Case Study 4: Analysis of Lending Club's Issued Loans

By

Hang Ding, Fangling Zhang, Qingquan Zhao, Yihao Zhou, Tongge Zhu

1. Business Target

Online loan industry is growing sharply recently, thus a large amount of capital flooding into this field. In these lending companies, there is no doubt of the importance of risk controlling, including evaluating potential losses, identifying potential risks as well as reducing these threats.

Lending Club, headquartered in San Francisco, California, was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. This report is an analysis of data from Lending Club, in order to detect and predict bad debt and distinguish the risk level of different kinds of loans with the help of machine learning, statistical methods and data visualization techniques.

In this report, we have explored statistical learning methods in role of detecting potential default clients, we believe by doing so, the loan company is easier to:

- Decrease default rate of loan
- Increase benefit
- Make business decisions
- Improve loan issue time cycle.

We used the dataset to predict 'bad behaviors' (late payment, default, etc.) by using several classification methods. We also tried tuning parameters of these methods to enhance their performance. And at last, we compared the performance among all the classification methods we used.

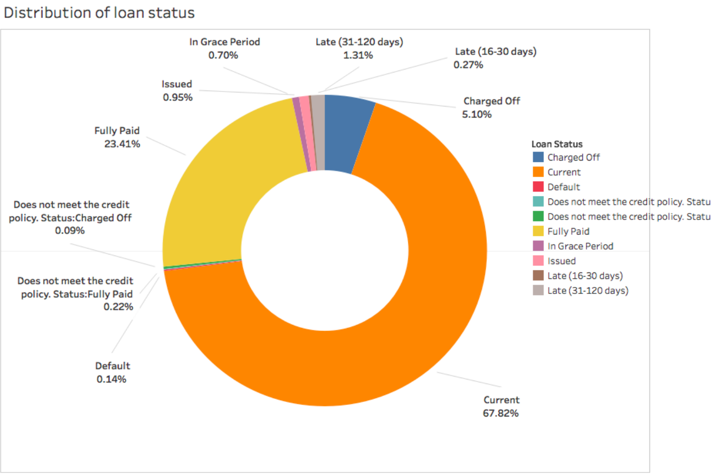
2. Data Statistical Description

2.1 Data Overview

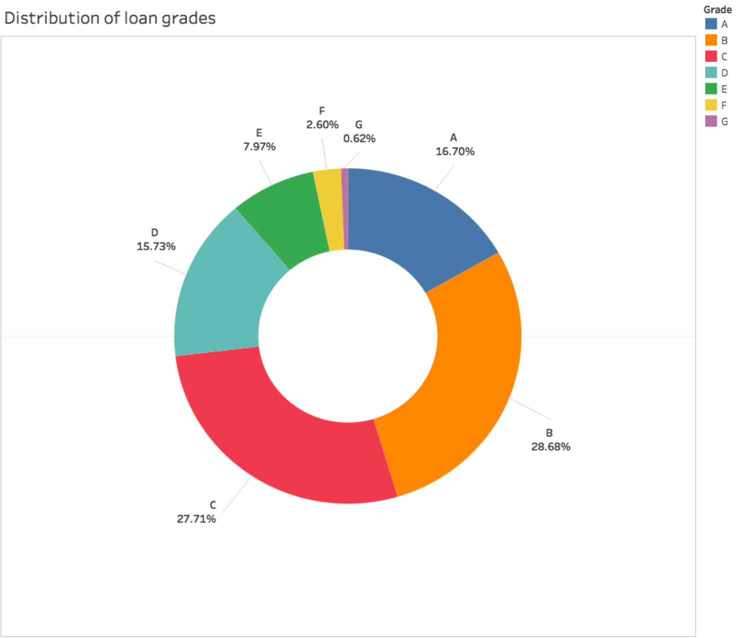
The original dataset includes 880000 samples and 73 features, roughly including three types: customer information such as the addresses of customers, loan application information such as grade of customers and loan purpose, and the third part post-loan information, such as time and amount of payment.

2.2 Data Visualization

The first chart below shows the distribution of dataset. As shown in the chart, 68% of loans are in progress, approximately 24% of loans are paid in full, and approximately 8% of loans have some type of defaults or late payment.

