal retail banking a company must operate on a long term customer strategy, young customers are recognized as being unprofitable in the early stage in lifecycle but will become profitable later on. So as the customer relationships last, maybe decades, the company must address the value of a potential loss of a customer. The customer lifetime value analysis will help to face this challenge.

2.1. The customer lifetime value concept

The customer lifetime value is usually defined as the total net income from the customer over his lifetime [13]. This type of customer analysis is done under several terms: customer value, customer lifetime value, customer equity, and customer profitability. The underlying idea in LTV concept is simple and measuring the lifetime value is easy after the customer relationship is over. The challenge in this concept is to define and measure the customer lifetime value during, or even before, the active stage of customer relationship.

For example Hoekstra et al. [13] defines a conceptual LTV model as follows:

> *LTV is the total value of direct contributions and indirect contributions to overhead and profit of an individual customer during the entire customer life cycle, that is from start of the relationship until its projected ending.*

Most LTV models stem from the basic equation, although there are also many other LTV models having various application areas. The components of the basic LTV model are [3]:

- The customer net present value over time (revenue and cost).
- Retention rate or length of service (LoS).
- Discount factor.

Each component can be measured or estimated separately and then combined for the LTV model.

The benefits of better understanding the customer lifetime value are numerous. The company can measure the present and the future income from the customers. The company can also foster customer retention and loyalty which will lead to higher customer profitability. The LTV analysis can also help the company on their customization of products and services. This understanding of the customer value helps the company to focus on revenue productive customers and yield the customer segment with potential negative impacts to the revenue. And last, the customer lifetime value is not a fixed value it can be influenced by marketing efforts.

2.2. Customer churn

The focus on customer churn is to determinate the customers who are at risk of leaving and if possible on the analysis whether those customers are worth retaining. The churn analysis is highly dependent on the definition of the customer churn. The business sector and customer relationship affects the outcome how churning customers are detected. Example in credit card business customers can easily start using another credit card, so the only indicator for the previous card company is declining transactions. On the other hand for example in Finnish wireless telecom industry a customer can switch one carrier to another and keep the same phone number. In this case the previous carrier will get the signal right at the churning moment.

The customer churn is closely related to the customer retention rate and loyalty. Hwang et al. [14] defines the customer defection the hottest issue in highly competitive wireless telecom industry. Their LTV suggest that churn rate of a customer has strong impact to the LTV value because it affects the length of service and the future revenue. Hwang et al. also defines the customer loyalty as the index that customers would like to stay with the company. Churn describes the number or percentage of regular customers who abandon relationship with service provider [14].

$$\text{Customer loyalty} = 1 - \text{Churn rate}$$

Modeling customer churn in pure parametric perspective is not appropriate for LTV context because the retention function tends to be “spiky” and non-smooth, with spikes at the contract ending dates [18]. And usually on the marketing perspective the sufficient information about the churn is the probability of possible churn. This enables the marketing department so that, given the limited resources, the high probability churners can be contacted first [2].

Table 1 examples of the churn prediction in literature.

article	market sector	case data	methods used
Au et al. [2]	Wireless telecom	100 000 subscribers	DMEL - method (data mining by evolutionary learning)
Buckinx et al. [4]	Retail business	158 884 customers	Logistic regression, ARD (automatic relevance determination), decision tree
Buckinx et al. [5]	Daily grocery	878 usable responses	MLR (multiple linear regression), ARD, and decision tree
Ferreira et al. [10]	Wireless telecom	100 000 subscribers	Neural network, decision tree, hierarchical neuro-fuzzy systems, rule evolver
Gatland [11]	Retail banking	1 100 customers	Multiple regression
Hwang et al. [14]	Wireless telecom	16 384 customers	Logistic regression, neural network, decision tree
Mozer et al. [16]	Wireless telecom	46 744 subscribers	Logistic regression, neural network, decision tree

Table 1 presents examples of the churn prediction studies found in literature. The methods used for churn analysis are presented in the table along with a case data size and market sector information. Buckinx et al. measures the loyalty and churn rate differently in retail setting. The loyal customers are those who shop frequently and at the same time exhibit a regular buying pattern [4]. In this retail setting the customer churn is defined as customers who switch their purchases to another store. This is hard to detect because customers may still have transactions in the previous store. So Butnix et al. classify the customer a partial defective if he deviates from his established buying behavior [4]. This is possible because in their setting they focus only on loyal clients.

