Statistical modeling and forecasting analysis of credit card default

ll

2021/4/29

Abstruct

The purpose of this report is to study which customers will default on credit card repayment and which customers will repay normally, to analyze the data, establish a logistics regression model and make classification prediction, and to predict that customers with those characteristics will default on repayment through customer information characteristics, so as to provide relevant models and effective information for financial institutions, thus reducing financial risks.

Introduction

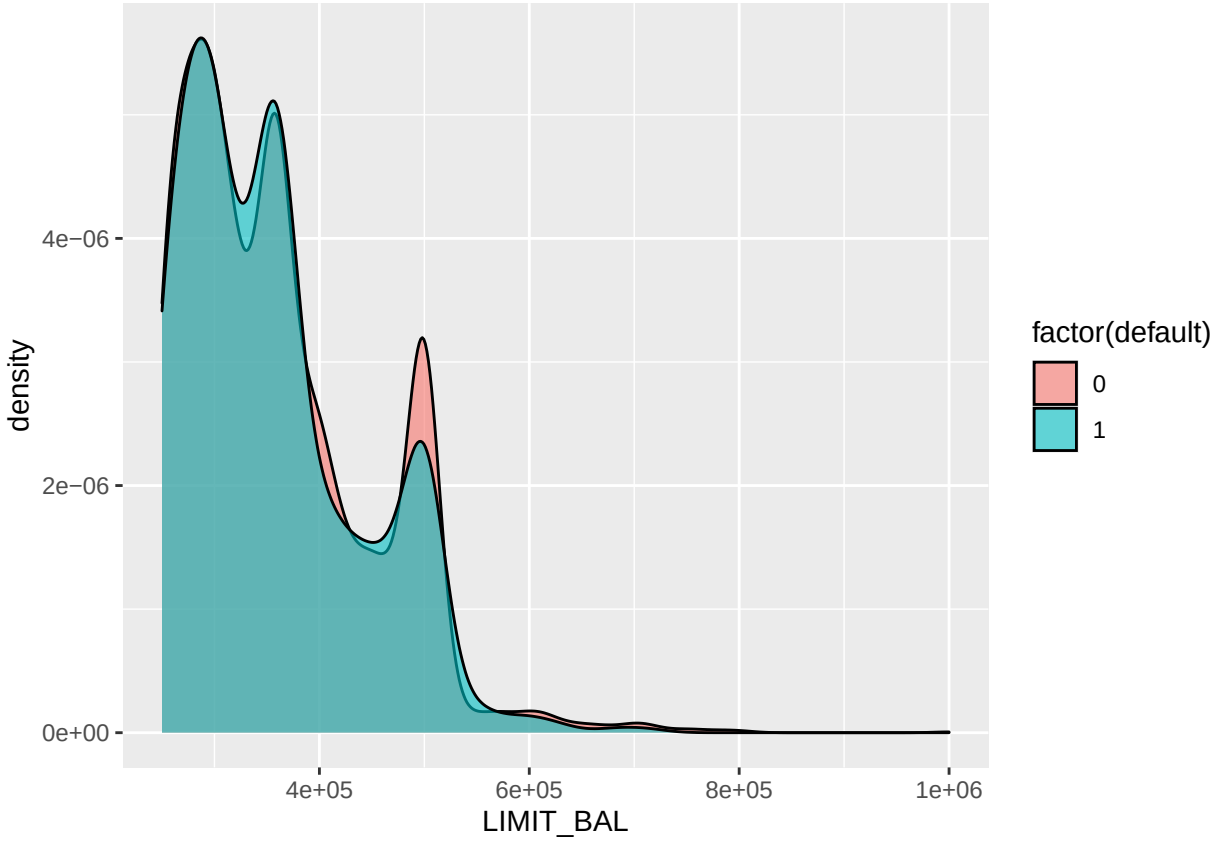
Credit Risk, also known as default risk, refers to the possibility that the borrower or counterparty is unwilling or unable to perform the contract conditions for various reasons, which constitutes a default, resulting in losses to the bank or counterparty. This is the main risk in the process of bank credit granting, and it is also a common risk in the process of bank credit card approval. The premise of credit card risk assessment is to obtain the credit information of the applicant, evaluate the obtained information of the applicant, quantify its performance ability, and finally assess whether to grant credit. In the past, credit granting to credit card applicants mainly depended on the evaluation of loan officers, and the evaluation results were influenced by subjective factors, or banks set up special credit decision-making committees to comprehensively evaluate applicants. However, the expansion of consumer loans and micro-loans in recent years has made the previous methods of relying on manual credit evaluation have limitations, and the manual evaluation also has some impersonality and incompleteness. At present, there are many problems in consumer credit, such as customers are difficult to keep, the loan amount is small, the default rate is high, there is no mortgage guarantee, and adjustment is difficult. It is necessary to cooperate with a more efficient, convenient, accurate, objective and lower cost credit analysis and evaluation scheme. At present, the financial market is facing great uncertainty, and both banks and even all financial institutions and investors are facing the test of huge credit risks. It is an urgent problem for all major financial institutions to establish a sound and effective investor credit evaluation system. Studies have shown that the bank's profit is closely related to the effectiveness of the credit evaluation model, and every 1% increase in the accuracy of the model, the bank can generate billions of returns. In addition, practical research shows that machine learning plays a positive role in the practice of establishing credit risk assessment model, which not only improves its accuracy but also expands its application scope. This paper tries to find a more accurate credit risk assessment system by applying the research concept of machine learning to the establishment of credit risk assessment model for credit card applicants, so as to help relevant financial institutions establish a more complete customer credit risk assessment system. The credit card risk studied in this paper is mainly to predict whether the trustee will have overdue repayment or non-repayment behavior for more than a large amount of time. The traditional credit card risk assessment is mainly through artificial judgment or discriminant analysis. Nowadays, most of the research uses Logistic regression model. This paper will model customers' credit characteristics and self-information to predict those customers who will repay on time and those who will be in arrears, hoping to provide relevant information for financial institutions to reduce financial risks.

