Predicting loan payback

1. Introduction

Whether borrowers will fully pay back loan is one of the most important information lenders are eager to know, for they lend money to borrowers in order to make profit from the interest on a loan. However, not all borrowers are able to fully pay off their loan, and in this case, lenders will incur a loss, not to mention gain interest from the loan. Thus, predicting who will pay back is a necessary and valuable practice.

To reduce potential risk, lenders could use borrower's personal information to predict the likelihood of fully payback. The relevant information may include borrowers' demographical information, credit history and something related to specific loan type. Taking all related information into consideration, I proposed an initial hypothesis that borrower's FICO credit score, annual income of borrowers, monthly installment and purpose of loan will affect whether a borrower could fully pay the money back. To be specific:

Hypothesis 1: Borrowers with lower FICO will be unlikely to fully pay back the loan.
Hypothesis 2: Borrowers with higher monthly installment will be unlikely to fully pay back loan.
Hypothesis 3: Borrowers with lower annual income will be unlikely to fully pay back the loan.
Hypothesis 4: Borrowers' purpose of loan has a significant effect on whether loan is paid in full.

Since investors would like to avoid borrowers who may fail to fully payback the loan, thus I would like to pay more attention to people who are less likely to fully payback, and try to make the prediction of this class more accurate than the fully pay-off class.

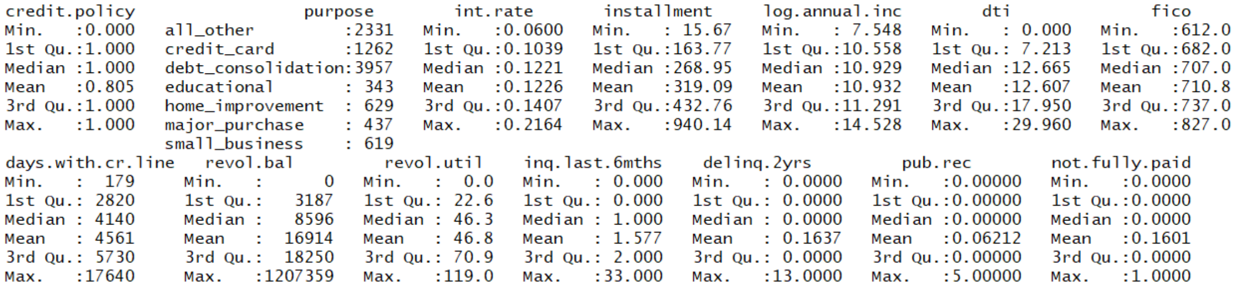
2. Data description

The loan data is available on Lendingclub website, which offers loan data for all loans over a long time period, and it records information about borrower's payment history and loan detail, making it available to generate prediction and assess the result. And I get a cleaner dataset from Kaggle dataset with removal of all missing values, and the dataset covers relevant information for about 10,000 borrowers, and cover the data from 2007 to 2010.

DV	**not.full.paid:** 1 if that the loan is not fully paid, 0 otherwise	
IV	**credit.policy** (1 if the customer meets the credit criteria of LendingClub.com, and 0 otherwise)	**inq.last.6mths**: The borrower's number of inquiries by creditors in the last 6 months.
	purpose: The purpose of the loan	**installment:** The monthly installments
	int.rate: The interest rate of the loan	**dti:** The debt-to-income ratio of the borrower
	fico: The FICO credit score of the borrower	**revol.bal**: The borrower's revolving balance
	pub.rec: The number of derogatory public records	**revol.util**: The revolving line utilization rate
	log.annual.inc: The natural log of the self-reported annual income of the borrower.	**days_with_cr_line:** The number of days the borrower has had a credit line.
	delinq_2yrs: The number of times borrower had been 30+ days past due on a payment in the past 2 years	

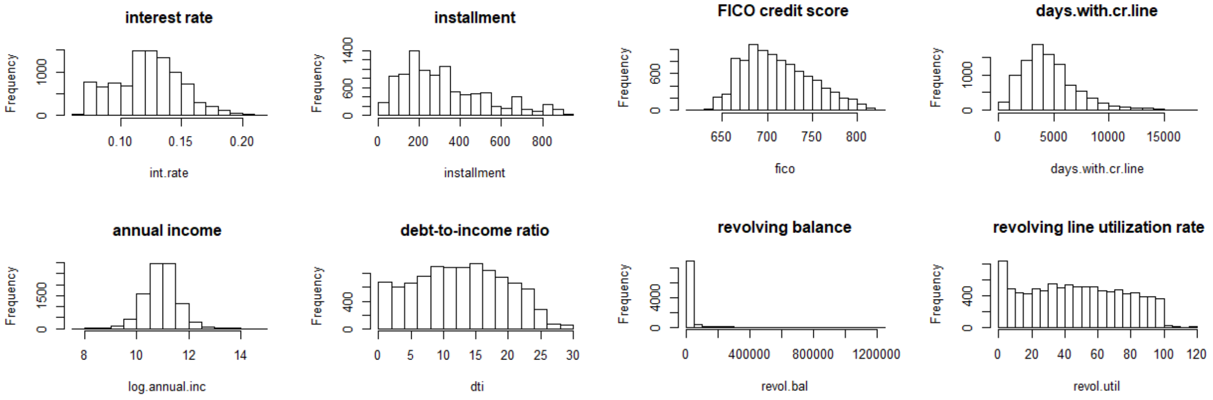
(The data description reference: https://www.lendingclub.com/info/download-data.action)

