Statistica Sinica Preprint No: SS-2017-0043.R1

Title	RFMS Method for Credit Scoring Based on Bank Card Transaction Data
Manuscript ID	SS-2017-0043.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0043
Complete List of Authors	Jing Zhou Danyang Huang and Hansheng Wang
Corresponding Author	Jing Zhou
E-mail	zhoujing_89@126.com
Notice: Accepted version subject to English editing.	

RFMS Method for Credit Scoring Based on Bank Card Transaction Data

Danyang Huang[1], Jing Zhou[1] and Hansheng Wang[2]

[1] *School of Statistics, Renmin University of China, Beijing, China;*
[2] *Guanghua School of Management, Peking University, Beijing, China*

Abstract

Microcredit refers to small loans to borrowers who typically lack of collateral, steady employment or a verifiable credit history. It is designed not only for start-ups but also for individuals. Microcredit industry is experiencing fast growing in China. In contrast with traditional loans, microcredit is typically lack of collateral, which makes credit scoring important. Due to the fast development of on-line microcredit platforms, there are various sources of data that could be used for credit evaluation. Among them, the bank card transaction records play an important role. How to conduct credit scoring based on this type of data becomes a problem of importance. The key issue needs to be solved is feature construction. That is how to construct meaningful and useful features based on the bank card transaction data. To this end, we propose here a so-called RFMS method. Here "R" stands for recency, "F" stands for frequency, and "M" stands for monetary value. As one can see, our method can be viewed as a natural extension of the classical RFM model in marketing research. However, we make a further extension by taking "S" (Standard Deviation) into consideration. The performance of the method is empirically tested by a real data example from a Chinese microcredit company.

KEY WORDS: Credit Scoring; Frequency; Logistic Regression; Microcredit; Monetary Value; Recency; Standard Deviation.

[†]The research of Danyang Huang is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (Grant No. 16XNLF01), in part by Ministry of Education Humanities Social Science Key Research Institute in University Foundation (Grant No. 14JJD910002) and the Center for Applied Statistics, School of Statistics, Renmin University of China. Zhou's research is supported in part by China Postdoctoral Science Foundation (Grant No.2016M600155), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No.17XNLF08). The research of Hansheng Wang is supported in part by National Natural Science Foundation of China (Grant No. 71532001, 11525101), China's National Key Research Special Program (Grant No. 2016YFC0207700) and Center for Statistical Science at Peking University.

1. INTRODUCTION

Microcredit refers to small loans to borrowers who typically lack of collateral, steady employment or a verifiable credit history. It is designed not only for start-ups but also for individuals. In China, microcredit industry is experiencing rapidly expanding in recent years for two reasons. First, besides the lack of collateral, start-ups often have lower probability to provide complete financial statements, but higher risk to default. Thus they could hardly get financial support from traditional financial institutions. Second, microcredit for individuals are mainly applied for consumption purposes (e.g., traveling), which are typically not supported by traditional loans and thus the approval rate is low. Consequently, more and more microcredit companies are springing up intending to provide more financial products to more people with lower cost. According to the report of the People's Bank of China, by the middle of 2016, there are 8,810 microcredit companies in China. The resulting loan balance is 9,364 billion RMB (Approximately 1,348 billion in USD !)

To implement microcredit properly, applicants' credit should be evaluated carefully. Practically, this means to assign for each applicant a credit score. This is a process typically referred to as credit scoring and is the key for successful microcredit. Typically, larger score implies better credit, and thus higher likelihood to be approved for a small loan and at a lower interest rate. For traditional loan, applicants with good credit scores are typically those who have valuable collaterals. However, as mentioned before, applicants for microcredit typically cannot meet this standard. Then, how to conduct credit scoring without collateral becomes a problem of importance. In this case, credit scoring is highly relied on an applicant's background information (without collateral). This background information typically includes, but are not limited to, educational background, working experiences, bank card transaction records. Due to the advance of information technology, particularly mobile Internet, various valuable and accurate data sources are becoming increasingly available. Then, how to make a good use of these data for accurate credit scoring becomes a problem of great interest. To this end, statistical models for credit scoring are inevitably needed.

To develop a statistical model for credit scoring, each applicant is typically treated as a sample. Applicant refers to individual who apply the loan product and has been actually approved. However, not all of the applicants successfully return the principal and interest. Thus we could observe their default behavior and they become the sample we study. Next, define for each sample i ($1 \leq i \leq n$) a binary response $Y_i \in \{0, 1\}$ according to whether the applicant eventually defaults (e.g., $Y_i = 0$) or not (e.g., $Y_i = 1$), where n is the sample size. Moreover, a set of covariates (i.e., applicants' characteristics) is needed, which is collected by a covariate vector $X_i \in \mathbb{R}^p$ and p is the dimension of the vector. Therefore, a credit scoring model could be developed by investigating the relationships between Y_i and X_i. It is remarkable that researchers show great interest in statistical models of credit scoring for more than 70 years. A variety of models have been proposed, which include but are not limited to discriminant analysis (Durand, 1941; Eisenbeis, 1977, 1978), ordinary linear regression (Orgler, 1970), logistic regression (Wiginton, 1980; Srinivasan and Kim, 1987; Leonard, 1993; Copas, 1999), k-nearest neighbors (Hand, 1986; Henley and Hand, 1996), and graphical model (Stanghellini et al., 1999). Even in recent years, researchers continuously develop new methodologies in credit scoring analysis. Antonakis and Sfakianakis (2009) adopt the spirit of naive Bayes for screening credit applicants. Lieli and White (2009) examine the econometric implications of credit analysis by solving a profit/utility-maximizing problem. Capotorti and Barbanera (2012) analyze the credit score based on the methodologies of rough sets, partial conditional probability assessments and fuzzy sets. For small and medium enterprise loan defaults, Calabrese and Osmetti (2013) use a generalized extreme value regression model.

