Expert Systems with Applications 38 (2011) 15273–15285

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Credit card churn forecasting by logistic regression and decision tree

Guangli Nie [a,e], Wei Rowe [c,1], Lingling Zhang [a,b], Yingjie Tian [a], Yong Shi [a,d,*]

[a] Research Center on Fictitious Economy and Data Science, CAS, Beijing 100190, China
[b] Management School, Graduate University of Chinese Academy of Sciences, Beijing 100190, China
[c] Department of Finance, Banking & Law, University of Nebraska at Omaha, NE 68182-0048, USA
[d] College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA
[e] Postdoctoral Programme of Agricultural Bank of China, Beijing 100005, China

ARTICLE INFO

Keywords:
Credit card
Customer churn
Logistic regression
Decision tree
Data mining

ABSTRACT

In this paper, two data mining algorithms are applied to build a churn prediction model using credit card data collected from a real Chinese bank. The contribution of four variable categories: customer information, card information, risk information, and transaction activity information are examined. The paper analyzes a process of dealing with variables when data is obtained from a database instead of a survey. Instead of considering the all 135 variables into the model directly, it selects the certain variables from the perspective of not only correlation but also economic sense. In addition to the accuracy of analytic results, the paper designs a misclassification cost measurement by taking the two types error and the economic sense into account, which is more suitable to evaluate the credit card churn prediction model. The algorithms used in this study include logistic regression and decision tree which are proven mature and powerful classification algorithms. The test result shows that regression performs a little better than decision tree.

1. Introduction

Data mining refers to discover knowledge from a large amount of data. In this paper, we discuss the application of data mining including logistic regression and decision tree to predict the churn of credit card users. The banks can take corresponding actions to retain the customers according to the suggestion of the models.

With today's cost-cutting and intensive competitive pressure, more companies start to focus on Customer Relationship Management (CRM). The unknown future behaviors of the customers are quite important to CRM. Hence, it is of crucial importance to detect the customers' future decision then the company can take corresponding actions early (Glady, Baesens, & Croux, 2008). The customers who stop using the company's products are usually called churners. Finding the churners can help companies retain their customers. Gustafsson, Johnson, and Roos (2005) studied telecommunication services to examine the effects of customer satisfaction and behavior on customer retention. Results indicated a need for CRM managers to more accurately determine customer satisfaction in order to reduce customer churn.

One of the major reasons for this is that it costs less to retain existing customers than to acquire new customers (Roberts, 2000). It costs up to five times as much to make a sale to a new customer as it does to make an additional sale to an existing customer (Dixon, 1999; Floyd, 2000; Slater & Narver, 2000). And, it is becoming more evident that the only way to remain a leader in this industry is to not only be customer-driven but also focus on building long-term relationships.

Due to the development of information technology, many companies have accumulated a large amount of data. Analyzing this data can help the manager make the right marketing decision and pinpoint the right customer to market. Because of the large amount of accumulated data and serious churn related to credit card holders, it is a very good field in which to predict churn.

Several studies have proved the effectiveness of the power of customer retention. A bank is able to increase its profits by 85% due to a 5% improvement in the retention rate (Reichheld & Sasser, 1990). Van den Poel and Larivière (2004) calculated the financial impact of a one percent increase in customer retention rate. The power of the model can stay for a relatively long time. According to the research of Neslin, the churn models in the data typically still perform very well if used to predict churn for a database compiled 3 months after the calibration data (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

As the economy develops in China, a large amount of credit cards are issued. As of the third quarter of 2008, 132 million cards have been issued in China.[2] But many of the card holders are not active (or called churn holders). With increasing bank competition,

* Corresponding author at: Research Center on Fictitious Economy and Data Science, CAS, Beijing 100190, China. Tel./fax: +86 10 82680698.
E-mail addresses: sdungl@163.com (G. Nie), wrowe@mail.unomaha.edu (W. Rowe), zhangll@gucas.ac.cn (L. Zhang), tianyingjie1213@163.com (Y. Tian), yshi@gucas.ac.cn (Y. Shi).
[1] Tel.: +1 402 554 2812.

[2] http://www.chinavalue.net/Article/Archive/2009/1/20/155619.html.

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.06.028

customers are able to choose among multiple service providers and easily exercise their right of switching from one service provider to another. If banks can predict future behaviors before the customers close their account or stop using the card to pay, they can market to retain these customers.

The main purpose of this paper is not to provide a new data mining algorithm, but to focus on the application of the churn prediction, to provide a framework of understanding the knowledge of the card holders' hidden pattern using the data of Chinese banks. From the data preparation to useful knowledge, the goal is application of churn prediction. In this paper, we introduce a way to complete churn prediction considering profit.

The rest of the paper is organized as follows. The definition of churn and the summary of the algorithms and criteria are introduced in Section 2. The data used in the research is described in Section 3, and the modeling process based on logistic regression and decision tree are presented in Section 4 and 5, respectively. In Section 6, we conclude.

2. Definition of churn and algorithm evaluation criteria

In different application fields, the definitions of churn differ. Churn means the customer shift from one service provider to another (Lu, 2002). Customer churn is defined as the propensity of customers to cease doing business with a company in a given time period (Neslin et al., 2006).

Many of the previous definitions of churn use the behaviors related to product and a threshold fixed by a business rule. Once the transactions of the customer is lower than the threshold, the customer would be regarded as a churner (Glady et al., 2008). Van den Poel and Larivière (2004) regard the customer who closed his accounts as a churner. Buckinx and Van den Poel (2005) define a partial defector as someone with the frequency of purchases below the average and the ratio of the standard deviation of the interpurchase time to the mean interpurchase time above the average. Glady et al. (2008) defines a churner as a customer with less than 2500 Euros of assets (savings, securities or other kinds of products) at the bank. Glady, Baesens, and Croux (2009) claimed that the threshold is not always relevant and one should observe the evolution of the customer activity instead.

In the telecommunications industry, the broad definition of churn is the action that a customer's telecommunications service is canceled which includes both a service provider-initiated churn such as a customer's account being closed because of payment default and a customer-initiated churn. In one study, only customer-initiated churn is considered and it is defined by a series of cancel reason codes. Examples of reason codes are: unacceptable call quality, more favorable competitor's pricing plan, misinformation given by sales, customer expectation not met, billing problem, moving, change in business, etc. (Lu, 2002).

The relationship between churn rate and average lifetime is also studied. If no new customers are acquired then the average lifetime of an existing customer is equal to $1/c$, where c is the annual churn rate (Neslin et al., 2004). Gustafsson et al. (2005) use customer satisfaction (CS_t), affective commitment (AC_t), calculative commitment (CC_t), a situational trigger condition (ST_t), and a reactional trigger condition (RT_t), all in time t, to predict churn in time $t+1$ ($Churn_{t+1}$).

In our application, we define that the customer who did not do any transaction with the bank on his own initiative during the observation period (explained later) is a churner.

In order to predict the churn of the customer effectively, it is crucial to build effective models which fulfill some evaluation criteria. To accomplish this, many predictive modeling techniques are available. These data mining algorithms can help to select variables and build models (Hung, Yen, & Wang, 2006). Researchers use a variety of approaches to develop churn models including a combination of estimation techniques, variable selection procedures, time allocations to various steps in the model-building process, and a number of variables included in the model (Neslin et al., 2006). The techniques include GA, Regression, Neural Networks, Decision Tree, Markov Model, Cluster Analysis (Hadden, Tiwari, Roy, & Ruta, 2005), and optimization (Better, Glover, Kochenberger, & Wang, 2008; Mclain & Aldag, 2009).

