Credit Risk Statistical Learning

Group 3

Fangling Zhang

Huayi Zhang

Jiexuan Sun

Jun Dao

Ziqi Lin

Final project

DS502 Statistical Methods for Data Science

Worcester Polytechnic Institute

December 6, 2016

Abstract

Online loan industry is very popular recently. A large amount of capital flood into this field. The Lending Club, as the leading company in online loan industry, grows very quickly. There is no doubt that risk control must play an important role in company such as the Lending Club. In order to detect/predict bad debt and distinguish the risk level of different kinds of loans in the Lending Club, we try many different classification methods to predict 'bad behaviors' (late payment, default, etc.) and distinguish each loan's grade by using the dataset provided by the Lending Club company. We try a variety method to improve each method's performance. And at last, we compare the performance among all the classification methods we used.

Key words: risk control, credit grade, classification

Introduction

Purpose and Importance of the Study

The purpose of this project is to explore what statistical learning methods could improve the ability of managers makes business decisions. Specifically, how statistical learning method could help financial companies to improve benefit.

We did two experiments in this project. First, we want to help the Lending Club to detect the potential default clients. If the Lending Club could identify potential default clients, they could make reactions earlier and avoid default lost. Second, we want to assist the Lending Club to develop the ability to re-assign client's interesting rate according to client's behavior after the loan is issued. In the traditional mode, the Lending Club gives client a fix interesting rate at the beginning of the loan is issued. Floating interest rate could encourage clients have better behavior like returning money earlier, which will improve lending club's profit.

Information about the Dataset

The dataset is from the Kaggle website. The original dataset including more than 880,000 observations and 70 variables. Considering the large amount of computation consumption, we randomly take 40,000 observations as our dataset to analyze. Deducing the variables have lots of NA, we finally have 50 variables, including 30 numerical variables and 20 categorical variables. Among these 50 variables, we divided them into three categories: customer information, loan information and post-loan information.

Numerical variables (30): loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment, grade, sub_grade, annual_inc, dti, delinq_2yrs, inq_last_6mths, open_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, pub_rec, revol_bal, revol_util, total_acc, last_pymnt_amnt, collections_12_mths_ex_med, policy_code, acc_now_delinq, tot_coll_amt, tot_cur_bal, total_rev_hi_lim

Categorical variables (20): term, emp_title,emp_length, home_ownership, issue_d, verification_status, loan_status, pymnt_plan, purpose, title, addr_state, earliest_cr_line, initial_list_status, last_pymnt_d, next_pymnt_d, last_credit_pull_d, application_type, verification_status_joint

Exploratory the Dataset

The Figure 1 shows the distribution of the loan status at the time of the data collecting. As we can see that approximately 73.3% of loans are still under processing, 19% of loans are already fully paid and 5% of loans seems have some ‘trouble’ like default, late payment and etc.

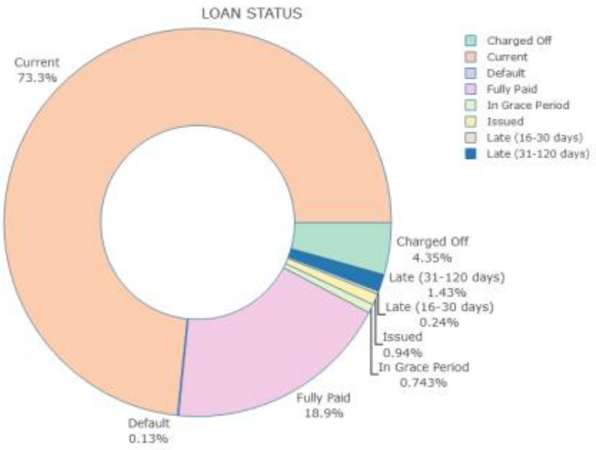


Figure 1 Loan Status

The Figure 2 is the pie chart about the loan grade. Each loan in the Lending Club will be graded based on its risk. For example, loan graded by ‘A’ means it has the lowest risk.

So, from the following pie chart, we can tell most of the loans are among grade A, B, and C. Only ¼ of the loans are graded by D, E, F, and G. So, the proportion of ‘bad’ loan is not too big.

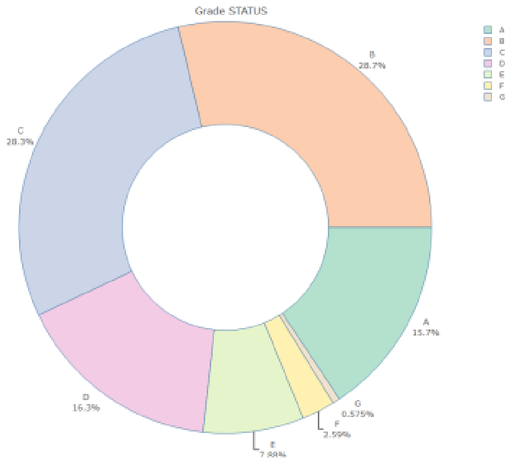


Figure 2 Grade Status

The Figure 3 is about Geographic Distribution of Default Rates. From this figure, we could clearly check each state's default rate. Interestingly, we notice that Nevada has the highest default rate. As all we know, the world-famous gambling-town Las Vegas is located in Nevada. We guess maybe there are some relationships between the city and Nevada's default rate.