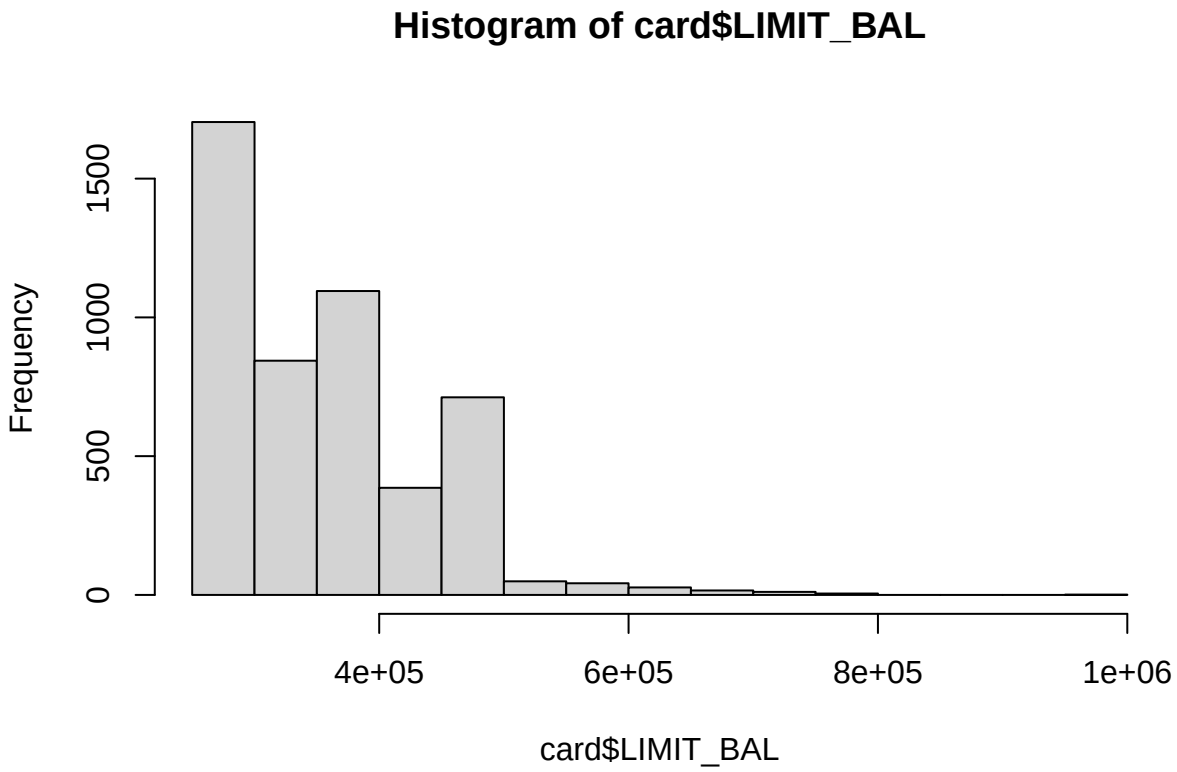
Data

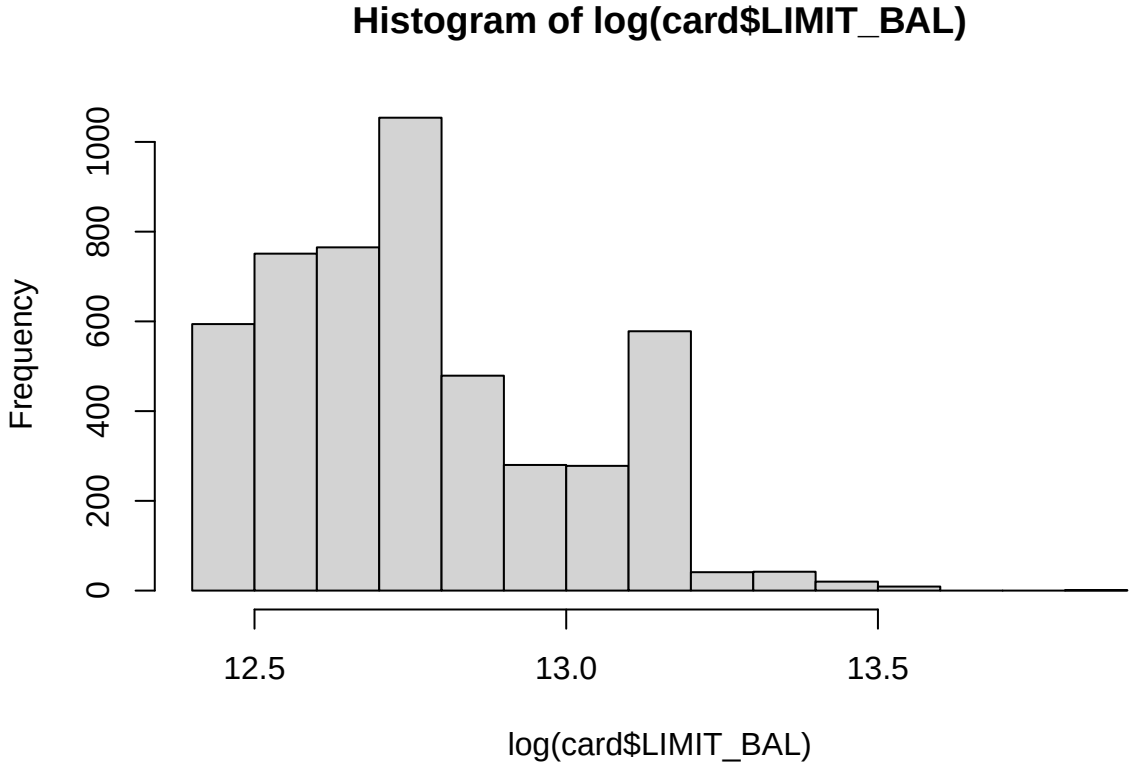
Data is extracted from "default of credit card clients Data Set" donated by I-Cheng Yeh and from Dua, D. and Graff, C. (2019). UCI Machine Learning Repository http://archive.ics.uci.edu/ml Irvine, CA: University of California, School of Information and Computer Science.The variables used in this report include credit limit, education level, Marital status, Gender, Age in years, The repayment status in date, Amount of bill statement in date, Amount paid in date and dependent variable Default payment(0=No,1=Yes)

