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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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 collateral, steady employment, or a verifiable credit history. It is designed not only for start-ups but also for individuals. The microcredit industry is experiencing fast growth in China. In contrast with traditional loans, microcredit typically lacks 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, 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 to be solved is feature construction: how to construct meaningful and useful features based on 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. 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 on a data example from a Chinese microcredit company.

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

[†]Huang's research was supported in part by National Natural Science Foundation of China (Grant No.11701560), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (Grant No.16XNLF01),the Beijing Municipal Social Science Foundation (Grant No.17GLC051) and the Center for Applied Statistics, School of Statistics, Renmin University of China.

Zhou is the corresponding author and her research is supported in part by National Natural Science Foundation of China (Grant No.71702185), China Postdoctoral Science Foundation (Grant No.2016M600155), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (Grant No.17XNLF08).

Wang's research 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

1. INTRODUCTION

The microcredit industry, experiencing rapid growth in China, makes small loans to individuals and start-ups that typically lack collateral, steady income, or a verifiable credit history. Start-ups often have a lower probability of providing complete financial statements and higher risk of defaulting. Financial support from traditional financial institutions is unlikely. Microcredit for individuals is often for consumption purposes (e.g., traveling), and not supported by traditional loans. Many microcredit companies are springing up to provide more financial products to people at a 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 !)

Applicants' credit needs to evaluate carefully. Practically, this means assigning each applicant a credit score. Credit scoring is the key for successful microcredit. Here, credit scoring relies heavily on an applicant's background information, including, but are not limited to, educational background, working experiences, bank card transaction records. Various valuable and accurate data sources are increasingly available, so how to make good use of these data for accurate credit scoring is 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 individual who apply for the loan product and has been approved is referred to an applicant. However, not all applicants successfully return the principal and interest. Thus we observe their default behavior and they become the sample we study. For each sample i $(1 \leq i \leq n)$ a binary response $Y_i \in \{0, 1\}$ is formed 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 is needed; we collected them in a covariate vector $X_i \in \mathbb{R}^p$ with p for the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

dimension of the vector. A credit scoring model can then be developed for investigating the relationships between Y_i and X_i. Researchers have shown great interest in statistical models of credit scoring for more than 70 years. A variety of models has been proposed, including but 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). This research continues. 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.

These methods are very useful, but, they mainly focus on general methodology. They assume that applicants' characteristics have been well summarized, and thus could be used directly for regression or statistical learning. However, the fast development of information technology has provided access to various useful datasets. Often, the structure of the collected datasets cannot be fitted well by classical models. This is particularly true for bank card transaction data that widely exist and are considered as an important data source for credit analysis (Till and Hand, 2003; Chehrazi and Weber, 2015). For a given applicant i, the 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. Bank card transaction records are rather complicated (or even unstructured) data. Different applicants have different number of transaction

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

records, which are made at different time points with different cash amounts. How to construct meaningful X-variables from this complex, useful information sources is the major concern of this article.

We 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 previous research, we focus on the variable construction process instead of the general modeling methodology. Our approach is inspired by the popularly 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 of a customer's purchases; "F" stands for frequency, how often a customer purchases; "M" 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, "S" for the volatility of an applicant's activities. We call it the RFMS approach. The performance of the method is empirically tested using a data example from a Chinese microcredit company. The results show that our approach can significantly improve the accuracy of existing credit scoring in applications.

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

2. DATA DESCRIPTION

To implement the proposed method, we collaborated with a major microcredit

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

company in mainland China. One of the company's businesses finances consumption loans to individuals. To assess their applicants' creditworthiness, a basic credit score has been developed for general purposes for all the products in the company. Although the basic score is useful in many cases, the efficiency can be low for a particular product. We propose a new score, specifically designed for the target loan product.

The company operates a major payment platform in mainland China, and thus can track applicants' transactions through their bank card transaction records. This means every transaction of an applicant through this platform is 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: over-simplified portrayal of individuals who lack of a multidimensional evaluation; high risk of financial fraud such as exaggeration of employee payments, impossible for companies to detect; lack of cross validation that impairs the reliability of credit evaluation.

We focus on those applicants whose loans have been approved. However, by the time of maturity, some of them successfully return the principal and interest, while others fail. For those applicants, we have their complete data records and these can be used for modeling. This carries the risk of over-evaluation and is a problem faced by the entire industry. We have a sample of 26,513 applicants and almost 4,500,000 transaction records and other related information over the period of 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.