The methods mentioned in previous literatures are very useful. However, they mainly focus on general methodology. It is assumed that applicants' characteristics (i.e., X-variables) have been well summarized, and thus could be used for regression or statistical learning directly. However, the fast development of information technology has made us access to various useful datasets. Very often, the structure of the collected datasets cannot be fitted by the classical models completely. This is particularly true for bank card transaction data, which widely exist and are considered as one of the most important data sources for credit analysis (Till and Hand, 2003; Chehrazi and Weber, 2015). For a given applicant i, its binary response Y_i (whether defaults or not) is well defined. However, its $X-$variables are not naturally defined and have to be extracted from bank card transaction records. In the meanwhile, bank card transaction records are rather complicated (or even unstructured) data. Different applicants have different number of transaction records, which are made at different time points with different cash amounts. Then, how to construct meaningful X-variables from this complex but useful information source becomes the major concern of this article.

This motivates us to develop a method for X-variable construction for credit scoring, based on applicants' bank card transaction records (from both debit cards and credit cards). In contrast to the previous researches, we focus on the variable construction process instead of general modeling methodology. Our approach is inspired by the popularly used RFM model for analyzing customer value in database marketing (Shepherd, 1990; Fader et al., 2005; Blattberg et al., 2008). In the RFM method, "R" refers to recency, which means how recently a customer purchases. "F" stands for frequency, which means how often a customer purchases. "M" is the monetary value, which indicates how much a customer spends. Except for redefinition of the three attributes in the classical RFM model in the scenario of microcredit, the standard deviation of an applicant's transaction amount is also considered. We refer to it as "S", which is used to evaluate the volatility of an applicant's activities. Then the proposed method is named as RFMS approach. The performance of the method is empirically tested using a real data example from a Chinese microcredit company. The results show that our approach can significantly improve the accuracy of existing credit scoring in the real data example.

The rest of the article is organized as follows. In Section 2, we give a detailed description about the data we use, including the data collection process and data structure. Next, we present the RFMS method and feature construction in Section 3. Model results and prediction accuracy are reported in Section 4. We then conclude this article with some business implications and some future topics in Section 5.

2. DATA DESCRIPTION

To implement the proposed method, we collaborate with a major microcredit company in mainland China. One of the company's business is to finance consumption loans to individuals. To assess their applicants' creditworthiness, a basic credit score has been developed for general purpose for all the products in the company. Although the basic score is practically useful in many cases, the efficiency could be low for a particular product. In that case, it is preferable to have a new score, which is specifically designed for the target loan product.

The company operates a major payment platform in mainland China, and thus could track applicants' transactions through their bank card transaction records. This means every transaction of an applicant through this platform would be recorded. With the approval of applicants, their transaction history can be traced using the registered phone number. Compared to traditional data acquisition method (e.g., customer self-reported data), this data source has some unique advantages by avoiding three problems as follows: (1) Over-simplified portrayal of individual who lacks of a multidimensional evaluation; (2) High risk of financial fraud such as exaggeration of employee payments, which is impossible for companies to detect; (3) Lack of cross validation that impairs the reliability of credit evaluation.

If an applicant is rejected by the company and thus no loan is approved, we should then have no chance to observe its default behavior (i.e., the response Y). As a result, those applicants cannot be used for modeling. Accordingly, we should focus on those applicants, whose loans have been actually approved. However, by the time of maturity, some of them successfully return the principal and interest, while the others fail. For those applicants, we have their complete data records and thus can be used for modeling. This might leads to the risk of over-evaluation and is a challenging problem faced by the entire industry. In this work, we follow the industry common practice and focus on approved applicants only. This leads to a sample of 26,513 applicants. There are almost 4,500,000 transaction records and other related information over a year. The dataset contains 7,980 default applicants and 18,533 non-default ones. It is also worth noting that this is a sample for model analysis and thus it does not reflect the real default rate among the applicants who adopt the product.

For each applicant, we can collect four types of information, they are applicant information, merchant classification, bank card information and transaction records. See Figure 1 for illustration. These four sources of information can be linked by different keywords as follows: (1) Applicant information and transaction records are connected by phone number. Each phone number corresponds to several transaction records. (2)

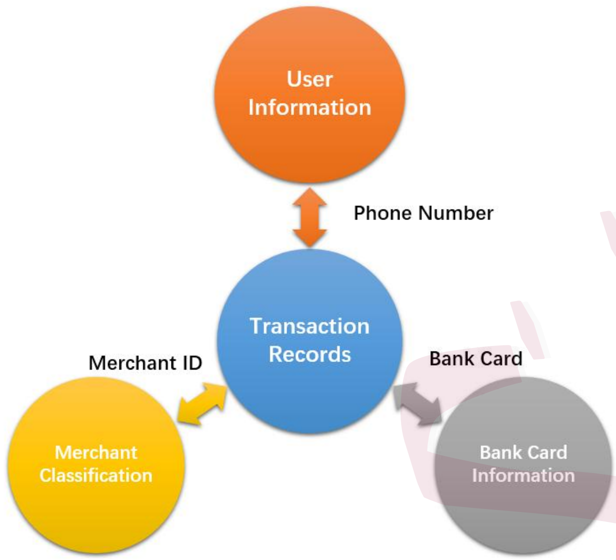

Figure 1: Data structure of a registered applicant for a specific product.

Merchant classification and transaction records are associated through merchant ID which is related to several transactions; (3) Bank card information can be linked to the bank card used in the transaction records. Because the first few digits of the bank card ID in the transaction record imply the bank card information. Variables extracted from these data are given as follows.