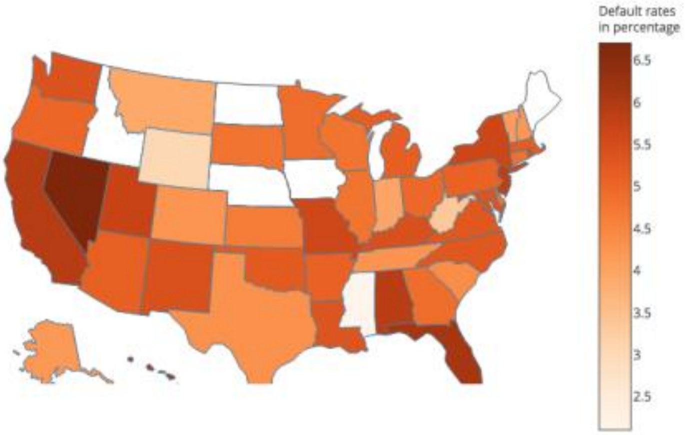


Figure 3 Geographic Distribution of Default Rates

The Figure 4 is the Distribution of Funded Amount by Grade. Obviously, as the grade level decreasing (A → G), the peak of the corresponding distribution shift to right. What's more, the right tail becomes 'fatter'. These two features show the relation between the loan amount and its grade. Generally speaking, as loan amount increasing, the grade level will be lower.

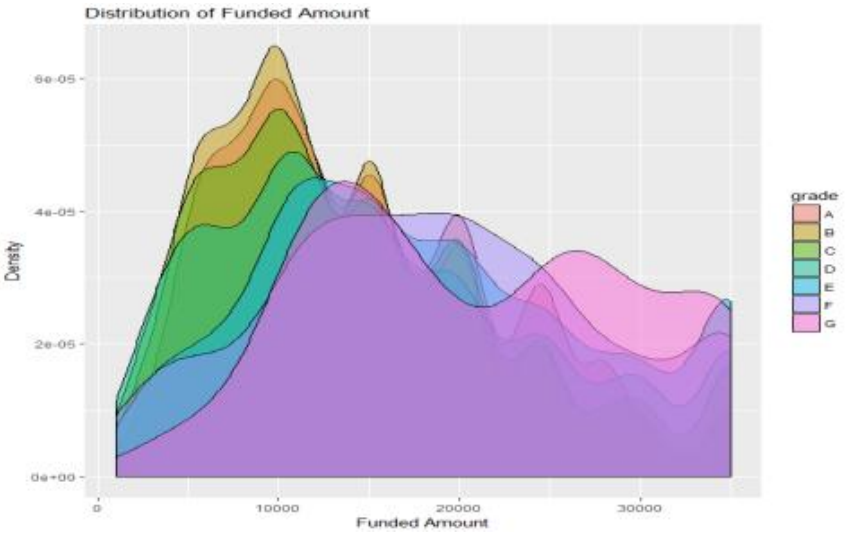


Figure 4 Distribution of Funded Amount by Grade

Methodology

‘Default’ is one of the final statues of a loan. Meanwhile the payment record of every month indicates the behavior of every account. Lending club is possible to detect the likelihood of default earlier. In this case, a new variable called ‘bad behavior’ is created. The loan status such as ’Late(16-30 days)’, ’Late(31-120 days)’, ‘Default’, and ‘Charged Off’ will be considered as ‘bad behavior’. Despite of this, the account has paid late fee record will also be treated as ‘bad behavior’.

Cleaning the Dataset

First, we remove the variables with Null value percentage larger than 10%. Then, we use an R package named ‘caret’ to clean the dataset more deeply, which include 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.

After the cleaning steps above, there are 27 variables left in the subset.

Prediction of Bad Behavior

The 'bad behavior' variable has two levels, one indicates it is a bad behavior, the other indicates it is not. Obviously, it is a classification problem. In the following, we use lasso to select features, then applied logistic regression and tree method to classify it.

Lasso

Lasso is a regression analysis method that can shrink the coefficient estimates towards zero. In order to obtain a simpler and more interpretablemodel, we first applied the lasso to perform variable selection. The model generated from the lasso selected 8 variables fordetectingbehaviors: funded_amnt,int_rate, installment,out_prncp,total_rec_prncp, total_rec_int, recoveries,last_pymnt_d.

Logistic Regression

Consider in this case, where the response 'bad behavior' falls into one of two categories, Yes or No. Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category.

In logistic regression, we use the logistic function,

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

Using those variables selected by lasso as predictors, we use maximum likelihood to fit the model. We separated the dataset into 80% as training set and 20% as test set. In prediction, we set the cutoff as 0.5 and calculate the confusion matrix:

	NOT Bad Behavior	Bad Behavior
Pred NOT Bad Behavior	7655	122
Pred Bad Behavior	1	185

As a lending company, the Lending club does not that much care how much the accuracy the model can achieve. The 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 is 60.26% and the precision is 99.46%.

The sensitivity is not high, so we want to improve it. At the same time, we don't want to lose too much precision. We plot the relationship between the value of cutoff, and precision and sensitivity (Figure 5).