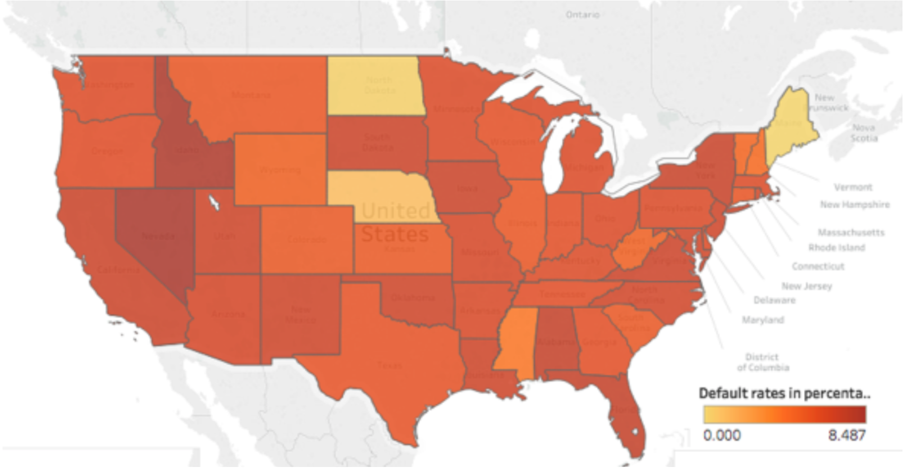


The second chart below shows the distribution of loan grade. Based on the its risk, each loan is graded from A to G in which grade A means the loan has lowest risk and grade G means the loan has highest risk. From the chart, we can 75% of the loans are in grade A to C and only 25% of the loans are in grades below C, which could be ‘bad’ loan.

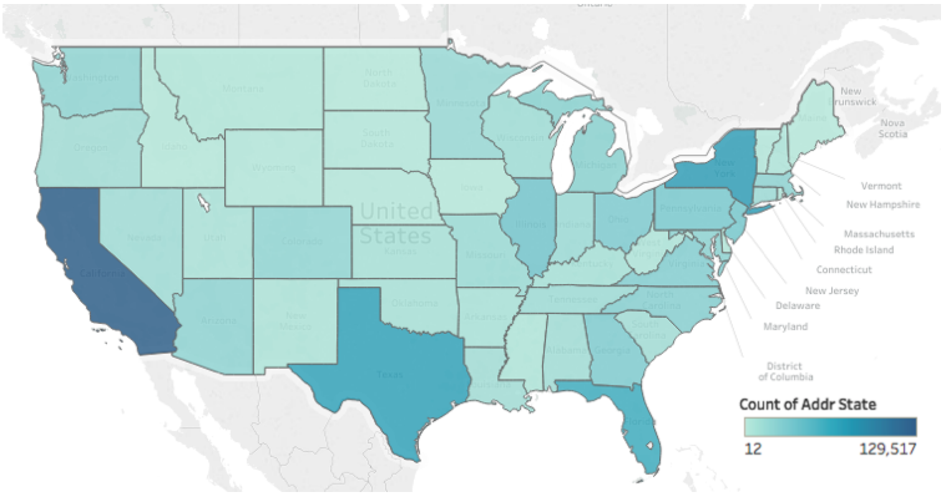


The third chart below shows the geographic distribution of default rate. The default rate is the percentage of the default loans in the state. The default rate maps to color from yellow to red.

The most default rate a state has, the darker red it shows. We can see in Nevada where Las Vegas located, the default rate is among the highest in US. It might have something to do with the gambling industry in Las Vegas.



The fourth chart below shows the geographic distribution of number of loans in each state. The more number of loans in the state, the higher the color is. In the chart, we can see California has the most number of loans. We believe it is because there are many startups in California and they need initial capital to run the businesses.



3. Machine learning

3.1 Preprocessing Data

3.1.1 Cleaning Dataset

The first operation we did on data is removing those variables with more than 10% missing values. With the help of a r package named 'caret', we further cleaned the dataset with following several steps: 1. Center and scale the data values; 2. Remove the variables with near zero variance; 3. Delete high correlation; 4. Delete linear combos. Finally, considering the large amount of computation consumption, we randomly take 40000 observations as our dataset for the following research. Eventually, 27 valuables are left over from 73 original features in the subset after following the cleaning steps as described above.

