Modeling Grade Distribution

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT
FOR THE REQUIREMENTS OF THE DEGREE OF

Bachelors of Science in Mathematics

By

Hanwen Chen

under the guidance of

Jebessa Mijena, Phd.

Department of Mathematics
Georgia College and State University 2020

Table of Contents

Abstract.....4

Introduction.....6

Methodology.....7-8

1. Type of Research.....7

2. Data Collection and Clean-Up.....7

3. Analysis.....7-8

Statistical Methods.....9-21

1. Linear Regression.....9

1. 1 Multiple Linear Regression's Model.....9

1. 2 Test Data RMSE.....9

2. Subset Selection.....9-13

2. 1 Forward Stepwise Selection.....9-10

2. 2 Backward Stepwise Selection.....10

2. 3 Choosing the Optimal Number of Predictor.....10-11

2. 3. 1 Bayesian Information Criterion.....10-11

2. 3. 2 Adjusted R^2.....11

2. 4 Results and Discussion (Forward Stepwise Selection).....11

2. 5 Results and Discussion (Backward Stepwise Selection).....12

2. 6 Test Data RMSE.....12-13

3. Shrinkage Methods.....13-15

3. 1 Ridge Regression.....13

3. 1. 1 Ridge Regression's Model.....13

3. 2 Lasso 13-14

3. 2. 1 Lasso's Models 13-14

3. 3 Choosing the Optimal λ 14

3. 4 Test Data RMSE 14-15

4. Shrinkage Methods 15-17

4. 1 Dimension Reduction Methods 15

4. 2 Choosing the Optimal Number of Principal Components 15

4. 3 Results and Discussion 15-17

4. 4 Test Data RMSE 17

5. Tree-Based Methods 17-21

5. 1 Regression Trees 17-19

5. 1. 1 Test Data RMSE 19

5. 2 Boosting 19-21

2. 3. 1 Results and Discussion 19-21

2. 3. 2 Test Data RMSE 21

Conclusion 22

References 23

Appendix 24-27

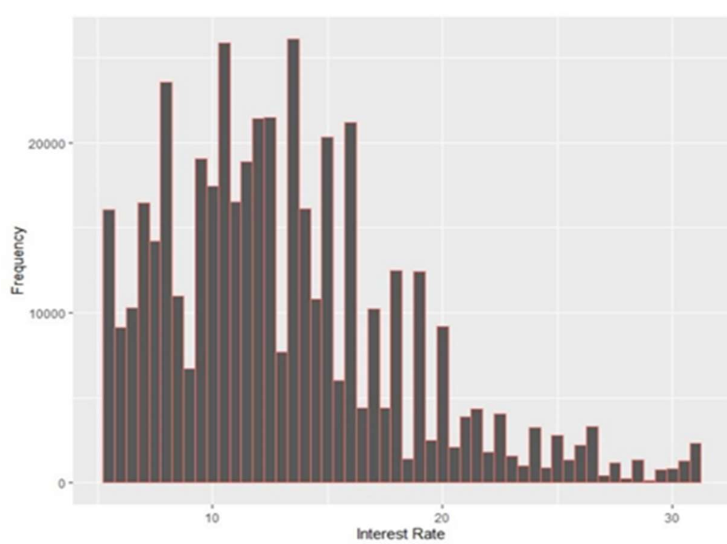
Abstract

The goal of this research is to develop a model that could predict the interest rate on loans with attention to accuracy based on the information provided by clients. We collected financial data from LendingClub, which is an American peer to peer lending company, and took out of uncorrelated predictors and missing values in the database. We applied different statistical methods to construct a predictive model with the highest accuracy. These methods were linear regression, shrinkage methods, dimension reduction methods, and tree-based methods. We evaluated the performance of these predictive models by comparing the difference between the predicted interest rate and the actual interest rate on the test data. We studied the association between the interest rate and the remaining predictors. We found that four predictors: the term of the loan, the last FICO scores, the total open-to-buy budget on revolving bankcards, and the initial listing status of the loan recorded as a whole or fractional loan, were most critical in predicting the interest rate. The best statistical method in predicting the interest rate was boosting. All model computations were done on R statistical software.

Keywords: Interest Rate, Pricing Methods, R Statistical Software

Introduction

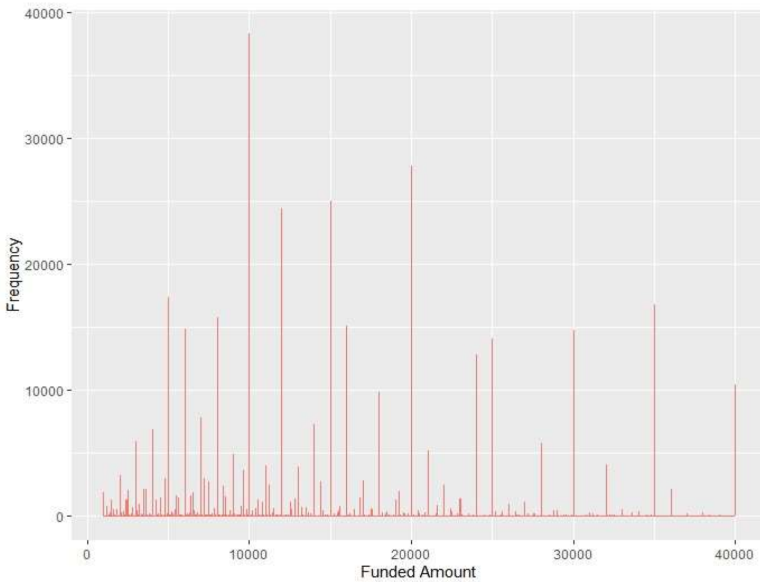
Financial institutions prefer to give loans to large, secured, and low-risk enterprises for the consideration of profitability and risk management. Therefore, the credit needs of small businesses, individuals are usually suppressed. However, small businesses and individuals sometimes require urgent cash investments for certain circumstances. Lending companies are a kind of financial institution that could quickly and comfortably solve most of these problems. For lending companies, the company's methods of loan pricing are critical to maintaining the operation and management. The motivation behind the study is to see whether or not there is a correlation between the interest rate and other predictors, which predictors are the essential variables in the construction of the predictive model, and how these critical variables affect the interest rates. The goal is to construct a simple predictive model, which determines the clients' interest rate on the loans through various information provided by clients. The following is the summary of the interest rate on the loans



Interest Rate
Min. : 5.31
1st Qu. : 9.44
Median : 12.59
Mean: 13.09
3rd Qu. : 16.02
Max.: 30.99

Figure: Interest Rate's Distribution and Summary

Note that the lowest interest rate on the loans is 5.31%, and the highest interest rate on the loans is 30.99%. On average, the interest rate on the loans is 12.59 %. The median interest rate on the loans is 13.09%. From the figure of the distribution of the interest rate, we find that the interest rate on the loans is mainly between 5% to 20%. The following is the summary of the total amount committed to the loan.