Carry out preliminary data cleaning to see if the data has missing values. Because the sample size is too large, delete the data group with missing values, and get a total of 4892 data in the training set. The variable credit limit is a continuous variable, the maximum credit limit is 1,000,000, the minimum credit limit is 250,000, and the average credit limit is 360,627. The density chart shows that the larger the credit limit, the smaller the possibility of arrears, indicating that customers with large credit limits have better credit. Data conversion is carried out on the credit limit to make its distribution more uniform.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  250000  290000  350000  360627  410000 1000000
```



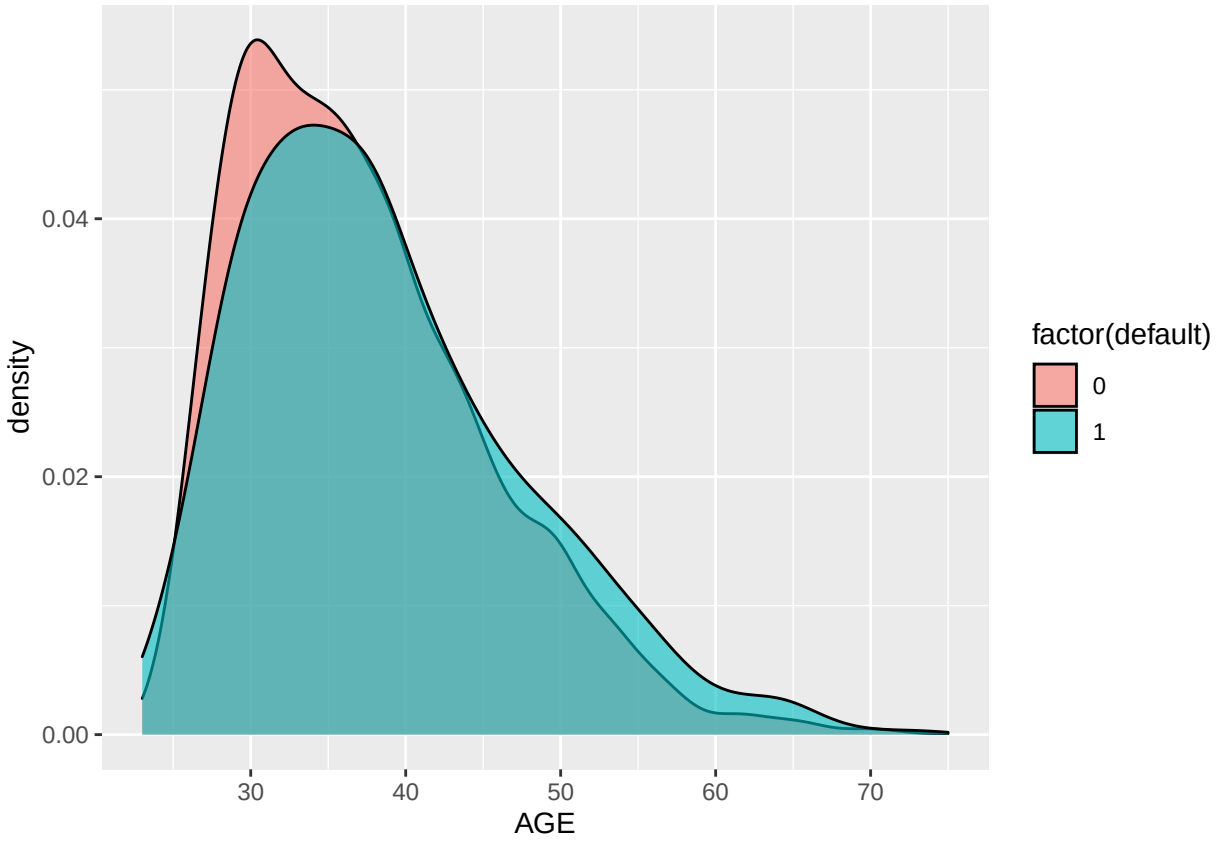




By analyzing the age data, we can see that the youngest customer is 23 years old, the oldest customer is 75 years old, and the average age of customers is 37.38 years old. The density map shows that young people have less probability of defaulting on payment than older people. With the increase of age, the probability of defaulting on payment will increase, probably because older customers will forget to pay on time.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   23.00   31.00   36.00   37.38   42.00   75.00
```



Continue to make descriptive statistical analysis of the three classified variables of the customer's marital status, education level and gender. Among the customers who defaulted on repayment, 406 were married and 285 were single; There are 307 men and 384 women. There are 364 graduate students, 254 undergraduate students and 73 high school students. Among the customers who repay on time, there are 2,194 married and 2007 single; There are 1636 men and 2565 women; There are 2204 graduate students, 1583 undergraduate students and 414 high school students. For the data from April to September, 2005 with the variable The repayment status in date, the variables no consumption, pay duly, the use of revolving credit are defined as 0. Through data analysis, it is found that very few customers have defaulted on repayment for 4 months or more, and the definition of arrears for more than 3 months is 3. The variables Amount of bill statement in date and Amount paid in date are not processed.

```
##
##     married single Other
##   0    2194   2007     0
##   1     406    285     0
```

```
##
##     GradSch  Uni HighSch Other
##   0    2204 1583     414     0
##   1     364  254      73     0
```

```
##
##     male female
##   0 1636   2565
##   1  307    384
```