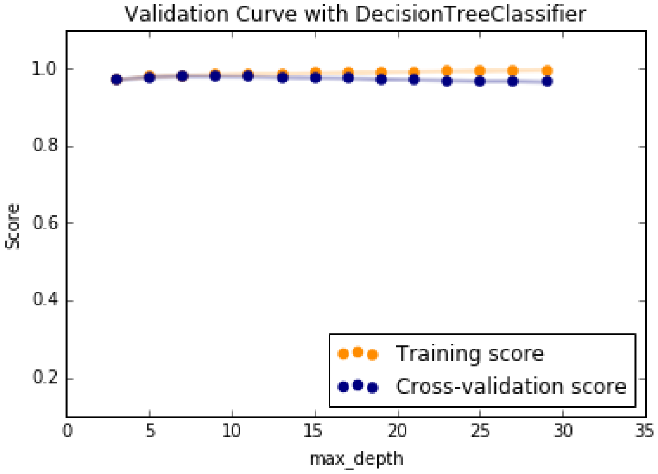
3.1.2 Encoding Categorical Features

Since some of the features are not given as continuous values but categorical. Instead of simply coding them as integers, which will be categorized as being ordered, we applied the get_dummies function in Pandas package to convert these categorical variables into dummy/indicator variables. After this step, the shape of dataset was changed to (40000, 54) for future analysis.

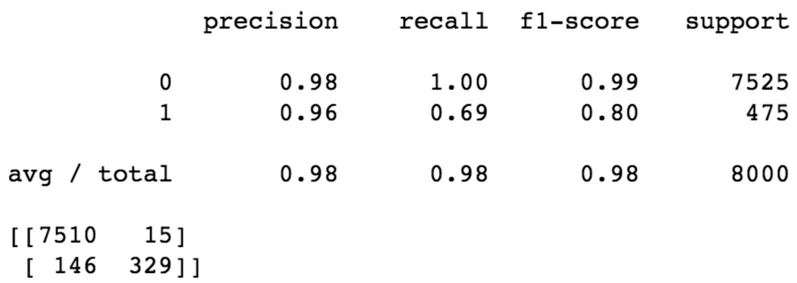
3.2 Model Selection

3.2.1 Decision Tree

Decision tree uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Tree-based methods can be applied for both regression and classification by stratifying or segmenting the predictor space into a number of simple regions.



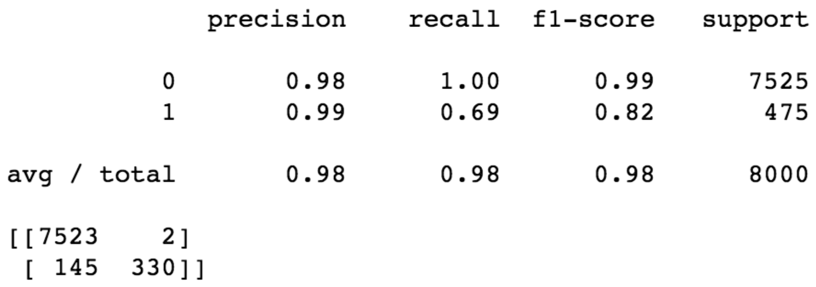
In our case, by separating the dataset into 80% as training and 20% as testing data, decision tree classifier is applied in predicting whether a customer will default or not. The maximum tree depth was determined to be 7 based on the result of validation curve as shown above. The confusion matrix as shown in below indicate a 70% sensitivity for "bad" behavior and 98% of sensitivity for "good" behavior.



```
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      7525
          1       0.96      0.69      0.80       475

avg / total       0.98      0.98      0.98      8000

[[7510   15]
 [ 146  329]]
```

3.2.2 Random Forest

A single decision tree might have several drawbacks such as high variance and possible overfitting to training set. Therefore, a random forest approach was explored for further improvement. Random forest is an ensemble of decision trees, which construct a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The confusion matrix as shown below indicates that random forest classifier could slightly improve the sensitivity for "good" behavior, whereas very little enhancement of the sensitivity for "bad" behavior.