According to the research of Hadden et al. (2005), regression and decision tree are the two most popular algorithms used in the research and perform well. Neslin et al. (2006) categorized the approaches as "Logit," "Trees," "Novice," "Discriminant," and "Explain." After comparison they found that the Logit and Tree approaches perform the best and result in that firm achieving a relatively good level of predictive ability. The Novice approach is associated with middle-of-the-road predictive performance, while the Discriminant and Explain approaches are associated with lower predictive performance (Neslin et al., 2004). Multiple-criteria quadratic programming approach has been used to credit card analysis and the models perform well (Li, Shi, & He, 2008; Peng, Kou, Shi, & Chen, 2008; Shi, Peng, Kou, & Chen, 2005).

In our research, we also use logistic regression and decision tree which are mature data mining algorithms to build models and predict the churn of credit card users. We will compare the performance of these two algorithms in credit card churn prediction.

After building a predictive model, marketers will use these classification models to predict future behaviors of customers. It is essential to evaluate the performance of the classifiers. Percentage of correctly classified (PCC) and receiver operating curve (ROC) are usually used as criteria. PCC, which computes the ratio of correctly classified cases to the total number of cases to be classified (also known as accuracy), is undoubtedly the most commonly used evaluation metric of a classifier. The ROC is a graphical plot of the sensitivity – i.e. the number of true positives versus the total number of events – and 1-specificity – i.e. the number of true negatives versus the total number of non-events. The ROC can also be represented by plotting the fraction of true positives versus the fraction of false positives (Coussement & Poel, 2008).

Another two criteria related to accuracy are also used; they are top-decile lift and Gini coefficient. Lift is a way to quantify the accuracy of a predictive model. e.g., among the 10% of customers predicted as most likely to churn, what percentage of them actually do relative to the percentage of all customers who churn. Gini coefficient is also related to lift which measures the area between an entry's cumulative lift curve and the random lift curve.

Beside PCC, the two types of errors, i.e. the Type I error which means a customer who did not churn is misclassified as a churner and Type II error which means a customer who churned is misclassified as an un-churner are also studied. The loss caused by Type II error is generally regarded as 5–20 times higher than the loss caused by Type I error (Lee, Chiu, Chou, & Lu, 2006)

Loss function is also used to compare the performance of the classifiers. The loss function is calculated on the basis of life value and life value has been discussed in several papers. Gupta et al. define the value of customer as "the value of a customer as the expected sum of discounted future earnings" (Gupta, Lehmann, & Stuart, 2004). Glady et al. designed a loss function based on previous definitions (Glady et al., 2009). In this research, we design a misclassification cost measurement to indicate the loss caused by the error of the model.

First, we examine a demo example to show the necessary to consider these two types error respectively. Consider the following misclassification table as shown in Table 1.

The overall error rate of the model is 10%. According to the prediction result of the model, there are 190 customers who may churn in the next period. However only 90 of the 190 are real churners; the remaining 100 are not real churners who are

Table 1
Confusing matrix of an example.

Observed		Predicted				
		Good/bad		Sum	Percentage correct (%)	Error (%)
		Good	Bad			
Good/bad	Good	900	100	1000	90	10
	Bad	10	90	100	90	10
	Sum	910	190	1100	90	10

Table 2
Confusing matrix of a changed example.

Observed		Predicted				
		Good/bad		Sum	Percentage Correct (%)	Error (%)
		Good	Bad			
Good/ bad	Good	880	120	1000	88	12
	Bad	8	92	100	92	8
	Sum	888	212	1100	90	11.64

misclassified by the model. Although the loyalty of these customers would be better after marketing, this payment is unnecessary when the customers have not trended to stop the service. The marketing budget cost for these customers would be wasted.

If the second error changes to 8%, the first error is 12%, and the overall error would be 11.64%. The misclassification matrix would be as follows.

As shown in Table 2, there would be 212 customers to be marketed if the first type error increases by 2% and the second type error also decreases by 2%. Only 2 of the new 22 predicted customers are real churners; the remaining 20 are not churners. It is difficult to say which example is better just based on the two types of errors.

Let's discuss an extreme example to demonstrate the shortcoming of the overall error measure (PCC). If all of the churners are misclassified, i.e. 100 churners are misclassified as good customers as Table 3 shows the accuracy is 81.82% which is still higher than 80%. Apparently, all of the100 churners given by the model are good customers. The model not only does not catch any churner, but also misleads the company to market to good customers.

From the above example, we can see that the accuracy is not good enough to be used to valuate the model without considering the economic cost. It is difficult to valuate the model only based on PCC and the two types of errors. Hence, we designed a new cost function to valuate the models built by the data mining algorithms taking the economic factors into consideration. In order to compare the classifiers in this paper, we designed a cost function taking the churn rate, Type I error, Type II error and economic factor into consideration.

The formula of the loss function is as follows:

$$\text{Cost} = Be_s P_{ave} + \frac{M}{Ge_f + B(1-e_s)} Ge_f \tag{1}$$

B is the number of churners (bad guys); G is the number of customers who did not churn (good guys); M denotes the marketing budget; P_{ave} means the average profit brought by one customer; e_f and e_s, respectively, denote Type I error and Type II error.

The first part of the cost function, i.e. $Be_s P_{ave}$, is the loss caused by the second type error. The misidentification of the model makes the service providers lose the opportunity to retain the customers who have the trend to leave.

The second part of the loss function, i.e. $\frac{M}{Ge_f+B(1-e_s)} Ge_f$, is the loss caused by the first type error. The left part $\frac{M}{Ge_f+B(1-e_s)}$ denotes the average marketing cost per predicted customers. Ge_f is the number of misidentified good customers.

Table 3
The confusing matrix of an extreme example.

Observed		Predicted				
		Good/bad		Sum	Percentage correct (%)	Error (%)
		Good	Bad			
Good/bad	Good	900	100	1000	90	10
	Bad	100	0	100	0	100
	Sum	1000	100	1100	81.82	18.18

As mentioned above, the cost caused by Type II error is generally regarded as 5–20 times higher than the cost caused by Type I error. In this example, we set the marketing budget 20 per customer and the average profit of the customer is 100. That is to say the marketing budget, i.e. M is 22,000($1100 * 20$), and the average profit, i.e. P_{ave}, brought by the customer is 100. The misclassification cost of the original model shown in Table 1 is as follows:

$$\text{Cost} = Be_s P_{ave} + \frac{M}{Ge_f + B(1-e_s)} Ge_f = 100 * 10\% * 100 + \frac{22,000}{1000 * 10\% + 100 * (1-10\%)} * 100 * 10\% = 12578.95$$

The cost of the changed model which is shown in Table 2 is 13252.82, and the cost of the extreme example is 32,000. The cost of the extreme example is nearly 3 times more than that of the original one. We can say that the cost measure better reflects the difference of the models than the accuracy measure.

3. Data

3.1. The data source

An anonymous China commercial bank provides the data for this study and is extracted from a data warehouse. All of the data is integrated at the level of the customer. No matter when the customers open the card accounts at any branch of the bank, the data warehouse can identify the customers by name and identification number of the customers. All of the customers in the warehouse are indexed by an unique customer number. The data warehouse records all the past changes to the card. Taking the balance of the card, for example, once the balance of the card changed, there will be an additional row for the new balance and the last balance is retained.