(1) Applicant Information. Phone number, registration channel (mobile application or website), applicant ID (encrypted data), time of last log-in, register time, etc.

(2) Merchant Classification. Created time, merchant number, merchant name, category code and category name, etc.

(3) Bank Card Information. Bank card number identifier (the first few digits of account number are used to identify the bank name and card type), length of identifier, account type(debit or credit), bank code, bank name, etc.

(4) TRANSACTION RECORDS. Transaction serial number, cell phone number, transaction time, merchant number, bill number, bill amount, amount of payment, etc.

To proceed with these complicated datasets, we first merge them by different keywords. Then each observation indicates one transaction, which includes an applicant's personal information, bank card information and merchant classification information. This leads to a total of about 4,500,000 observations. Such a big amount of data can hardly be analyzed straightforwardly. One of the possible solutions is to transform this transaction level data to applicant level data. It is equal to say the analysis unit is no longer a transaction record but an applicant. That means we have to integrate multiple transaction records from a single applicant into one observation by deriving applicant level variables. Then the sample size will be substantially reduced. It should be noted that deriving applicant related variables is a crucial and interesting part in this project. Because we have little prior knowledge about what kind of applicants would default. In the next subsection, we will explain how to generate applicant related variables. Some of the variables can be obtained directly from the dataset, while others are generated following a guidance.

3. VARIABLE CONSTRUCTION

We have two sets of variables. The first is applicant basic information variables which can be derived directly from the provided dataset. The second is applicant classification information variables which are generated under the guidance of the proposed RFMS method.

1. Basic Information Variables.

(1) BASIC SCORE. The basic score is a credit score developed by the company. It is carefully designed for a general purpose and is expected to be useful for all the loan

products. However, because it is a general credit score, it is unlikely to produce best prediction accuracy for one particular loan product.

(2) Length of Registration. The number of days past since the first time the applicant registered the company's product.

(3) Number of Transactions. Total number of transactions in a period of time. This reflects the frequency of the applicant's usage of bank cards.

(4) Mean of Transaction. Average amount of transactions. This is related to the concept of "monetary" value (i.e., the "M") in the RFMS model.

(5) Maximum Transaction. Maximum amount of transactions. This measures the extreme behavior of an applicant. An extremely high value in maximum transaction might be related to credit card's cash scandal. Thus, higher value might suggest higher possibility of default.

(6) Credit to Debit Ratio. The proportion of transactions using credit cards to those of using credit and debit cards. This ratio could reflect an applicant's consumption habit. For instance, some people like to pay with debit cards while others prefer to use credit cards first and then pay off. These are two different consumption patterns.

(7) Number of Bank Cards. Number of bank cards owned by an applicant. For example, applicants with too many credit cards might be more likely to default, because the repayment pressure is relatively high.

2. Classification Information Variables.

As we mentioned before, a very important task in feature construction is to generate applicant level data. An applicant has many transaction records of different behaviors.

Based on the merchant types, we can classify the transaction into different categories of behaviors, such as grocery shopping, buying game cards or paying electricity bills. For each category of behavior, we need a standard criterion to generate variables that could measure applicants' characteristics. To this end, we adopt the spirit of the RFM method (Shepherd, 1990; Fader et al., 2005; Blattberg et al., 2008), which is widely used in marketing field.

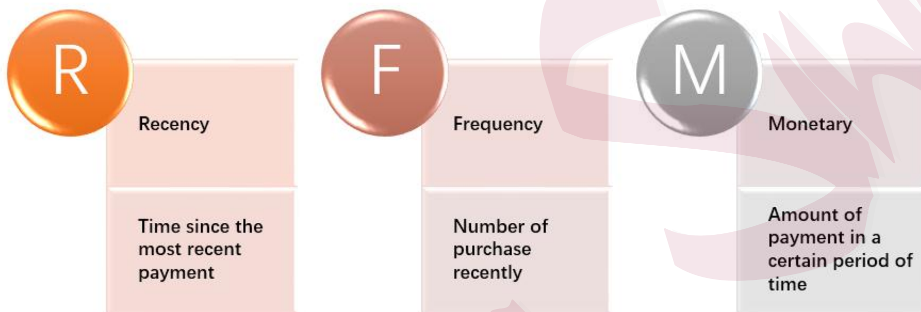

Figure 2: RFM model.

In marketing, the RFM method is an important tool to measure customer value and profitability. The model uses three characteristics to describe customer value. They are, recency, frequency and monetary value, respectively.

(1) R (Recency). "R" refers to recency, which calculates the time since the last purchase. Customers with a small "R" are often regarded as relatively active applicants. As we can see, in practice, customer who has an R value of less than 6 months interacts with salespeople more frequently than those whose "R" value is around 31 to 36 months.

(2) F (Frequency). "F" refers to frequency, which calculates the total number of purchases in a certain period of time. Typically, the higher the frequency of purchase, the more loyal the customers are.

(3) M (Monetary). "M" refers to the monetary value that a customer spends in a certain period of time.

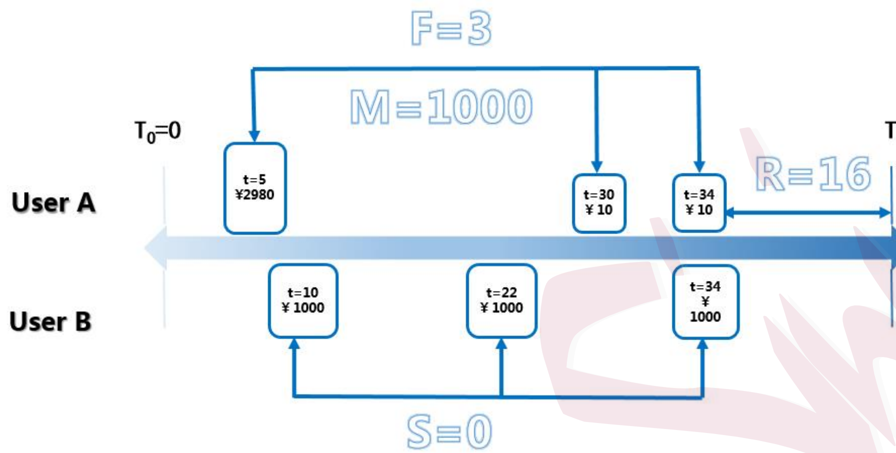

Figure 3: An Example to Explain S (Standard Deviation).