Personal retail banking sector is a typical market sector where a customer is not regularly switching from one company to another. Customers usually give their banking business to one or two companies for long periods of time. This makes customer churn a priority for most companies in the banking sector. Garland has done research on customer profitability in personal retail banking [11]. Although their main focus is on the customers' value to the study bank, they also investigate the duration and age of customer relationship based on profitability. His study is based on customer survey by mail which helped him to determine the customer's share of wallet, satisfaction and loyalty from the qualitative factors.

3. Methods

3.1. Logistic regression

Binomial (binary) logistic regression is a form of regression which is used in a situation when dependent is not a continuous variable but a state which may or may not happen, or a category in a specific classification [8]. Logistic regression can be used to predict a discrete outcome on the basis of continuous and/or categorical variables. Multinomial logistic regression exists to handle the case of dependents with more classes than two.

Although logistic regression has been used in variety of areas, for example in childhood ADHD context [19], logistic regression has also been used in customer analysis. For example Buckinx et al. have used logistic regression for predicting partially defect customers in retail setting [4]. Multinomial regression has been used for predicting the customer's future profitability, based on his demographic information and buying history in the book club [1].

In the logistic regression there can be only one dependent variable. Logistic regression applies maximum likelihood estimation after transforming the dependent into a logistic variable [8]. Unlike the normal regression model the dependent variable in logistic regression is usually dichotomous: the dependent variable can take value 1 with probability q and value 0 with probability 1-q.

The logistic regression is presented here as it is presented on the book by J.S. Cramer [8]. The logistic regression model has history in biological science sector. The normal regression model may be briefly revived by specifying

$$P(X) = \alpha + \beta X \text{, where } X = (x_1, x_2, \mathrm{K}, x_n),$$

which is the linear probability model. It leads to the solution estimation by linear regression methods. In order to restrict the P(X) to the observed values of 0 and 1, complex properties must be attributed to the disturbance ε. If we wish to hold the probability P(X) between the bound 0 and 1 and to vary monotonically with X, we have to use other functions than linear functions. One of these functions that meet the requirements is logistic function, that is

$$P(X) = \frac{e^{-(\alpha+\beta X)}}{1+e^{-(\alpha+\beta X)}}$$

$$Q(X) = 1 - P(X) = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

So in the normal regression model with n inputs the output is obtained by using the formula:

$$P(X) = \alpha + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n,$$

in the logistic regression model the corresponding formula is:

$$Q(X) = \frac{1}{1+e^{-(\alpha+b_1x_1+b_2x_2+\ldots+b_nx_n)}}.$$

The results of the continuous probabilities that are produced by the logistic regression model will be discriminated into two groups by using a threshold value. Usually this threshold value is 0.5, and in this paper the threshold value will separate the churners from nonchurners.

3.2. Lift Curve

In this paper we use binary prediction, ‘churn’ and ‘no churn’. We will analyze the estimation results of the logistic regression by using lift curve. The lift curve is related to the ROC curve of signal detection theory and precision-recall curve in the information retrieval literature [16]. The lift is a measure of a predictive model calculated as the ratio between the results obtained with and without the predictive model.

The Figure 1 shows a lift curve indicating perfect separation of types ‘churn’ and ‘no churn’: all churning customers are detected by the prediction model. The figure also represents a situation where no separation between customers has been done. This type of situation occurs when the churn probabilities are random.

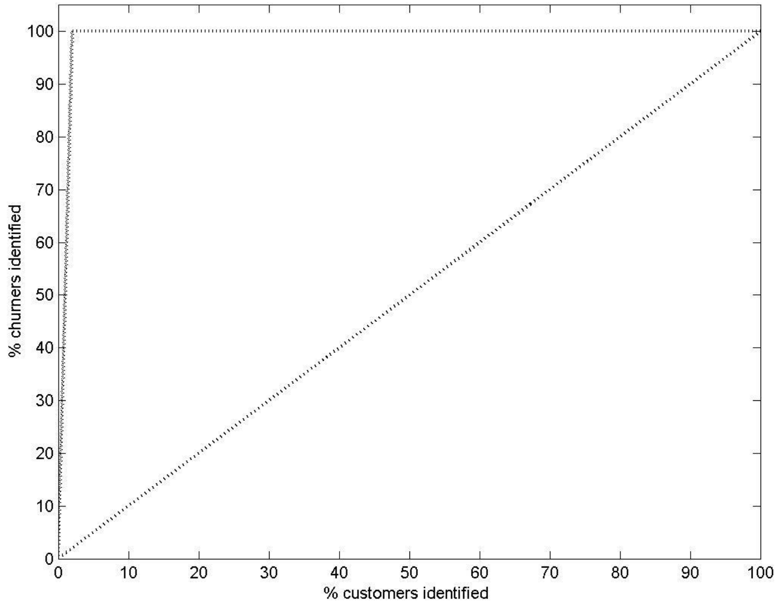


Figure 1 Lift curve for indicating perfect discrimination and no discrimination of churners and nonchurners.