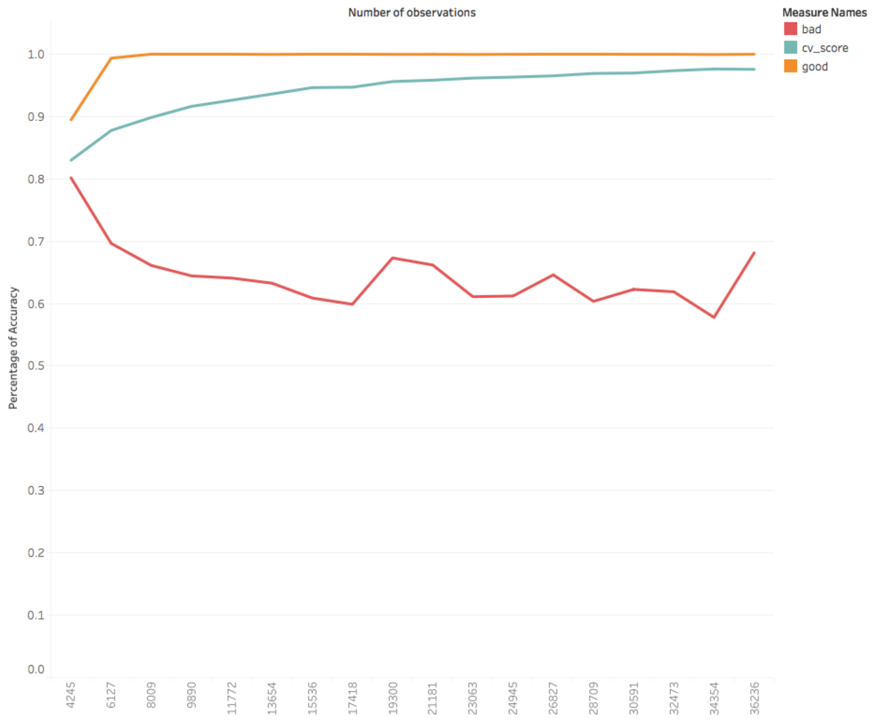


```
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      7525
          1       0.99      0.69      0.82       475

avg / total       0.98      0.98      0.98      8000

[[7523    2]
 [ 145  330]]
```

3.2.3 Further Improvement with balanced Data

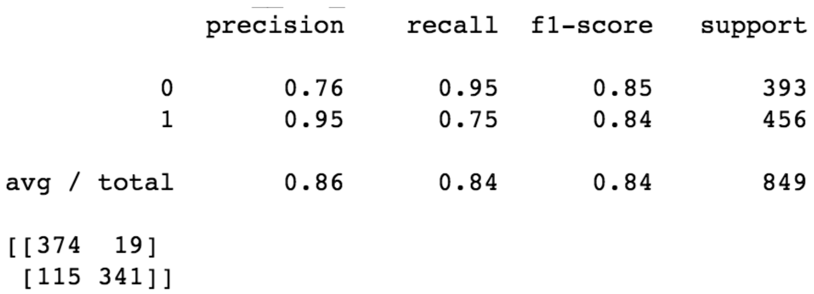
In our cleaned subset, there are 40000 observations, in which just 2363 observations have 'bad behavior'. The bad behavior percentage is just 5.91%. Therefore, our data is typically imbalanced. Supposedly, if we used a particular sampling method to get balanced data, we might be able to improve our prediction.

To reduce the sample amount of customers who have "good behavior" to an optimal ratio to the number of customers who behave "bad", a cross-validation of sample ratio belonging to each class was explored, together with the corresponding predicting scores based on random forest classifier. The result shown below suggests that the cross validation score and the prediction of good behavior are increasing when data becomes unbalanced. However, the prediction score of "bad" behavior class is decreasing dramatically in the unbalanced data. Eventually, we choose a ratio of roughly 3: 1 (6127:2363) as a happy point for the future analysis.



3.2.4 Decision Tree After Balanced Data

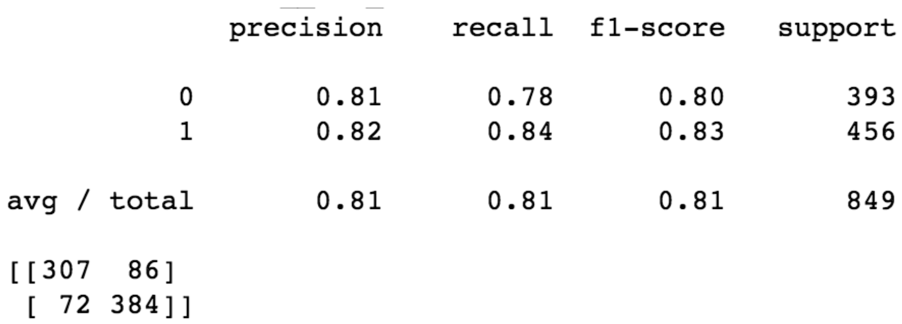
With the balanced data in hand, a grid search of max_depth parameters of decision tree classifier was performed and offers a best score of 0.83 with the optimal maximum tree depth of 3. The confusion matrix based on decision tree classifier with optimized parameter is shown below. An obvious improvement was observed in terms of 75% of sensitivity of “bad” behavior.



```
             precision    recall  f1-score   support

          0       0.76      0.95      0.85       393
          1       0.95      0.75      0.84       456

avg / total       0.86      0.84      0.84       849

[[374  19]
 [115 341]]
```