3. Analysis

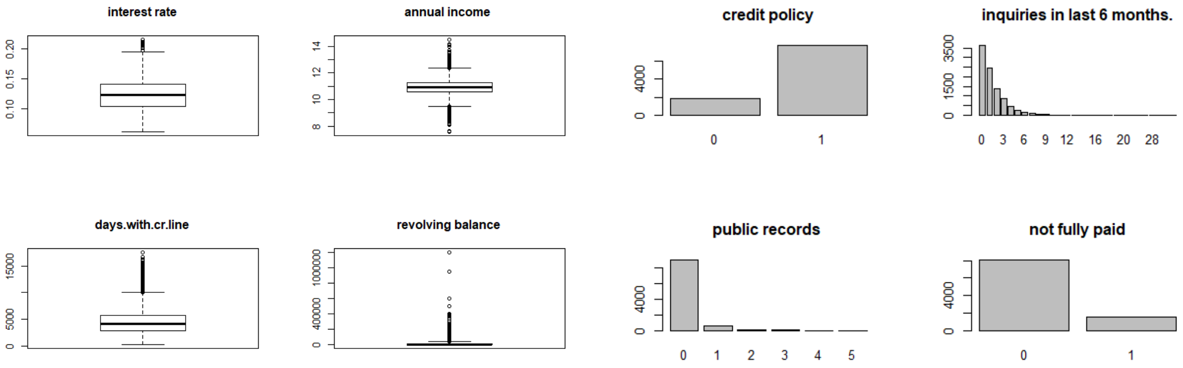


```
 credit.policy                 purpose       int.rate        installment     log.annual.inc        dti              fico      
 Min.   :0.000   all_other         :2331   Min.   :0.0600   Min.   : 15.67   Min.   : 7.548   Min.   : 0.000   Min.   :612.0  
 1st Qu.:1.000   credit_card       :1262   1st Qu.:0.1039   1st Qu.:163.77   1st Qu.:10.558   1st Qu.: 7.213   1st Qu.:682.0  
 Median :1.000   debt_consolidation:3957   Median :0.1221   Median :268.95   Median :10.929   Median :12.665   Median :707.0  
 Mean   :0.805   educational       : 343   Mean   :0.1226   Mean   :319.09   Mean   :10.932   Mean   :12.607   Mean   :710.8  
 3rd Qu.:1.000   home_improvement  : 629   3rd Qu.:0.1407   3rd Qu.:432.76   3rd Qu.:11.291   3rd Qu.:17.950   3rd Qu.:737.0  
 Max.   :1.000   major_purchase    : 437   Max.   :0.2164   Max.   :940.14   Max.   :14.528   Max.   :29.960   Max.   :827.0  
                 small_business    : 619                                                                                      
 days.with.cr.line   revol.bal          revol.util     inq.last.6mths    delinq.2yrs         pub.rec         not.fully.paid 
 Min.   :  179     Min.   :      0   Min.   :  0.0   Min.   : 0.000   Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.: 2820     1st Qu.:   3187   1st Qu.: 22.6   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median : 4140     Median :   8596   Median : 46.3   Median : 1.000   Median : 0.0000   Median :0.00000   Median :0.0000  
 Mean   : 4561     Mean   :  16914   Mean   : 46.8   Mean   : 1.577   Mean   : 0.1637   Mean   :0.06212   Mean   :0.1601  
 3rd Qu.: 5730     3rd Qu.:  18250   3rd Qu.: 70.9   3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :17640     Max.   :1207359   Max.   :119.0   Max.   :33.000   Max.   :13.0000   Max.   :5.00000   Max.   :1.0000  
```

From the descriptive statistics, we can find that the dataset includes 14 variables. The data type of most variables are continuous, and there is only one categorical variables that is the purpose of the loan. The target variable should be not.fully.paid, which indicates that whether borrowers fully pay off the loan. Since the minimum and maximum of credit policy and not fully paid are 0 and 1, thus we can conclude that these two variables should be binary variables so I change their data type into factor.

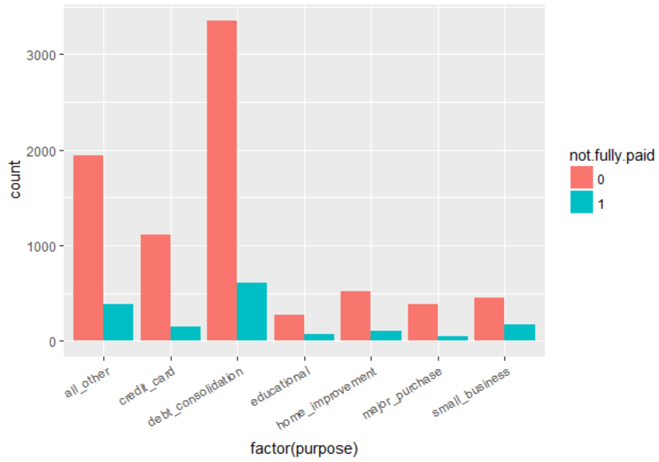


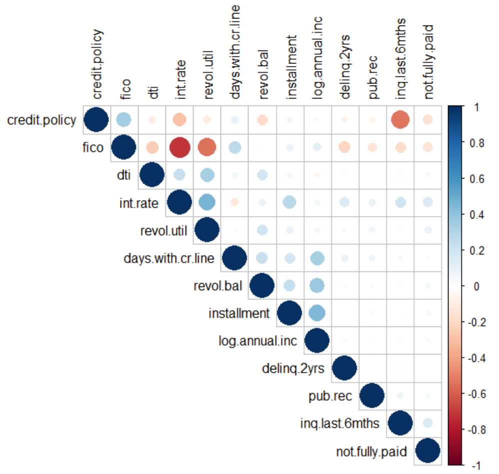
From the histogram, we can find that distribution of most variables are near symmetry, except revolving balance. Also from the descriptive statistics we can find that the median of revolving balance is 8596 while the mean is 16914, which indicates that the distribution of revolving balance is highly skewed, then I decided to use a box plot to detect outliers.



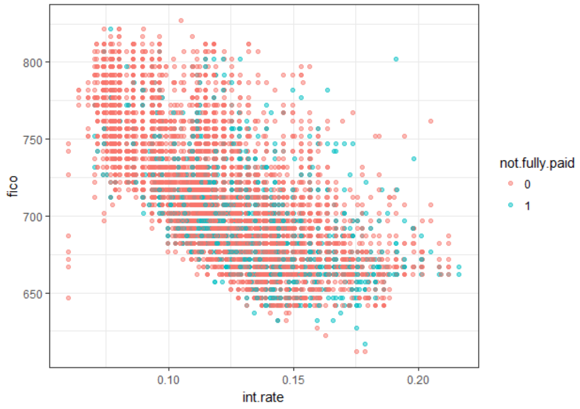
From the box plot, we can find that most of the value in revolving balance are considered to be outliers. But I cannot remove these outliers for the revolving balance, for the value of revolving balance shows the amount unpaid at the end of the credit card billing cycle, and the larger amount unpaid indicates a likelihood of losing more money. Thus, I decide to use a log function to transform the distribution to a normal distribution.