Similarly, we redefine our RFM attributes in the scenario of microcredit, see Table 1 for more details. However, even for applicants with exact the same "R", "F" and "M", they could have very different behaviors. See for example in Figure 3. For applicant A's and applicant B's consumption behavior, which are observed from $T_0 = 0$ to T_1=50, their total expense are the same. If we adopt the classical RFM method, all of the three attributes are the same. But it is obvious that applicant B behaves more regularly than applicant A. This can be reflected by their standard deviations, which is referred to as "S". This leads to the proposed RFMS approach and detailed explanation of each attribute is given in Table 1. For example, in the behavior of game cards purchase, we define "R" as the time since the last purchase date till the time data was extracted, "F" as the number of times an applicant has bought game cards within a year. "M" is defined as average amount of monetary value an applicant spent on game cards in a year. And "S" is the standard deviation of amount paid per purchase. We name the variables in a simple way as a combination of behavior type and attribute abbreviation, such as "Game R" means the "R" value of purchasing game cards. Given different business scenarios, we summarize here ten kinds of applicant behaviors. For each behavior, we derive the corresponding RFMS attributes. These behaviors are described as follows.

Table 1: RFMS Definition.

Variable	Definition
R	Time since the last purchase for a certain behavior.
F	Total number of purchases for a certain behavior in a year.
M	Average amount of expense for a certain behavior in a year.
S	Standard deviation of amount spent for a certain behavior in a year.

(1) Debit. An applicant's transactions with debit cards.

(2) Consumption. An applicant's daily expense. The amount and frequency of daily consumption may indicate an applicant's repayment ability.

(3) Consumption Loan. An applicant's previous loan behavior (not necessarily adopt the same loan product). The lending history may indicate a credit record of a customer.

(4) Transfer. An applicant transfers money through the company's channel.

(5) Phone bill. An applicant pays his or her mobile phone bills.

(6) Utility bill. An applicant pays bills of water, electricity, gas or other infrastructures.

(7) Game. An applicant buys game cards or spends in computer games.

(8) State-owned Bank Card. An applicant's behavior of using state-owned bank cards. These banks include the Bank of China, the Agricultural Bank of China, the Industrial and Commercial Bank of China, and the China Construction Bank. Credit cards issued by these banks are typically more conservative, which means applicants might have a relatively higher qualification.

(9) Medium Bank Card. An applicant's behavior of using medium bank cards. These banks include the China Merchants Bank, the Shanghai Pudong Development Bank, the Industrial Bank and the PingAn Bank, etc. Unlike state-owned banks, these banks are more flexible when issuing credit cards.

(10) VIP card. The behavior of using VIP cards. A bank typically issues different types of cards according to the applicants' qualification. For example, some applicants with more deposits in the bank can obtain a VIP card. Applicants with VIP cards might have a lower default rate.

In conclusion, for each of the 10 behavior categories, we calculate the corresponding RFMS attributes. This leads to a total of 40 new variables, combining with the 7 applicant basic information variables, we have constructed 47 variables. With these variables, we can characterize applicant's transactional behaviors in a multidimensional manner. It is remarkable that the behaviors are generated according to the dataset. In real practice, other independent variables could be considered if more applicant behaviors could be observed, such as official credit registries, P2P loans and multi-lending behaviors.

4. EMPIRICAL RESULTS

4.1. Descriptive Analysis

The final data set includes 26,513 applicants who register a specific microcredit product, among which 7,980 are default applicants and 18,533 are non-default ones. The data are collected from T_0 (July 1st, 2015) to T_1 (December 31st, 2016). We illustrate the sample observation period in Figure 4. Consider, for example, a hypothetical applicant Mark. Mark registered the product at T_i, which is November 1st, 2015. Thus the registration length of Mark is $T_i - T_1$. It is remarkable that whenever Mark register the product, his transaction records could be collected from T_0 to T_1 in the platform. We could observe detailed information for each transaction, for example, the merchant name, card number, date and time and the amount of money. The transactions could be classified into different behaviors.

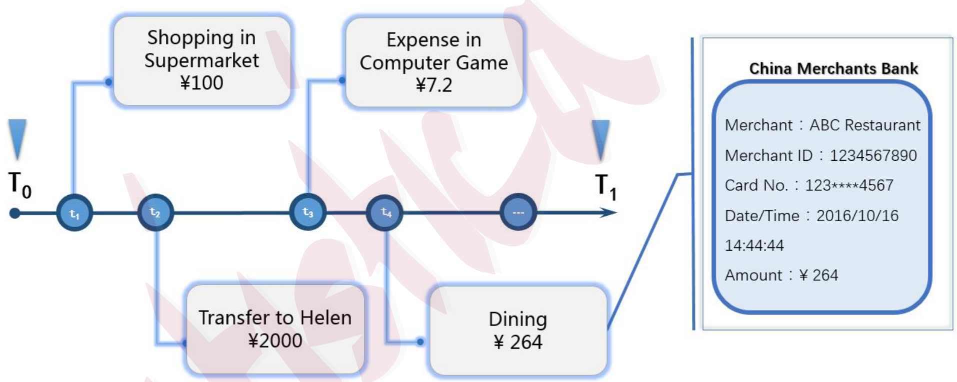

Figure 4: Example of Sample Observation Period.