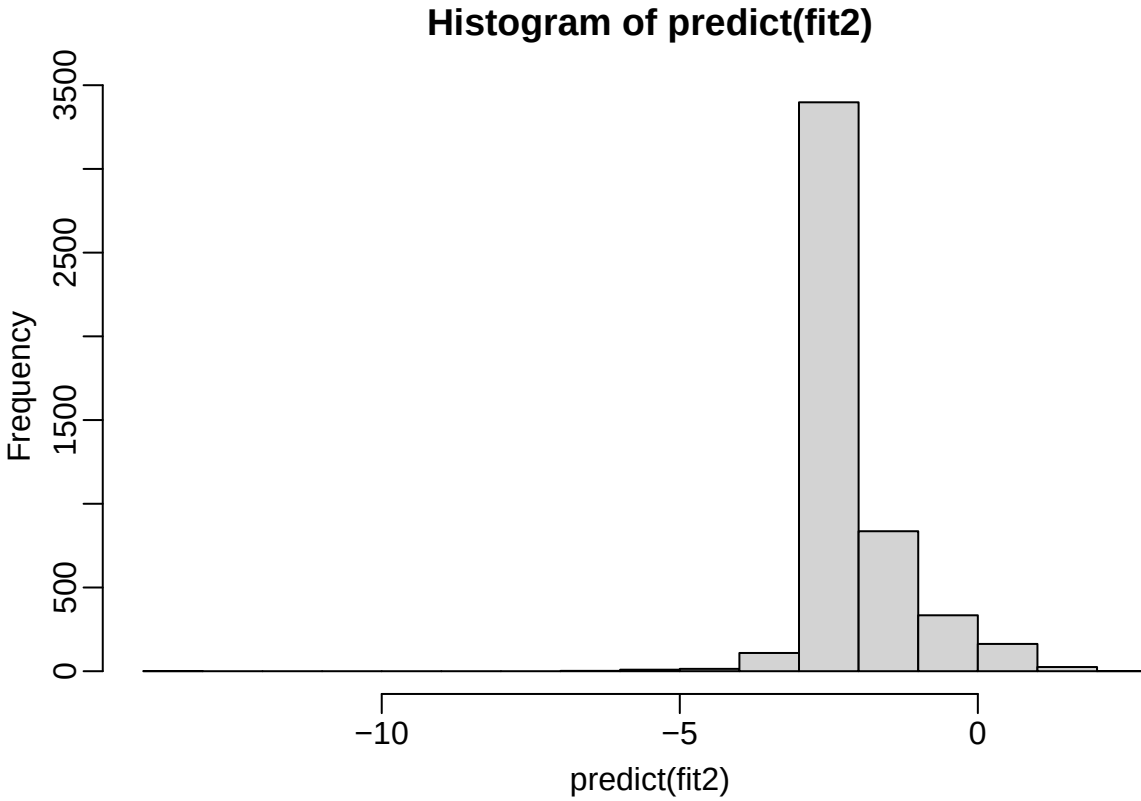
Method

In this report, the Logistic regression model will be used for regression analysis, and a prediction classification model will be established. The Logistic regression model is mainly used to analyze the relationship between independent variables and discrete dependent variables. Dependent variables are generally classified variables of "0-1" type. In this study, the main research content is personal credit risk evaluation, and the dependent variable y is binary variable with values of 0 and 1 respectively; Y=1 represents a customer with default behavior, and y=0 represents a customer without default record. Logistic is essentially a discriminant model based on conditional probability. In actual credit approval classification, a threshold is set for classification, so Logistic regression model can also be regarded as a probability estimation, that is, an estimation of the default probability of the user. ROC curve is used to judge the validity of the model, which represents the result combination of multiple confusion matrices. assuming that the threshold definition in the above model is unsuccessful, the prediction results of the model are simply sorted in descending order, and the threshold is defined by each probability value in sequence, so that many confusion matrices can be generated. Confusion matrix is the basis of ROC curve drawing, and it is also the most basic, intuitive and simple method to measure the accuracy of classification model. Confusion matrix is to count the number of observed values of the wrong and right classes of the classification model respectively, and then display the results in a table, which is called confusion matrix.

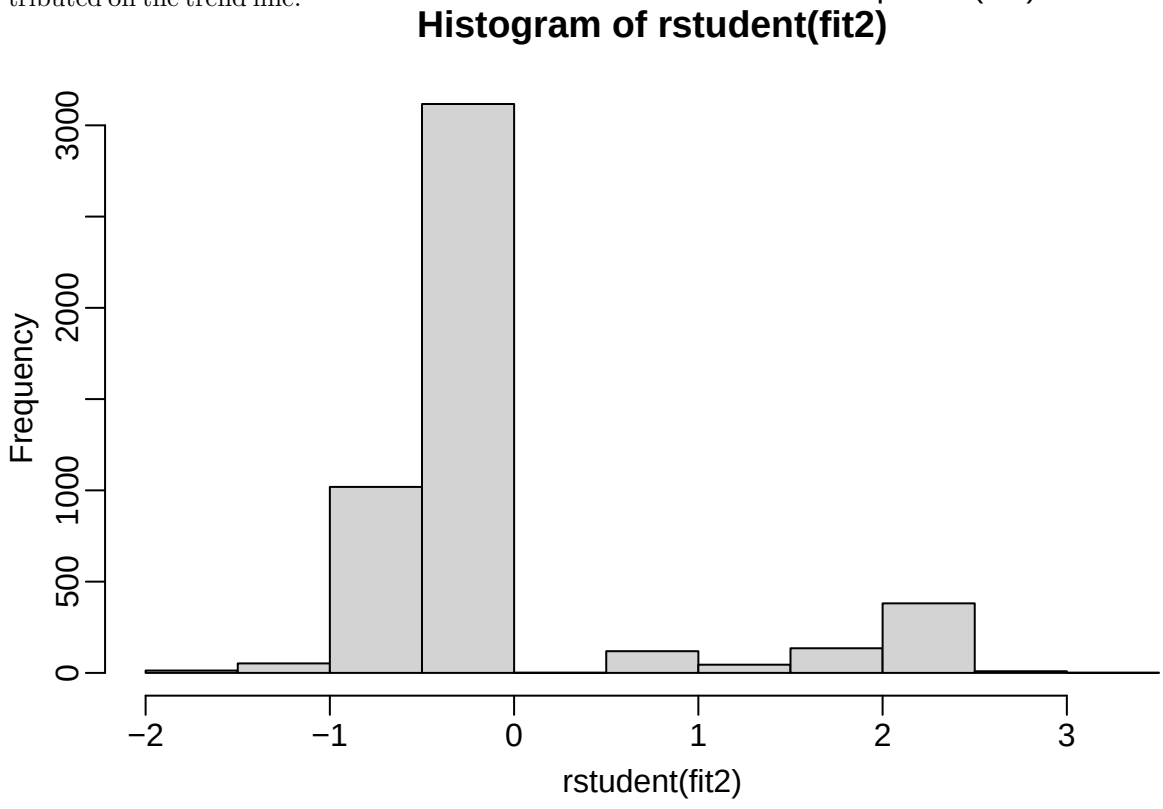
Results

Model fitting: After describing the variables, we carry out regression analysis. Firstly, we carry out logistics regression on all the variables, and get Model 1. In Model 1, age, September payment is delayed by 1 month, 2 months and 3 months. The amount in April bill, the amount in May bill and the amount paid in September have significant influence on the dependent variables, among which age, September payment is delayed by 1 month, 2 months and 3 months. The less the amount paid, the positive effect on default payment is produced. However, because many independent variables have no significant influence, it is necessary to optimize the model. Through a series of variable selection attempts, the stepwise regression method is adopted, and the trial process is omitted, and Model 2 is obtained. In Model 2, sex(p-value=0.0926) has a negative effect on default payment, while age(p-value=0.0052) has a positive effect on default payment. In September, every month's increase in repayment delay has a positive impact on default payment. The more money in the bill in April has a negative impact on default payment. The more money in the bill in May has a positive impact on default payment. The more money paid in September and October has a negative impact on default payment.