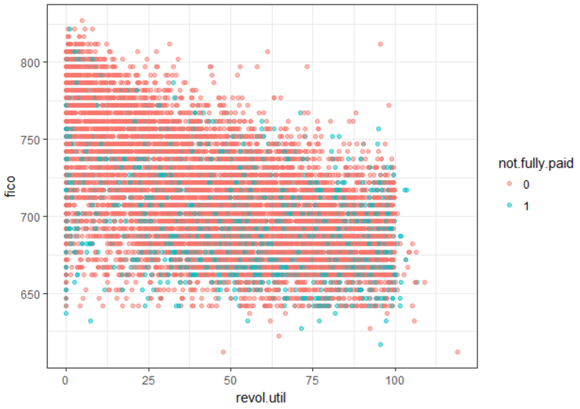
And the bar plot shows that there is an imbalanced classification of the target variable. 84% of the borrowers are fully paid off while 16% of the borrowers are not. While the imbalanced class distribution may lead to a problem in the modeling, because if the classifier predict that all borrowers could fully pay back the loan, the accuracy of the model could be 84%, which is overly optimistic. To deal with this uncertainty, I try to use both oversampling and undersampling method to the dataset in the modeling part.





From the left plot, we can find that for each type of loan purpose, the ratio of fully repayment and not fully repayment is almost same. But it seems that when the loan is used for small business, the likelihood of not fully payback is higher than others. And most borrowers needs loans for the debt consolidation. And from the heatmap ,we can find that fico score has a strong negative correlation with interest rate and revolving line utilization rate, and credit policy is highly positive correlated to borrower’s number of inquiries by creditors in the last 6 months. The multicollinearity problem may have an effect on the model accuracy, so I need further explore the relationship between these variables.

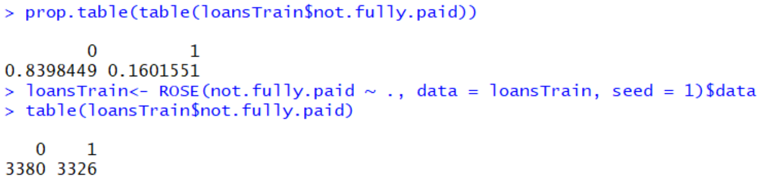




From the above scatter plots, we can clearly find that there is a linear relationship between fico and interest rate, but the relationship between fico and revolving line utilization rate is not very strong. And the left scatter plot shows that the intensity of not fully payback class is higher in the lower right corner, which indicates that when the fico score is lower and interest rate is higher, the possibility of not fully payback increase. And the right scatter plot shows that intensity of not fully payback class is higher in the lower level, while the intensity seems to be same in the left corner and right corner. So it seems that revolving line utilization rate has no strong correlation with target variable.

4. Model Development and Application of models

4.1 Data split



```
> prop.table(table(loansTrain$not.fully.paid))

        0         1 
0.8398449 0.1601551 
> loansTrain<- ROSE(not.fully.paid ~ ., data = loansTrain, seed = 1)$data
> table(loansTrain$not.fully.paid)

   0    1 
3380 3326 
```

As mentioned before, the distribtuion of target variable is highly imbalanced, and to deal with this uncertainity, I try two methods:

1) Apply stratified random split of the data (By using caret package)

Since there is no available test dataset, I need to split the data into train and validation set to evaluate the model fit. Due to the imbalanced classification in target variable, a ramdom split may lead to a bias result for the percentage of target variable in training and validation set is also imbalanced. Thus, a stratified random split technique is a good choice because it could keep the percentage of samples for each target class as the complete dataset in both sets, reducing the bias in the data and making the predction more reliable.

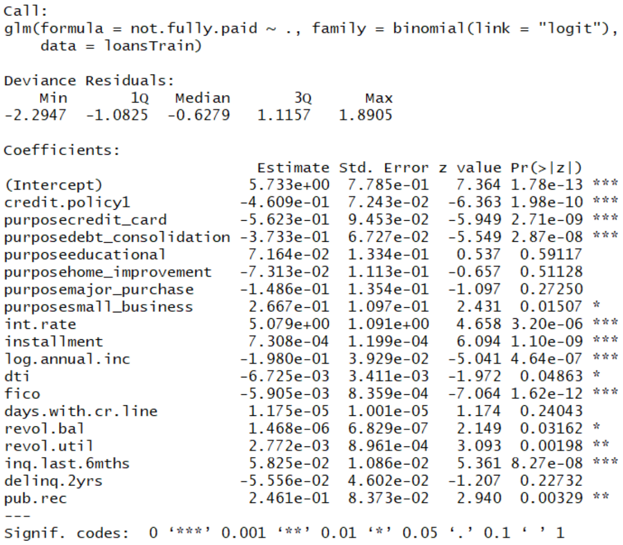
2) Apply oversampling and undersampling on the training dataset (By using ROSE package)

The oversampling method is helpful to increase the records of minority class, but may give rise to a overfitting problem. And the undersmapling method, on the other hand, is used to balance the class size by reducing the number of majority class, and this method may lead to a information loss. Thus, considering that the data size is either too large or small, I apply both oversampling and undersampling methods at same time on the training dataset to get a better balanced class size.