For each applicant, we can collect four types of information: applicant information; merchant classification; bank card information; transaction records. See Figure 1 for

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

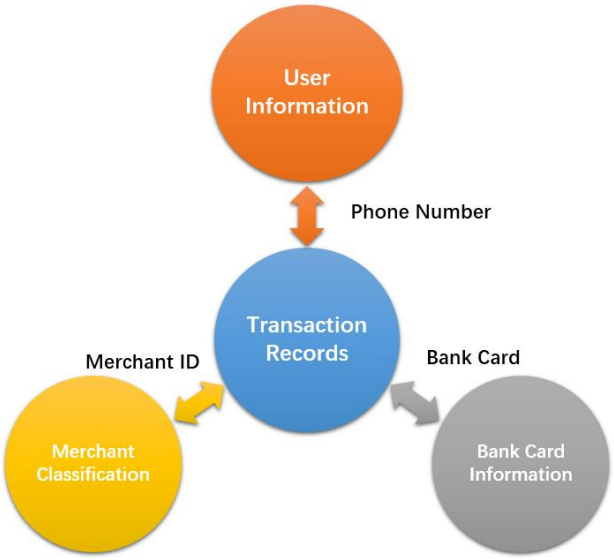

Figure 1: Data structure of a registered applicant for a specific product.

an illustration. These sources of information can be linked by keywords. Applicant information and transaction records are connected by phone number. Each phone number corresponds to several transaction records. Merchant classification and transaction records are associated through merchant ID which is related to several transactions. Bank card information can be linked to the bank card used in the transaction records. The first few digits of the bank card ID in the transaction record yield the bank card information. Variables extracted from these data are given as follows.

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

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

Bank Card Information. Bank card number identifier (the first few digits of

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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.

Transaction Records. Transaction serial number, cell phone number, transaction time, merchant number, bill number, bill amount, amount of payment, etc.

To proceed, 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. We integrate multiple transaction records from a single applicant into one observation by deriving applicant level variables, and the sample size is substantially reduced. Deriving applicant related variables is an interesting part of this project. Because we have little prior knowledge about what kind of applicants would default. In the next section, we explain how to to this.

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.

Basic Score. A credit score developed by the company, carefully designed for a general purpose and expected to be useful for all loan products. Being general, it is unlikely to produce best prediction accuracy for a particular loan product.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

Number of Transactions. Over a period of time, this reflects the frequency of the applicant's usage of bank cards.

Mean of Transaction. This is related to the concept of "monetary" value in the RFMS model.

Maximum Transaction. This to measure the extreme behavior of an applicant.

Credit to Debit Ratio. The proportion of transactions using credit cards to those of using credit and debit cards.

Number of Bank Cards. The number of bank cards owned by an applicant.

2. Classification Information Variables.

An 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. 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), that is widely used in marketing field.

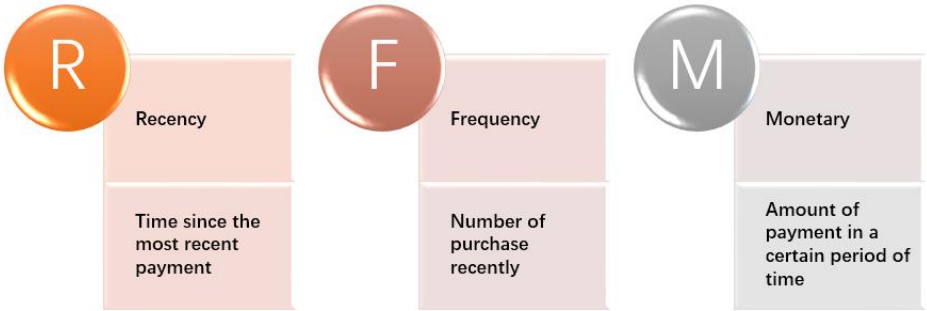

Figure 2: RFM model.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

The RFM method measures customer value and profitability using recency, frequency and monetary value. A typical RFM model is displayed in Figure 2. Here, "R" refers to the time since the last purchase; "F" refers to the total number of purchases in a certain period of time; "M" refers to the monetary value that a customer spends in a certain period of time.