The lift curve will help to analyze the amount of true churners are discriminated in each subset. This will be extremely helpful in a marketing situation where a group of customers are to be contacted. Thus a company can count how many customers to contact if an example of 25 % of potential churners is to be contacted. Or if the marketing effort has a limit of 5 000 customer contacts, how many churners are then reached.

4. Case data

In this study a customer database from a Finnish bank is used and analyzed. The data consisted only of personal customer. The data at hand was collected from time period December 2001 till September 2005. Each time step consisted data from previous 3 month period, so for this study there was 12 relevant data points. The notation used for the rest of the paper is t = {0, 1, 2,…, 11}, which means that March 2003: t=1, and June 2003: t=2.

In total, 75 variables were collected from the customer database. These variables are related to the topics as follows: (1) account transactions IN, (2) account transactions OUT, (3) service indicators, (4) personal profile information, and (5) customer level combined information. Transactions have volumes in both in and out channels, out channels have also frequency variables for different channels.

The database had 251 000 customers overall. The data was divided in three groups randomly, two samples with 50 000 customers each and one sample with 151 000 customers. One of the samples with 50 000 customers was used in descriptive analysis of the customers. The sample with 151 000 customers was used in logistic regression model analysis. The third sample of customers (n=50 000) was initially discriminated because of potential future use in validation purpose. But the sample was left alone in this study because the validation was made using later time steps from the second sample.

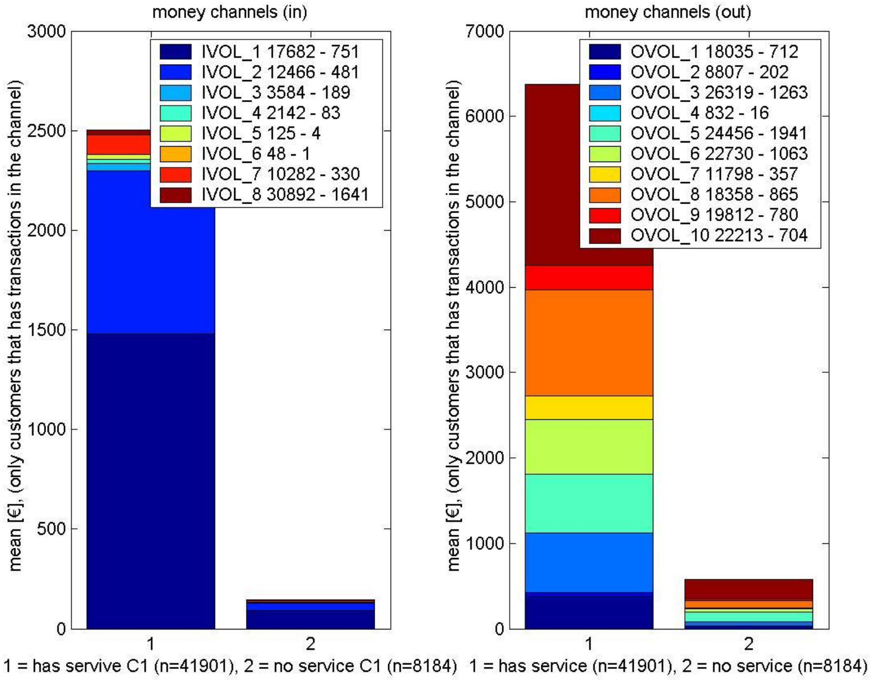


Figure 2 Customer with and without a current account and their average in/out money is presented in different channels. Legend shows the number of customers that has transactions in each channel. (A test sample was used n=50 000).

The customer database had 30 service indicator overall, for example whether customer has housing loan or not. One of these indicators C1 described the current account. The Figure 2 shows the average money volumes in different channels in two groups of customers on sample 1 as customers are divided into discriminated based on the current account indicator. The rest of these indicators (sum of all the rest) were used as a predictive variable in logistic regression.

5. Case focus

As mentioned previously customers' value to a company is at the heart of all customer management strategy. In retail banking sector the revenue is generated by both from the margins of lending and investment activities and revenues earned form service/transactions/credit card/etc. fees. And as Garland noted [11], retail banking is characterized by many customers (compared to wholesale banking with its few customers), many of whom make relatively small transactions. This setup in retail banking sector makes it hard to define customer churn based on customer profitability.

