QingheGao
2659024
Wenlin Chen
2664925

Assignment 4
Default prediction mode

2020/06/15

1 Introduction

Loan default prediction is a vital task for banks or financial institutes. In this report we build a default prediction model based on logistic regression. Firstly, 60 features are roughly selected from the Lending club dataset, to solve the imbalance in the loan status ”default” and ”non-default”, 5000 recordings are selected from each status as the training set by downsampling. Secondly, the properties of new dataset is explored by visualization. Features whose values are highly concentrated located, are not convincing for prediction. Thirdly, in the feature selection, only 12 features are left after univariate model selection and information test. And only 4 features left after deleting highly correlated features according to the correlation map. Fourthly, the best model is arrived by stepwise fit function. Fifthly, the prediction accuracy of test dataset is verified by odds ratio, confusion matrix and ROC. At last, the prediction accuracy is improved further when outliers are removed after the diagnostic of residuals. The accuracy score of model arrives 83%.

2 Feature engineering

The data in this project is from Lending Club, which is a peer-to-peer lending company. We download data of 2012-2013 years, which contains 150 features and 188181 observations. And the target feature is loan_status. We set 1 → Defaulter and 0 → Non-Defaulter for prediction.

Furthermore, we deal with the features of dataset. First, we delete all features which have more than 85% missing values. Additionally, we carefully investigate the description of left features[1] and delete the redundant and irrelevant features, for example, *zip_code*. Then, there are 60 features left.

Additionally, we normalize these 60 features. For string and categorical features has been transformed into integers. For example, *grade* features: *A→1, B→2*. And all the missing values are filled by the mean of corresponding features.

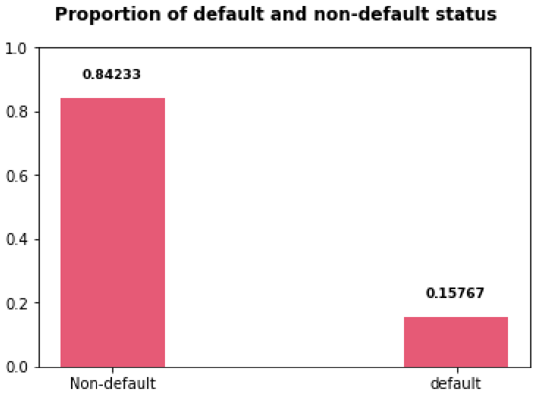


Figure 1: Imbalance proportion of default and non-default data.

After normalization the feature, we explore the whole dataset. We found there are 158510 Non-Defaulter observations and 29671 Defaulter observations, which is a extremely imbalanced dataset. Thus, we perform undersampling the Non-Defaulter data and sample 10000 observations of Non-Defaulter data and 10000 observations of Defaulter data. Finally, we have the dataset with 20000 observations and 60 features.

3 Data visualization

After roughly selecting 60 features and undersamping 5000 recordings from both default and non-default status, data visualization is helpful to explore property of data and for feature selection.

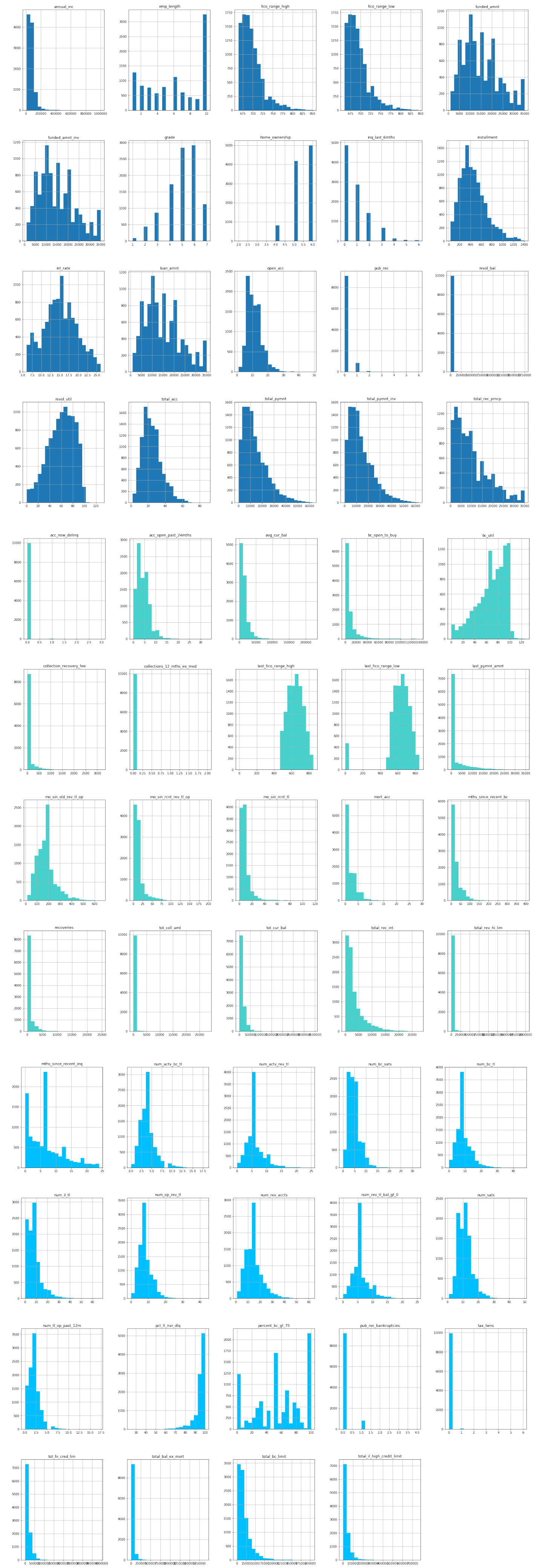


Figure 2: Distribution of 58 features except *Grade and loan status*

From the histgrams of 58 features, there are dozen of features have highly concentrated distribution on certain values, which will makes the evaluation or prediction of model one-sided and incomplete. Like ”*tot_coll_amt*” in the first row and first column (1,1), ”*home_ownership*” in the second row and fourth column (2,3).