3.2.5 Random Forest After Balanced Data

The similar grid search of max_depth parameters of random forest classifier was performed the offers a best score of 0.82 with the optimal maximum tree depth of 9. The confusion matrix shown below indicates a further improvement of sensitivity to 84% for “bad” behavior.



```
             precision    recall  f1-score   support

          0       0.81      0.78      0.80       393
          1       0.82      0.84      0.83       456

avg / total       0.81      0.81      0.81       849

[[307  86]
 [ 72 384]]
```

3.2.6 Logistic Regression

Using the balanced data, we applied logistic regression and the following is the confusion matrix:

	Pred NOT Bad Behavior	Pred Bad Behavior
NOT Bad Behavior	1115	8
Bad Behavior	121	358

As a lending company, Lending club does not that much care how much the accuracy the model can achieve. Lending club puts more attention on how much bad behavior the model can predict and how much bad behavior the model predicts is correct, in other words, they are sensitivity and precision in statistical learning. From this model, the sensitivity of bad behavior is 75% and the sensitivity of good behavior is 99%.

3.2.7 Comparison between the models

Sensitivity: the percentage of specificity true defaulters that are identified.
Specificity: the percentage of non-defaulters that are correctly identified.
Key Point = Sensitivity * 0.2 + Specificity * 0.8

	Sensitivity	Specificity	Key Point	Selection
Decision Tree	0.69	1.00	0.75	
Random Forest	0.69	1.00	0.75	
Decision Tree with balanced data	0.75	0.95	0.79	

Random Forest with balanced data	0.84	0.78	0.83	
Logistics Regression	0.75	0.99	0.80	
SVM	0.66	1.00	0.73	
K-NN	0.54	0.92	0.62	
ExtraTrees	0.63	0.97	0.70	
Gradient Boosting	0.75	0.99	0.80	
MLP	0.70	0.98	0.76	

4. Mathematical Rationalization

In this part, we will introduce related mathematical problems behind classification methods we used: classification tree, random forest and logistic regression.

4.1. Classification Tree

In this case, we have 27 feature vectors such as (interest rate, loan amount) denoted by $\{x_i\}_{i=1}^{n}$ with outcomes $y_1 = 1,\ y_2 = 0$ that means a people who loan money from bank will default or not respectively. So whole data could be denoted by $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$. in this case, n is the number of observation, which is 880,000. And each feature vector can be denoted by $x_k = (x_{k1}, \cdots, x_{k27})$.

A classification tree is a decision tree in which each node has a binary decision based on whether $x_i < a$ or not for a fixed a. a is a kind of classification criteria that could filter or allocate a numerical data. In this case, if a customer feature data $x_{10} <- 1.074$ then they will be allocated in the first group, otherwise they will be allocated in the second group. The top node contains all of the data, and the set at second layer is subdivided among the children of each node according to the classification at that node. At each node, feature x_i and threshold a are chosen to minimize resulting in the children nodes. And resulting is measured by Gini criterion. So, what is Gini criterion? As we can see in the picture, we choose layer 2 as an example. $C_1 =$ default $C_2 =$ not default. And we already know the set is divided to two subsets by classification criteria a. denoted by $S = S_1 \cup S_2$. And each set is partitioned into two classes C_1 , C_2 .

Define $\hat{P}(S_j) = |\frac{S_j}{S}| = proportion\ of\ S_j\ in\ S$.

$\hat{P}(C_i|S_j) = |\frac{S_j \cap C_i}{S}| = proportion\ of\ S_i\ which\ is\ in\ C_i$

Variation $g(S_j)$ in set S_i to be : $g(S_j) = \sum_{1}^{2} \hat{P}(S_j)(1-\hat{P}(S_j))$

$g(S_j)$ is largest if set S_i is equally divided, so as picture shows above, in the second layer, S_1 =7478 and S_2 =24522, so $g(S_1)$ is 0.1106. and it is smallest when all of S_i is just one of the C_i , we can see in the right bottom of the picture, $g(S_{16}) = 0.0166$, it divided into two set, and one of them only have 5 observes.