The definition of the churn in this case is based on the current account indicator (C1). This indicator C1 tells whether the customer has a current account in the time period at hand or not. This simple definition makes it easy to detect the exact moment of churn. Those customers without a C1 indicator before the time period were not included in the analysis. The volume of these customers, without the current account, in the dataset is small. And in banking sector a customer who does churn, may leave an active customer id behind because bank record formats are dictated by legislative requirements. This definition made possible to focus on active customers and detect the churners from these customers.

As mentioned earlier the case data also had indicators indicating for example an existence of an ATM-card or a housing loan. In this study these indicators weren’t used in the churn definition because the current account indicator based churn definition was simple. This definition did not produce for example partially defective customers nor numerous different types of churning customers. And in the case of loan contracts the churn before the contract ending date is, if not impossible, very rare.

5.1. Altered data sample

This definition of churn, presented above, has strong impact on predictive analysis. First each of the data samples shrank a little because the customers without a C1 indicator were filtered away. The second impact was on the amount of churners in each sample. The data sample 3 (115 000 customers with C1 indicator) only had 4 275 customers considered churners in the time period t(1) - t(11). So on average there were less than 500 customers in each time step to be considered churners. The number of churning customers is also presented in the Figure 3. The number of churners has very little variation during the time period t(1)-t(10).

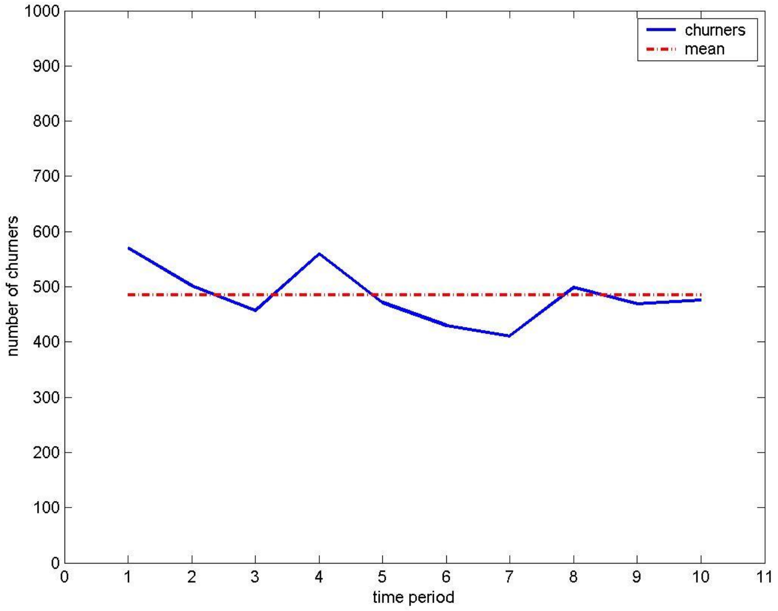


Figure 3 the number of churning customers is presented in each corresponding time period.

This problem has been identified in the literature under term class imbalance problem [15]. The problem of imbalanced data sets occurs when the one class is represented by a large number of examples while the other is represented by only a few. And the problem of imbalanced data sets is particularly crucial in application where the goal is to maximize recognition of the minority class [9]. The issue of class imbalance problem has been actively studied and it is handled in number of ways.

Two methods for dealing with class imbalances are: over-sampling (up-sampling) and down-sizing [15]. The over-sampling method has been widely used in signal detection theory [6] and it consists of re-sampling the small class at random until it contains as many examples as the other class. In this study a down-sizing method was used to avoid all predictions turn out as nonchurners. The down-sizing (under-sampling) method consists of the randomly removed samples from the majority class population until the minority class becomes some specific percentage of the majority class [7]. This produced two different datasets for each time step: one with a churner/nonchurner ratio 1/1 and the other with a ratio 2/3.

In the predictive analysis the datasets from time steps t(1) - t(8) was used. The datasets from time periods t(9) - t(11) was left for validation. This made possible to the predictions in the future. This predictive model could thus be used for example to predict churning customers in Q1 in 2006 based on the data collected from Q3 and Q4 in 2005. Although the predictions were made based on the down-sized dataset the validation for each model was made using the sample of 115 000 customers dataset.