Funded Amount
Min: 1000
1st Qu: 8000
Median: 13425
Mean: 15487
3rd Qu: 20850
Max: 40000

Figure: Total Amount Committed to The Loan's Distribution and Summary

The lowest total amount committed to the loan is \$1000, and the highest total amount committed to the loan is \$40000. On average, the total amount committed to the loan is \$13425. The median value of the total amount committed to the loan is \$15487. From the distribution of the total amount committed to the loan, we see that the majority of clients' lending needs are between \$10000 to \$20000.

The analysis and model were carried out in RStudio version 3.6.2.

Methodology

1. Type of Research

Quantitative approaches focus on the analysis of variables by leveraging numerical values to bring meaning to the variables. (Leedy & Ormrod, 2013). This research seeks to use numerical values to find the correlation between the interest rate on loans and associated predictors.

2. Data Collection and Clean-Up

We collected financial data from LendingClub, which is an American peer to peer lending company. The original database had 97 predictors and 1048575 rows of data. Then, we began the process of data cleaning. We took out of the predictors that were not associated with the interest rate, such as the amount of received principal and received late fees. We also took out of the predictors that missed more than 100000 rows of data, such as the number of open trades in the last 6 months and the number of personal finance inquiries. After we took out of uncorrelated predictors and predictors with a large amount of missing values, the database had 66 predictors remained. We cleaned the missing value in these 66 remaining predictors, which left 66 predictors and 454653 rows of data in the database.

3. Analysis

We used 8 different statistical methods to develop a predictive model that could predict the interest rate on loans with attention to accuracy. These 8 statistical methods were the multiple linear regression, ridge regression, the lasso, principal components regression, forward stepwise selection, backward stepwise selection, regression trees, and boosting. We divided the data into training and test data. The training data was 70 percent of the data in the database, which has a sample of 318257 people's financial data. The test data was the remaining 30 percent of the data, which has a sample of 136396 people's financial data. We applied these 8 statistical methods to study the association between the interest rate and remaining predictors on R statistical software. We evaluated the performance of these models by comparing the difference between the predicted interest rate and the actual interest rate on the test data. The evaluation of the predictive models' performance was done by value of the test root-mean-square error (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Statistical Methods

1. Linear Regression

1.1 Multiple Linear Regression's Model:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{66} X_{66} + \varepsilon$$

Where X_i is the ith predictor, β_i is the association between that variable and the response, and β_0 is the intercept term

1.2 Test Data RMSE

We fitted the multiple linear regression model by these 66 predictors. 56 predictors were statistically significant. The Adjusted R^2 value was 0.454. The test RMSE value of the multiple linear regression model was 3.842987.

2. Subset Selection

2.1 Forward Stepwise Selection

Forward stepwise selection is a method that creates the null model with no predictors, and then augments one predictor to the model until all significant predictors are in the model.

Algorithm 2.1 Forward Stepwise Selection

---

1. Let M_0 represents the null model with no predictors.
2. For $k = 0, 1, \ldots, p-1$:

(a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.

(b) Choose the best model among these $p-k$ models, and call it M_{k+1}.

3. Select a single best model from among $M_0, M_1, \cdots, M_n$ using cross-validated

---

prediction error, C_P , BIC or adjusted R^2.

2.2 Backward Stepwise Selection

Backward stepwise selection is a method that begins with the full least squares model with all predictors, and then remove one the least useful predictor out of the model until only significant predictors are in the model.

Algorithm 2.2 Backward Stepwise Selection

1. Let M_p represents the full model, which contains all p predictors.
2. For $k = p$, $p-1$, … , 1 :

(a) Consider all k models that contain all but one of the predictors in M_k , for a total of $k-1$ predictors.

(b) Choose the best model among these k models, and call it M_{k-1} .

3. Select a single best model from among M_0 , M_1 , ⋯ , M_n using cross-validated prediction error, C_P , BIC or adjusted R^2.

2.3 Choosing the Optimal Number of Predictors

We used the BIC value and the adjusted R^2 value to choose a model with the optimal number of predictors.

2.3.1 Bayesian Information Criterion

Bayesian information criterion (BIC) derived from a Bayesian point of view. BIC tends to have a smaller value when the model tends to have a lower test error. Thus, we generally choose a model with a small BIC value (Gareth, Daniela, Trevor & Robert, 2017, 212).

$$BIC = \frac{1}{n\,\hat{\sigma}^2}\,(RSS + \log_n d\hat{\sigma}^2)$$

Where $\hat{\sigma}^2$ is an estimate of the variance of the error, n is the number of observations, d is the number of predictors, and RSS is the *residual sum of squares*.

2.3.2 Adjusted R^2

Theoretically, a model with the largest adjusted R^2 value only has correct variables and no noise variables. A large adjusted R^2 value indicates the model has a small test error (Gareth, Daniela, Trevor & Robert, 2017, 212)

$$\text{Adjusted } R^2 = 1 - \frac{RSS/\,(n-d-1)}{TSS/(n-1)}$$

Where n is the number of observations, d is the number of predictors, TSS is the *total sum of squares*, and RSS is the *residual sum of squares*.