Then, Gini criterion $G = \hat{P}(S_1)g(S_1) + \hat{P}(S_2)g(S_2)$.

4.2. the second part is Random forest method

In this part, the first thing we need to clarify is that the random vector $X = (X_1, \cdots, X_d)$ (d is the number of features that is randomly picked from 27 features.) has a joint distribution which is the same as $(X_1, \cdots, X_d, y)$. And our goal is to build a classifier which predicts y from x based on the data set of D. and ensemble of classifiers denoted $h = \{h_1(x), \cdots, h_k(x)\}$. If each $h_k(x)$ is a decision tree, then the ensemble is a random forest. We define the parameters of the decision tree for classifier $h_k(x)$ to be $\theta_k = (\theta_{k1}, \theta_{k2}, \cdots, \theta_{kp})$. So, in random forest method, we choose which features appear in which nodes of the k^{th} three at random, according to parameters θ_k , which are randomly chosen from a model variable. So, it means θ_k determines subset x_θ of the full set of features. For the final classification $f(x)$, each tree casts a vote for the most popular class at input x, and the class with the most votes wins.

We use ensemble methods (multiple learning algorithms) to obtain better predictive performance, which we will not discuss at this time.

4.3. Logistic regression classification

How should we model the relationship between p(X) = Pr(Y = 1|X) andX? In this case, 1 coding for the bad behavior (default or late pay in loan status). we must model p(X) using a function that gives outputs between 0 and 1 for all values of X. Many functions meet this description.
In logistic regression, we use the logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

To fit the model, we use a method called maximum likelihood. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β0 and β1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, corresponds as closely as possible to the individual's observed default status. In other words, we try to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging these estimates into the model for p(X), yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not. This intuition can be formalized using a mathematical equation called a likelihood function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

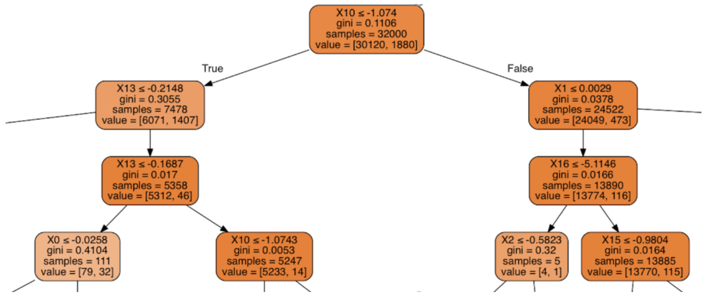
The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

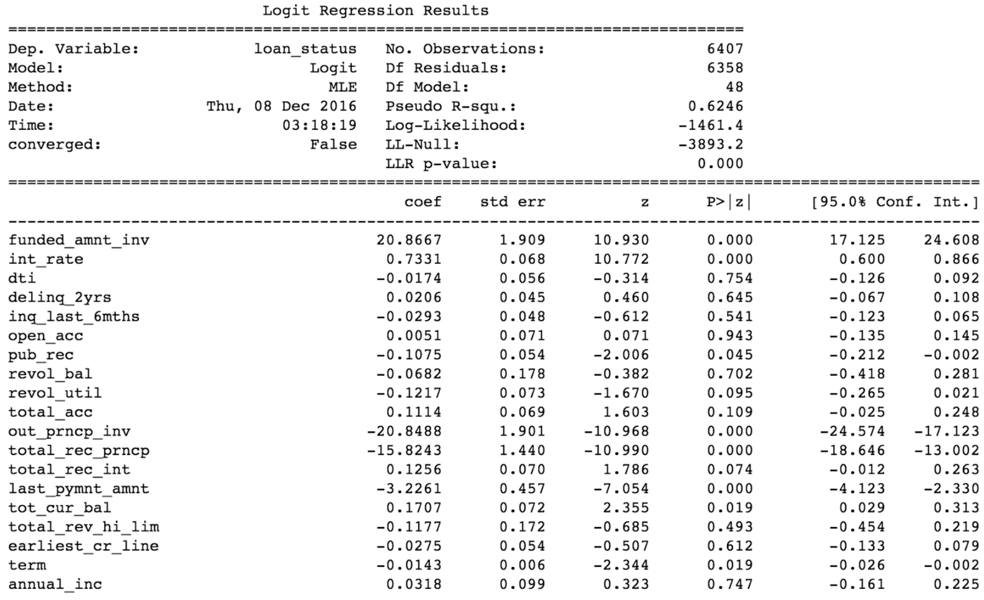
5. Business Meaning

5.1 Prediction

For online loan platform, the best model to predict default or bad behavior is random forest classifier with balanced data. Using our random forest model, the percentage of specificity true defaulters that are identified (sensitivity) is as high as 84%. From the perspective of a credit card company that is trying to identify high-risk individuals, an error rate of 16% among individuals who default may well be acceptable. On the other hand, the percentage of non-defaulters that are correctly identified (specificity) here can be up to 100 %. Decision tree, random forest and SVM can all get that high specificity.