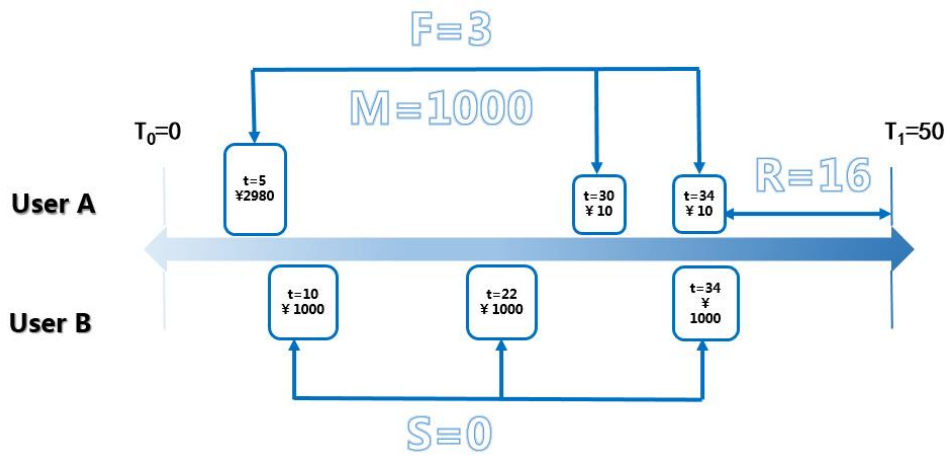

Figure 3: An Example to Explain S (Standard Deviation).

In this case of microcredit, Table 1 shows the details. Figure 3 shows different consumption behaviors of applicant A and applicant B with the same RFM score. The difference can be captured in part by their standard deviations, "S". This leads to the proposed RFMS approach, see Table 1. We derive the corresponding RFMS attributes based the following behaviors.

Debit is applicant's transactions with debit cards; Consumption is applicant's daily expense; Consumption Loan is applicant's previous loan behavior; Transfer is how an applicant transfers money through the company's channel; Phone Bill for when applicant pays his or her mobile phone bills; Utility Bill for when applicant pays bills of water, electricity, gas or other infrastructures; Game for applicant who buys game cards or spends in computer games; State-owned Bank Card measures

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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.

applicant's behavior in using state-owned bank cards; Medium Bank Card measures applicant's behavior in using other bank cards; VIP Card measures the behavior of applicant in using VIP cards.

For each of the ten behavior categories, we calculate the corresponding RFMS attributes. This leads to 40 new variables. By which, with 7 applicant basic information variables, we have constructed 47 variables. In practice, other independent variables could be considered if more applicant behaviors are 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 registered a specific microcredit product, among which 7,980 are default applicants and 18,533 are non-default ones. The data we collected from T_0 (July 1st, 2015) to T_1 (December 31st, 2016). See Figure 4. From transaction records collected from T_0 to T_1 in the platform, we can observe detailed information for each transaction: merchant name, card number, date, time and amount of money.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

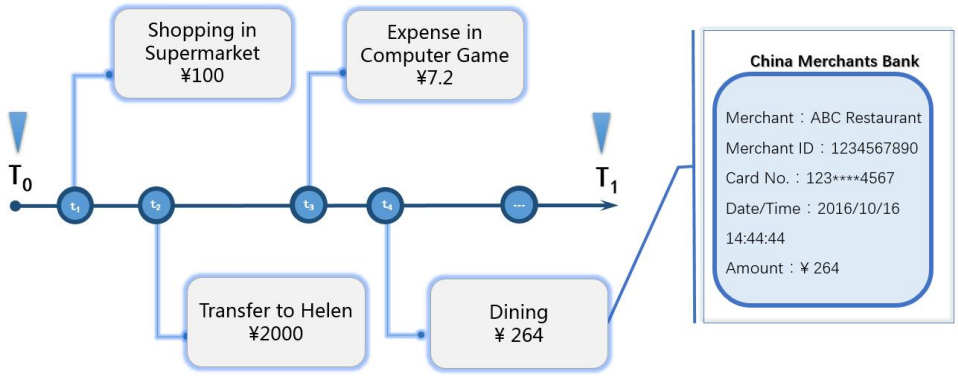

Figure 4: Example of Sample Observation Period.