2.4 Results and Discussion (Forward Stepwise Selection)

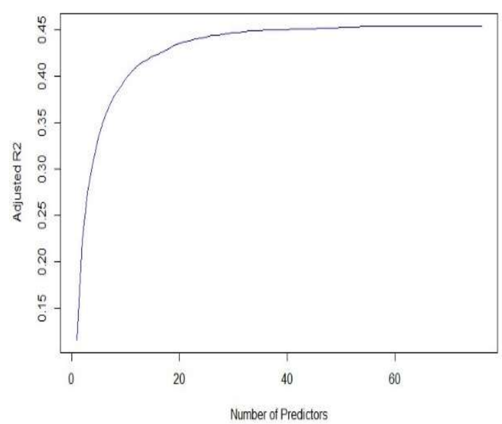


Figure: Prediction Error: Adjusted R^2

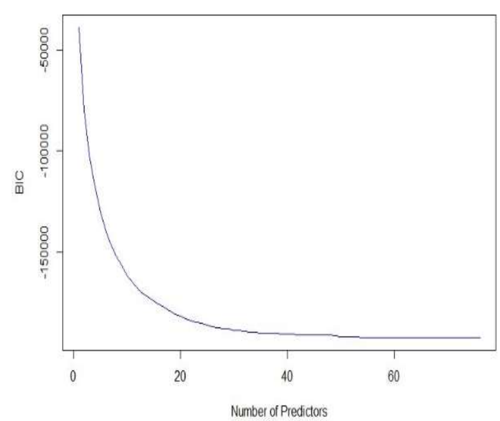


Figure: Prediction Error: BIC

From the above two figures, we see that the best model is to choose a model with 26 predictors. We also built a model with a different number of predictors. For example, we found that using forward stepwise selection, the best two predictor model contained: the term of the loan, the last FICO scores, and the best four predictor model contained: the term of the loan, the last FICO scores, the total open-to-buy budget on revolving bankcards, and the initial listing status of the loan recorded as a whole or fractional loan.

2.5 Results and Discussion (Backward Stepwise Selection)

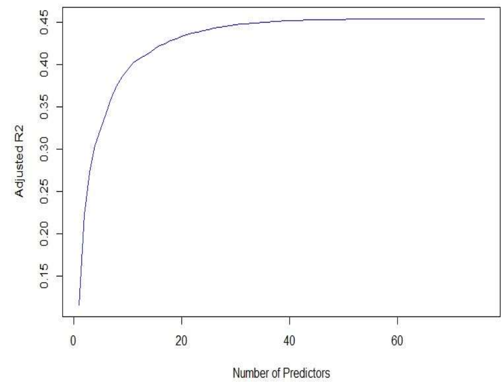


Figure: Prediction Error: Adjusted R^2

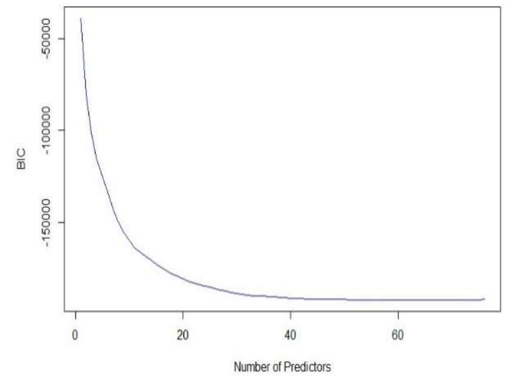


Figure: Prediction Error: BIC

From the above two figures, we see that the best model is to choose a model with 30 predictors. We also built a model with a different number of predictors. For example, we found that using backward stepwise selection, the best two predictor model contained: the term of the loan, the last FICO scores, and the best four predictor model contained: the term of the loan, the last FICO scores, the balance to the credit limit on all trades, and the initial listing status of the loan recorded as a whole or fractional loan.

2.6 Test Data RMSE

The test RMSE value of the forward stepwise selection was 3.875595, and the test RMSE value of the backward stepwise selection was 3.865191. Thus, the backward stepwise selection method, which contained 30 predictors in the model, was slightly better the forward stepwise selection method in the predictive accuracy. However, there was no significant difference in accuracy between the predictive models generated by backward & forward stepwise selection methods.

Although the predictors that were chosen by the backward selection method and the forward selection method were slightly different, the accuracy of the predictive models generated by these two statistical methods was similar.

3. Shrinkage Methods

3.1 Ridge Regression

3.1.1 Ridge Regression's Model

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 = RSS + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

Where λ is the *tuning parameter*, β_j is the regression coefficient

Ridge regression aims to make the regression coefficient estimates fit the data well by reducing the RSS value. The second term is called a shrinkage penalty. The tuning parameter λ is used to control the relative impact of these two terms on the regression coefficient estimates. When $\lambda = 0$, the impact of shrinkage penalty does not exist, and ridge regression produces the least squares estimates. When $\lambda \to \infty$, the effect of the shrinkage penalty increases, and the ridge regression coefficient estimates approach zero (Gareth, Daniela, Trevor & Robert, 2017, 215).

3.2 Lasso

3.2.1 Lasso's Models

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where λ is the *tuning parameter*, β_j is the regression coefficient

Ridge regression has a distinct disadvantage. Since ridge regression contains all predictors in the model, the shrinkage penalty shrinks all the regression coefficients towards zero, but it does not set any of these regression coefficients to zero. The increase of the value of the shrinkage penalty decreases the magnitudes of the regression coefficients, but the exclusion of any useless predictors is not possible in ridge regression. The lasso is an alternative solution to ridge regression. The lasso and ridge regression are very similar that the lasso also uses the shrinkage penalty term to shrink the coefficient estimates towards zero. However, the term of shrinkage penalty is replaced from β_j^2 to $|\beta_j|$. When the tuning parameter λ is infinitely large, the coefficient estimates would be equal to zero. Therefore, the lasso could make a variable selection. (Gareth, Daniela, Trevor & Robert, 2017, 219)

3.3 Choosing the Optimal λ

Ridge Regression and the lasso produce a different set of coefficient estimates for different values of λ. Therefore, to choose the best set of coefficient estimates, we used the ten-fold cross-validation method to choose the optimal values of λ. We created a grid of 1000 possible values of λ ranging from $\lambda = 10^{-2}$ to $\lambda = 10^{10}$ in R statistical software. We found the best tuning parameter λ by using the ten-fold cross-validation function in R statistical software. We saw that the value of tuning parameter λ of ridge regression that results in the smallest cross-validation error was 0.5934529 , and the value of tuning parameter λ of the lasso that results in the smallest cross-validation error was 0.03904524.

3.4 Test Data RMSE

We refitted the ridge regression model using $\lambda = 0.5934529$, and the test RMSE of ridge regression was 3.862053. We refitted the lasso model using $\lambda = 0.03904524$, and the test RMSE of the lasso was 3.863146. The test RMSE difference of the predictive models generated by ridge regression and the lasso was not significant;

however, the lasso had a considerable advantage over ridge regression that the number of predictors in the predictive model was reduced from 66 predictors to 47 predictors. Therefore, the lasso was a better method than ridge regression when constructing the model to predict the interest rate on the loans.