5.2. Input variables

In the logistic regression model a set of 12 variables were used as input variables. The Table 2 presents all of these variables with a description. Some of the variables (OVOL8, ONO3, ONO6, and ONO9) had two representations in different time steps. The first representative is from two time steps before the time period at focus, the second is one time step before. The other four variables (age, bank age, E_ind, and osuus_tr1) were collected from one time step before the focused time period. In this case the volume and frequency variables could be used as such because the analysis was carried out in different time periods independently. The salary was used in the logarithmic scale because the impact it has to the customers' life is stronger among the customers earning small amounts than among the customers that earn huge salaries.[†]

[†] For example the difference between the salaries 1 300€ and 1 600€ in a month is same as between the salaries 5 000€ and 5 300€ in a month. But the effect that the former difference has on the customer's behavior is stronger than the effect on the latter case. The logarithmic scale helps to address this problem.

Table 2 the input variables that were used in the logistic regression model.

C age	Customer age
C bank age	Customer bank age
OVOL8_1	Volume of (phone) payments in t=i-1
OVOL8_2	Volume of (phone) payments in t=i
ONO3_1	Number of transactions (ATM) in t=i-1
ONO3_2	Number of transactions (ATM) in t=i
ONO6_1	Number of transactions (card payments) t=i-1
ONO6_2	Number of transactions (card payments) t=i
ONO9_1	Number of transactions (direct debit) t=i-1
ONO9_2	Number of transactions (direct debit) t=i
E_IND	Number of services, other than the current account
OSUUS_TR1	Salary on logarithmic scale in t=i

6. Results

In this study a collection of six different regression models were estimated and validated. As presented above each time step had two samples, one with a churner/nonshurner ratio 1/1 and the other with a ratio 2/3. Three time periods (t = 4, 6, 8) were selected for the logistic regression analysis. This produced six regression models which were validated by using data sample 3 (115 000 customers with the current account indicator). The data for validation was collected from time periods t(9) – t(11).

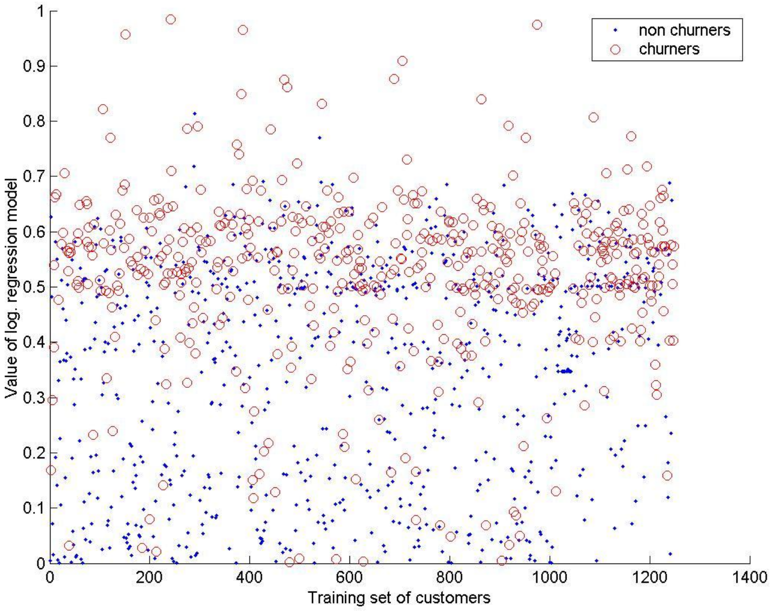


Figure 4 Example of the logistic regression model (82). Value of the regression model is presented for each customer and the marker color represents the real state of the customer (red=churn, blue=non-churn).

In the Figure 4 is presented an example of the logistic regression prediction when the training set 8_1 (t=8 and churn/nonchurn ratio is 1/1) is used. In the figure the value of logistic regression is presented for every customer in the training set. The marker color represents the reality of the customer in churn/nonchurn situation. It is clearly seen that the majority of churners are above the threshold value 0.5 but there are still churner below the threshold value. And in case of nonchurners, the majority of these customers receive a value below 0.5 but there is also a portion of customers that will be considered as churners based on the model.

Logistic regression estimation procedure was carried out using the SPSS software [21]. The SPSS software provides a comprehensive set of statistical tools. To the logistic regression it provides statistics of every input variable in the model along with the statistic of the whole model performance. Every model was estimated using all the variables in the Table 2 but those variables that weren't significant were left out of the model. The variables used in each model are presented in the Table 3. This variable selection process was done by iterative manner. The iteration follows from the usual statistical behavior that only the variables having significance value less than 0.05 are considered to be significant. A variable can thus be left out of the model if the significance value is over 0.05 and a new regression can be estimated using the remaining variables. This creates an iterative process until each variable has the significance value below the level 0.05.