As these observations are not enough to represent stable behavior, we give descriptive statistics of predictors, as in Table 2 and Table 3. Money-related variables are counting in Renminbi (RMB). For confidentiality reasons, the summary statistics of basic scores are not reported. Too further explore the relationship between the predictors and the response, we use boxplots to compare the difference between default applicants ($Y_i = 0$) and non-default ones ($Y_i = 1$) in each variable. See Figure 5. From them, we summarize as follows.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

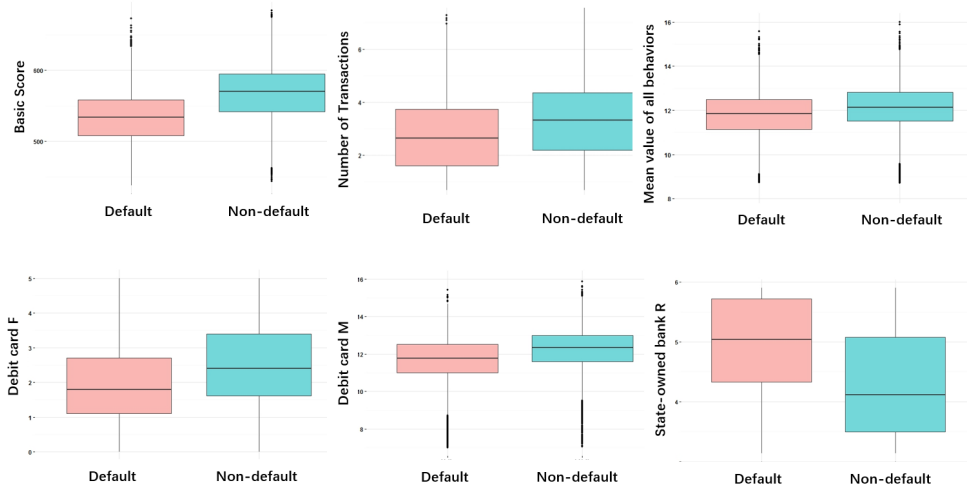

Figure 5: Boxplots of representative variables.

Next, we analyze 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, the recency value of an applicant's state-owned bank cards' usage behavior. 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.

Thus, there are indeed differences between default applicants and non-default applicants in terms of the constructed features. Whether these variables are significant or not in explaining the default behavior, we use a model to comprehensively analyze the impact of each variable on credit scoring.

4.2. Model Results

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

To examine the impact of each predictor on the response variable, we conducted a logistic regression. Due to the large number of variables, it is difficult to interpret all the coefficients, so we applied the BIC criterion to select the best predictors. All continuous variables were transformed via logarithmic treatment and standardization.

The estimated coefficients chosen by BIC 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 divided the coefficients into positive and negative parts. According to the absolute value of each coefficient, we rearranged them by descending order, see Figures 6 and 7. 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. First, non-default applicants tend to have a higher credit to debit ratio. Second, the mean value of all behaviors for a non-default applicant is typically higher than a default applicant. Debit card F has a negative impact on the probability of default. Non-default applicants tend to have a higher debit card F value. Compared with default applicants, non-default applicants also have more transactions and a higher value of debit card M. Other characteristics could be similarly summarized.

From Figure 7, we can conclude some features for those default applicants when controlling all other variables. First, credit loan R has a negative coefficient. 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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 does not show up in a 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 significant variables in terms of explaining applicants' default behavior. These interpretations are made conditional on the existence of one important X-variable, that is the basic score. Due to confidentiality reasons, we are not aware of the construction details of the basic score. It is likely that some of the information here is included in the basic score in 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. In practice, the model is intended to be used with basic score and prediction accuracy.

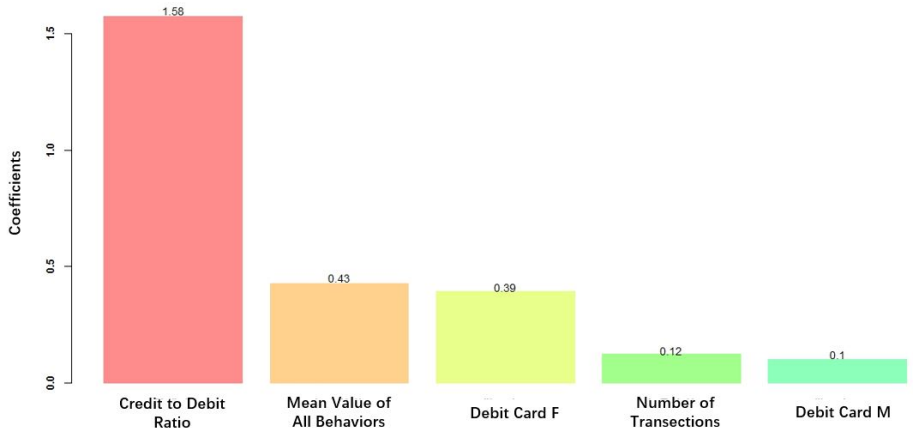

Figure 6: Positive Estimators of Regression Chosen by BIC.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