4. Dimension Reduction Methods

4.1 Principal Components Regression

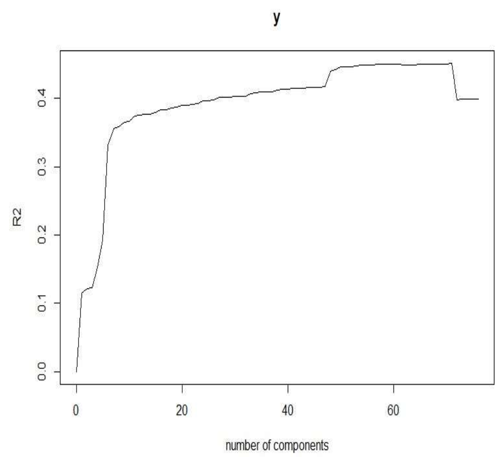
Principal components regression is a method, which aims to reduce the dimension of a data matrix. Principal components regression builds M principal components $Z_1, Z_2, \cdots, Z_M$, and these components are used as predictors in a linear regression model. The idea of the principal component regression is only to use a small number of principal components to explain most of the variability in the data, and the relationship to the response (Gareth, Daniela, Trevor & Robert, 2017, 233).

Since the raw data in different predictors spanned different range, and the high-variance variables could have a significant impact on the objective functions, we scaled these data to make the objective functions work correctly.

4.2 Choosing the Optimal Number of Principal Components

We computed the ten-fold cross-validation error for each possible value of the number of principal components. We chose the number of principal components that results in a small cross-validation error.

4.3 Results and Discussion



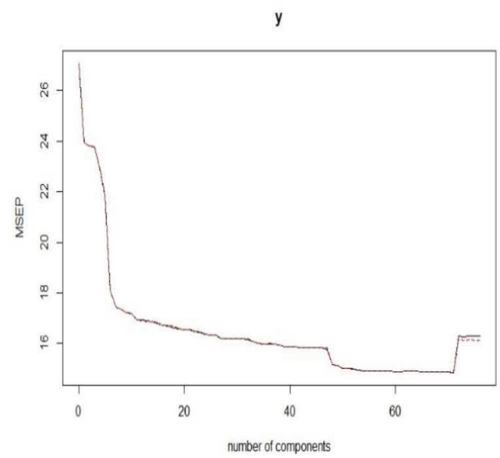


Figure: Adjusted R^2 **Figure:** Cross-Validation MSE

From the above two figures, we find that the smallest cross-validation error occurs when we use 71 principal components in the model. The cross-validation error of 71 principal components in the model is slightly less than using 76 principal components; however, there is almost no dimension reduction occurs. We see from these two figures that the model containing 10 principal components or 76 principal components have roughly the same cross-validation error, which shows that a model using 10 principal components is sufficient. The following is the percentage of variance explained in the predictors and the response.

```
TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
X    63.87    69.56    73.35    76.48    79.14    81.12    82.62    84.01    85.12     86.09     86.92     87.70     88.43
y    11.50    12.10    12.22    15.35    19.29    33.13    35.64    35.88    36.45     36.58     37.46     37.56     37.66
   14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
X     89.14     89.77     90.36     90.91     91.45     91.97     92.46     92.93     93.38     93.81     94.23     94.64
y     37.67     37.87     38.26     38.28     38.57     38.74     38.96     38.96     39.21     39.22     39.62     39.63
   26 comps  27 comps  28 comps  29 comps  30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
X     95.04     95.39     95.73     96.03     96.32     96.59     96.85     97.11     97.36     97.60     97.82     98.02
y     39.70     40.23     40.23     40.25     40.27     40.28     40.29     40.68     40.89     40.96     40.96     41.01
   38 comps  39 comps  40 comps  41 comps  42 comps  43 comps  44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
X     98.20     98.38     98.56     98.71     98.83     98.96     99.07     99.17     99.27     99.35     99.43     99.49
y     41.18     41.36     41.40     41.47     41.50     41.50     41.58     41.59     41.59     41.60     43.99     44.19
   50 comps  51 comps  52 comps  53 comps  54 comps  55 comps  56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
X     99.55     99.60     99.65     99.69     99.73     99.76     99.80     99.83     99.85     99.88     99.90     99.92
y     44.61     44.61     44.61     44.84     44.97     44.97     44.98     45.00     45.01     45.01     45.08     45.12
   62 comps  63 comps  64 comps  65 comps  66 comps  67 comps  68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
X     99.93     99.94     99.95     99.96     99.97     99.98     99.98     99.99     99.99     99.99    100.00    100.00
y     45.13     45.19     45.20     45.24     45.25     45.25     45.25     45.28     45.29     45.30     45.34     45.34
   74 comps  75 comps  76 comps
X     100.0     100.0    100.00
y      45.4      45.4     45.41
```

Figure: Percentage of Variance Explained in The Predictors and The Response

From the above figure, we could see that when only 1 principal component is used in the model, the predictors capture 63.87% of the information. 10 principal components could capture 86.19% of the information. If we use 72 principal components, all information is captured.

Therefore, we performed principal components regression with 10 principal components and evaluated its performance by test data.

4.4 Test Data RMSE

We fitted the principal components regression model with 10 principal components, and the test RMSE of the principal components regression model was 4.142822. However, the predictive model was challenging to interpret because the model did not generate any coefficient estimates and select predictors.