Because these observations are not enough to represent stable attributes of applicants' behavior. Descriptive statistics of all the predictors are given in Table 2 and Table 3. Money-related variables are counting in Renminbi (RMB). Due to confidentiality reason, the summary statistics of basic score are not reported. Next, to further explore the relationship between the predictors and the response, we use boxplot to compare the difference between default applicants and non-default ones in each variables. For illustration purpose, variables are analyzed with boxplots in Figure 5. We use $Y_i = 0$ to indicate default applicants and $Y_i = 1$ to indicate non-default ones. We can summarize the following findings from these boxplots.

The first variable is the basic score which is developed by the company itself. From the boxplot, we can see that on average, a non-default applicant has a higher basic score than default applicant. This indicates that the basic score could roughly discriminate these two groups of applicants. The second variable is the number of transactions. It has a similar pattern as the basic score. It can be observed from the figure that non-default applicants make more transactions compared to their counterparts. The third variable is the mean of transactions. Although the difference is not obvious, we can still find that non-default applicants have a relative higher mean value of all behaviors than default applicants.

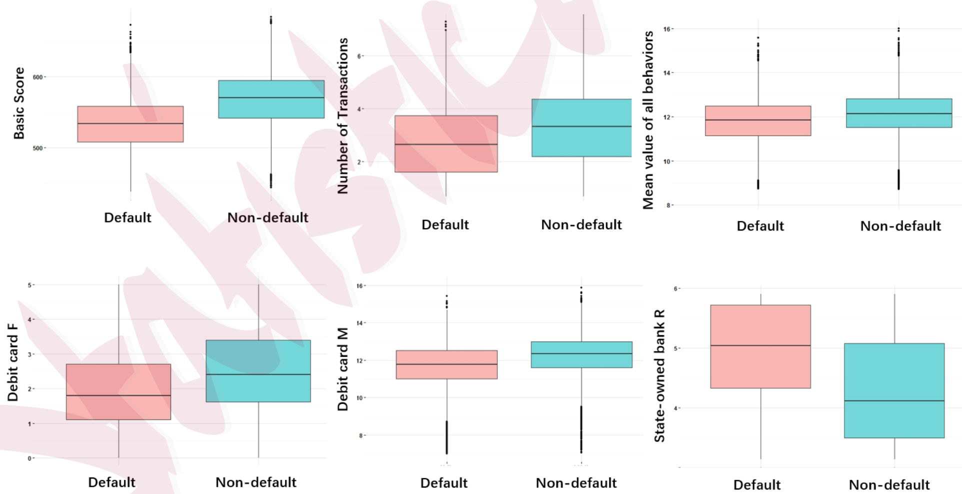

Figure 5: Boxplots of representative variables.

Next, we analysis applicant classification information variables. The debit card F means the frequency value of debit card usage behavior. It is clearly shown that on average, non-default applicants use their debit cards more frequently than default ones.

The debit card M stands for average monetary value of transactions with debit cards. From the boxplot, we can find non-default applicants have a relatively higher M value than default ones. The last illustrated variable is the state-owned bank R, which is the recency value of an applicant's state-owned bank cards' usage behavior. It is shown that the R value of default applicants is relatively higher than that of non-default applicants, suggesting that non-default applicants might pay with state-owned bank cards more often.

The descriptive analysis reveals that there are indeed some differences between default applicants and non-default applicants in terms of the constructed features. However, we still do not know whether these variables are significant or not in explaining the default behavior. Therefore, in the next subsection, we will use a model to comprehensively analyze the impact of each variable on credit scoring.

4.2. Model Results

To examine the impact of each predictor on the response variable, we conduct a logistic regression. However, due to the large number of variables, it is difficult to interpret all the coefficients as one may expect. Then we apply BIC criterion to select the most powerful predictors. All the continuous variables are transformed via logarithmic treatment and standardization.

The estimated coefficients chosen by BIC criterion are shown in Table 4. The standard errors, Z-values and P-values of the regression model are reported as well. To better summarize the features between default and non-default applicants, we divide the coefficients into positive parts and negative parts. According to the absolute value of each coefficient, we rearrange them by descending order and show them in Figure 6 and Figure 7. It is worth noting that, in comparison, coefficients displayed in Figure 6 indicate the features of non-default applicants, while Figure 7 summarizes the characteristics of default applicants. These two figures give us a relatively intuitive conclusion about default and non-default applicants.

From Figure 6, we can summarize some characteristics for those non-default applicants under the control of all other variables. For example, first of all, non-default applicants tend to have a higher credit to debit ratio. Applicants who frequently use credit cards and repay on time are often identified as good applicants by the bank. Therefore, a higher credit to debit ratio suggests more frequent use of credit cards and better credit records. Second, the mean value of all behaviors for a non-default applicant is typically higher than a default applicant. Next, debit card F also has a negative impact on the probability of default. It is shown that higher frequency of debit card usage indicates better customer's credit record. So a non-default applicant tends to have a higher debit card F value. Besides, compared with those default applicants, non-default applicants also have more transactions and a higher value of debit card M. Other characteristics for non-default applicants compared with default ones could be similarly summarized.