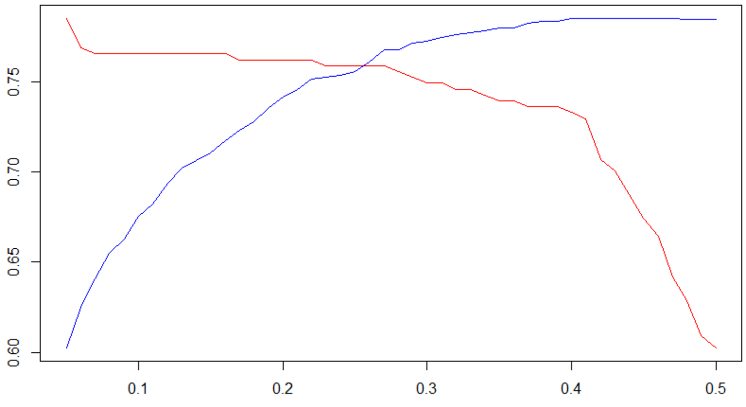


Figure 5 relationship between the value of cutoff, and precision and sensitivity

In Figure 5, the blue line indicates the change of precision and the red line represents the change of sensitivity. As the value of cutoff goes up, the precision increases and the sensitivity decreases. We add them up to find the maximum point, which means the sensitivity and the precision are high at the same time. Figure 6 suggests that when the cutoff equal to 0.4, the result will be better.

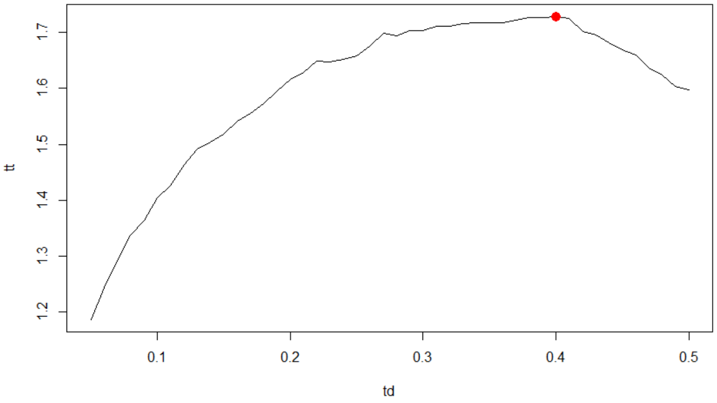


Figure 6 relationship between the value of cutoff and the sum of precision and sensitivity

Using maximum likelihood to fit the model and made predictions on the test set, we gained the confusion matrix as follow:

	NOT Bad Behavior	Bad Behavior
Pred NOT Bad Behavior	7655	82
Pred Bad Behavior	1	225

The sensitivity of the model is 73.29% and the precision is 99.56%. Both of them increase, comparing to the previous model. The ROC curve of the model is shown in Figure 7 and the area under ROC curve is 0.92. We would consider it as a good model.

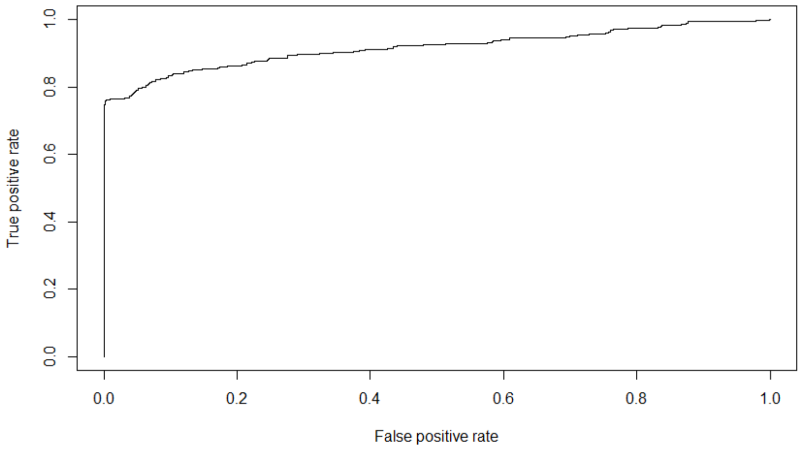


Figure 7 ROC curve

Tree and Random Forests

Tree-based methods can be applied for regression and classification. These involve stratifying or segmenting the predictor space into a number of simple regions. In this case, we applied tree on classification problem (James, 2013).

Using all the variables in the dataset, separated 80% as training set and 20% as test set, the tree model can be seen in Figure 8. The sensitivity of the tree model is 59.9% and the precision is 100%. The confusion matrix of the tree model:

	NOT Bad Behavior	Bad Behavior
Pred NOT Bad Behavior	7540	184
Pred Bad Behavior	0	275