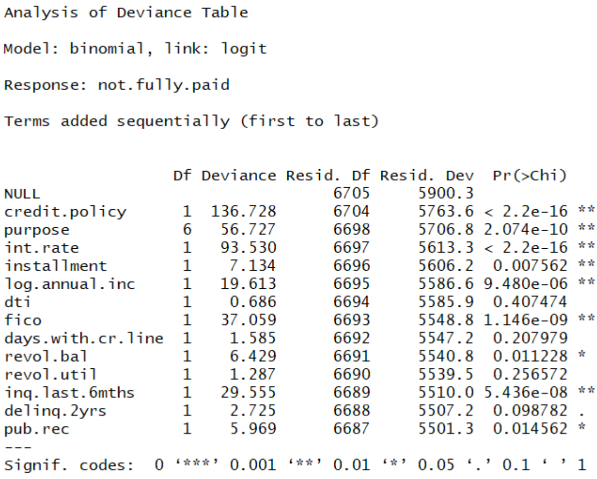
4.2 Modeling

Since the value of dependent variable is binary, I select three classification models to train the data. The first one is Logistics Regression, for it could be used to select important features and explain the relationship between independent variables and dependent variables. And the second model I used is SVM, for the SVM usually has a good performance on small dataset, and since the relationship between variables may be complex, I could choose a non-linear kernel for classification. The third model I applied is Decision Tree, because the model could automatically select the most important variables as predictors, and it is also not sensitive to outliers. And the fourth model I use is Random Forest for the ensemble model usually does well in reducing misclassification.

4.2.1 Model 1——Logistics Regression



```
Call:
glm(formula = not.fully.paid ~ ., family = binomial(link = "logit"),
    data = loansTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2947  -1.0825  -0.6279   1.1157   1.8905

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               5.733e+00  7.785e-01   7.364 1.78e-13 ***
credit.policy1           -4.609e-01  7.243e-02  -6.363 1.98e-10 ***
purposecredit_card       -5.623e-01  9.453e-02  -5.949 2.71e-09 ***
purposedebt_consolidation -3.733e-01  6.727e-02  -5.549 2.87e-08 ***
purposeeducational        7.164e-02  1.334e-01   0.537  0.59117
purposehome_improvement  -7.313e-02  1.113e-01  -0.657  0.51128
purposemajor_purchase    -1.486e-01  1.354e-01  -1.097  0.27250
purposesmall_business     2.667e-01  1.097e-01   2.431  0.01507 *
int.rate                  5.079e+00  1.091e+00   4.658 3.20e-06 ***
installment               7.308e-04  1.199e-04   6.094 1.10e-09 ***
log.annual.inc           -1.980e-01  3.929e-02  -5.041 4.64e-07 ***
dti                      -6.725e-03  3.411e-03  -1.972  0.04863 *
fico                     -5.905e-03  8.359e-04  -7.064 1.62e-12 ***
days.with.cr.line         1.175e-05  1.001e-05   1.174  0.24043
revol.bal                 1.468e-06  6.829e-07   2.149  0.03162 *
revol.util                2.772e-03  8.961e-04   3.093  0.00198 **
inq.last.6mths            5.825e-02  1.086e-02   5.361 8.27e-08 ***
delinq.2yrs              -5.556e-02  4.602e-02  -1.207  0.22732
pub.rec                   2.461e-01  8.373e-02   2.940  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

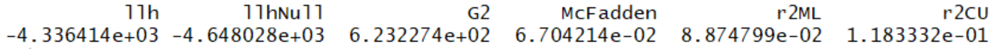


```
Analysis of Deviance Table

Model: binomial, link: logit

Response: not.fully.paid

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              6705     5900.3
credit.policy      1  136.728      6704     5763.6 < 2.2e-16 ***
purpose            6   56.727      6698     5706.8 2.074e-10 ***
int.rate           1   93.530      6697     5613.3 < 2.2e-16 ***
installment        1    7.134      6696     5606.2  0.007562 **
log.annual.inc     1   19.613      6695     5586.6 9.480e-06 ***
dti                1    0.686      6694     5585.9  0.407474
fico               1   37.059      6693     5548.8 1.146e-09 ***
days.with.cr.line  1    1.585      6692     5547.2  0.207979
revol.bal          1    6.429      6691     5540.8  0.011228 *
revol.util         1    1.287      6690     5539.5  0.256572
inq.last.6mths     1   29.555      6689     5510.0 5.436e-08 ***
delinq.2yrs        1    2.725      6688     5507.2  0.098782 .
pub.rec            1    5.969      6687     5501.3  0.014562 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the Logistics Regression result, most of the independent variables have significant effects on the target variable. When credit policy equals to 1, which means when the borrowers meets the criteria of LendingClub.com, it is highly possible that the borrower will fully pay back the loan due a negative correlation between credit policy1 and not fully payback. Thus, to find how predictors could lead to a not fully repayment result, I can mainly focus on the coefficient and the p-value of the variables.

The credit policy, purpose, interest rate, annual income, fico, and borrower’s number of inquiries by creditors in the last 6 months are the most significant predictors. Among these predictors, only the coefficients of interest rate, installment and number of inquiries are positive, indicating the higher value of these variables, the higher possibility that the borrowers’ will not fully pay back the loan.

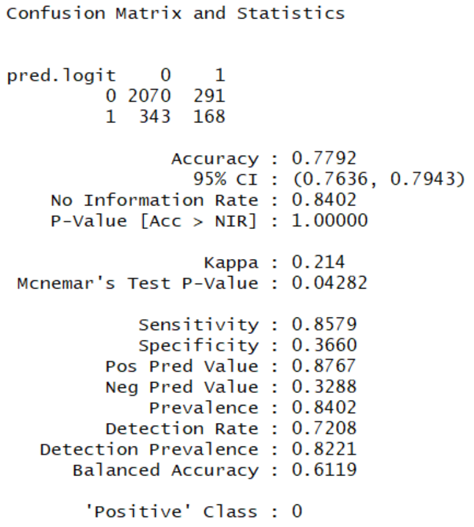
And by looking at the difference between null deviance and the residual deviance in the ANOVA table, I could know how well the model performs compare to a null model with only intercept. The result is in accordance with the outcome of the model, because credit policy, purpose, interest rate, annual income, fico, and borrower’s number of inquiries by creditors significantly reduce the residual deviance. And that means these six variables are the most important predictors to improve the fitness of the model.