From Figure 7, we can conclude some features for those default applicants when controlling all other variables. First of all, credit loan R has a negative coefficient, which implies an applicant with a higher value of credit loan R has a higher probability to be a default applicant. Second, the more the number of cards owned by an applicant, the more likely he or she will be a default applicant. The maximum value of all behaviors is also negatively correlated with the predicted non-default probability, meaning that the more extreme an applicant's behavior is, the more likely he or she is to default. The length of registration is another indicator to distinguish default applicants from non-default applicants. For a new applicant, his default behavior will not show up in a very short period of time due to the repayment deadline. However, as the length of registration becomes longer, the likelihood for an applicant to default increases. Other variables displayed in Figure 7 (e.g., those RFMS related variables) can be explained in a similar way.

In conclusion, we have identified those significant variables in terms of explaining applicants' default behavior. With these variables, we can tell which applicant is more likely to default. It is remarkable that all these interpretations are made conditional on the existence of one important X-variable, that is the basic score. Due to the confidentiality reason, we are not aware of the construction details about the basic score. It is likely that some of the information considered in our analysis is also included by the basic score in some other ways. It is also possible that some information included by the basic score is not given in our dataset. Thus, all the regression coefficients obtained should be interpreted with caution. Nevertheless, this causes no problem for practical implementation. For real practice, the model is intended to be used with basic score and the prediction accuracy is the primary concern.

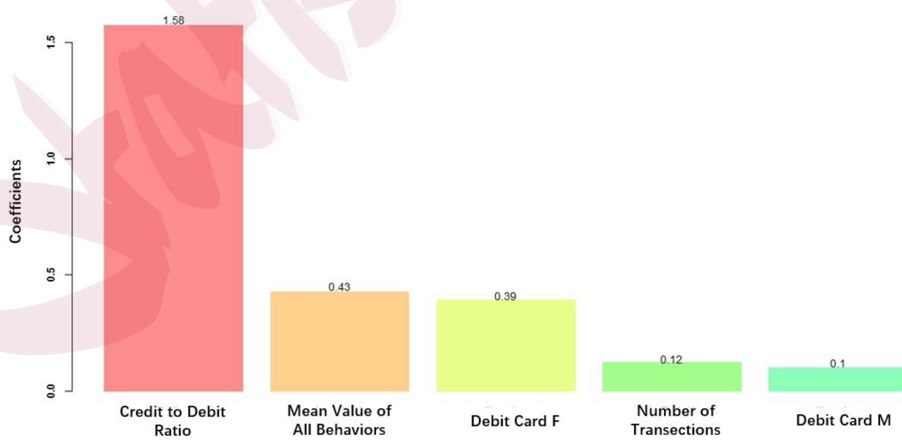

Figure 6: Positive Estimators of Regression Chosen by BIC.

4.3. Model Accuracy

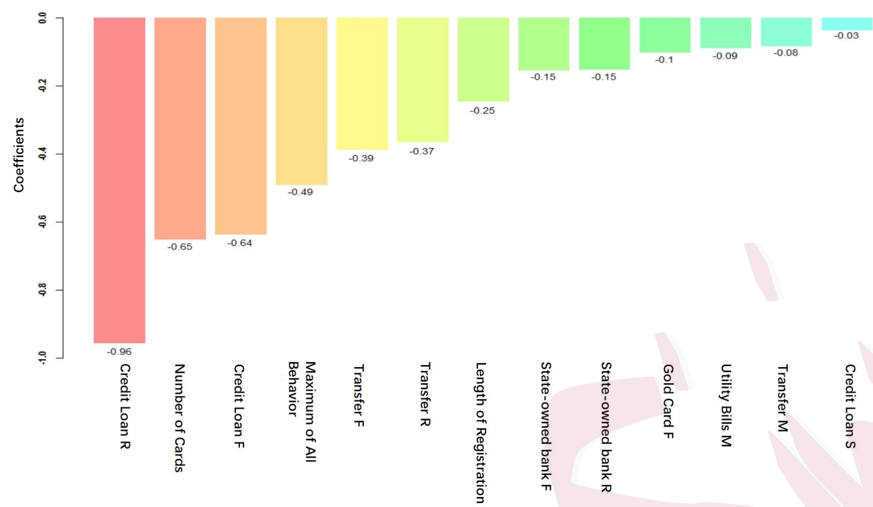

Figure 7: Negative Estimators of Regression Chosen by BIC.

To demonstrate the performance of the proposed model, we compare the prediction accuracy of the proposed model with the other two models. Model A, logistic regression model based on the basic score only. Model B, logistic regression model with all 47 variables. Model C, the proposed model, which based on model B after variable selection by BIC criterion.

To assess model accuracy, ROC (Receiver Operating Characteristic) curve and the value of AUC (Area Under Curve) are applied. The horizontal coordinate of ROC curve represents false positive rate (FPR), which is also known as 1 - specificity. It is calculated as the ratio between the number of negative events wrongly categorized as positive ones and the total number of actual negative events. The vertical coordinate is true positive rate (TPR), also known as sensitivity, referring to the proportion of positive events that are correctly identified as such. The closer ROC curve is to the upper left corner, the better the prediction is. AUC is the area under the ROC curve, whose value is positively related to the prediction accuracy.

For the convenience of model comparison, we randomly divide the data into training set (80%) and testing set (20%). We estimate parameters on the training set and apply the estimated coefficients on testing set. In this way, we obtain the predicted probability of non-default. This process is randomly repeated for 100 times. Figure 6 shows the ROC curves at once a time for illustration. In the figure, "score" stands for the prediction accuracy of Model A, "full model" presents the prediction accuracy of Model B, "BIC" shows the result of Model C. From the figure, we can draw a conclusion that the prediction accuracy of Model B and Model C are almost the same, which are much better than that of Model A. Furthermore, we compute the average value of AUC for 100 times. The absolute improvement on AUC is omitted because of commercial confidentiality. But compared to Model A, the prediction accuracy of Model B and Model C is relatively improved by 13.6%. This helps the company to evaluate an applicant's creditworthiness in a more precise way.