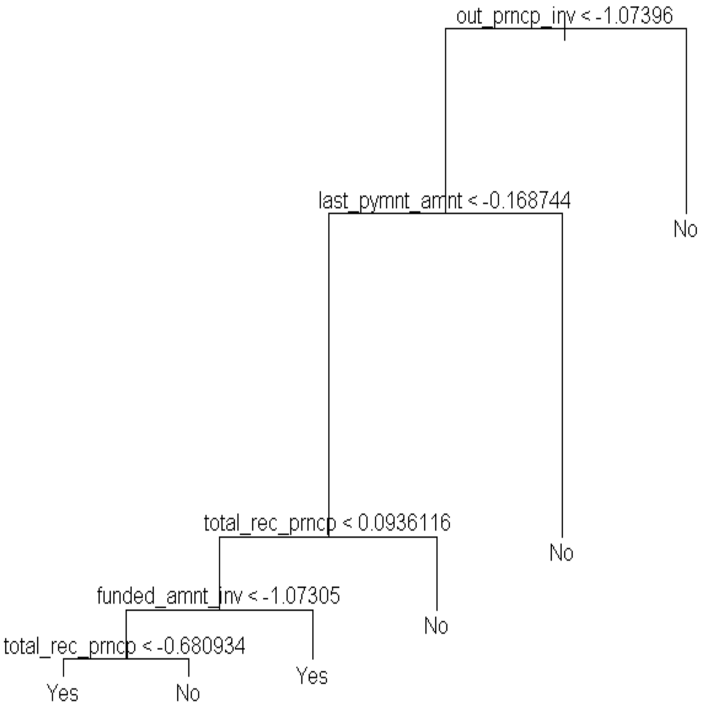

out_prncp_inv < -1.07396
last_pymnt_amnt < -0.168744
No
total_rec_prncp < 0.0936116
No
funded_amnt_inv < -1.07305
No
total_rec_prncp < -0.680934
Yes
Yes
No

Figure 8 Tree for predict bad behavior (before pruning)

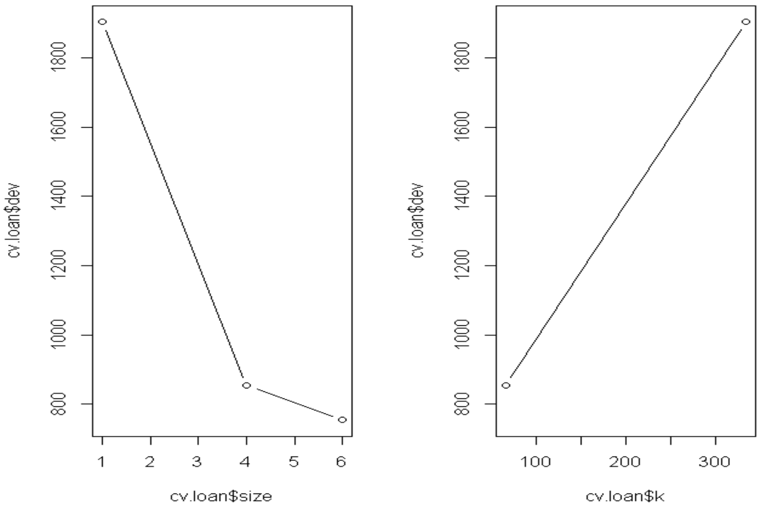


Figure 9 cross validation to choose the tree size

As mentioned before, the Lending club focuses on how much bad behavior the model can predict. The model doesn’t perform well here. Since our cleaned dataset has 27 variables, using all of them to build the model is likely to result in overfitting problem. Then we consider to prune the tree. We then built a large regression tree on the training data and varied values, in order to create subtrees with different numbers of terminal nodes. Finally, we performed six-fold cross validation and find out the tree size equal to 4 is a good option (Figure 9).

We build the pruning tree model and gained the confusion matrix as follow:

	NOT Bad Behavior	Bad Behavior
Pred NOT Bad Behavior	7492	160
Pred Bad Behavior	49	299

The sensitivity is 65.14% and the precision is 85.92%. Comparing to the previous tree model, the sensitivity increases and the precision decreases. Both of them are more acceptable now. The pruning tree model is shown in Figure 10.

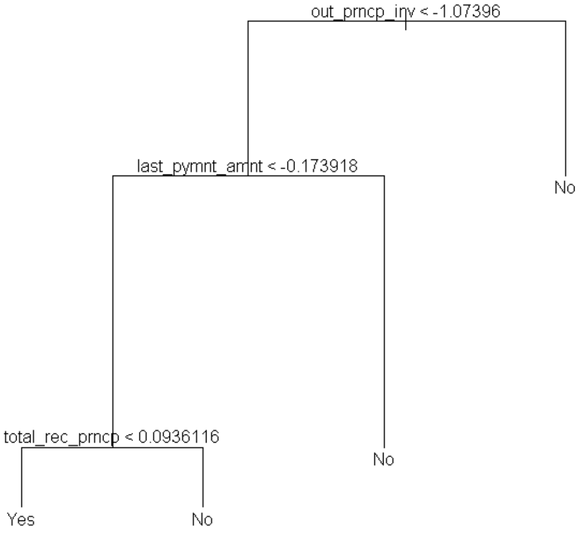


Figure 10 Tree for predict the bad behavior (after pruning)

Random forests provide an improvement of decision tree by way of a random small tweak that de-correlates the trees. When building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors, and a fresh sample being generated every time (James, 2013).

Applying random forest in this case, the confusion matrix of this model can be seen as below:

	NOT Bad Behavior	Bad Behavior
Pred NOT Bad Behavior	7540	126
Pred Bad Behavior	1	333

The sensitivity is 72.55% and the precision is 99.7%. Both of them increases, comparing to the previous pruning tree model. Indeed, random forests method is a more powerful prediction model.

Prediction of Loan Grade

The grade is a qualitative variable which has 7 levels: A, B, C, D, E, F,G. The loan graded as A indicates that the loan has a high grade and relatively low risk. The loan graded as G indicates that the loan is highly risky. Each grade is divided into several subgrade, but in this project, we are only interested in predicting the loan grade rather than subgrade. Due to the fact that each subgrade is correspond to a specific number of interest rate, we need to remove interest rate this variables when we apply statistical methods on predicting the loan grade. We want to find out if there are any significant variables other than interest rate could predict the loan grade.