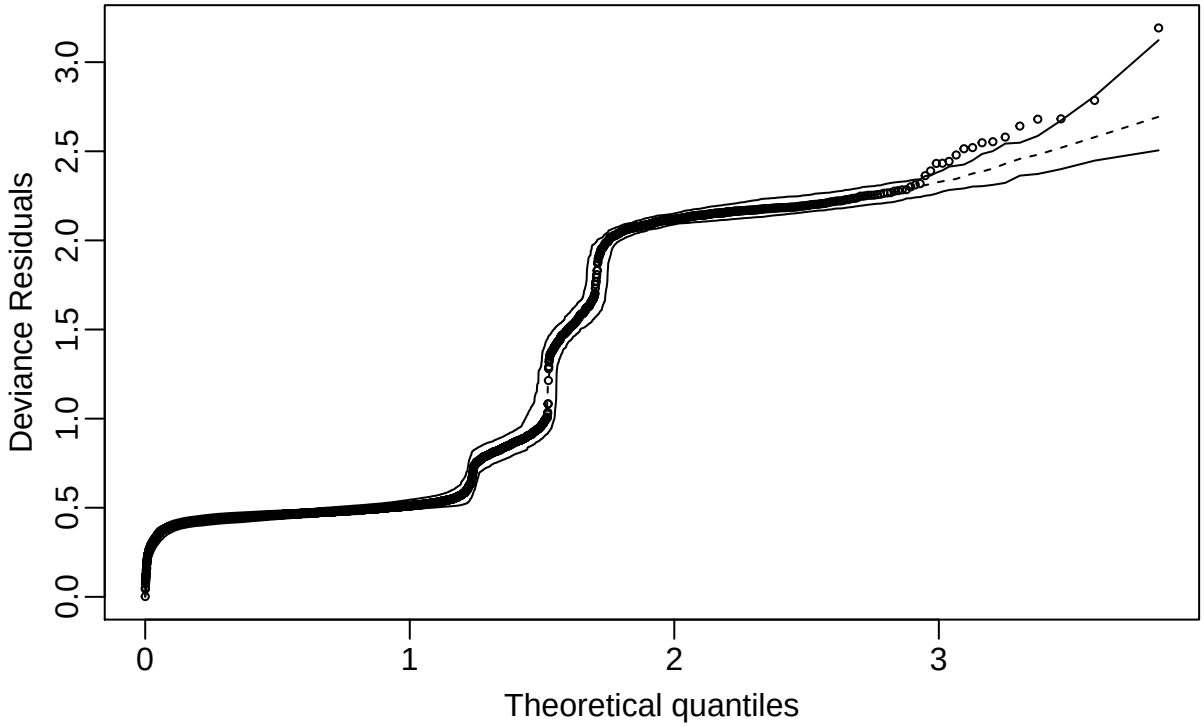
Model diagnosis: Compared with model one (full model), model two has a better goodness of fit, and AIC is smaller than 3553.471. fitting the residual diagram of model two, deviance residuals or the student residuals shows that there are some trends in the fitting residual of the model, but the overall fitting is good. Added Varaible Plots each variable is evenly distributed, and the residual of partial residual graph is evenly dis-



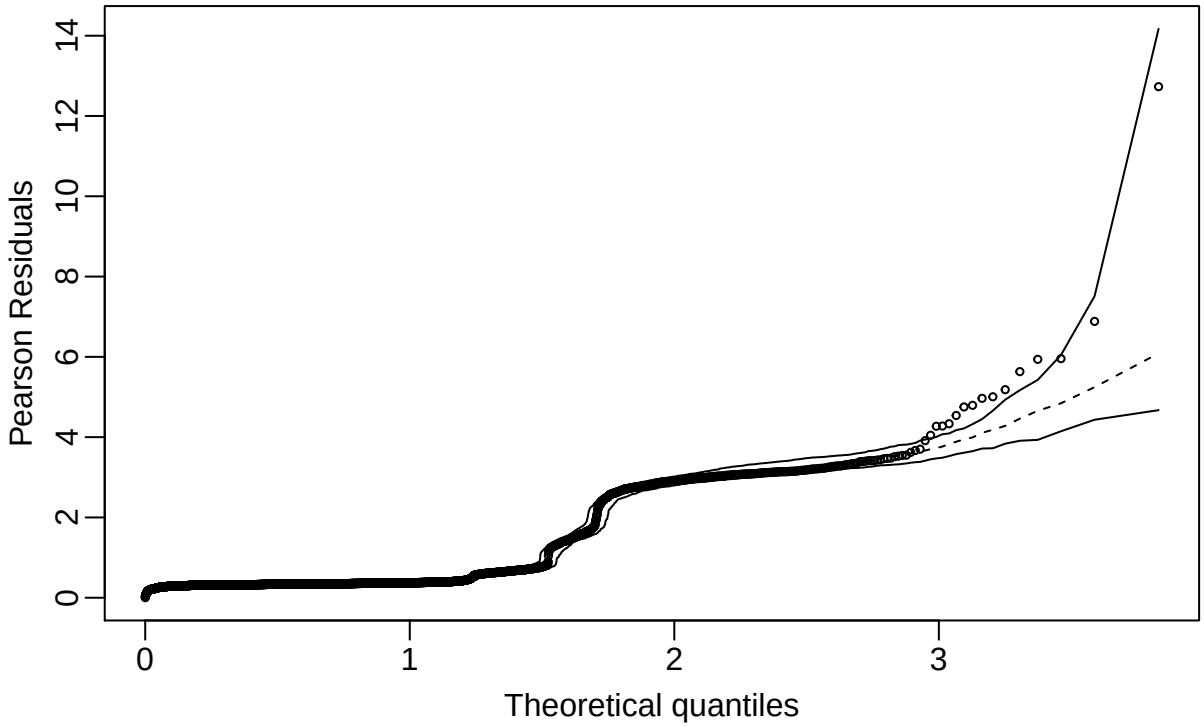
tributed on the trend line.