5. Tree-Based Methods

5.1 Regression Trees

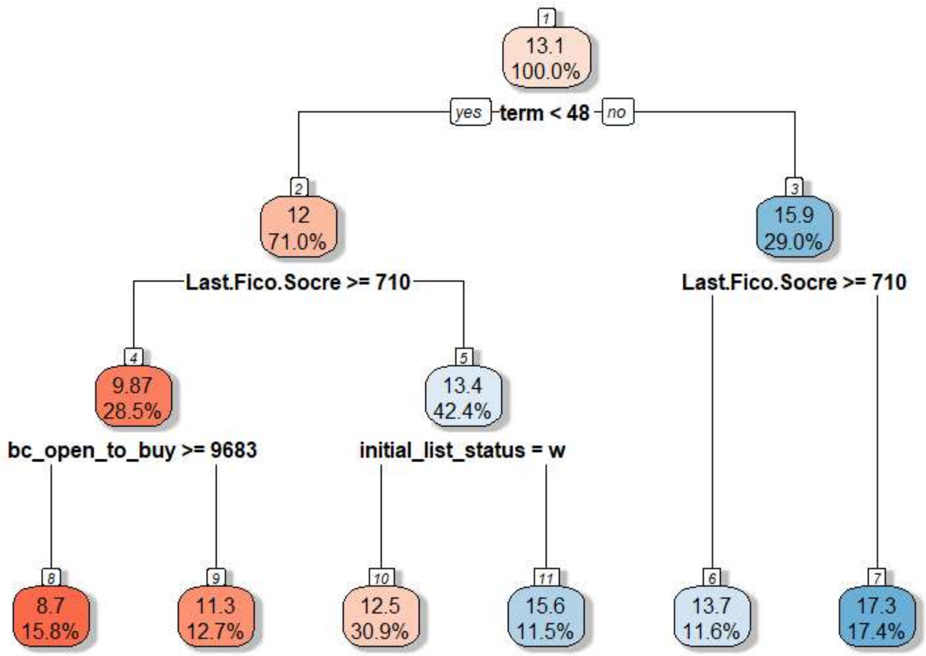


Figure: Regression Trees

Note that only 4 predictors have been used to build the regression tree. These 4 predictors are the term of the loans, the last FICO scores, the initial listing status of the loan recorded as a whole or fractional loan, and the total open-to-buy budget on revolving bankcards. The top split assigns observation with the term of the loans less than 48 months to the left branch and the term of the loans more than 48 months to the right branch. The last FICO scores further subdivide both groups. The group of the term of the loans less than 48 months and the last FICO scores more than 710 is further subdivided by the initial listing status of the loan recorded as a whole or fractional loan, and the total open-to-buy budget on revolving bankcards. The tree segments the loans into six regions of predictor space. The first region of predictor space is the loans with the term less than 48 months, the last FICO scores more than 710, and the total open-to-buy budget on revolving bankcards more than 9683. The second region of predictor space is the loans with the term less than 48 months, the last FICO scores more than 710, and the total open-to-buy budget on revolving bankcards less than 9683. The third region of predictor space is the loans with the term less than 48 months, the last FICO scores less than 710, and the initial listing status of the loan recorded as the whole loan. The fourth region of predictor space is the loans with the term less than 48 months, the last FICO scores less than 710, and the initial listing status of the loan recorded as a fraction loan. The fifth region of predictor space is the loans with the term more than 48 months, the last FICO scores more than 710. The sixth region of predictor space is the loans with the term more than 48 months, the last FICO scores less than 710. The mean predicted interest rate for these six groups are 8.7%, 11.3%, 12.5%, 15.6%, 13.7%, and 17.3%, respectively.

5.1.1 Test Data RMSE

The regression tree indicated that the lower value of the term and higher value of the last FICO scores corresponded to a lower interest rate. For example, the regression tree predicted that when the loan had term more than 48 months and the last FICO scores less than 710, the mean response value of the interest rate on the loans would be 17.3% ; however, if the FICO scores were more than 710, the mean response value of the interest rate on the loans would be reduced to 13.7%.

Regression trees are easy to interpret to people. However, the accuracy of the prediction was not as good as other regression approaches. The test RMSE of the predictive model generated by the regression tree was 4.436356.

5.2 Boosting

Boosting involves creating multiple copies of the training data set using the modified version of the original data set, fitting a separate decision tree to each copy, and then combining all of the trees to create a single predictive model. The trees are grown sequentially: using information from previously grown trees to grow trees (Gareth, Daniela, Trevor & Robert, 2017, 321).

We set the number of trees was 5000 trees, and the depth of each tree was 2.

5.2.1 Results and Discussion

We found that the term of the loans, the last FICO scores, the total open-to-buy budget on revolving bankcards, and the initial listing status of the loan recorded as a whole or fractional loan were the most important predictors.

The following is the partial dependence plots for these four variables.

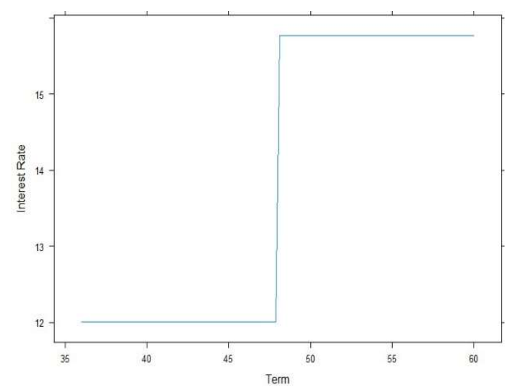


Figure: Term of the Loans

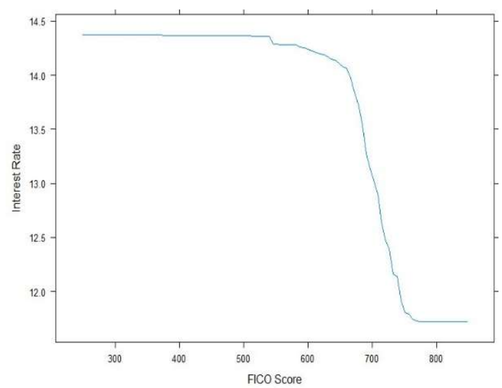
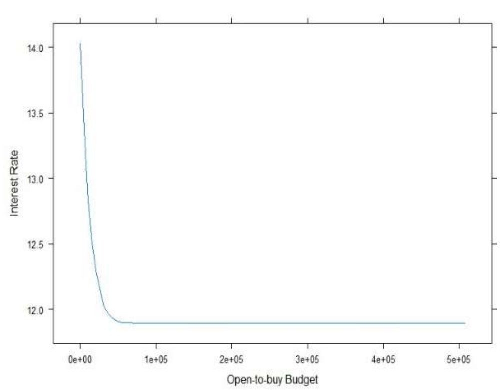


Figure: Last FICO Scores



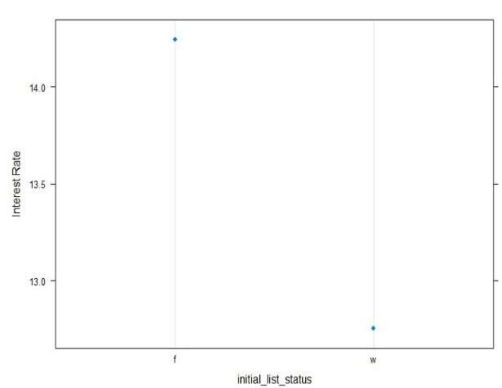


Figure: Total Open-To-Buy Budget **Figure:** Initial Listing Status

Note that the interest rate on the loans raises with an increasing term of the loans, and the interest rate on the loans decreases as the last FICO Scores and the total open-to-buy budget increase. The interest rate goes up if the initial listing status of the loan recorded as a fraction loan.