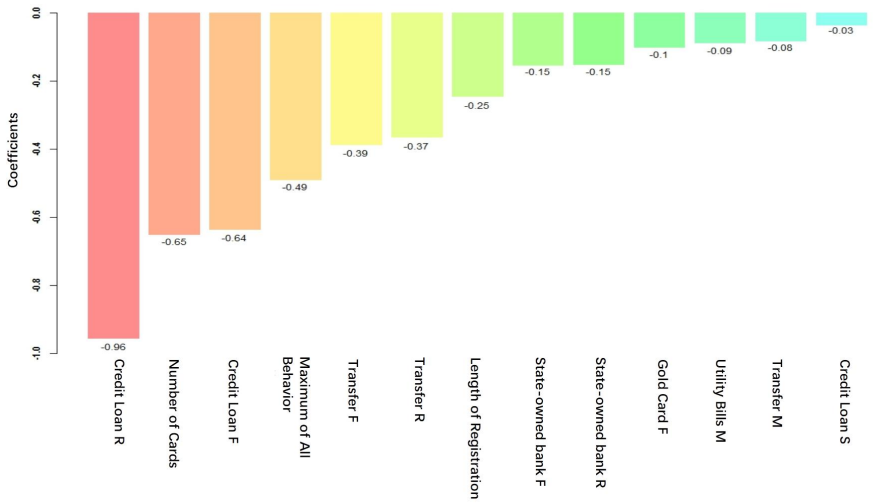

Figure 7: Negative Estimators of Regression Chosen by BIC.

4.3. Model Accuracy

To demonstrate the performance of the proposed model, we compared 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 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) were applied. The horizontal coordinate of ROC curve represents false positive rate (FPR). 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), 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 convenience of model comparison, we randomly divided the data into training

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

set (80%) and testing set (20%). We estimated parameters on the training set and applied the estimated coefficients on the testing set, obtaining the predicted probability of non-default. This process was randomly repeated for 100 times. Figure 8 shows the ROC curves for illustration. In the figure, "score" is 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, the prediction accuracy of Model B and Model C are almost the same, and much better than that of Model A. The absolute improvement on AUC is omitted here 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.

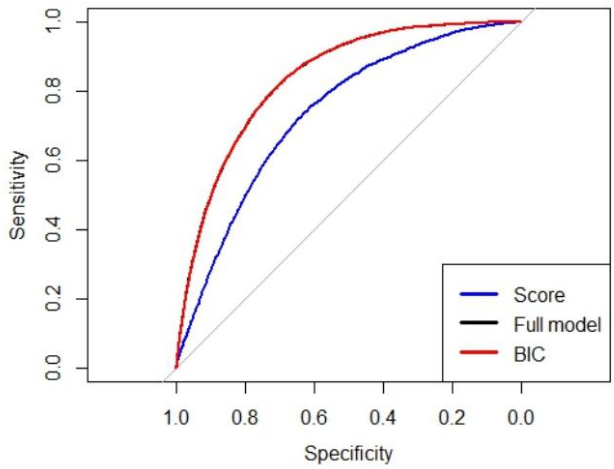

Figure 8: ROC curve of the 3 models.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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 microcredit market is maturing, the competition intensifies. In the near future, return rate will decrease and default rate will increase. How to keep the default rate to lower levels is a primary concern for microcredit companies. The proposed model and prediction method has implications in practice.

Analysts can use the predicted non-default rate to determine the approval of loan applications. Thus, with our model, one could choose applicant A with a predicted non-default rate of 0.85, over applicant B at 0.27.

We can use the predicted non-default probability to improve basic score developed by the company. Thus, a linear transformation could convert the predicted probability P to applicant score Q ranging from 400 to 800, where Q=400+400×P, and the basic score could be updated.

We discuss here three future topics. First, we can adopt the RFMS method for feature construction due to the characteristics of the transaction data, collected in a third-party payment platform. Competing theories and models in finance can be explored if given different data sources and structures. Second, we took the default behavior to be a binary variable. With addition of days to default, the performance of the model could be reevaluated. Finally, the merging of data from different platforms is a challenging issue.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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*.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0043

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	