The data warehouse was built in 2005, and thus the data can be tracked back to January 2005. There are 60 million customers in the data warehouse. We randomly sampled from the system and the time interval of data is from January 2005 through the end of performance period (i.e. censoring on April 30, 2008) or the end of the customer relationship (i.e. churn). This is only raw data and we will refine the data when we calculate the variables.

The data is related to all the aspects of the credit card holder including personal information of the card holder, basic information of the card, detailed transaction information, abnormal usage information of the card, etc. There are nine tables related to each credit card.

Dealing with the time in the right way is quite important in this research. Research in the field of telecommunication is defined as the total number of months over a given period that the customer was not retained over a 9-month period. This period ensures that there was more variation using the 9-month cumulative churn measure (Gustafsson et al., 2005). Larivière and Van den Poel (2004) deals with the time by customer's lifecycle.

The final aim of the churn study is to predict what will happen to the customer in the future. Thus, we must split the time window

into two phases, i.e. the observation period and the performance period. In the observation period, we designed variable to observe the transaction behaviors of a customer, and then we check whether the customer became a churner or not during the performance period. The independent variables ($X1X2\cdots$) are calculated from the data recording the information obtained during the observation period and the dependent variable (Y) is calculated from the data recording the information during the performance period. In order to avoid the affect of the festival or season, we set a whole year (12 months) as the observation period (the interval between t_1 and t_2) and a whole year (12 months) as the performance period (the interval between t_2 and t_3). In our study, the observation period is from January 1, 2006 to December 31, 2006 and the performance period is from January 1, 2007 to December 31, 2007. After observing the behaviors of the customers for a whole year, our model will predict whether the customer becomes a churner or not in the performance period (see Fig. 1).

3.2. Variable design and refinement

In research of wireless telephone companies (Neslin et al., 2004), 171 database variables fell into three main categories: customer behavior such as minutes of use, revenue, handset equipment, and trends in usage; company interaction data such as customer calls to the customer service center; and customer household demographics, including age, income, geographic location, and home ownership. In another study of the wireless industry, the input data are categorized into four types: Demographics, Usage Level, Quality of Service (QOS), Features/Marketing Paging (Zhao, Li, Li, Liu, & Ren, 2005). Different types of data may have different influences on the model. The good derivable variables can help to build a good model. Future research could also tunnel in on the types of data that are important for churn prediction models.

The aim of this research is to find out the reason why customers churn, but we don't only rely on the theory of churn when we are designing variables. We do not give any assumption before data mining. We would not suppose that some factors would affect the dependent variable in advance. The task of this phase is to design as many variables as possible. We designed 135 derivative variables from four perspectives: customer personal information, card basic information, risk information and transaction information. The detailed information of the variables is listed in Appendix 1.

We designed the above four types of variables from frequent, time and extreme perspectives. The frequent variables reflect the frequency of the transactions of the customer. The customer with lower transaction frequency is more prone to churn. The variables related to frequency include the trade times in the last year, the overdraft in the last year, the trade times in the last 3 months, and so on. The variables of the total amount reflect the information related to the total amount, such as the trade amount in the last year, the overdraft amount in the last year, the trade amount in the last 3 months and so on. The time-related variables include the time period the account has been open, the longest period of time the cardholder did not have any transactions, etc. The extreme variables have two functions. The first function is to reflect the extreme information such as the maximum times that the holder trade 1 month in recent year. The second function is to integrate the data into customer level.

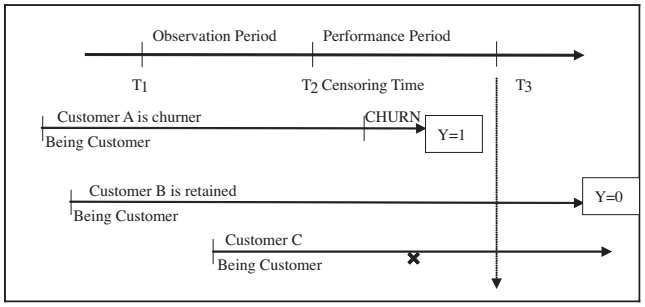

Fig. 1. The time window of analysis (Nie et al., 2009).

Interval is quite an important factor to consider when calculating the derivative variables. In our study, we have three principles to deal with data. For the data in terms of current status, we use the last row of the record. For the data related to the trade (such as the trade detailed data), we just intercept the data recorded during the observation time (i.e. from January 1, 2006 to December 31, 2007). For the data related to the integration performance (such as the number of cards the customer has), we need to extract the data from the time that the person became customer of the bank to the time censoring (t_2).

The customer information refers to the basic information of cardholders such as age, sex, foreign customers, staff of bank, marriage status, occupation, income, etc. Fifteen variables related to demographic information are designed first. The quality of three variables is poor. Most values of $X4$, $X5$ and $X8$ are the same. These variables are deleted because of poor contribution to the classification of the churn of the customers. This leaves 12 variables related to the demographic information.

Basic card information refers to the basic information related to the credit cards. We designed 41 variables ($X16$–$X56$) in this sector. The age of the card reflects how long the holder has used the credit card; the number of the new card reflects whether the holder is apt to apply for new cards.

We test the multicollinearity with the help of a variance inflation factor (VIF) which is listed in Appendix 1. According to the value of VIF, we can see that there is serious multicollinearity among the variables. The VIFs of $X18$, $X19$, $X20$, $X33$, $X42$, $X43$, $X68$, $X69$, $X70$, $X71$, $X72$, $X73$, and $X74$ are quite large which means these variables are seriously correlated with each other.

$X18$, $X19$ and $X20$, respectively, reflect the longest, shortest and average age of card. Because most the holders own only one card, these variables are highly correlated with each other. We retain $X18$ and delete $X19$ and $X20$.[3]

$X33$ which is *the interval length between first card and censor time* also has a high VIF. This variable is not only highly correlated with $X18$, $X19$ and $X20$, but also highly correlated with the variable $X49$ which denotes *the max age of the cards*. Therefore, we delete variable $X33$. The reason why $X42$ is highly correlated with $X43$ may be because the longest useful-life is nearly the same as the shortest useful-life. The maximum variable is more suitable for predicting the churn; hence we keep $X42$ and delete the variable $X43$.

The variables $X44$ through $X53$ reflect the new card information of the cardholder. Due to the rare application of the new cards, these variables are highly correlated with each other. We can conclude from this fact that Chinese holders are unwilling to apply for a new card once they obtain a bank credit card. We keep $X45$, $X48$, $X49$, and $X52$ to reflect the general information on the age of the card and delete $X46$, $X47$, $X50$, and $X51$.

In this sector, 8 of 41 variables related to basic card information are deleted. We keep 33 variables in this part to reflect the general basic information of the cards.

The variables from $X57$ to $X72$ are related to risk. The bank will valuate the risk level of the card every month. We calculate the times of different risk levels the bank has evaluated the card. Because the times of risk are similar in different time windows

[3] We decide which variables should be kept based not only on economic sense but also the correlation between the variables. Because of the space, the correlation of the variables is not posted here, but is available upon request from the authors. The correlation between $X18$, $X19$, and $X20$ is quite large, so we leave one of them.

(3 months, 6 months, 9 months), the VIFs of the risk-related variables are quite large. We just keep the times of a special risk level in the last 3 months, i.e. $X27$, $X31$, $X35$, and $X39$ and delete the rest.