5.2 Statistical reference





```
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:                 6407
Model:                          Logit   Df Residuals:                     6358
Method:                           MLE   Df Model:                           48
Date:                Thu, 08 Dec 2016   Pseudo R-squ.:                  0.6246
Time:                        03:18:19   Log-Likelihood:                -1461.4
converged:                      False   LL-Null:                       -3893.2
                                        LLR p-value:                     0.000
===================================================================================================
                                         coef    std err          z      P>|z|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------------------
funded_amnt_inv                       20.8667      1.909     10.930      0.000        17.125    24.608
int_rate                               0.7331      0.068     10.772      0.000         0.600     0.866
dti                                   -0.0174      0.056     -0.314      0.754        -0.126     0.092
delinq_2yrs                            0.0206      0.045      0.460      0.645        -0.067     0.108
inq_last_6mths                        -0.0293      0.048     -0.612      0.541        -0.123     0.065
open_acc                               0.0051      0.071      0.071      0.943        -0.135     0.145
pub_rec                               -0.1075      0.054     -2.006      0.045        -0.212    -0.002
revol_bal                             -0.0682      0.178     -0.382      0.702        -0.418     0.281
revol_util                            -0.1217      0.073     -1.670      0.095        -0.265     0.021
total_acc                              0.1114      0.069      1.603      0.109        -0.025     0.248
out_prncp_inv                        -20.8488      1.901    -10.968      0.000       -24.574   -17.123
total_rec_prncp                      -15.8243      1.440    -10.990      0.000       -18.646   -13.002
total_rec_int                          0.1256      0.070      1.786      0.074        -0.012     0.263
last_pymnt_amnt                       -3.2261      0.457     -7.054      0.000        -4.123    -2.330
tot_cur_bal                            0.1707      0.072      2.355      0.019         0.029     0.313
total_rev_hi_lim                      -0.1177      0.172     -0.685      0.493        -0.454     0.219
earliest_cr_line                      -0.0275      0.054     -0.507      0.612        -0.133     0.079
term                                  -0.0143      0.006     -2.344      0.019        -0.026    -0.002
annual_inc                             0.0318      0.099      0.323      0.747        -0.161     0.225
```

To get the important variable for default prediction, we compared the result of random forest with balanced data and logistic regression. At last, we found that there were 5 significant variables: int_rate, installment, total_rec_int, total_rec_prncp and out_prncp. The coefficients of int_rate, installment, total_rec_int are positive, whereas the coefficients of total_rec_prncp and out_prncp are negative.

In this case, the above result means that given the other variables, the default risk will increase significantly with the improvement of interest rate on the loan, the monthly payment owned by the borrower if the loan originates, interest received to date. That can be understand easily, when Lending Club issue the loan, it asks higher interest rate for people with lower grade. These people are more easier to default. If the other variables are constant, the more the borrower need to pay monthly, the more likely he/she is to default. Given the other variables, the default risk will decrease significantly with the improvement of principal received to date and remaining outstanding principal for total amount funded. How to interpret that? The more principal One borrower repays until now, the less remaining principal he/she need to pay. Therefore, the default risk of this loan will be lower. The remaining outstanding principal for total amount funded means remaining principal receivable of his/her investment when the borrower is also as an investor. The more money other people own one, his/her default risk is higher.

6. Conclusion

In summary, we have explored statistical learning methods in role of detecting potential default clients. We used the dataset containing 880000 samples and 73 features to predict ‘bad behaviors’ (late payment, default, etc.) by using several classification methods. We also tried tuning parameters of these methods to enhance their performance. And at last, we compared the performance among all the classification methods we used. We believe by applying our strategies, the loan companies will be able to easier in making decisions to decrease default rate of loan, make business decisions, improve loan issue time cycle, and eventually Increase benefit.