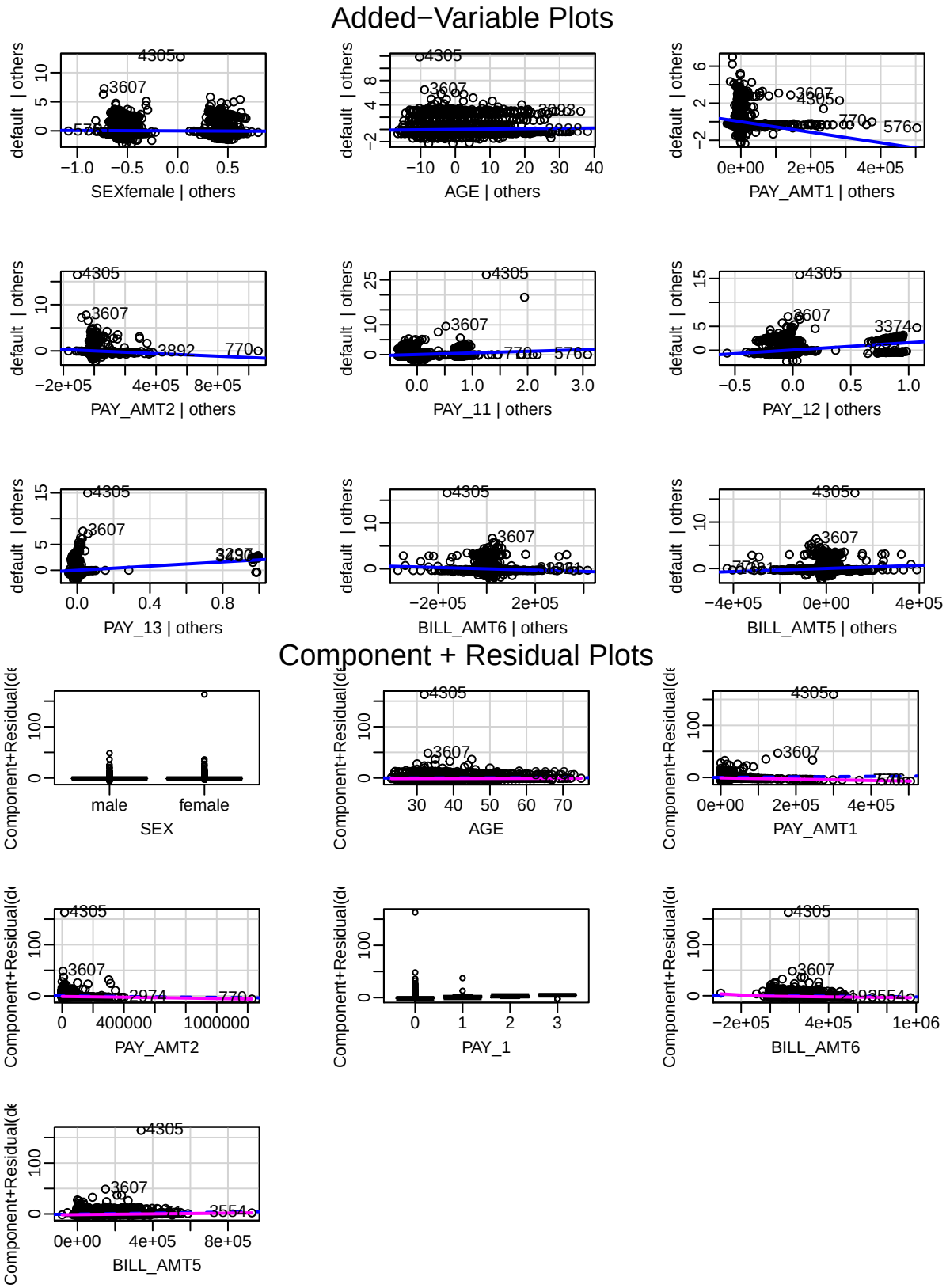



```
## Binomial model
```




```
## Binomial model
```

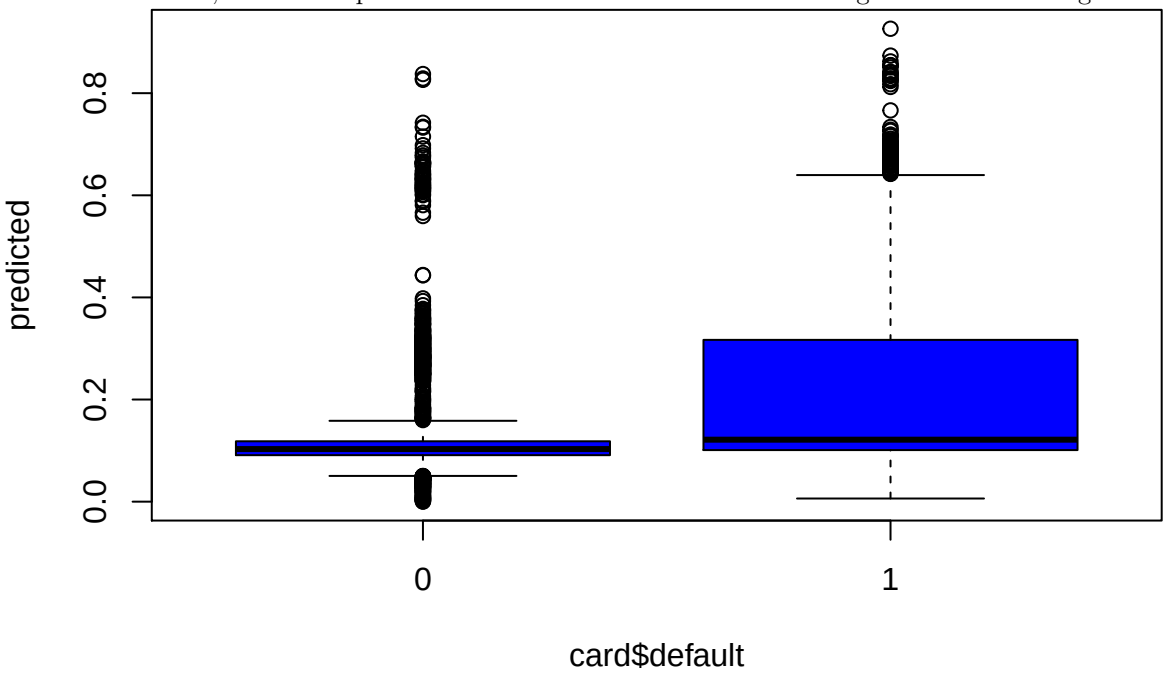




Model prediction and classification:

The purpose of establishing the model is to use the model to predict and classify credit card customers, and

check whether customers have the possibility of default payment by fitting variables. The model 2 is used for logistics regression classification, and the confusion matrix is obtained by calculation. There are a large number of first-class error data and second-class error data. By selecting the cutoff point several times, the cutoff point is finally set at 0.46, so as to get the maximum accuracy. Then the accuracy, false positive rate and false negative rate are calculated, and the accuracy is 87.26. In order to check the accuracy of classification and fit the ROC curve, we first fit the traditional ROC curve, which can easily find out the influence of any threshold on the generalization performance of the learner. It is helpful to select the best threshold. The closer the ROC curve is to the upper left corner, the higher the recall rate of the model. The point on the ROC curve closest to the upper left corner is the best threshold for the least classification errors, and the total number of false positive cases and false negative cases is the least. Then calculate AUC, and get AUC = 0.7. The confidence interval of AUC 95% is [0.673,0.719]. It can be concluded that AUC is average but not ideal. AUC is the area under ROC curve, which is a performance index to measure the advantages and disadvantages of learners.