The rest variables we designed are related to the active transactions information initiated by the cardholders. Sixty-three variables are designed and calculated in this part, but some of them are with low quality. $X83$, $X96$, $X97$, $X98$, $X101$, $X102$, $X111$, $X112$, $X128$, $X129$, $X130$, $X131$, $X132$, and $X133$ are the liner combination of other variables. This means the information on these variables has been completely reflected by other variables and we therefore deleted them. $X82$, $X90$, $X123$ and $X127$ with a large VIF are also deleted. Eighteen variables are deleted and 46 variables remain.

After considering the multicollinearity, there are 95 variables in the dataset to build a model including 12 personal variables, 33 card basic variables, 4 risk related variables and 46 transaction variables.

3.3. Data description

The original data in our study consists of nine tables and the largest table has 8 million records. There exist churners (without any activity during the performance period) and ordinary customers. All transactions are aggregated at the customer level. One customer may have several cards, but the objective is to predict the churn of a customer instead of a card, thus it is necessary to aggregate the data at the customer level. There are many methods to aggregate the data to customer level. One of the functions of the extreme variables is to aggregate the records to customer level. After the preparation, we get a data mart consisting of 5456 samples, 440 are churners (91.1%) and 5416 have not churned (8.1%).

According to the design of Section 4.2, we calculated the derivative variables. There are 135 variables reflecting the complete information of the customer. The independent variables are calculated from the data during the observation period of January 2006 through December 2006 (12 months). The independent variable (Y) is calculated from the data during the performance period of January 2007 through December 2007 (12 months). According the definition of Section 2, we define that the customer who did not do any transaction with the bank on his own initiative during the observation period is a churner

$$y_i = \begin{cases} 1 & \text{if customer } i \text{ has no transaction in performance period} \\ 0 & \text{if customer } i \text{ has at least one transaction in performance period} \end{cases}$$

We can see that the real proportion between ordinary customers and churners is 1:9, which is unbalanced. In order to get the suitable proportion of the samples to build a model, we try three proportions: 1:1, 1:2 and 1:4. The proportion 1:1 was repeated 10 times, 1:2 and 1:4 were repeated for 5 times. The test accuracy of the model of the proportions is shown in Fig. 2. The proportion 1:1 performs the best and 1:4 performs the worst. This means that 1:1 is the most suitable proportion to the existing data.

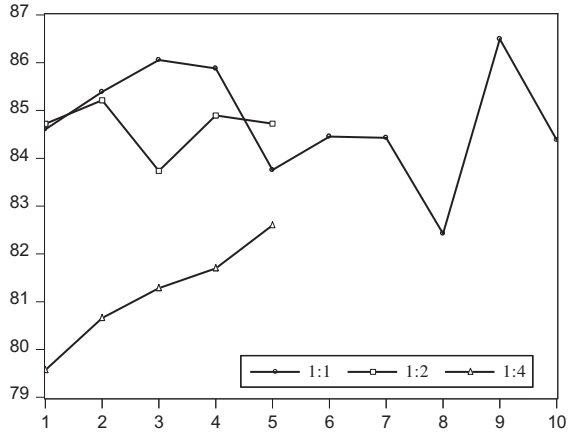

Fig. 2. The accuracy of the model with different proportion.

Now that the balanced samples perform the best, in the following research, we will use the balanced data to build a model. To be specific, we randomly sample fifty percent of the churners (215 samples) and five percent of the non-churners (244 samples) to be a training set. The rest of the 4997 customers would be a test set to validate the model and test the predict performance of the model.

4. Logistic regression

4.1. Modeling

Logistic regression, which is a widely used statistical modeling technique, could build a model with dichotomous outcome and has been proven as a powerful algorithm (Lee et al., 2006). Logistic regression has been well studied and used in a lot of applications such as agriculture (Bielza, Barreiro, Rodríguez-Galiano, & Martın, 2003), overweight and obesity (Martın, Nieto, Ruiz, & Jiménez, 2008), credit scoring (Lee et al., 2006), accident analysis and prevention (Al-Ghamdi, 2002).

The specific form of the logistic regression model is

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2}$$

The transformation of the $\pi(x)$ logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{3}$$

The traditional method of estimation that leads to the least squares function under the linear regression model is called maximum likelihood which provides the foundation for estimating the parameters of a logistic regression model. This paper is related to the knowledge discovery based on logistic regression instead of a new approach. The detail of fitting the logistic regression can be found in other research (Hosmer & Lemeshow, 1989).

In order to find the power of different variables, we build models with different variable combinations. As stated above, there are 135 variables designed in all, it would mislead the model if we put all of the designed variables into the algorithms. Also, the cost would be quite high when the model is used if we build the model with all of the variables because it will be time-consuming to calculate all of the variables. Hence, stepwise is used to select variables during the process of model building. We build six models with different variable combinations to examine the power of the different kind information.

The first model is built based on customer information. Model 2 is built on the variables of basic card information. Model 3 is built on the variables of customer information and basic card information. The variables used in this model are the combination of the variables of model 1 and model 2. Model 4 only takes the transaction information into consideration. Model 5 consists of the variables related to customer information, card basic information and risk related information. Model 6 consists of the variables related to customer information, card basic information, risk related information and transaction information.

As stated above, we randomly sample fifty percent of the churners (215 samples) and five percent of the non-churners (244 samples) to be a training set. The result of the model is displayed in Table 4.

Table 4
Logistic regression based prediction model.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dependent variable: dummy churn						
Intercept	−0.9462	−0.6153	1.4692	−1.6186	1.0500	−5.0277
	(6.76)[a]***	(0.2603)	(2.51)	(33.05)***	(0.74)	(17.96)***
*X*1					0.6779	1.2142
					(4.22)**	(8.30)***
*X*2	0.0236					
	(6.29)**					
*X*6	−0.7698		−1.7211		−1.7610	
	(4.24)**		(12.29)***		(9.13)***	
*X*15			0.0978		0.1488	0.1657
			(7.27)***		(11.48)***	(9.74)***
*X*17		2.1619	2.2477		2.2601	
		(22.63)***	(22.44)***		(15.61)***	
*X*27			0.7080		1.1461	
			(5.07)**		(8.76) ***	
*X*30		3.3547	3.4746		3.2906	2.3042
		(15.87)***	(15.50)***		(8.74)***	(4.89)**
*X*37		1.0696				
		(4.90)**				
*X*39		−1.2304	−1.5450		−1.6478	
		(16.79)***	(20.52)***		(15.59)***	
*X*40		−4.75E−7	−5.8E−7		−7.2E−7	−6.12E−7
		(21.18)***	(27.13)***		(29.77)***	(15.11)***
*X*48		−2.9458	−2.9864		−3.1612	−2.6024
		(41.58)***	(40.58) ***		(32.99)***	(18.34)***
*X*54		−0.00181	−0.00226		−0.00268	−0.00367
		(13.95)***	(16.33)***		(16.49)***	(16.95) ***
*X*57					1.2309	2.2523
					(8.33)***	(16.77) ***
*X*61					−2.8038	−2.7327
					(25.38)***	(16.92) ***
*X*65					−3.5874	−2.3251
					(11.04)***	(4.30)**
*X*73				−5.5734		−6.9575
				(19.18) ***		(12.43)***
*X*74				0.00906		0.0108
				(21.15)***		(15.15) ***
*X*79				0.00702		0.00444
				(19.83)***		(5.16)**
*X*81				−0.0654		−0.0580
				(15.90)***		(8.34)***
*X*85				0.3256		0.2976
				(26.67)***		(12.10) ***
*X*106						2.2168
						(5.06)**
AIC	629.157	394.207	376.623	338.506	304.122	208.159
SC	641.545	427.239	417.914	363.280	361.929	286.611
−2 Log L	623.157	378.207	356.623	326.506	276.122	170.159
Deviance	600.9767	378.2067	356.6231	326.5057	276.1218	170.1587
	(<.0001)[c]	(0.9945)	(0.9995)	(1.0000)	(1.0000)	(1.0000)
Pearson	442.8810	439.3798	424.6991	1146.9453	316.4246	222.3051
	(0.3610)	(0.6436)	(0.7892)	(<.0001)	(1.0000)	(1.0000)
HL[b]	5.4125	2.1581	7.5576	34.1927	2.2130	3.6644
	(0.7127)	(0.9758)	(0.4778)	(<0.001)	(0.97)	(0.8861)