Table 3 Predictive variables that were used in each of the logistic regression models. Notation "X_1" marks for dataset with a churner/nonchurner ratio 1/1 and "X_2" for dataset with a ratio 2/3. The coefficients of variable in each of the models are presented in the table.

Model	4_1	4_2	6_1	6_2	8_1	8_2
Constant	-	-	0,663	-	0,417	-
C age	0,023	0,012	0,008	0,015	0,015	0,013
C bank age	-0,018	-0,013	-0,017	-0,014	-0,013	-0,014
OVOL8_1	-	-	-	-	0,000	0,000
ONO3_1	0,037	0,054	-	-	0,053	0,062
ONO3_2	-0,059	-0,071	-	-	-0,069	-0,085
ONO6_1	0,011	0,013	-	0,016	0,020	0,021
ONO6_2	-0,014	-0,017	-	-0,017	-0,027	-0,026
ONO9_1	0,296	0,243	0,439	0,395	-	-
ONO9_2	-0,408	-0,335	-0,352	-0,409	-	-
E_IND	-1,178	-1,197	-1,323	-1,297	-0,393	-0,391
OSUUS_TR1	0,075	0,054	-	-	-	-

Although all the variables in each of the models presented in the Table 3 were significant there could still be correlation between the variables. For example in this study the variables ONO*X*_1 and ONO*X*_2 are correlated in some degree because they represent the same variable only from different time period. This problem that arises when two or more variables are correlated with each other is known as multicollinearity. Multicollinearity does not change the estimates of the coefficients, only their reliability so the interpretation of the coefficients will be quite difficult [17]. One of the indicators of multicollinearity is high standard error values with low significance statistics. A number of formal tests for multicollinearity have been proposed over the years, but none has found widespread acceptance [17].

It is seen in the Table 3 that all the variables have very little variance between the models. The only larger difference is in the variable's E_ind value in the models 8_1 and 8_2 compared to the value in the rest of the models. Overall behavior in the coefficients is that the coefficient half year before the churn (ONO3_1, ONO6_1, and ONO9_1) has a positive sign and the coefficient three months before the churn has a negative sign. This indicates that the churning customers are those that have declining trend in their transaction numbers. Also a greater customer age and a smaller customer bank age have both positive impacts on the churn probability based on the coefficient values.

6.1. Predictive performance

The logistic regression model will generate a value between bounds 0 and 1 as presented in the chapter 3.1 based on the estimated model. By using a threshold value on the discrimination of the customers there will be both error types of classification made. A churning customer could be classified to a nonchurner and a nonchurning customer could be classified as a potential churner. In the Table 4 the number of correct predictions is presented in each model. In the validation the sample 3 was used with the churners before the time period t=9 removed.

Table 4 Number and % share of the correct predictions (mean from the time periods t=9, 10, 11). In the validation sample there were a 111 861 cases. The results were produced by the models when the threshold value 0.5 was used.

Model #	Number of correct predictions	% correct predictions	% churners in the predicted set	% true churners identified as churners
model 4_1	69 670	62	0,8	75,6
model 4_2	81 361	72	0,9	60,5
model 6_1	66 346	59	0,8	79,5
model 6_2	72 654	65	0,8	73,4
model 8_1	15 384	14	0,5	97,5
model 8_2	81 701	73	0,9	61,3

The values in the Table 4 are calculated using a threshold value 0.5. If the threshold value would be for example set to 1 instead of 0.5 the % correct predictions would be 99.5 because all the predictions would be nonchurners and because there were only 481 churners (0.45%) on average in the validation set. The important result in the Table 4 is the column '% churners in the predicted set' which tells the percentage of the true churners in the predicted set when the threshold value 0.5 is used. The effect of the different threshold value is analyzed in the next chapter.

The important result found in the Table 4 is the proportional share of true churners to be identified as churners by the model. It is also seen in the table that the models with a good overall prediction performance won't perform so well in the predictions of churners. The previously discussed class-imbalance problem has an impact here.