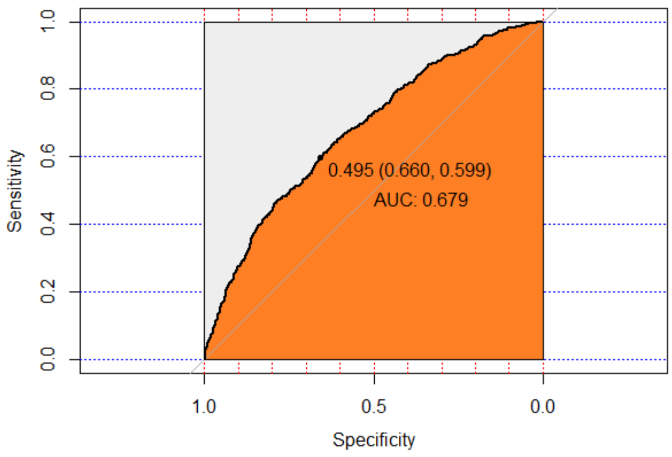
```
          llh       llhNull            G2      McFadden          r2ML          r2CU
-4.336414e+03 -4.648028e+03  6.232274e+02  6.704214e-02  8.874799e-02  1.183332e-01
```

I use McFadden's pseudo r-squared to evaluation how the model fit. While a range between 0.2 and 0.4 indicates that the model is a perfect fit to the data, the 0.067 of pseudo r-squared indicates that the model is not perfectly fit. Then I use validation set to assess the prediction result.

Reminding that the data is imbalanced, and to tune the model, I try to select a cut-off value by finding the optimal cut off of ROC curve. And the result shows that 0.495 is a perfect cut-off value instead of 0.5.



```
Confusion Matrix and Statistics

pred.logit    0    1
         0 2070  291
         1  343  168

               Accuracy : 0.7792
                 95% CI : (0.7636, 0.7943)
    No Information Rate : 0.8402
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.214
 Mcnemar's Test P-Value : 0.04282

            Sensitivity : 0.8579
            Specificity : 0.3660
         Pos Pred Value : 0.8767
         Neg Pred Value : 0.3288
             Prevalence : 0.8402
         Detection Rate : 0.7208
   Detection Prevalence : 0.8221
      Balanced Accuracy : 0.6119

       'Positive' Class : 0
```

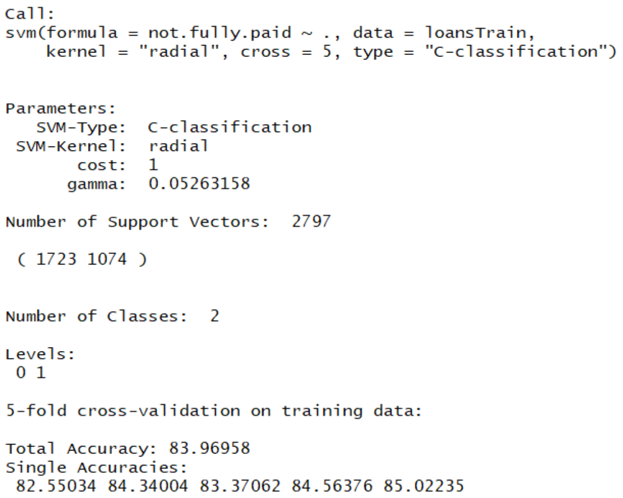


From the confusion matrix table, we can find that the overall accuracy of the model is 77.9%, which is fair. The sensitivity and positive predictive value are above 85%, and that means the model does well in correctly predicting the borrowers who fully pay off the loan. The specificity and negative predictive value are about 36%, which means that the model has a relative poor performance on the minority class.

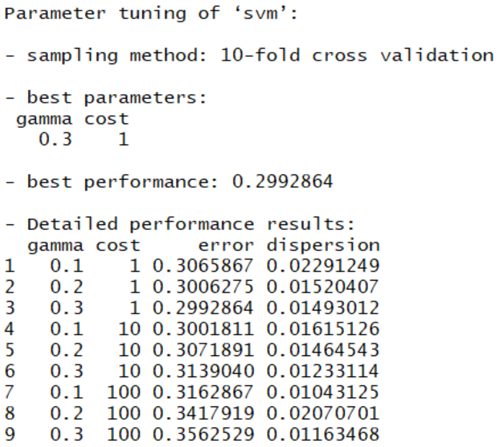
And the ROC curve and the AUC value are the better measurements than accuracy to evaluate the performance of binary classification. The ROC curve shows that the curve is slightly higher than the diagonal line, and the area under the curve in 0.679, which means that the model performs much better than a random classification, but the accuracy still needs improvement.

4.2.2 Model 2——SVM

I choose all variables except y as predictors, and select Gaussian kernel for non-linear classification. I also use 5 fold cross validation in the model to calculate the accuracy in each fold.



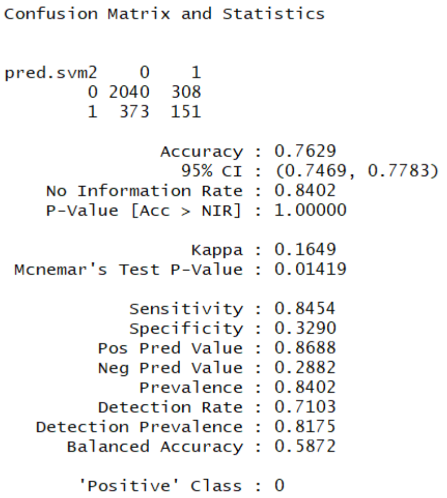
```
Call:
svm(formula = not.fully.paid ~ ., data = loansTrain,
    kernel = "radial", cross = 5, type = "C-classification")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.05263158

Number of Support Vectors:  2797

 ( 1723 1074 )


Number of Classes:  2

Levels:
 0 1

5-fold cross-validation on training data:

Total Accuracy: 83.96958
Single Accuracies:
 82.55034 84.34004 83.37062 84.56376 85.02235
```