[a] The value in the parameter is Wald Chi-Square except the row of HL.
[b] Hosmer and Lemeshow Goodness-of-Fit.
[c] Values in the parameter in the row of Deviance, Pearson, and HL are *P* value.
* Represent significance at the 10% level.
** Represent significance at the 5% level
*** Represent significance at the 1% level.

The measures related to the fit of the model are also listed at the bottom of the model result table. All measures show that model 6 fits the data best.

4.2. Prediction performance

This section will discuss the prediction performance of model 6, which we determined best fits the data. The main goal of the model is to predict the churn of the customers, so it is quite important for the model to have a strong predicting ability. The model is tested on the test set which is comprised of 4997 samples (4772 usual customers and 225 churners).

As reviewed above, the first type error, the second type error, and average error are usually used to test the model. In our research, we propose a new measure to show the performance of the model. According to formula (1), we take model 6 as an example to demonstrate the calculation process of the misclassification cost. Table 5 is the confusing matrix of model 6.

There are 4997 customers in the test set. 225 of them are churners and 4772 are ordinary customers. From Table 5, we can see the first type error, i.e. $e_f = 12.43\%$, second type error, i.e. $e_s = 18.22\%$,

Table 5
The prediction performance of logistic regression based model 6.

Observed		Predicted				
		No churn/churn		Sum	Percentage correct (%)	Error (%)
		No churn	Churn			
Good/bad	Good	4179	593	4772	87.57	12.43
	Bad	41	184	225	81.78	18.22
	Sum	4220	777	4997	84.675	15.325

the number of churners, i.e. $B = 225$, and the number of non-churners, i.e. $G = 4772$. Consuming the marketing budget, i.e. M is 99,940($4997 * 20$) and the average profit, i.e. P_{ave} brought by one customer is 100, we can calculate the loss as follows:

$$\begin{aligned}\text{Loss} &= Be_sP_{ave} + \frac{M}{Ge_f + B(1-e_s)}Ge_f \\ &= 225 * 18.22\% * 100 + \frac{99,940}{4772 * 12.43\% + 225 * (1 - 18.22\%)} \\ &\quad * 4772 * 12.43\% \\ &= 80377.25\end{aligned}$$

Table 6 displays the predictive performance of the six models.

To show the relationship in a more direct way, we display the performance of the six models in Fig. 3.

4.3. Model interpretation

From the above, we can see that good-of-fit and prediction performance of the six models are different. The models are with increasing accuracy from model 1 to model 6.

In model 1, there are only two variables; $X2$, the age of the customer, and $X6$, whether the customer is a staff member of the bank selected from the 12 variables. The coefficient of $X6$ is negative which means if the customer is a staff member of the bank, then the customer has less trend to churn. This result also makes sense in the real world.

Although the customer basic information is useful to predict the churn of the customer, we do not support building a model with too much customer personal information for the following reasons. First of all, the data quality of the personal information is not very good. Most of the personal data is incomplete. Second, even if the data is complete; it is still difficult to say whether the data offered by the customers is real or not because of privacy issues. Taking the variable income, for example, the customer is prone to offer a higher number when they apply for the credit card. Third, the rules with customer basic information are not so useful to make decision to retain the customers. Taking the result of this model for example, the age ($X2$) has a positive effect on the churn which means the older the customer is the higher probability of customer churn. This rule is useless for the bank to make decision to market. Therefore, we do not think it is beneficial to take much of the fixed personal information into account (Nie, Wang, et al., 2009).

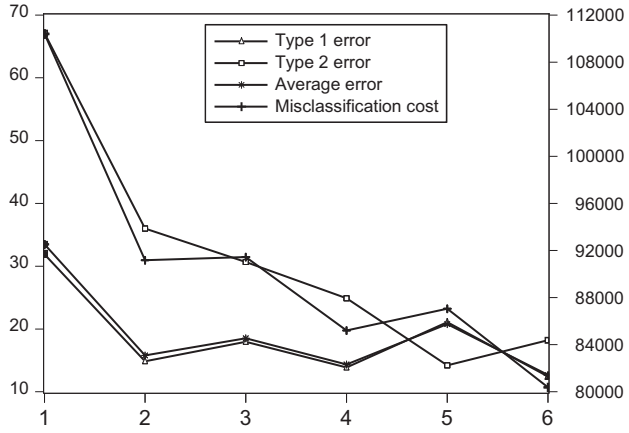

Fig. 3. The prediction performance of logistic regression-based model.

The predictive performance of model 1 is poor. The Type II is as high as 67.11% which is higher than 50%.

Model 2 is built on the variables of basic card information. Seven variables are shown to be powerful enough to discriminate between the churner from the ordinary customers. The AIC and SC are lower than in model 1 which means the card information is more powerful than customer information. The two types of errors, average error and misclassification cost of model 2 are significantly lower than model 1.

Model 3 is built on the variables of customer information and basic card information. This model is built from the variables of model 1 and model 2. The first type error of model 3 is higher than model 2. Sometimes, more information may mislead the model to classify the good customer correctly. However, model 3 works better from the perspective of average error and misclassification cost.

Model 4 only takes the transactions information into consideration. Generally speaking, when the error is lower than 25%, the model is powerful enough to be used in a commercial application. The average error of this model is 24.3% which is lower than 25%. Nonetheless, the misclassification cost of this model is higher than the following two models. That is to say the model just based on transaction behaviors is good enough to predict but with high cost.

Model 5 consists of the variables related customer personal information, basic card information and risk-related information. The difference between model 5 and model 3 is the risk related variables. The risk-related variables decrease the average error by 2%.

Model 6 consists of the variables related to customer information, basic card information and risk-related information and transaction information. Fifteen variables are used in the model. Although the second type error of this model is higher than model 5, other measures indicate that model 6 is the best model of the six models. The average error of model 6 is less than model 5 by 8%, and the misclassification cost of model 6 is less than model 5 by 7%. Model 6 is far better than model 5 from the perspective of accuracy and misclassification cost.

Table 6
The prediction performance of logistic regression based models.

Error	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Type I error (%)	31.89	14.86	17.96	13.87	21.10	12.43
Type II error (%)	67.11	36.00	30.67	24.89	14.22	18.22
Average error (%)	33.48	15.8	18.53	14.36	20.79	12.69
Misclassification cost	110405.2	91170.89	91451.59	85212.72	87064.02	80377.25

Table 7
Decision tree based models' prediction performance.