6.2. Lift curve and predicted churners

The lift curve will help to analyze the amount of true churners that are discriminated in each subset of customers. In the Figure 5 the % identified churners are presented based on each logistic regression models. It can be seen from the Figure 5 that the model based on the set 8_1 isn't predicting the churning customers very efficiently. This model is actually as effective as random numbers in the prediction. The rest five models are predicting the churners almost identically. This is seen in the Figure 5 as well as in the Table 4. In the Table 4 the models 4_1, 6_1, and 6_2 have correct predictions close to 60% where models 4_2 and 8_2 have above 70% of correct predictions. This difference between the five models has vanished when amount of correct predictions is analyzed in the subsets as is presented in the Figure 5.

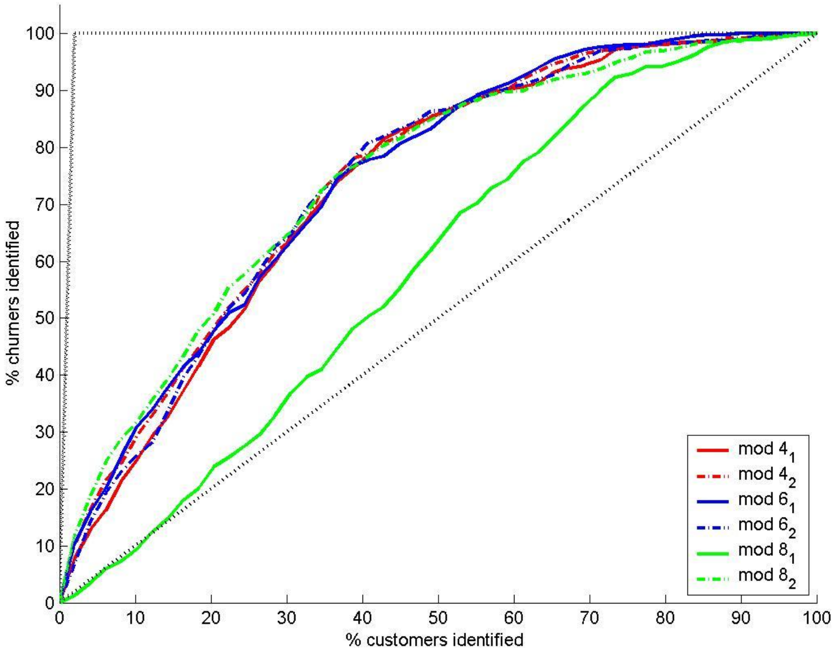


Figure 5 Lift curves from test-set (t=9) performance of six logistic regression model presented in the chapter 4.

In the Figure 6 the predicted churners that actually churn is presented as the function of number of customers in case of each model. A threshold value 0.5 was used. In this figure the same behavior is seen as is presented in the Table 4: the models 8_2, 6_1, and 4_2 have the best predictive efficiency while the model 8_1 closely follows the random probabilities. Both of the figures (Figure 5 and Figure

6) show that the logistic regression models have good predictive performance in the relatively small subset of customers, based on the logistic regression models.

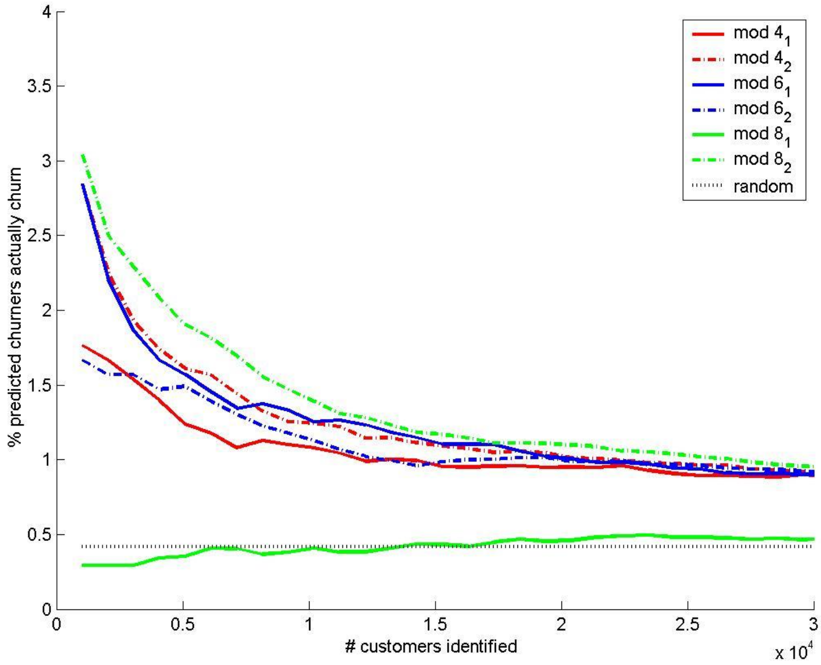


Figure 6 Predicted churning customers that actually churn presented with the number of customers.