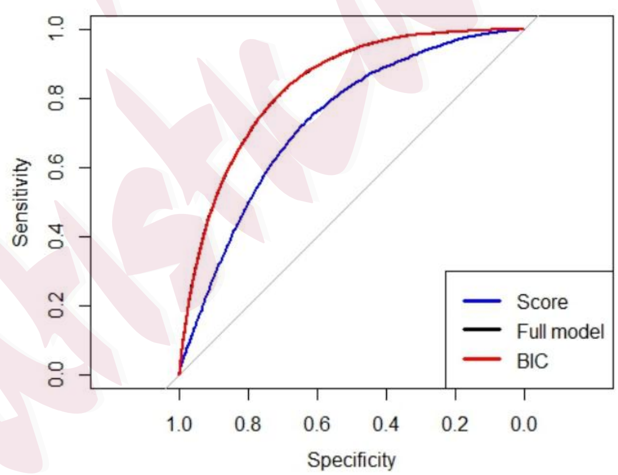

Figure 8: ROC curve of the 3 models.

5. BUSINESS IMPLICATION AND CONCLUDING REMARKS

For microcredit companies, the profit is reflected in return rate and the cost comes from default rate. The emergence of microcredit companies is mostly due to the inequality between profit and cost. To keep the default rate at a relatively low level, the microcredit companies usually use a set of strict rules to select qualified applicants. It means microcredit companies can achieve very high profits without complex data analysis. But as the market is mature, there are too many companies entering this field and the competition becomes intensified. As a consequence, in the near future, the return rate will decrease and default rate will increase. By that time, how to keep the default rate at a lower level becomes the primary concern for each microcredit company. The proposed model and prediction method have the following implications in practice.

(1) Analysts can use the predicted non-default rate to determine the approval of loan applications. For instance, there are two applicants who want to apply a loan from the comany. However, due to the limited budget, only one application could be approved. Then with our model, we can calculate that applicant A has a predicted non-default rate of 0.85 and applicant B is 0.27. It is clear that the company should approve the application of A. In practice, this model could help a company to target their applicants and control bad debt rate.

(2) We can use the predicted non-default probability to improve basic score developed by the company. For example, a linear transformation could convert the predicted probability P to applicant score Q ranging from 400 to 800, where Q=400+400×P. Then the basic score could be updated to get a more precise result.

In conclusion, this paper propose an RFMS method for credit scoring modeling in the background of microcredit industry in China. We demonstrate the significant effect of applicants' historical bank card transaction data on credit evaluation. It is remarkable that prior business knowledge plays a crucial role in variable construction. We discuss here three future topics as follows. First, we adopt the RFMS method for feature construction due to the characteristics of the transaction data, which are collected in one particular third-party payment platform. In practice, other competing theories and models in finance could be explored if we have different data sources and structures. Second, we define the default behavior as a binary variable in this model. If the actual days of default could be treated as a continuous variable, then the performance of the model would be reevaluated. Third, data from various platforms might have different structures. It is worthwhile exploring how to merge the data in different platforms. In different platforms, an applicant may have different credit scores. But a comprehensive credit evaluation model based on results from various platforms would lead to more fair results.

References

Antonakis, A. C. and Sfakianakis, M. E. (2009), "Assessing naïve Bayes as a method for screening credit applicants," *Journal of Applied Statistics*, 36:5, 537–545.

Blattberg, R. C., Kim, P. D., and Neslin, S. A. (2008), "Database marketing : analyzing and managing customers," *Springer*.

Calabrese, R. and Osmetti, S. A. (2013), "Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model," *Journal of Applied Statistics*, 40(40), 1172–1188.

Capotorti, A. and Barbanera, E. (2012), "Credit scoring analysis using a fuzzy probabilistic rough set model," *Computational Statistics & Data Analysis*, 56(4), 981–994.

Chehrazi, N. and Weber, T. A. (2015), "Dynamic valuation of delinquent credit-card accounts," *Management Science*, 61 (12), 3077–3096.

Copas, J. (1999), "The effectiveness of risk scores: the logit rank plot," *Applied Statistics*, 48(2), 165–183.

Durand, D. (1941), "Risk Elements in Consumer Instalment Financing," *New York: National Bureau of Economic Research.*

Eisenbeis, R. A. (1977), "Pitfalls in the application of discriminant analysis in business, finance, and economics," *The Journal of Finance*, 32, 875–900.

— (1978), "Problems in applying discriminant analysis in credit scoring models," *Journal of Banking and Finance*, 2, 205–219.

Fader, P. S., Hardie, B. G., and Lee, K. L. (2005), "RFM and CLV: Using iso-value curves for customer base analysis," *Journal of Marketing Research*, 42(4), 415–430.

Hand, D. J. (1986), "New instruments for identifying good and bad credit risks: a feasibility study," *Report. Trustee Savings Bank, London.*

Henley, W. E. and Hand, D. J. (1996), "A k-nearest-neighbour classifier for assessing consumer credit risk," *Statistician*, 45, 77–95.

Leonard, K. J. (1993), "Empirical Bayes analysis of the commercial loan evaluation process," *Statistics & Probability Letters*, 18, 289–296.

Lieli, R. P. and White, H. (2009), "The construction of empirical credit scoring rules based on maximization principles," *Journal of Econometrics*, 157(1), 110–119.

Orgler, Y. E. (1970), "A credit scoring model for commercial loans," *Journal of Money Credit & Banking*, Nov., 435–445.

Shepherd, D. (1990), "The New Direct Marketing," *Homewood, IL: Business One Irwin.*

Srinivasan, V. and Kim, Y. H. (1987), "Credit granting: a comparative analysis of classification procedures," *The Journal of Finance*, 42, 665–683.