Error	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Type I error (%)	34.1	26.99	25.02	19.61	18.32	16.89	15.86
Type II error (%)	61.78	20.89	27.56	12	23.56	17.78	20.44
Average error (%)	35.34	26.72	25.14	19.27	18.55	16.93	16.07
Misclassification cost	108822.6	92505.4	94135.42	85190.16	88806.27	85283.19	85427.5

From the above analysis, we can see that average error is not good enough to reflect the prediction ability of the model; the cost caused by the misclassification should be taken into consideration. Logistic regression is powerful enough to predict the churn of credit card customers.

5. Decision tree

5.1. Decision tree algorithms

Decision tree is a well-known technique and has had many successful applications to real-world problems (Tsai & Chiou, 2009). Decision tree is a symbolic learning technique that organizes information extracted from a training dataset in a hierarchical structure composed of nodes and ramifications. Because the output of the decision tree can be organized in the form of a tree or rules, it is easy to understand the results for decision trees (Mitchell, 1997). What's more, a decision tree has the ability to build models using datasets including numerical and categorical data (Lorena & Carvalho, 2007).

There are two main types of decision trees (Osei-Bryson, 2004): classification trees and regression trees. The target variable takes its values from a discrete domain, and for each leaf node the decision tree associates a probability (and in some cases a value) for each class (i.e. value of the target variable) in a classification tree. The decision tree algorithm family includes classical algorithms, such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and CART (Breiman, Friedman, Olshen, & Stone, 1984).

Because they are easily understood, decision trees have been widely used in many fields such as supplier selection (Wu, 2009), churn of email user (Nie, Zhang, Li, & Shi, 2006). To the best of our knowledge, there is no application to the credit card churn prediction.

5.2. Model and discussion

This paper doest not intend to improve the existing decision tree algorithms. We use Weka (Witten & Frank, 2005) which is a well known open source data mining software to build models.[4] In order to compare the two algorithms, we also build six models with the same variables used in the logistic regression.

The decision tree algorithm used in our research is J48 and Table 6 is the result of our model. In the decision tree we try another model, i.e. model 7 using all of the 135 variables without considering the multicollinearity (see Table 7).

The results are also displayed in Fig. 4. Because of the unbalanced sample, the average error is nearly the same as the first type error. From the perspective of accuracy, model 7 performs the best, however, model 4 works best evaluated from misclassification cost. Because model 4 and model 6 perform the best from the perspective of misclassification cost, in order to compare the result of logistic regression and decision tree, we will deeply analyze model 6 of the tree.

[4] http://www.cs.waikato.ac.nz/ml/weka/.

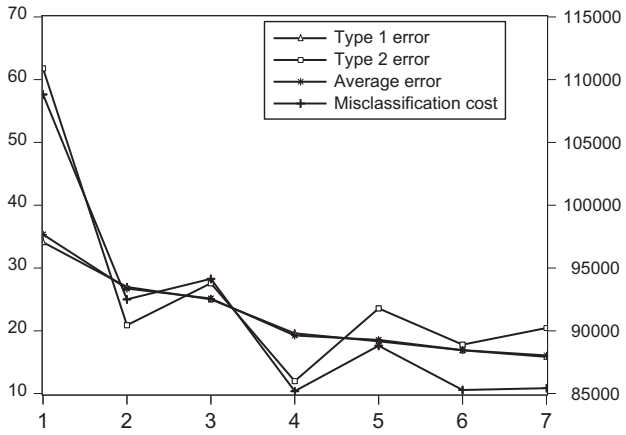

Fig. 4. The prediction performance of decision tree based model.

The following are the rules extracted from the decision tree. The rules of the tree are formed as: If A then churn (m/n). This means among the samples which satisfy condition A, m samples are churners and n samples are not churners. Here we list the main rules of the tree.

If $X117 \leqslant 397$ $X42 \leqslant 4$ $X55 \leqslant 956$ $X110 \leqslant 0.347245$ $X17 \leqslant 1$ $X40 \leqslant 730$ then churn (43.0)
If $X117 \leqslant 397$ $X42 \leqslant 4$ $X55 \leqslant 956$ $X110 \leqslant 0.347245$ $X17 \leqslant 1$ $X40 > 730$ $X42 \leqslant 3$ $X38 \leqslant 2$ then churn (35.0/3.0)
If $X117 \leqslant 397$ $X42 \leqslant 4$ $X55 \leqslant 956$ $X110 \leqslant 0.347245$ $X17 > 1$ then 1 (65.0/1.0)
If $X117 \leqslant 397$ $X42 > 4$ $X17 \leqslant 1$ $X41 \leqslant 4$ then 1 (2.0)
If $X117 \leqslant 397$ $X42 > 4$ $X17 > 1$ then 1 (7.0/1.0)
If $X117 > 397$ $X17 > 1$ $X61 \leqslant 0$ then 1 (8.0/1.0)

According to the above rule, we can see $X117$, $X42$, $X110$, and $X17$ are the most powerful variables. From Appendix 1, $X117$ reflects the amount of the last transaction, the smaller this value is the more probable the customer churns. $X42$ is the number of times the risk level of the card is above 3 in the last 12 months. $X55$ is the interval length between issuing time of the first card and censor time (i.e. December 31, 2006 in our research); the rules show that the older customers are more unlikely to churn. $X110$ is ratio of debit transaction via Internet to all channels. If most of the transactions are finished via internet, the customer is unwilling to stop using the card.

Most of these rules are consistent with fact. The model reflects the real business objectively. These results will not only help the bank to detect the churners but also develop marketing strategies. For example, the bank knows those customers frequently using the internet to finish transactions are loyal customers; the bank can focus on the promotion of the internet bank. The rules also tell us that the longer the customer is a customer, the more loyal they are.

The best model based on decision tree is model 6 and the error rate of this model is 16.93%. This rate is higher than that of the logistic regression based model 6. The cost of a decision tree model is

also higher than that of logistic regression. From the comparison, we can say that a logistic regression performs better than a decision tree.

6. Conclusions

In this research, we have proposed a process of churn prediction of credit card in China's banking industry. The purpose of this research is not to propose a new algorithm, but focuses on the execution and the understanding of the model. The suitable design of derivable variables and the systematical way to build a model could be helpful to execute the rules.

The two types of errors are not good enough to reflect the fit of the model. If the evaluation is only relied on accuracy, the result may mislead the choice of the model users. In our research, we have developed a new measure criterion called misclassification cost which takes the economic cost into the evaluation of the model. The empirical results of the case study in the paper have shown that the cost coefficient is an effective measure for the model's performance.

We have designed 135 variables to summarize behaviors and choices of the credit card users. After considering the multicollinearity, 95 variables are chosen to build the model. They are variables related to the categories of customer personal information, basic card information, risk information and transaction information. In the best model (model 6) of this paper based on logistic regression, there are two customer personal information variables, four card basic information variables, three risk information variables and six transaction information variables. The selected variables have shown that the demographic information makes little contribution to the churn prediction. The card information and the transaction information which relate to behavior work very well in the model.

Decision tree algorithm has been also used to build models. The test results of the model have shown that the logistic regression performs better than the decision tree. In the existing researches, multicollinearity has not been discussed in decision tree application. In this paper, one more model using all of the 135 variables is also tried. The results show that the variables without multicollinearity work better. However, decision tree-based models can provide rules in the rule form easy to understand; the rules can guide banks in making marketing strategies.