5.2.2 Test Data RMSE

The test RMSE of the predictive model generated by the boosting method was 3.60744.

Conclusion

When the financial institutions decided the interest rate on the loans to clients, including more clients' information could improve the accuracy of the prediction. In the above statistical methods, boosting was the best statistical method to construct a model to predict the interest rate. Other predictive models, except the predictive models generated by the regression trees and principal components regression, had no significant difference in predictive accuracy. The forward stepwise selection was a better method than the multiple linear regression or the lasso in constructing a model to predict the interest rate because the predictive model generated by the forward stepwise selection was more straightforward than the predictive model generated by the multiple linear regression or the lasso.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2017). An introduction to statistical learning : with applications in R. New York :Springer,

Leedy, P. &,Ormrod, J. (2013). *Practical Research: Planning and Design*. New Jersey: Pearson Education.

Appendix

1. Linear Regression

1.1 Multiple Linear Regression's Model:

```
Call:
lm(formula = int_rate ~ ., data = database.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-36.955  -2.551  -0.564   1.892  37.127 

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 4.716e+00  1.864e+00    2.530 0.011419 *  
funded_amnt                 4.623e-05  9.174e-07   50.387  < 2e-16 ***
term                        1.709e-01  7.023e-04  243.267  < 2e-16 ***
emp_length                  3.240e-03  1.944e-03    1.667 0.095530 .  
home_ownershipOWN           5.111e-01  2.403e-02   21.273  < 2e-16 ***
home_ownershipRENT          3.895e-01  1.917e-02   20.325  < 2e-16 ***
annual_inc                 -2.560e-06  1.079e-07  -23.721  < 2e-16 ***
dti                         3.253e-02  5.744e-04   56.625  < 2e-16 ***
delinq_2yrs                 3.956e-01  1.163e-02   34.013  < 2e-16 ***
inq_last_6mths              5.484e-01  1.084e-02   50.599  < 2e-16 ***
open_acc                   -1.298e-01  2.781e-02   -4.669 3.02e-06 ***
pub_rec                     2.100e-03  3.840e-02    0.055 0.956389    
revol_bal                   1.365e-05  1.491e-06    9.156  < 2e-16 ***
revol_util                  1.775e-02  7.640e-04   23.237  < 2e-16 ***
total_acc                   4.852e-02  1.158e-02    4.190 2.79e-05 ***
initial_list_statusw       -1.966e+00  1.810e-02 -108.641  < 2e-16 ***
tot_coll_amt                2.826e-06  7.877e-07    3.588 0.000333 ***
tot_cur_bal                -5.022e-07  1.647e-07   -3.049 0.002297 ** 
open_acc_6m                 4.065e-02  9.173e-03    4.432 9.36e-06 ***
open_act_il                 8.008e-02  1.257e-02    6.369 1.90e-10 ***
open_il_12m                 5.000e-01  2.071e-02   24.146  < 2e-16 ***
mths_since_rcnt_il         -3.203e-03  5.094e-04   -6.288 3.22e-10 ***
total_bal_il               -4.385e-06  1.133e-06   -3.871 0.000108 ***
il_util                     6.873e-05  5.269e-04    0.130 0.896220    
open_rv_12m                 1.331e-01  1.998e-02    6.658 2.78e-11 ***
max_bal_bc                 -4.602e-05  2.388e-06  -19.275  < 2e-16 ***
all_util                    2.421e-02  8.352e-04   28.987  < 2e-16 ***
total_rev_hi_lim           -2.056e-05  9.460e-07  -21.730  < 2e-16 ***
inq_fi                      1.787e-01  5.370e-03   33.283  < 2e-16 ***
total_cu_tl                -3.706e-02  2.763e-03  -13.413  < 2e-16 ***
inq_last_12m                2.163e-02  4.082e-03    5.299 1.17e-07 ***
acc_open_past_24mths        1.813e-01  3.732e-03   48.568  < 2e-16 ***
avg_cur_bal                -1.556e-05  1.011e-06  -15.398  < 2e-16 ***
bc_open_to_buy             -2.596e-05  1.637e-06  -15.860  < 2e-16 ***
bc_util                     1.191e-03  6.827e-04    1.745 0.081030 .  
mo_sin_old_il_acct         -3.727e-03  1.492e-04  -24.979  < 2e-16 ***
mo_sin_old_rev_tl_op       -2.419e-03  1.371e-04  -17.652  < 2e-16 ***
mo_sin_rcnt_rev_tl_op      -7.327e-04  6.511e-04   -1.125 0.260486    
mo_sin_rcnt_tl             -5.061e-03  1.611e-03   -3.142 0.001676 ** 
mort_acc                   -1.950e-01  1.223e-02  -15.944  < 2e-16 ***
mths_since_recent_bc       -6.315e-03  3.085e-04  -20.470  < 2e-16 ***
mths_since_recent_inq      -1.421e-02  1.520e-03   -9.353  < 2e-16 ***
num_accts_ever_120_pd       3.804e-02  6.337e-03    6.003 1.94e-09 ***
num_actv_bc_tl              1.064e-02  8.818e-03    1.206 0.227752    
num_actv_rev_tl            -6.114e-02  1.177e-02   -5.193 2.07e-07 ***
num_bc_sats                -3.520e-02  6.816e-03   -5.165 2.41e-07 ***
num_bc_tl                  -1.152e-02  4.515e-03   -2.551 0.010751 *  
num_il_tl                  -7.757e-02  1.163e-02   -6.671 2.54e-11 ***
num_op_rev_tl               2.230e-01  1.311e-02   17.011  < 2e-16 ***
num_rev_accts              -8.739e-02  1.186e-02   -7.372 1.69e-13 ***
num_rev_tl_bal_gt_0         4.953e-02  1.231e-02    4.024 5.73e-05 ***
num_sats                   -3.330e-02  2.821e-02   -1.180 0.237842    
num_tl_90g_dpd_24m         -2.411e-01  1.858e-02  -12.977  < 2e-16 ***
num_tl_op_past_12m         -6.474e-04  1.980e-02   -0.033 0.973912    
pct_tl_nvr_dlq             -1.945e-02  1.092e-03  -17.806  < 2e-16 ***
percent_bc_gt_75            1.854e-02  3.655e-04   50.712  < 2e-16 ***
pub_rec_bankruptcies        2.512e-01  4.261e-02    5.896 3.73e-09 ***
tax_liens                   1.808e-02  4.327e-02    0.418 0.676100    
tot_hi_cred_lim             1.436e-07  1.010e-07    1.421 0.155282    
total_bal_ex_mort           1.292e-05  9.577e-07   13.487  < 2e-16 ***
total_bc_limit              1.899e-05  1.360e-06   13.969  < 2e-16 ***
total_il_high_credit_limit -9.621e-06  6.110e-07  -15.748  < 2e-16 ***
TITLECAR FINANCING         -1.722e+00  9.526e-02  -18.073  < 2e-16 ***
TITLECREDIT CARD REFINANCING -3.054e+00  6.919e-02  -44.141  < 2e-16 ***
TITLEDEBT CONSOLIDATION    -1.813e+00  6.816e-02  -26.596  < 2e-16 ***
TITLEGREEN LOAN            -2.367e-01  2.815e-01   -0.841 0.400473    
TITLEHOME BUYING           -2.634e-01  1.084e-01   -2.429 0.015136 *  
TITLEHOME IMPROVEMENT      -1.762e+00  7.211e-02  -24.430  < 2e-16 ***
TITLEMAJOR PURCHASE        -1.456e+00  8.059e-02  -18.062  < 2e-16 ***
TITLEMEDICAL EXPENSES      -1.098e+00  9.020e-02  -12.177  < 2e-16 ***
TITLEMOVING AND RELOCATION -7.184e-01  1.042e-01   -6.894 5.43e-12 ***
TITLEOTHER                 -6.956e-01  7.262e-02   -9.579  < 2e-16 ***
TITLEVACATION              -1.290e+00  1.022e-01  -12.621  < 2e-16 ***
Last.Fico.Socre            -8.743e-03  9.325e-05  -93.760  < 2e-16 ***
APPLICATION_TYPEJOINT APP   1.661e-01  2.663e-02    6.237 4.45e-10 ***
issue_date.month.           7.018e-02  2.577e-03   27.229  < 2e-16 ***
earliest_cr_line.year.      4.275e-03  9.135e-04    4.680 2.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.845 on 318180 degrees of freedom
Multiple R-squared:  0.4541,	Adjusted R-squared:  0.454 
F-statistic:  3483 on 76 and 318180 DF,  p-value: < 2.2e-16
```