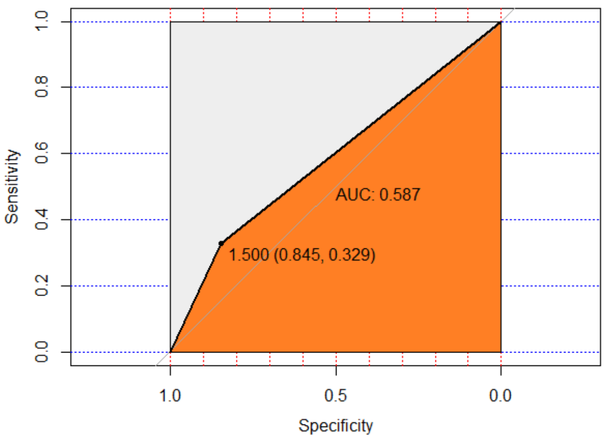


```
Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
   0.3    1

- best performance: 0.2992864

- Detailed performance results:
  gamma cost     error dispersion
1   0.1    1 0.3065867 0.02291249
2   0.2    1 0.3006275 0.01520407
3   0.3    1 0.2992864 0.01493012
4   0.1   10 0.3001811 0.01615126
5   0.2   10 0.3071891 0.01464543
6   0.3   10 0.3139040 0.01233114
7   0.1  100 0.3162867 0.01043125
8   0.2  100 0.3417919 0.02070701
9   0.3  100 0.3562529 0.01163468
```

From the SVM model summary result, we can find that the average accuracy of SVM model on train dataset is 83.1%, and the model find 2797 support vectors in total, including 1723 support vectors in class fully paid off and 1074 support vectors in class not fully paid off. And the model result also shows that the cost of the model is 1 and gamma is 0.05. While the value of gamma controls the standard deviation of the Gaussian function, and cost controls how much a misclassification will affect the result. To optimize the model, I use the tune.svm function to look for the best parameter.

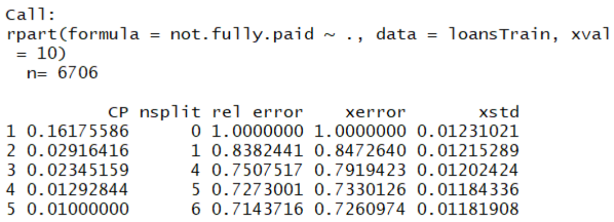
By given a cost of 1, 10 and 100, and gamma of 0.1, 0.2, 0.3 and applied a 10 fold cross validation, the function shows that when the cost equals to 1 and gamma equals to 0.3, the model will have the best performance. Then, I use the optimal parameter to generate the prediction on the validation set and assess the performance.

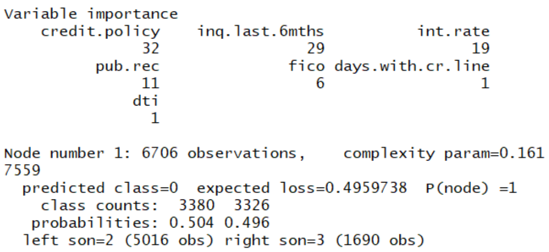

```
Confusion Matrix and Statistics


pred.svm2    0    1
        0 2040  308
        1  373  151

               Accuracy : 0.7629
                 95% CI : (0.7469, 0.7783)
    No Information Rate : 0.8402
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.1649
 Mcnemar's Test P-Value : 0.01419

            Sensitivity : 0.8454
            Specificity : 0.3290
         Pos Pred Value : 0.8688
         Neg Pred Value : 0.2882
             Prevalence : 0.8402
         Detection Rate : 0.7103
   Detection Prevalence : 0.8175
      Balanced Accuracy : 0.5872

       'Positive' Class : 0
```



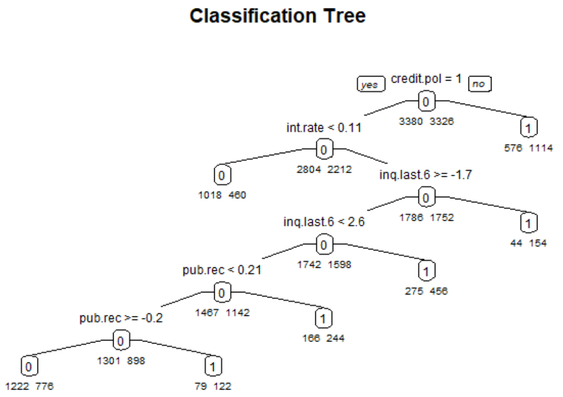
After applying the model on validation set, I calculate the accuracy and apply confusion matrix and contingency table to validate the result. From the confusion and statistics graph, we can find that the overall accuracy of test set is 0.76, which is lower than the Logistics Regression model. The sensitivity and positive predictive value are about 85%, while the specificity and negative predictive value are about 30%. From the contingency table, we can find that the value of False negative and False positive are almost equal, and 308 borrowers who fail to fully pay back are misclassified as fully paid off, and this mistake may increase the risk of losing money. Thus, to minimize the value of false positive is important, for the misclassification will increase the potential risk for investors.

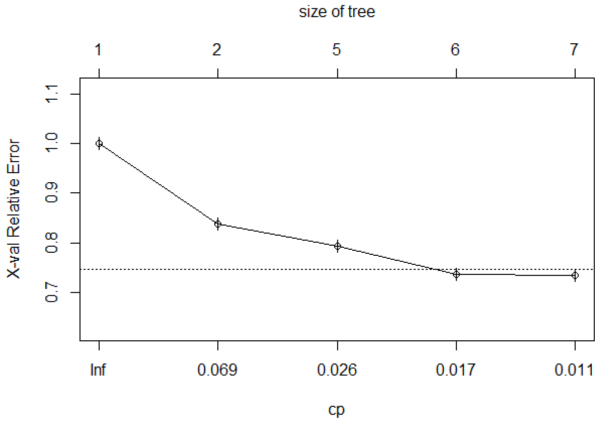
4.2.3 CART


```
Call:
rpart(formula = not.fully.paid ~ ., data = loansTrain, xval
 = 10)
  n= 6706

          CP nsplit rel error    xerror       xstd
1 0.16175586      0 1.0000000 1.0000000 0.01231021
2 0.02916416      1 0.8382441 0.8472640 0.01215289
3 0.02345159      4 0.7507517 0.7919423 0.01202424
4 0.01292844      5 0.7273001 0.7330126 0.01184336
5 0.01000000      6 0.7143716 0.7260974 0.01181908
```


```
Variable importance
   credit.policy    inq.last.6mths          int.rate
              32                29                19
         pub.rec              fico days.with.cr.line
              11                 6                 1
             dti
               1

Node number 1: 6706 observations,    complexity param=0.161
7559
  predicted class=0   expected loss=0.4959738  P(node) =1
    class counts:  3380  3326
   probabilities: 0.504 0.496
  left son=2 (5016 obs) right son=3 (1690 obs)
```