Lasso

Before applying classification methods to predict the loan grade, we fit the lasso to select the predictor variables. The lasso results in a model that contains 11 highly related variables: home_ownership, purpose, term, emp_length, verification_status, loan_status, dti_n, pub_rec_n, initial_list_status, delinq_2yrs_n, inq_last_6mths_n

KNN

K-Nearest Neighbors is an approach that predicts the class of a given test observation by identifying the observations that are nearest to it. KNN method is sensitive to the scale of different variables. Any variables that are on a large scale will have a much larger effect on the loan grade between the observations than variables that are on a small scale. Therefore, we normalized all variables so that they will be on a comparable scale.

We used 80% of the data to train KNN classifier and the rest we use to test our predictions. The accuracy rates of KNN classifier with different values of K are as follow: (Figure 11)

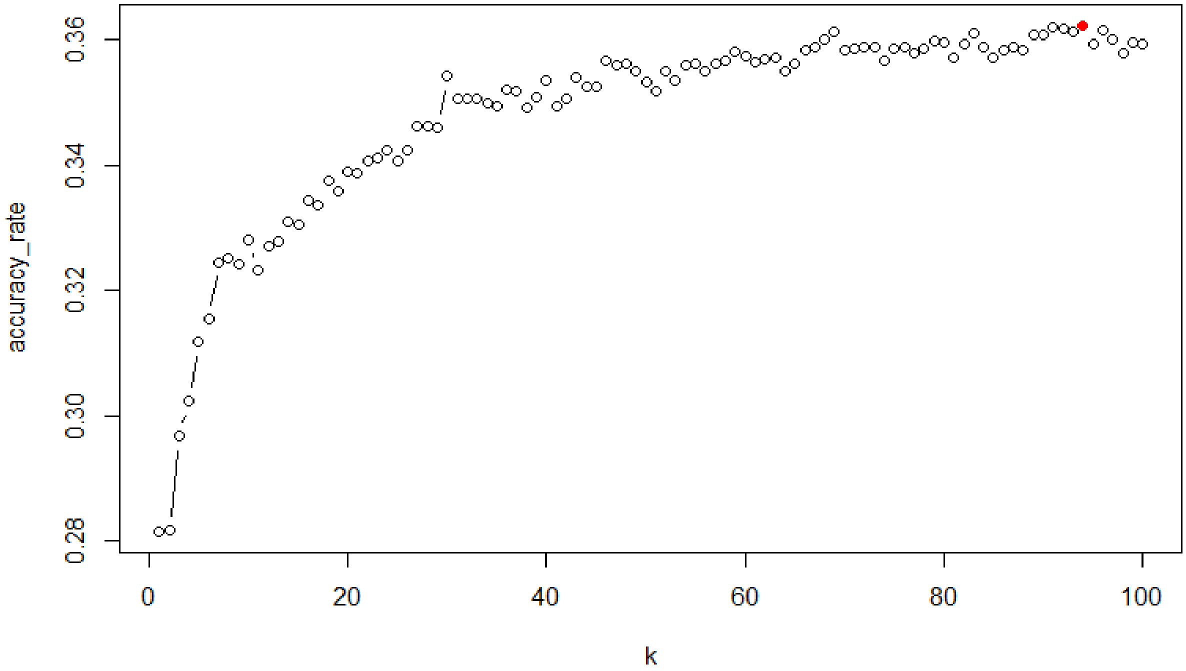


Figure 11 The accuracy rates of KNN classifier with different values of K

From the Figure 11 we can see that the accuracy rate increases as the value of K increases. From K = 32, by increasing K, the accuracy rate has improved only slightly. Increasing K further turns out to provide no further more improvements. Given the computational cost, instead of choosing K with highest accuracy rate, we select 32 as the optimal value for K. The accuracy we got is 35.25%. The results are not very good, since only 35.35% of the observations are correctly predicted. The confusion matrix for KNN using K=32 is as follow:

KNN pred		True Grade						
		A	B	C	D	E	F	G
	A	451	407	188	64	13	4	1
	B	624	1202	903	438	96	29	3
	C	187	567	925	629	316	107	22
	D	14	78	192	199	121	48	19
	E	2	11	26	47	42	17	2
	F	0	0	0	0	3	1	0
	G	0	0	0	0	0	0	0

Next we decided to classify 7 loan grades into 2 categories: high and low. We grouped A, B as high loan grade, and grouped C, D, E, F, G as the low grade. Then we fit a KNN model again, and evaluate its performance on the test data. The accuracy rates of KNN classifier with different values of K are as follow: (Figure 12)