2. Subset Selection

2.1 Forward Stepwise Selection

```
> coef(regfit.forward,26)
           (Intercept)            funded_amnt                   term                          dti
          1.221217e+01           3.948957e-05           1.707784e-01                 3.464055e-02
            delinq_2yrs        inq_last_6mths             revol_util                    total_acc
          2.890977e-01           6.258828e-01           1.441841e-02                -2.766092e-02
   initial_list_statusw           open_act_il            open_il_12m                     all_util
         -2.002582e+00          -8.114313e-02           3.445685e-01                 2.949038e-02
                 inq_fi  acc_open_past_24mths          bc_open_to_buy            mo_sin_old_il_acct
          1.987139e-01           1.883561e-01          -2.497067e-05                -4.470868e-03
   mo_sin_old_rev_tl_op              mort_acc    mths_since_recent_bc            num_tl_op_past_12m
         -2.840267e-03          -1.964820e-01          -7.906672e-03                 1.641305e-01
           pct_tl_nvr_dlq      percent_bc_gt_75         tot_hi_cred_lim TITLECREDIT CARD REFINANCING
         -2.253055e-02           1.924324e-02          -2.272428e-06                -1.304243e+00
            TITLEOTHER        Last.Fico.Socre      issue_date.month.
          1.010277e+00          -9.371349e-03           7.043924e-02
```

2.2 Backward Stepwise Selection

```
> coef(regfit.backward,30)
           (Intercept)            funded_amnt                         term                  annual_inc
          1.335547e+01           4.634842e-05                 1.705416e-01               -2.553657e-06
                    dti            delinq_2yrs               inq_last_6mths                    open_acc
          3.178977e-02           2.794495e-01                 6.260184e-01               -1.077049e-01
              revol_bal   initial_list_statusw                  open_il_12m                 open_rv_12m
          3.529432e-05          -1.992822e+00                 5.280588e-01                1.689021e-01
               all_util      total_rev_hi_lim                       inq_fi        acc_open_past_24mths
          3.460193e-02          -3.098939e-05                 1.934852e-01                1.735231e-01
            avg_cur_bal      mo_sin_old_il_acct         mo_sin_old_rev_tl_op                   mort_acc
         -1.939357e-05          -4.025658e-03                -2.803134e-03               -1.932319e-01
    mths_since_recent_bc              num_il_tl                num_op_rev_tl                num_rev_accts
         -6.601871e-03          -3.164732e-02                 1.638038e-01               -4.403764e-02
         pct_tl_nvr_dlq      percent_bc_gt_75 TITLECREDIT CARD REFINANCING    TITLEDEBT CONSOLIDATION
         -2.390482e-02           2.381914e-02                -2.127936e+00               -9.052365e-01
 TITLEHOME IMPROVEMENT        Last.Fico.Socre            issue_date.month.
         -9.193011e-01          -9.206802e-03                 7.200450e-02
```