The decision tree rules have shown that the behaviors of the customer can better reflect future customer decisions. Even if it is impossible to access the personal information of the customers, it is acceptable to build a model based only on the transaction data of the users.

Acknowledgements

The authors are very grateful to the anonymous bank that supplied the data to perform the analysis. This research has been partially supported by The CAS Special Grant for Postgraduate Research, Innovation and Practice, President Fund of (GUCAS) (A) (Grant No. 085102HN00), National Natural Science Foundation of China (Grant Nos. 70840010, 70921061, 70871111), Beijing Natural Science Foundation (No. 9073020) and the BHP Billiton Cooperation of Australia.

Appendix 1

Var.	Description	MIN	MAX	VIF	Var.	Description	MIN	MAX	VIF
*X*1	Sex	1	2	1.12	*X*69	The times that the risk level of card is above 3 in recent 3 months	0	5	9.28
*X*2	Age	17	82	1.52	*X*70*	The times that the risk level of card is above 6 in recent months	0	9	22.19
*X*3	Dummy, takes 1 if VIP	0	1	3.23	*X*71*	The times that the risk level of card is above 3 in recent 9 months	0	10	33.56
*X*4*	Dummy, takes 1 if foreigner	0	1	1.04	*X*72*	The times that the risk level of card is above 3 in recent 12months	0	13	19.76
*X*5*	Dummy, takes 1 if minority	0	1	1.04	*X*73	Time(Being overdraft)/365	0	1	3.79
*X*6	Dummy, takes 1 if bank's staff	0	1	2.82	*X*74	The largest interval between transactions	1	365	2.52
*X*7	Dummy, takes 1 if in blacklist	0	1	1.22	*X*75	The shortest interval between transactions	1	637	1.66
*X*8*	Dummy, takes 1 if card holder	0	1	1.03	*X*76	The average balance of the account	−1E+06	92,236,113	2.54
*X*9	Dummy, takes 1 if international card holders	0	1	1.22	*X*77	The average overdraft amount	0	32,520,846	9.94
*X*10	Dummy, takes 1 if loan client	0	1	1.14	*X*78	average(overdraft/limit)	0	1989.7	1.03
*X*11	Dummy, takes 1 if Intermediary Business	0	1	1.42	*X*79	The length from last transaction to censoring time	0	820	2.30
*X*12	Marriage status	0	5	1.32	*X*80	The loan transaction times via	0	2782	7.89

(*continued on next page*)

Appendix 1 (*continued*)

Var.	Description	MIN	MAX	VIF	Var.	Description	MIN	MAX	VIF
						card			
*X*13	Education status	0	4	1.17	*X*81	The number of loan transactions in time via card	0	719	16.61
*X*14	Occupation status	0	8	1.13	*X*82	The amount of loan transactions in time via card	0	37,977,770	8.52
*X*15	Affiliation status	0	13	3.23	*X*83*	The number of loan transactions overdue via card	0	2765	NA
*X*16	The age of the account(days)[a]	13	2726	1.55	*X*84	The amount of loan transactions overdue via card	0	80,182,193	10.04
*X*17	Dummy, takes 1 if account status is 1	1	2	2.49	*X*85	The times of loan for interest	0	45	6.60
*X*18	The largest age of the cards(days)	4	2286	L	*X*86	The amount of loan for interest	0	451,883	3.35
*X*19*	The shortest age of the card	0	2286	L	*X*87	The times of loan for fee	0	15	2.06
*X*20*	The average age of the card	4	2286	L	*X*88	The amount of loan for fee	0	149,875	1.34
*X*21	The card age of the first card	0	759	1.54	*X*89	The times of loan for cash	0	392	16.76
*X*22	The largest usage ratio of the cards	0	5250	1.02	*X*90*	The amount of loan for cash	0	27,832,364	L
*X*23	average(the interval between open card till censor time)	0	1003	4.40	*X*91	The times of loan for private card	0	160	1.80
*X*24	The number of cards till censor time	1	5	16.4	*X*92	The amount of loan for private card	0	20,326,314	L
*X*25	The number of using cards Till censor time	0	4	3.39	*X*93	The times of loan for affiliation card	0	196	14.02
*X*26	The number of unopened cards till censor time	0	2	11.0	*X*94	The amount of loan for affiliation card	0	25,897,172	L
*X*27	The number of unopened cards till censor time getting through change	0	3	3.21	*X*95	The times of loan for consume	0	855	4.41
*X*28	The number of frozen Cards till censor time	0	1	1.37	*X*96*	The amount of loan for consume	0	80,182,193	NA
*X*29	The number of Reporting a Lost Card till censor time	0	2	2.26	*X*97*	The times of loan for deposit	0	2782	NA
*X*30	The number of card getting back from holders till censor time	0	2	1.44	*X*98*	The amount of loan for deposit	0	0	NA
*X*31	The number of cards stop using till censor time	0	4	10.3	*X*99	Ratio of Credit transaction as medium/all channel	−0.1801	1	4.82
*X*32	The length that the first card used	1	1491	19.9	*X*100	Ratio of Debit transaction as medium/all channel	0	1	2.22
*X*33*	The interval length between first card and censor time	4	2286	L	*X*101*	Ratio of Credit transaction via counter/all channel	0	1	NA
*X*34	The revoke of the card by the bank	1	5	2.50	*X*102*	Ratio of Debit transaction via counter/all channel	0	1	NA
*X*35	Dummy, takes 1 if there is golden card	1	2	1.28	*X*103	Ratio of Credit transaction via ATM/all channel	0	1.18008	9.92
*X*36	Dummy, takes 1 if there is card with picture	1	2	1.10	*X*104	Ratio of Debit transaction via ATM/all channel	0	1	3.25
*X*37	Dummy, takes 1 if there is staff card	1	2	2.56	*X*105	Ratio of Credit transaction via POS/all channel	0	1	10.83

Appendix 1 (*continued*)