Stanghellini, E., McConway, K. J., and Hand, D. J. (1999), "A discrete variable chain graph for applicants for credit," *Applied Statistics*, 48(2), 239–251.

Till, R. and Hand, D. (2003), "Behavioural models of credit card usage," *Journal of Applied Statistics*, 30(10), 1201–1220.

Wiginton, J. C. (1980), "A note on the comparison of logit and discriminant models of consumer credit behaviour," *Journal of Financial and Quantitative Analysis*, 15, 757–770.

Table 2: Summary Statistics of Predictors

Variable	Mean	Median	SD	Min	Max
Length of Registration	455.1	512.0	229.0	10.0	729.0
Number of Transactions	61.0	26.0	99.5	2.0	2,194.0
Mean of Transactions	264,890.4	171,320.0	307,848.5	0.0	11,316,725.0
Maximum Transaction	1,254,409.5	750,000.0	1,466,762.9	0.0	40,000,000.0
Credit to Dedit Ratio	0.1	0.0	0.2	0.0	1.0
Number of Bank Cards	6.2	4.0	5.5	1.0	71.0
Debit R	99.5	56.0	95.4	22.0	365.0
Debit F	26.8	10.0	50.2	0.0	1397.0
Debit M	280,691.9	189,786.5	316,959.4	0.0	9,236,688.0
Debit S	240,320.4	156,003.5	306,607.1	0.0	9,801,692.0
Consumption R	319.4	365.0	95.5	22.0	365.0
Consumption F	1.1	0.0	5.0	0.0	317.0
Consumption M	5,479.8	0.0	43,913.5	0.0	1,716,190.0
Consumption S	2,115.0	0.0	24,809.5	0.0	1,056,125.0
Consumption Loan R	151.5	99.0	126.6	22.0	365.0
Consumption Loan F	3.7	2.0	4.5	0.0	178.0
Consumption Loan M	271,737.4	200,766.7	263,287.3	0.0	1,926,525.0
Consumption Loan S	90,811.9	28,867.5	124,982.7	0.0	2,225,230.0
Transfer R	207.1	198.0	135.6	22.0	365.0
Transfer F	15.6	3.0	34.8	0.0	1,302.0
Transfer M	244,484.5	101,600.0	405,649.4	0.0	7,200,000.0
Transfer S	193,374.5	51,316.0	341,762.0	0.0	12,931,014.0
Phone bill R	326.5	365.0	89.2	22.0	365.0
Phone bill F	0.9	0.0	3.5	0.0	138.0
Phone bill M	1,738.8	0.0	4,297.4	0.0	50,000.0
Phone bill S	320.5	0.0	1,540.3	0.0	28,284.0

Table 3: Summary Statistics of Predictors

Variable	Mean	Median	SD	Min	Max
Utility bill R	358.2	365.0	41.1	22.0	365.0
Utility bill F	0.1	0.0	2.2	0.0	255.0
Utility bill M	732.1	0.0	13,562.9	0.0	1,604,670.0
Utility bill S	243.7	0.0	4,907.2	0.0	429,533.0
Game R	364.0	365.0	15.1	22.0	365.0
Game F	0.0	0.0	0.7	0.0	62.0
Game M	29.0	0.0	469.7	0.0	22,000.0
Game S	7.6	0.0	189.8	0.0	14,142.0
State-owned Bank R	130.4	79.0	118.6	22.0	365.0
State-owned Bank F	18.8	7.0	36.9	0.0	953.0
State-owned Bank M	253,333.1	152,951.2	336,552.1	0.0	10,914,000.0
State-owned Bank S	209,243.8	124,047.5	309,527.9	0.0	15,496,251.0
Medium Bank R	243.5	352.0	137.3	22.0	365.0
Medium Bank F	4.8	1.0	13.1	0.0	603.0
Medium Bank M	100,987.1	0.0	255,608.5	0.0	7,000,000.0
Medium Bank S	72,225.8	0.0	216,016.0	0.0	11,258,779.0
VIP Cards R	297.2	365.0	116.2	22.0	365.0
VIP Cards F	2.7	0.0	11.8	0.0	522.0
VIP Cards M	68,004.5	0.0	277,760.4	0.0	10,523,830.0
VIP Cards S	47,833.0	0.0	217,594.4	0.0	9,864,043.0

Table 4: Regression Coefficients After Variable Selection via BIC Approach.

	Coefficient	SE	Z value	P-value
Intercept	-154.600	2.874	-53.802	0.000
Basic Score	26.370	0.469	56.220	0.000
Credit to Debt	1.576	0.153	10.331	0.000
Ratio	0.427	0.035	12.174	0.000
Mean Value of All Behaviors	0.392	0.044	8.956	0.000
Debit Card F	0.122	0.036	3.363	0.001
Number of Transactions	0.098	0.006	15.337	0.000
Debit Card M	-0.035	0.005	-7.592	0.000
Credit Loan R	-0.081	0.005	-16.956	0.000
Number of Cards	-0.088	0.011	-8.011	0.000
Credit Loan F	-0.101	0.023	-4.415	0.000
Transfer F	-0.152	0.035	-4.375	0.000
Transfer R	-0.153	0.032	-4.741	0.000
Length of registration	-0.246	0.013	-19.451	0.000
State-owned bank F	-0.365	0.035	-10.361	0.000
State-owned bank R	-0.387	0.037	-10.361	0.000
Gold Card F	-0.490	0.033	-14.856	0.000
Utility Bills M	-0.636	0.045	-14.119	0.000
Transfer M	-0.650	0.049	-13.145	0.000
Credit Loan S	-0.955	0.034	-27.874	0.000
P -value of Likelihood Ratio Test			0.000	