4. Dimension Reduction Methods

4.1 Principal Components Regression

```
Data:   X dimension: 318257 76
        Y dimension: 318257 1
Fit method: svdpc
Number of components considered: 76

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
CV           5.204    4.896    4.879    4.876    4.788    4.675    4.255    4.175    4.167    4.148     4.144     4.115     4.114
adjCV        5.204    4.896    4.879    4.876    4.788    4.675    4.254    4.175    4.167    4.149     4.144     4.115     4.111
       13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps
CV        4.109     4.108     4.100     4.088     4.086     4.078     4.073     4.066     4.066     4.058     4.056     4.044
adjCV     4.109     4.108     4.101     4.089     4.088     4.079     4.073     4.066     4.066     4.058     4.057     4.044
       25 comps  26 comps  27 comps  28 comps  29 comps  30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps
CV        4.043     4.040     4.024     4.024     4.023     4.023     4.022     4.020     4.009     4.001     3.999         4
adjCV     4.043     4.042     4.024     4.024     4.023     4.023     4.022     4.022     4.009     4.002     3.999         4
       37 comps  38 comps  39 comps  40 comps  41 comps  42 comps  43 comps  44 comps  45 comps  46 comps  47 comps  48 comps
CV        3.997     3.992     3.986     3.985     3.982     3.981     3.981     3.978     3.978     3.978     3.970     3.895
adjCV     3.998     3.992     3.986     3.985     3.982     3.981     3.981     3.978     3.978     3.978     3.978     3.895
       49 comps  50 comps  51 comps  52 comps  53 comps  54 comps  55 comps  56 comps  57 comps  58 comps  59 comps  60 comps
CV        3.888     3.874     3.874     3.873     3.867     3.862     3.862     3.861     3.861      3.86      3.86     3.858
adjCV     3.889     3.874     3.874     3.874     3.867     3.862     3.862     3.861     3.861      3.86      3.86     3.858
       61 comps  62 comps  63 comps  64 comps  65 comps  66 comps  67 comps  68 comps  69 comps  70 comps  71 comps  72 comps
CV        3.857     3.864     3.862     3.862     3.859     3.859     3.858     3.858     3.857     3.856     3.854     4.036
adjCV     3.857     3.863     3.861     3.861     3.858     3.858     3.857     3.857     3.857     3.856     3.854     4.018
       73 comps  74 comps  75 comps  76 comps
CV        4.034     4.036     4.036     4.036
adjCV     4.016     4.017     4.017     4.017
```

5. Tree-Based Methods

5.1 Regression Trees

```
Regression tree:
tree(formula = int_rate ~ ., data = database.train)
Variables actually used in tree construction:
[1] "term"                "Last.Fico.Socre"     "initial_list_status" "bc_open_to_buy"
Number of terminal nodes:  6
Residual mean deviance:  19.71 = 6274000 / 318300
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11.9900  -3.1080  -0.8536   0.0000   2.3460  22.2900
```

5.2 Boosting

```
                                                 var      rel.inf
term                                            term 2.219471e+01
Last.Fico.Socre                      Last.Fico.Socre 1.973408e+01
bc_open_to_buy                        bc_open_to_buy 1.080780e+01
initial_list_status              initial_list_status 5.593068e+00
dti                                              dti 5.354633e+00
all_util                                    all_util 4.405643e+00
TITLE                                          TITLE 3.057619e+00
funded_amnt                              funded_amnt 2.846829e+00
percent_bc_gt_75                    percent_bc_gt_75 2.634504e+00
acc_open_past_24mths            acc_open_past_24mths 2.492997e+00
num_tl_op_past_12m                num_tl_op_past_12m 2.466331e+00
inq_last_6mths                        inq_last_6mths 2.272387e+00
tot_hi_cred_lim                      tot_hi_cred_lim 2.058492e+00
issue_date.month.                  issue_date.month. 1.640871e+00
annual_inc                                annual_inc 1.098575e+00
total_acc                                  total_acc 1.041214e+00
mort_acc                                    mort_acc 8.188896e-01
mths_since_rcnt_il                mths_since_rcnt_il 7.809167e-01
inq_fi                                        inq_fi 7.435744e-01
mths_since_recent_bc            mths_since_recent_bc 7.028174e-01
mo_sin_old_rev_tl_op            mo_sin_old_rev_tl_op 6.976634e-01
inq_last_12m                            inq_last_12m 5.636101e-01
delinq_2yrs                              delinq_2yrs 5.606494e-01
mo_sin_old_il_acct                mo_sin_old_il_acct 5.211625e-01
bc_util                                      bc_util 5.121296e-01
total_rev_hi_lim                    total_rev_hi_lim 4.968878e-01
open_il_12m                              open_il_12m 4.827041e-01
revol_util                                revol_util 4.094233e-01
APPLICATION_TYPE                    APPLICATION_TYPE 3.805115e-01
total_il_high_credit_limit total_il_high_credit_limit 3.640017e-01
mths_since_recent_inq          mths_since_recent_inq 3.575889e-01
earliest_cr_line.year.        earliest_cr_line.year. 3.511118e-01
open_act_il                              open_act_il 2.516240e-01
num_il_tl                                  num_il_tl 2.387850e-01
il_util                                      il_util 1.613301e-01
pct_tl_nvr_dlq                        pct_tl_nvr_dlq 1.470125e-01
num_bc_sats                              num_bc_sats 1.445434e-01
mo_sin_rcnt_tl                        mo_sin_rcnt_tl 1.004851e-01
home_ownership                        home_ownership 9.901650e-02
total_cu_tl                              total_cu_tl 9.023456e-02
num_accts_ever_120_pd          num_accts_ever_120_pd 8.975925e-02
num_actv_bc_tl                        num_actv_bc_tl 8.620092e-02
total_bc_limit                        total_bc_limit 4.448514e-02
open_acc_6m                              open_acc_6m 4.032019e-02
pub_rec                                      pub_rec 2.421221e-02
tot_cur_bal                              tot_cur_bal 1.080108e-02
max_bal_bc                                max_bal_bc 9.073993e-03
total_bal_il                            total_bal_il 8.953200e-03
revol_bal                                  revol_bal 5.102824e-03
pub_rec_bankruptcies            pub_rec_bankruptcies 2.735399e-03
total_bal_ex_mort                  total_bal_ex_mort 1.295736e-03
open_rv_12m                              open_rv_12m 6.337975e-04
emp_length                                emp_length 0.000000e+00
open_acc                                    open_acc 0.000000e+00
tot_coll_amt                            tot_coll_amt 0.000000e+00
avg_cur_bal                              avg_cur_bal 0.000000e+00
mo_sin_rcnt_rev_tl_op          mo_sin_rcnt_rev_tl_op 0.000000e+00
num_actv_rev_tl                      num_actv_rev_tl 0.000000e+00
num_bc_tl                                  num_bc_tl 0.000000e+00
num_op_rev_tl                          num_op_rev_tl 0.000000e+00
num_rev_accts                          num_rev_accts 0.000000e+00
num_rev_tl_bal_gt_0              num_rev_tl_bal_gt_0 0.000000e+00
num_sats                                    num_sats 0.000000e+00
num_tl_90g_dpd_24m                num_tl_90g_dpd_24m 0.000000e+00
tax_liens                                  tax_liens 0.000000e+00
```