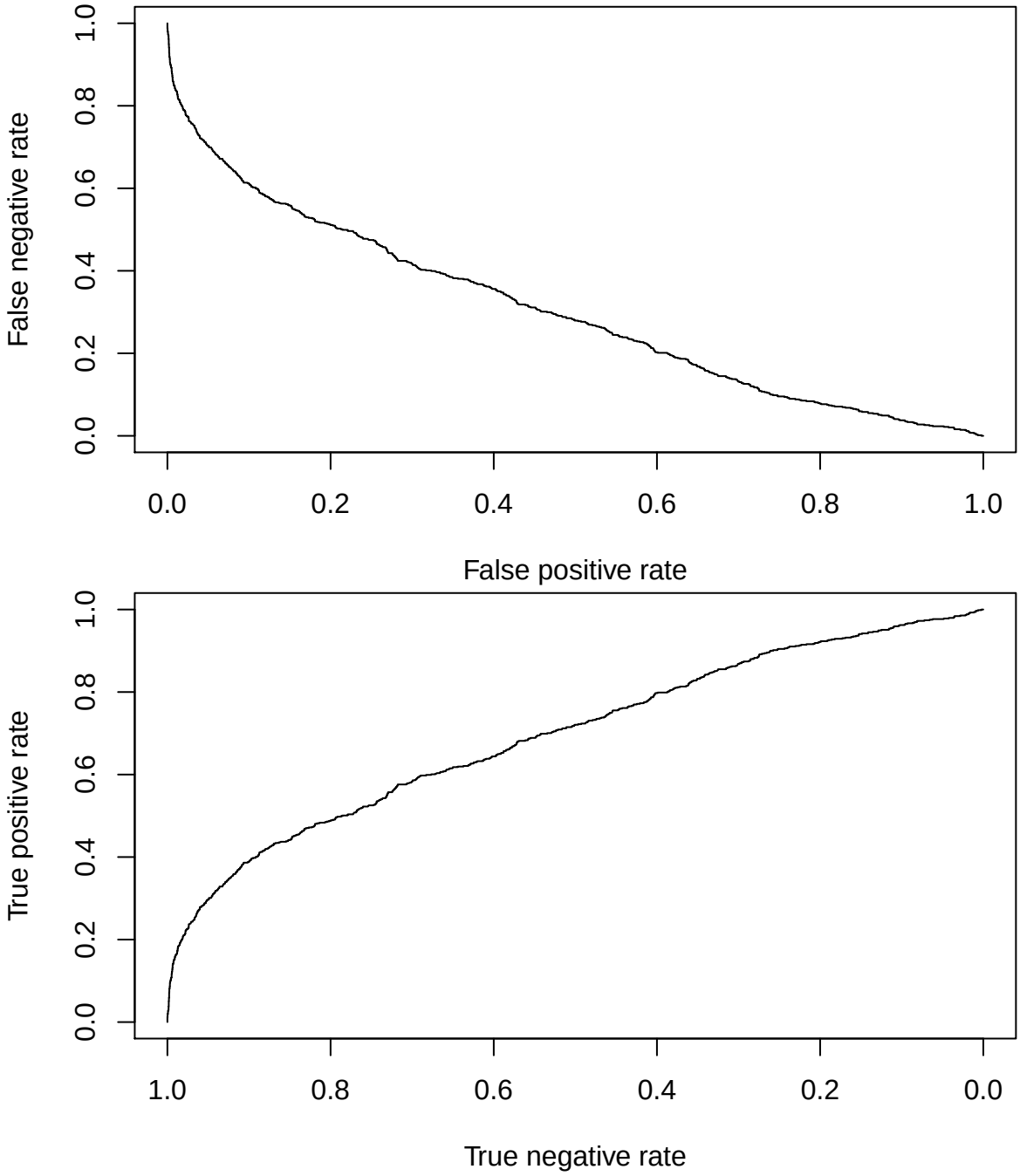

```
##      ypred
##       FALSE TRUE  Sum
##   0    4140   61 4201
##   1     562  129  691
##   Sum  4702  190 4892
```

```
## [1] 0.8726492
```

```
##
## Fold:  5 7 8 1 6 2 10 3 9 4
## Internal estimate of accuracy = 0.872
## Cross-validation estimate of accuracy = 0.872
```