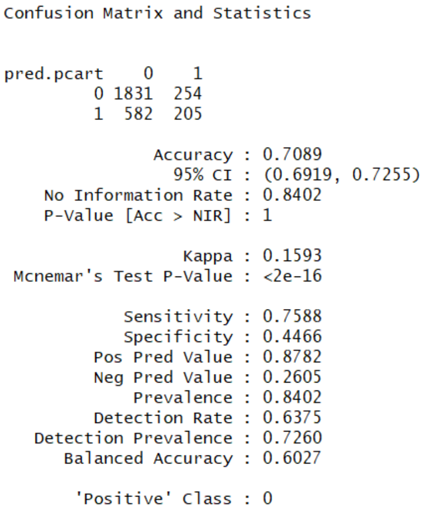
From the CART model result, we can learn that the credit policy is the most important predictor among all variables, and this result is in accordance with Logistic Regression model. And the rest 6 top predictors are borrower's number of inquiries by creditors in the last 6 months, interest rate, the borrower's number of derogatory public records, the number of days the borrower has had a credit line, and the debt-to-income ratio of the borrower.



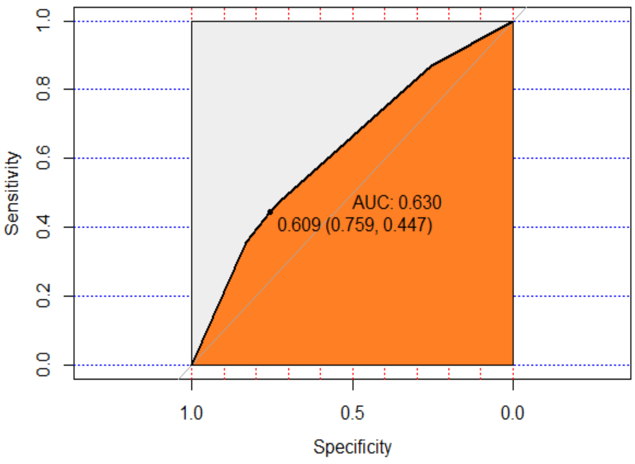


From the classification tree, we can find that the depth of the tree is 6 and it only selects 4 predictors, including credit policy, borrower's number of inquiries by creditors in the last 6 months, interest rate and the borrower's number of derogatory public records. However, we can find that the numerical predictors such as public records and number of inquiries have different values, and that means that the tree tends to uses the favor predictors with many potential split points. Thus, I decide to use complexity parameter table to find the optimal cp for tree pruning.

From above complexity parameter plot, we can learn that when the cp equals about 0.02, and the x-val relative error is about 0.75. Thus, when the value of relative error falls into the range of (0.75-0.02, 0,75+0.02), the complexity parameter has the optimal value. And by checking the value in cp table, we can learn that the cp equals to 0.2345, and then apply the parameter for the prediction on validation set. Also, I try to select a cut-off value by finding the optimal cut off of ROC curve, and the result shows that 0.609 is a perfect cut-off value to separate the two classes.

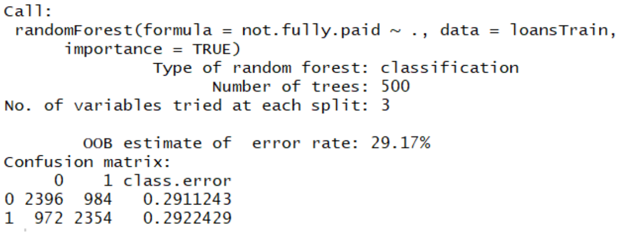


```
Confusion Matrix and Statistics


pred.pcart    0    1
         0 1831  254
         1  582  205

               Accuracy : 0.7089
                 95% CI : (0.6919, 0.7255)
    No Information Rate : 0.8402
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1593
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7588
            Specificity : 0.4466
         Pos Pred Value : 0.8782
         Neg Pred Value : 0.2605
             Prevalence : 0.8402
         Detection Rate : 0.6375
   Detection Prevalence : 0.7260
      Balanced Accuracy : 0.6027

       'Positive' Class : 0
```

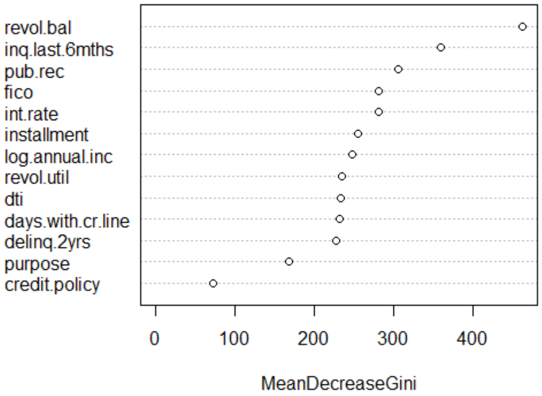


From the confusion matrix, we can learn the accuracy of decision tree model is only 71%, which is lower than the previous two models. However, this model has the highest value of specificity, because it has the lowest false positive and highest true negative among the three models. And this improvement is important for it values more for the correct prediction of the not fully payback class. For investors, they are more willing to know which borrower will not fully repay the loan, so that they can reduce the risk of losing money. Thus, the false positive value should be as small as possible. Even if the model misclassifies more fully paid borrowers as not fully payback borrowers, this model helps provide the most important information with highest accuracy. Thus, the CART model has the best performance so far.