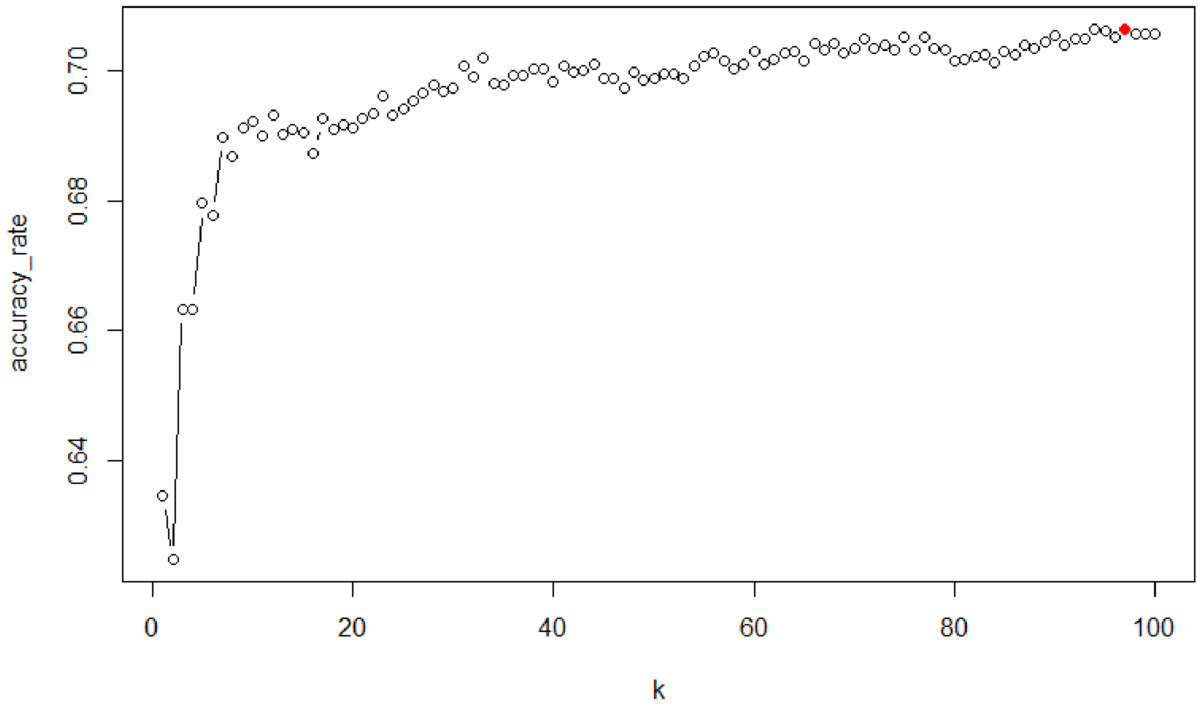


Figure 12 The accuracy rates of KNN classifier with different values of K

The accuracy rate improved a lot after we grouped the 7 grades in 2 grades. For the same reason, there is not much benefit in using higher value for K to get the highest accuracy rate. We chose K = 35 as the optimal value for K. The accuracy rate increased to 69.81%. If we predict high/low grades instead of predicting 7 grades, we get much better results: we are correct for about 70% of loan grade. The confusion matrix for KNN using K = 35 is as follow:

	High	Low
Pred High	2534	1405
Pred Low	1010	3051

LDA

We divide the cleaning dataset into two part. The 20% part is our testing dataset and the other part is our training dataset. Then we train the LDA model to distinguish loan into 7 grades. After that, we test the model by using the testing dataset. The final accuracy we got is 43.05%. The confusion matrix for LDA is as follow:

		True Grade						
		A	B	C	D	E	F	G
LDA pred	A	720	302	112	19	6	1	0
	B	770	1719	954	263	57	6	0
	C	80	811	1423	764	313	77	7
	D	6	58	245	271	188	72	9
	E	0	9	93	129	110	52	7
	F	1	2	18	65	69	47	18
	G	0	2	22	36	32	20	15

Tree (Rpart and Random Forests)

Ordinary tree in our textbook can just classify 2 levels variable. Our target variable ‘grade’ has 7 levels: A, B, C, D, E, F and G, so we used Rpart tree instead of ordinary tree. The Rpart tree is also a tree-based method which can classify multiple levels. The model yielded by the Rpart gave us a very complicate tree, which classified 7 levels grade. (Figure 13)

We then divided dataset into 80% training and 20% testing subset. The prediction with testing subset gave us a confusion matrix.

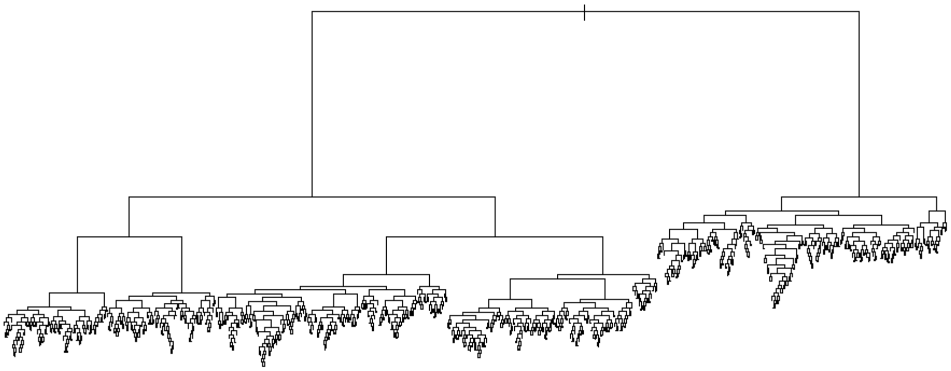


Figure 13 Therpart result

		True Grade						
		A	B	C	D	E	F	G
Rpart pred	A	1977	851	241	59	17	3	0
	B	963	3296	1290	315	95	10	2
	C	204	1355	3127	1014	259	56	7
	D	37	224	869	1400	505	101	14
	E	4	23	133	357	615	216	34
	F	0	4	25	34	78	124	39
	G	0	0	0	3	5	8	7

The Rpart's confusion matrix showed that the accuracy of prediction is 52.7%, which increased a lot compared to KNN and LDA classification. Pruning was applied and the result showed the accuracy improved to 53.2%.