Var.	Description	MIN	MAX	VIF	Var.	Description	MIN	MAX	VIF
X38	Dummy, takes 1 if there is card With year fee	1	4	1.26	X106	Ratio of Debit transaction via POS/all channel	0	1	1.81
X39	Dummy, takes 1 if there is card issued with other cooperation	1	3	1.80	X107	Ratio of Credit transaction via telephone/all channel	0	2.62	14.87
X40	The longest length between issue and first time use	0	3E+06	3.79	X108	Ratio of Debit transaction via telephone/all channel	0	1	1.31
X41	The shortest length between issue and first time use	0	3E+06	1.07	X109	Ratio of Credit transaction via Internet/all channel	0	1	1.95
X42	The longest useful-life length	0	6	25.0	X110	Ratio of Debit transaction via Internet/all channel	0	1	1.61
X43*	The shortest useful-life length	0	5	20.6	X111*	Ratio of Credit transaction via other channel/all channel	0	1	NA
X44	The times changing card	0	7	9.0	X112*	Ratio of Debit transaction via other channel/all channel	0	1	NA
X45	The number of new cards in recent 3Ms	0	2	7.4	X113	Debit amount during Labor's day/May	0	1	1.26
X46*	The number of new cards in recent 6Ms	0	2	15.2	X114	Credit amount during Labor's day/May	0	1	1.09
X47*	The number of new cards in recent 9Ms	0	3	19.6	X115	Debit amount during National Day/amount in October	0	1	1.21
X48	The number of new cards in recent 12Ms	0	3	25.9	X116	Credit amount during National Day/amount in October	0	1	1.08
X49	the max age of the cards	4	2225	43.7	X117	The amount of last transaction	0	1.33E+08	6.38
X50*	the min age of the cards	0	1491	23.1	X118	maximum(ratio of singal transaction to the amount of a month)	0.07	1.18	1.73
X51*	the average age of the cards	4	1858	61.2	X119	max of the average amount of the 12 Ms	1	3E+08	3.69
X52	The number of cards older than 1 Year	0	3	19.3	X120	min of the average amount of the 12 Ms	−10,000	1.33E+08	6.86
X53	The number of cards older than 2 Years	0	2	6.72	X121	The maximum times of all the months during censor time	1	570	19.74
X54	The interval length between validating card and first time using	0	2013	7.27	X122	The minimum times of all the months during censor time	1	21	3.18
X55	The interval length between card issued and first time used	0	2092	10.6	X123*	The maximum amount of all the months during censor time	1	4.57E+09	L
X56	The amount of the first transaction	0	9E+07	1.04	X124	The minimum amount of all the months during censor time	−11,495	4.6E+08	7.40
X57	The times that the risk level of card is 1 in recent 3 months	0	12	11.7	X125	The largest amount of the transaction	1	4.42E+08	12.28
X58*	The times that the risk level of card is 1 in recent 6 months	0	20	41.2	X126	The average transaction times of 12Ms	1	314.3	20.17
X59*	The times that the risk level of card is 1 in recent 9 months	0	28	83.6	X127*	The average amount of 12Ms	−1675	3.52E+09	L
X60*	The times that the risk level of card is 1 in recent 12 months	0	40	89.4	X128*	The average times of last 9Ms	1	314.3	NA

(*continued on next page*)

Appendix 1 (*continued*)

Var.	Description	MIN	MAX	VIF	Var.	Description	MIN	MAX	VIF
*X*61	The times that the risk level of card is 2 in recent 3 months	0	11	11.5	*X*129∗	The average amount of last 9Ms	−1675	3.52E+09	NA
*X*62∗	The times that the risk level of card is 2 in recent 6 months	0	20	37.4	*X*130∗	The average times of last 6Ms	1	314.3	NA
*X*63∗	The times that the risk level of card is 2 in recent 9 months	0	28	78.9	*X*131∗	The average amount of last 6Ms	−1675	3.52E+09	NA
*X*64∗	The times that the risk level of card is 2 in recent 12 months	0	38	97.5	*X*132∗	The average times of last 3Ms	1	314.3	.NA
*X*65	The times that the risk level of card is 3 in recent 3 months	0	2	2.8	*X*133∗	The average amount of last 3Ms	−1675	3.52E+09	NA
*X*66∗	The times that the risk level of card is 3 in recent 6 months	0	4	7.7	*X*134	The average times of last month	0	55	2.03
*X*67∗	The times that the risk level of card is 3 in recent 9 months	0	5	16.2	*X*135	The average amount of last month	0	3.85E+08	1.64

a. The times and amount in the table refers to the times and amount of transactions; the unit of time is day and the unit of the amount is.
b. In our paper, the censor time is from January 1, 2006 to December 31, 2006 and all of the data is generated during this period.
∗: This variable has been deleted because of the poor data quality or multicollinearity.
L: The value of this variable is larger than 100.
NA: This variable is the linear combination of other variables.

References

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention, 34*, 729–741.

Better, M., Glover, F., Kochenberger, G., & Wang, H. (2008). Simulation optimization: Applications in risk management. *International Journal of Information Technology and Decision Making, 7*(4), 571–587.

Bielza, C., Barreiro, P., Rodríguez-Galiano, M. I., & Martín, J. (2003). Logistic regression for simulating damage occurrence on a fruit grading line. *Computers and Electronics in Agriculture, 39*, 95–113.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. California, USA: Wadsworth.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research, 164*(1), 252–268.

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

Dixon, M. (1999). 39 Experts predict the future. *America's Community Banker, 8*(7), 20–31.

Floyd, T. (2000). Creating a new customer experience. *Bank Systems and Technology, 37*(1), R8–R13.

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

Glady, N., Baesens, B., & Croux, C. (2008). Modeling churn using customer lifetime value. *European Journal of Operational Research*. doi:10.1016/j.ejor.2008.06.027.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research, 41*(1), 7–18.

Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing, 69*(4), 210–218.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2005). Computer assisted customer churn management: State-of-the-art and future trends. *Computers and Operations Research, 34*, 2902–2917.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515–524.

Lu, J. (2002). *Predicting customer churn in the telecommunications industry –– An application of survival analysis modeling using SAS*. Sprint Communications Company.

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.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis, 50*, 1113–1130.

Li, A., Shi, Y., & He, J. (2008). MCLP-based methods for improving "bad" catching rate in credit cardholder behavior analysis. *Applied Soft Computing, 8*(3), 1259–1265.

Lorena, A. C., & Carvalho, A. C. P. L. F. d. (2007). Protein cellular localization prediction with support vector machines and decision trees. *Computers in Biology and Medicine, 37*, 115–125.

Martín, A. R. g., Nieto, J. M. M. n., Ruiz, J. P. N., & Jiménez, L. s. E. (2008). Overweight and obesity: The role of education, employment and income in Spanish adults. *Appetite, 51*, 266–272.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Mclain, D., & Aldag, R. (2009). Complexity and familiarity with computer assistance when making ill-structured business decisions. *International Journal of Information Technology and Decision Making, 8*(3), 407–426.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: Improving predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211.

Nie, G., Wang, G., Zhang, P., Tian, Y., & Shi, Y. (2009). Finding the hidden pattern of credit card holder's churn: A case of China. *Lecture Notes in Computer Science, 5545*, 561–569.

Nie, G., Zhang, L., Li, X., & Shi, Y. (2006). The analysis on the customers churn of charge email based on data mining – Take one internet company for example. In *Sixth IEEE international conference on data mining, Hong Kong, China* (pp. 843–847).

Osei-Bryson, K. M. (2004). Evaluation of decision trees: a multi-criteria approach. *Computers and Operations Research, 31*, 1933–1945.

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A multi-criteria convex quadratic programming model for credit data analysis. *Decision Support Systems, 44*, 1016–1030.

Quinlan, J. R. (1986). Induction on decision trees. *Machine Learning, 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Kluwer Academic Publishers.

Reichheld, F. F., & Sasser, W. E. Jr., (1990). Zero defections: Quality comes to service. *Harvard Business Review, 68*(5), 105–111.

Roberts, J. H. (2000). Developing new rules for new markets. *Journal of the Academy of Marketing Science, 28*(1), 31–44.

Shi, Y., Peng, Y., Kou, G., & Chen, Z. (2005). Classifying credit card accounts for business intelligence and decision making: A multiple-criteria quadratic programming approach. *International Journal of Information Technology and Decision Making, 4*, 581–600.

Slater, S. F., & Narver, J. C. (2000). Intelligence generation and superior customer value. *Journal of the Academy of Marketing Science, 28*(1), 120–127.

Tsai, C. F., & Chiou, Y. J. (2009). Earnings management prediction: A pilot study of combining neural networks and decision trees. *Expert Systems with Applications, 36*, 7183–7191.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufman.

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. *Expert Systems with Applications, 36*, 9105–9112.

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). *Customer churn prediction using improved one-class support vector machine*. Berlin: Springer.