4.2.4 Random Forest

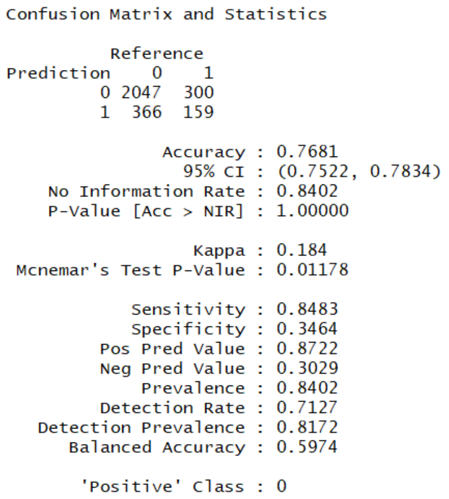


```
Call:
 randomForest(formula = not.fully.paid ~ ., data = loansTrain,
      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 29.17%
Confusion matrix:
     0    1 class.error
0 2396  984   0.2911243
1  972 2354   0.2922429
```

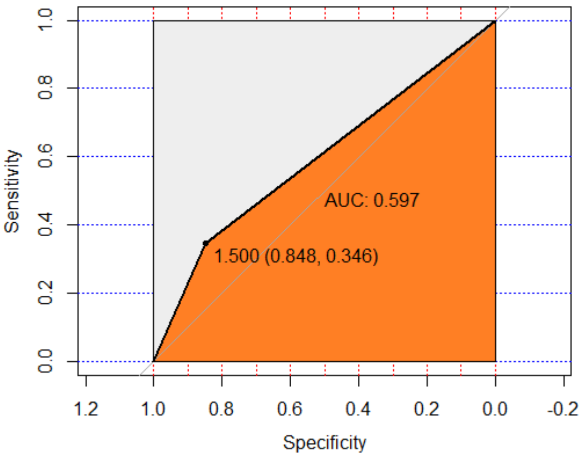


From the Random Forest model, we can learn that when the number of variables at each split equals to 3, the out-of-bag error rate will be 29.17%. The confusion matrix shows the model performance on training set, and due to the oversampling and undersampling method, the class size of fully paid off class and not fully paid off are almost same, and the misclassification rate are also similar. Thus, the accuracy of the model on training dataset is about 81%.

The plot of feature importance indicates that the most important feature is revolving balance, borrower’s number of inquiries by creditors in the last 6 months, public record, interest rate and fico credit score. Note that the feature importance is different from the result of Logistic Regression model. Compared to the Logistic Regression model and CART model, in which credit policy is the most important variable, the Random Forest model shows that the credit policy is the least importance feature, while the revolving balance is the top important feature. The reason of the inconsistence may due to the different algorithm. The logistic regression uses the maximum likelihood estimation method while the random forest model uses out of bag for evaluation. But except credit policy, the rest predictors including borrower’s number of inquiries by creditors in the last 6 months, public record, interest rate and fico credit score are top important variables in the previous models.

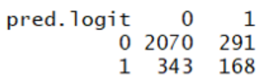
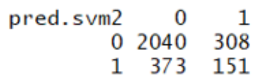
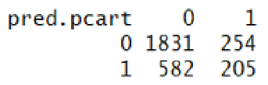
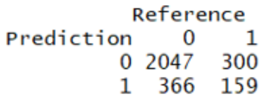


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2047  300
         1  366  159

               Accuracy : 0.7681
                 95% CI : (0.7522, 0.7834)
    No Information Rate : 0.8402
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.184
 Mcnemar's Test P-Value : 0.01178

            Sensitivity : 0.8483
            Specificity : 0.3464
         Pos Pred Value : 0.8722
         Neg Pred Value : 0.3029
             Prevalence : 0.8402
         Detection Rate : 0.7127
   Detection Prevalence : 0.8172
      Balanced Accuracy : 0.5974

       'Positive' Class : 0
```



From the confusion matrix, we can find that the classification result is similar to SVM model, for the balanced accuracy of the model is about 60%, and the model has a relative poor performance on the classification for minority class. Thus, I can conclude that among the four models, the Decision Tree model has the best performance. Even though the AUC value is higher in the Logistics Regression model, but the Decision Tree model has a lower misclassification on the not fully paid class, which is very important for lenders to reduce loan risk.

5. Conclusions and Discussion

Model	Contingency Table	Important Features	
Logistic Regression	 pred.logit 0 1 0 2070 291 1 343 168	• Credit policy • Purpose • Interest rate	• Annual income • Fico • Inq.last.6mths
SVM	 pred.svm2 0 1 0 2040 308 1 373 151		
Decision Tree	 pred.pcart 0 1 0 1831 254 1 582 205	• Credit policy • Interest rate	• Public record • Inq.last.6mths
Random Forest	 Reference Prediction 0 1 0 2047 300 1 366 159	• Revolving balance • Inq.last.6mths	• Public record • Fico • Interest rate

By looking at the important features in the four model, I can prove my four hypotheses are true. That is: Borrowers with lower FICO score, higher monthly installment and lower annual income will be unlikely to fully pay back the loan. Also, borrowers' purpose of loan has a significant effect on whether loan is paid in full.

Comparing the contingency tables among the four models, we can learn that the Decision Tree model has the highest True Negative value, which indicates that the model has the highest accuracy of correctly predicting the not fully paid class. The classification result of other three models are similar that they all do well in classifying the fully paid class, and has a relative poor performance on the classification of not fully paid class. Thus, these three models tend to have higher overall accuracy, but in this case, the overall accuracy is not a good measurement of the model performance due to the imbalanced class size.

At first, I want to use AUC and ROC curve to assess the model performance, for they are usually a better choice compared to overall accuracy. However, the result shows that the Logistic Regression has the highest AUC value, but the McFadden's pseudo r-squared indicates that the model is not a good fit to the data. And the AUC value among the four models only have a slight difference, so I think the AUC value cannot be selected as the only measurement for the prediction evaluation. Then by looking at the confusion matrix, I pay more attention to the False positive rate and True negative rate. These two indices suggests the misclassification rate of the not fully paid class and the correct classification rate of the not fully paid class. A lower false positive rate means that the model could help to filter more borrowers who cannot fully pay back the loan, then lenders could choose to only lend the money to the rest borrowers. Investors could reduce the risk of losing money by using the Decision Tree model for prediction.

To further improve the model, I would like to bin the numerical predictors with many values. That’s because in the current decision tree model, the important predictor are used for multiple split, so the tree is keep splitting by using the limited top important features. And by binning the numerical predictor could avoid this problem and the accuracy could be better. Also, I would like to combine the variables to new features. I think by reducing the dimension, the noise in the model could reduce and thus improve the model performance.

Reference:

Data Source: https://www.kaggle.com/sarahvch/predicting-who-pays-back-loans/data

Data description: https://www.lendingclub.com/info/download-data.action