We grouped 7 grade into two classes when performing KNN. Here we used the same method to classify 7 grade into high grade and low grade, and applied tree to predict the loan grade. 80% training and 20% testing subset are used for this method. The prediction with testing subset gave us a confusion matrix.

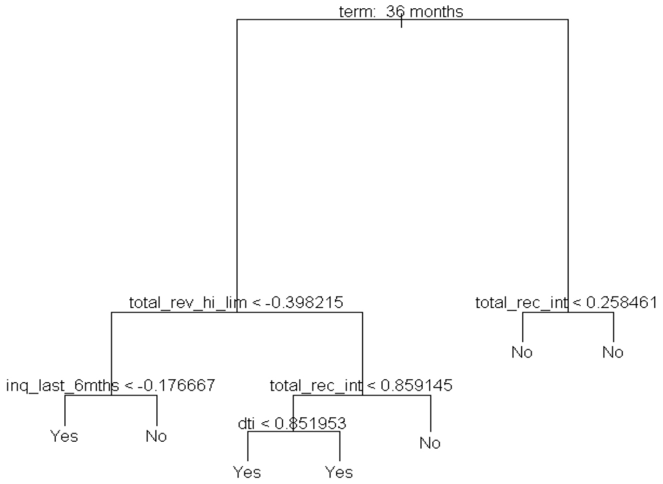


Figure 14 The 2 levels grade classification result

From Figure 14, the 2 levels grade classification result of full tree is a simpler tree. The important variables here are term, toal_rev_hi_lim, total_rec_int, inq_last_6mths. Those are similar to the result of 7 levels grade classification of random forest.

	True Low	True High
Pred Low	2935	823
Pred High	1495	2747

The full tree confusion matrix showed that the accuracy of prediction is 71.0%, which increased a little bit compared to result of KNN.

Pruning was applied, and the figure of deviation with size is as follows:

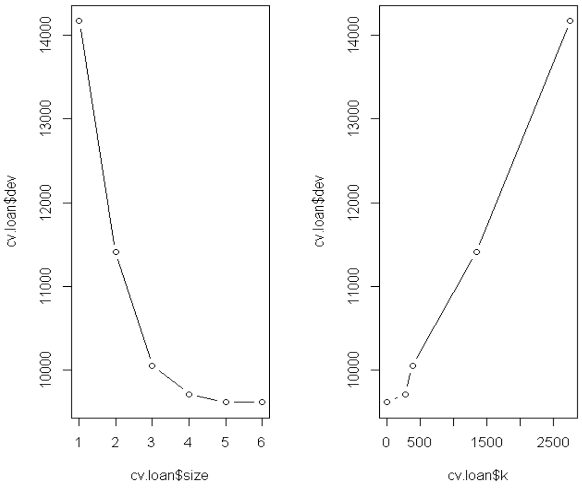


Figure 15 the cross valuation result

Figure 15 showed that when size=5 or 6, the deviation is minimum. We choose size=5 as the best size to prune the full tree. The accuracy did not change, but the tree became a little simpler.

Lastly, we applied random forest to classify 7 levels grade.

		True Grade						
		A	B	C	D	E	F	G
RF pred	A	720	302	112	19	6	1	0
	B	770	1719	954	263	57	6	0
	C	80	811	1423	764	313	77	7
	D	6	58	245	271	188	72	9
	E	0	9	93	129	110	52	7
	F	1	2	18	65	69	47	18
	G	0	2	22	36	32	20	15

By using random forest, we get the accuracy of 59.3%. And the important variables are term, total_rev_hi_lim, revol_util, last_pymnt_amnt and total_rec_int.

We also performed random forest to classify 2 levels grade. By using random forest, we get the accuracy of 88.0%. And the important variables are term, total_rev_hi_lim, purpose and total_rec_int.

	Low	High
Pred Low	3917	451
Pred High	513	3119

Results & Discussion

Prediction of Bad Behavior

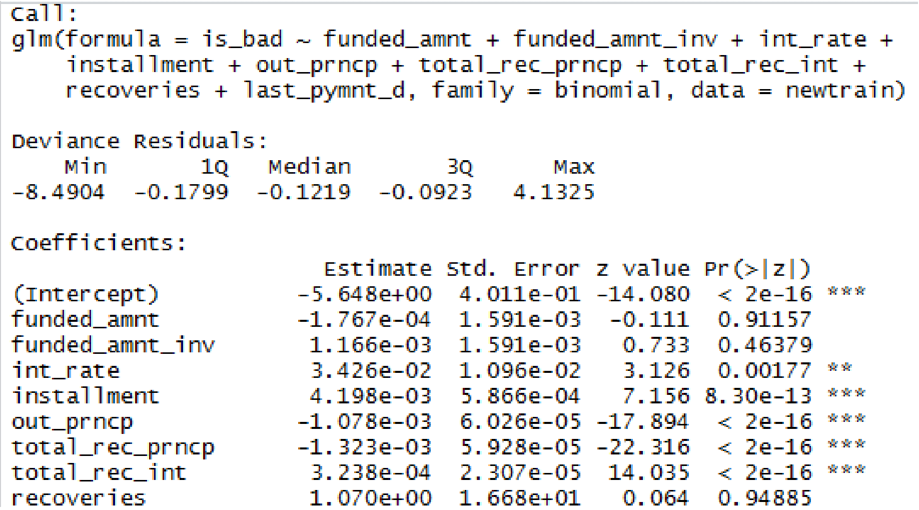
We applied logistic regression, tree and random forests on this problem. Since Lending club focus on how much bad behavior we can predict, we consider the model has the highest sensitivity performs the best. Comparing those models (Table 1), we find that logistic regression has the highest sensitivity.