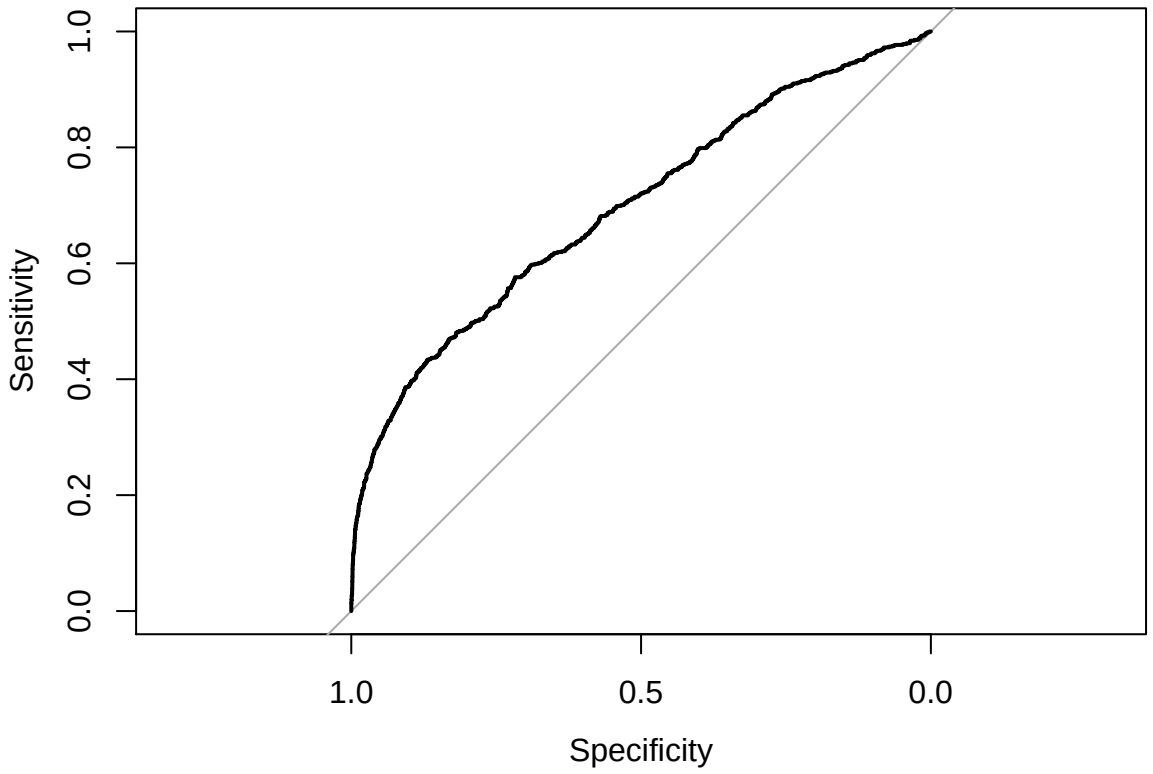
```
## [1] 0.01452035
```

```
## [1] 0.813314
```




```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



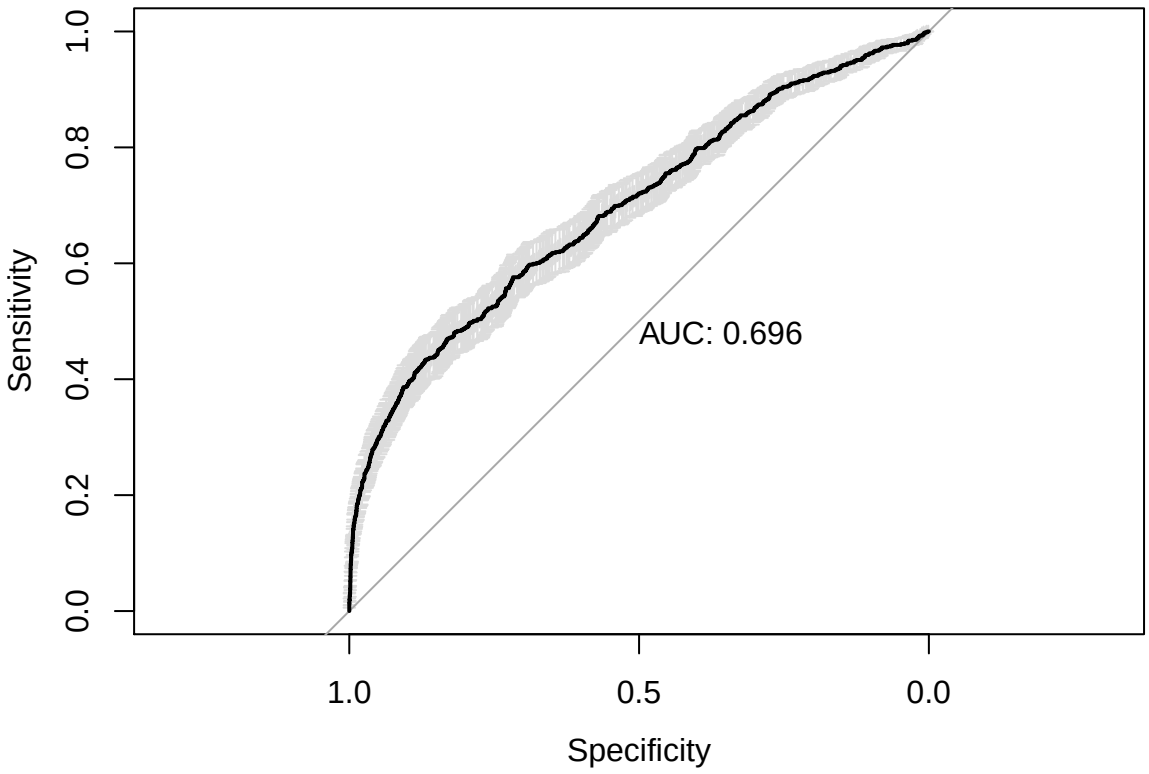

```
## Area under the curve: 0.6962

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

## Area under the curve: 0.6962

## 95% CI: 0.6732-0.7192 (DeLong)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Conclusion

In this paper, logistics regression is used to build the model, and the risk assessment of credit card applicants is carried out through data mining tools. In the process, the shortcomings of building the model are fully understood, and the model optimization is not good enough, and the prediction accuracy is not high enough. The data predicted by the test set shows that the classification results are not good enough, and the sampling method can be optimized, and stratified sampling and other methods can be considered. From the information of the variables in the article, it can be known that men, who have delayed repayment for more than one month, have a positive impact on default payment, and are more inclined to default payment, but these variables are not enough to build a complete model, so more effective variables can be considered to build a model. The credit card risk assessment method constructed in this paper cannot completely cover the credit assessment system, but is only a part of it. With the help of a more complete credit evaluation system, customer information can be obtained more effectively and comprehensively. With the popularization of credit card business and the continuous development of social credit, customer credit evaluation is becoming more and more complicated. At the same time, data mining tools are used more and more frequently, which is more helpful for us to complete the evaluation of customer credit. With the advantage of data mining, we can summarize the habits and laws of customer credit more efficiently, help banks to complete the mining of key elements, and help them to carry out customer credit information verification. At the same time, it also plays a guiding role in further perfecting the establishment of customer credit evaluation system.

Appendix

[1]Altman E I .Financial Ratios,Discriminates Analysis and the Prediction of Corporate Bankruptcy[J].Journal of Finance,1968.

[2]Ausubei and Lawrence M.The Failure of Competition in the Credit Card Market[J].The American Economic Review,1991(81):50-81.

[3]Balse Committee on Banking Supervision.Credit risk modeling:current practices and applications,1999.

[4]Chen T,Guestrin C.XGBoost:A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM,2016:785-794.

[5]Jackson J R.Simluation Research on Job Shop Production[J].Naval Res Log Quart,1957,4(3):287-295.

[6]Mays E Handbook of Credit Scoring[M]. Fizroy Dearborn,2004.