7. Conclusions and future work

In this paper a customer churn analysis was presented in a retail personal banking sector. The analysis focused on churn prediction based on logistic regression. The different models predicted the actual churners relatively well. One of the models (8_1) did work almost as the random probabilities. The differences between the models input data (the significance level in case of each of the variables) indicates the dynamic nature of the churning customer profile. This makes it hard to formulate one standard model that could be used as the predictive model in the future. The findings of this study indicate that, in case of logistic regression model, the user should update the model to be able to produce predictions with high accuracy.

The customer profiles of the predicted churners weren't included in the study. It is interesting for a company's perspective whether the churning customers are worth retaining or not. And also in marketing perspective what can be done to retain them. Is a 3 month period enough to make positive impact so that the customer is retained? Or should the prediction be made for example six months ahead?

The customer churn analysis in this study might not be interesting if the customers are valued based on the customer lifetime value. The churn definition in this study was based on the current account. But if the churn definition was based on for example loyalty program account or active use of the internet service. Then the customers at focus could possibly have greater lifetime value and thus it would be more important to retain these customers.

References

[1] Ahola J., Rinta-Runsala E., Data mining case studies in customer profiling. Research report TTE1-2001-29, VTT Information Technology (2001).

[2] Au W., Chan C.C., Yao X.: A Novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on evolutionary computation, Vol. 7, No. 6, Dec 2003.

[3] Bauer H., Hammerschmidt M., Braechler M.:The customer lifetime value concept and its contribution to corporate valuation. Yearbook of Marketing and Consumer Research, vol. 1 (2003).

[4] Buckinx W., Van den Poel D.: Customer base analysis: partial detection of behaviorally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research 164 (2005) 252-268

[5] Buckinx W., Verstraeten G., Van den Poel D.: Predicting customer loyalty using the internal transactional database. Expert Systems with Applications xxx (2005)

[6] Candy J., Temes G.: Oversampling delta-sigma data converters: theory, design, and simulation. IEEE-PC0274-1. New York (1992).

[7] Chawla N., Boyer K., Hall L., Kegelmeyer P. SMOTE: Synhetic minority over-sampling technique. Journal of Artificial Research 16 p321-357 (2002).

[8] Cramer J.S. The Logit Model: An Introduction. Edward Arnold (1991). ISBN 0-304-54111-3.

[9] Cohen G., Hilario M., Sax H., Hugonnet S., Geissbuhler A.: Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine (2005). *Article in press*.

[10]Ferreira J., Vellasco M., Pachecco M., Barbosa C.: Data mining techniques on the evaluation of wireless churn. ESANN2004 proceedings – European Symposium on Artificial Neural Networks Bruges (2004). ISBN 2-930307-04-8, p 483-488.

[11]Garland R.: Investigating indicators of customer profitability in personal retail banking. Proceedings of the Third Annual Hawaii International Conference on Business, June 18-21 2003. 13 pages.

[12]Hasan M.: Customer Churn: The Stealth Enemy. Sigillum Corporation, Thought Leadership Monograph. (White paper)

[13] Hoekstra J., Huizingh E.: The Lifetime Value Concept in Customer-Based Marketing. Journal of Market Focused Management, 3, 257-274 (1999).

[14] Hwang H., Jung T., Suh E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26 (2004) 181-188.

[15] Japkowicz N., Stephen S.:The class imbalance problem: A systematic study. Intelligent Data Analysis Vol 6 p.429-449 (2002).

[16] Mozer M. C., Wolniewicz R., Grimes D.B., Johnson E., Kaushansky H. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunication Industry. IEEE Transactions on Neural Networks, Special issue on Data Mining and Knowledge Representation (2000).

[17] Pindyck R., Rubinfeld D. Econometric models and econometric forecasts. Irwin/McGraw-Hill (1998). ISBN 0-07-118831-2-

[18] Rosset S., Neumann E., Eick U.,Vatnik N., Idan Y.: Customer lifetime value modeling and its use for customer retention planning. Proceedings of the eighth

ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Canada p.332-340 (2002).

[19] Soldin O., Nandedkar A., Japal K., Stein M., Mosee S., Magrab P., Lai S., Lamm S.: Newborn thyroxine levels and childhood ADHD. Clinical Biochemistry 35 (2002) 131-136.

[20] MATLAB software. http://www.mathworks.com

[21] SPSS software. http://www.spss.com