	Logistic Regression	Tree	Random Forest
Sensitivity	73.29%	65.14%	72.55%
Precision	98.96%	85.92%	99.70%
Selection	✓	✕	✕

Table 1 *Models comparison*

In figure 16 is a screenshot of part of the result of the logistic regression model. From the result, the coefficient of installment is 0.004198. When holding all the predictors other than installment constant, per unit increase in installment, the odds of bad behavior go up by 0.42%.

Also, the model shows that funded amount and recoveries are not significant in the model. Recoveries means post charge off gross recovery. The result makes sense, because funded amount is set at the very beginning, and most of non-bad behavior accounts do not have post charge off gross recovery fee. They are not significant predictors with respect to tell the likelihood of bad behavior among current accounts.



```
Call:
glm(formula = is_bad ~ funded_amnt + funded_amnt_inv + int_rate +
    installment + out_prncp + total_rec_prncp + total_rec_int +
    recoveries + last_pymnt_d, family = binomial, data = newtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.1799  -0.1219  -0.0923   4.1325

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.648e+00  4.011e-01 -14.080  < 2e-16 ***
funded_amnt          -1.767e-04  1.591e-03  -0.111  0.91157
funded_amnt_inv       1.166e-03  1.591e-03   0.733  0.46379
int_rate              3.426e-02  1.096e-02   3.126  0.00177 **
installment           4.198e-03  5.866e-04   7.156 8.30e-13 ***
out_prncp            -1.078e-03  6.026e-05 -17.894  < 2e-16 ***
total_rec_prncp      -1.323e-03  5.928e-05 -22.316  < 2e-16 ***
total_rec_int         3.238e-04  2.307e-05  14.035  < 2e-16 ***
recoveries            1.070e+00  1.668e+01   0.064  0.94885
```

Figure 16 Part of the result of the logistic regression model

What's more, given the other variables, the probability of default 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 the Lending Club issues the loan, it asks higher interest rate for people with lower grade. These people are 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 probability of default will decrease significantly with the improvement of principal received to date and remaining outstanding principal for total amount funded. The more principal a borrower repays until now, the less remaining principal he/she needs to pay. Therefore, the default risk of this loan will be lower. 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.

Prediction of Loan Grade

	7 Grades Accuracy	2 Grades Accuracy	Selection
KNN	35.25%	69.81%	✕
LDA	43.05%	/	✕
Rpart	53.2%	/	✕
Tree	/	71%	✕
Random Forest	59.3%	88%	✓

Table 2 Modes comparison

We applied KNN, LDA, Rpart, tree and random forest to predict the loan grade. The performance of each classification methods is assessed by the accuracy rate. Among these different methods, random forest outperformed the other methods, it has 59.3% accuracy rate in the 7-grades setting and 88% accuracy rate in the 2-grades setting. Each methods performed better in 2-grades setting, which indicates that we are more confident about the prediction of 2 levels loan grade than the prediction of 7 levels loan grade.

Using random forest to predict the 7 levels loan grade, we find out that the most important variables are term, total_rev_hi_lim, revol_util, last_pymnt_amnt and total_rec_int. In this case, we using one borrower's existing loan in other bank or Lending Club to predict his next loan's grade in the future. For one person, when his/her existing loan's term, total revolving high credit limit, the amount of credit the borrower is using to all available revolving credit, interest received to date, last total payment amount received are lower, his/her grade in the next loan will be higher. This indicates that the debt pressure of the borrower will be lowerwhen existing loan term is 36 other than 60, existing credit limit and using credit amount are lower, so it will be more likely to get higher grade (A or B) in the next loan. His/her low interest received and last payment received indicate low existing credit amount and low interest rate, thus thisborrower has higher quality.

Conclusion

In the first experiment, we try to predict potential default clients based on if the client has any bad behavior records. We finally choose logistic regression, because of its higher precision and sensitivity. We successfully detected 225 of 307 clients with bad behaviors among 10,000 test set. Considering about 1 million clients the lending club has, this model could save millions of dollars for the Lending Club. In the second experiment, we build a model to float client's interest rate by adjusting clients' grade. We finally choose random forest to accomplish this function, with 59.3% accuracy for 7 grade prediction and 88% accuracy for 2 grades prediction.

In this project, we utilized statistical methods we learned in class including lasso, cross-validation, tree methods, logistic regression, LDA. We also added new metrics and methods. For example, in first experiment, we used the sum of precision and sensitivity metric to evaluate the result of logistic regression to overcome the imbalance of bad behaviors among the whole dataset(10%). We also compared different K value to choose optimal value of K for KNN model. In second experiment, we extended binary tree we learned in class to RPart trees to predict the outcome of 7 levels loan grades.

For future research, to predict potential default customers, we believe that it is meaningful to predict the rest 82 clients with bad behaviors. We think anomaly detection methods can be powerful to solve this problem.

References

James, G., Witten, D., Hastie, T., &Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Landing Club Loan Data. (2016). *Kaggle* [Data file]. Retrieved from https://www.kaggle.com/wendykan/lending-club-loan-data

Machine Learning for Predicting Bad Loans. (2013, Aug 16). Retrieved from http://blog.yhat.com/posts/machine-learning-for-predicting-bad-loans.html