Another phenomenon is the long tail in data distribution, like ”*last_pymnt_amnt*” in the sixth row and fifth column (6,5) and ”*total_rec_int*” in the eighth row and fourth column (8,4).

Selecting such features which show such characteristics should be more carefully.

To avoid the effect of loan’s grades for default probability, the figure 3 plot the proportion of grades of loans in default or non-default separately. Fortunately, the distribution of grades has lower density in low grade loans, which may improve the probability of default. Meanwhile, the grade-distribution of default loans seems to be randomly.

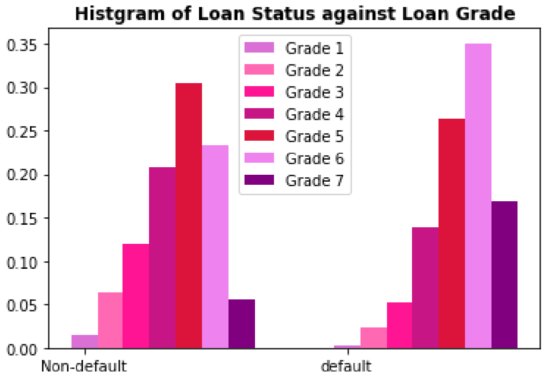


Figure 3: loan grades (1-7) against loan status, where 7 means the best grades scoring.

4 Feature selection

In this part, we do the feature selection for the model

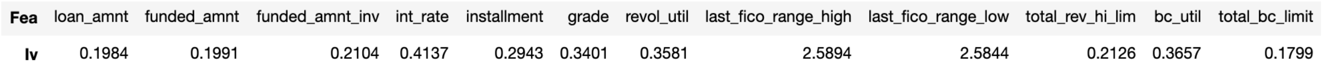
4.1 Univariate feature selection

We set the threshold p-value is 0.2. First, the process will creates 60 logistic regression models and each models use exactly one features. And the importance of features is explained by their individual ability to explain variation in the outcome. Besides, if the p-value is more than 0.2, this features will be deleted. Otherwise the feature would be included in subset features. The detailed results of step selection shows that 56 out of 60 features have been selected and we use these features to do the further feature selection.

4.2 Information value

Information value(IV) is a method that provides a measure of how well a special feature is able to distinguish between a binary response in some target features. And we select the features which IV is more than 0.15.

Figure 4 shows the results of IV. We can see that total 12 features are selected in IV. And two IV values are even more than 2, which shows strong classification power.



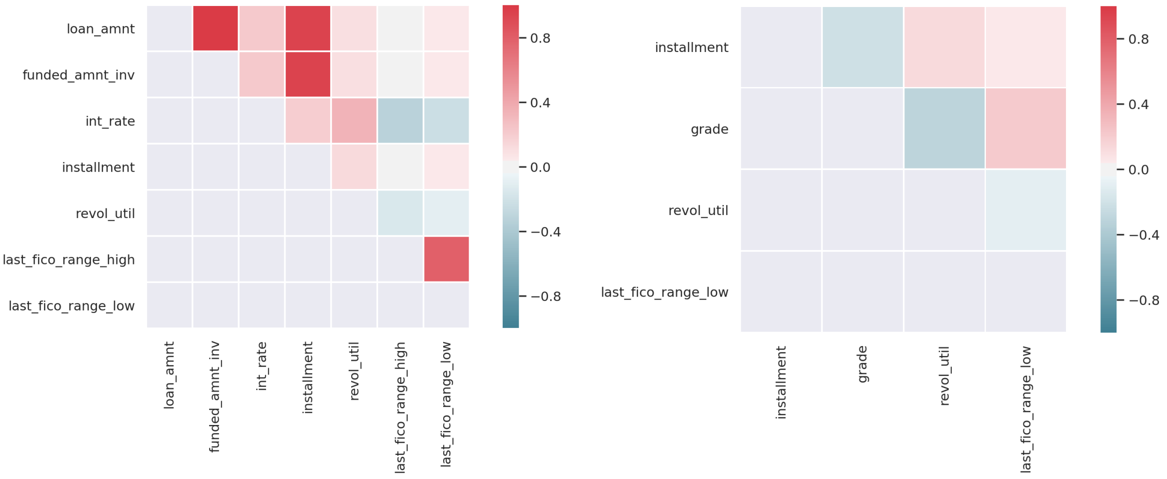
Fea	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	grade	revol_util	last_fico_range_high	last_fico_range_low	total_rev_hi_lim	bc_util	total_bc_limit
Iv	0.1984	0.1991	0.2104	0.4137	0.2943	0.3401	0.3581	2.5894	2.5844	0.2126	0.3657	0.1799

Figure 4: IV

4.3 Step selection and Final feature

Stepwise selection is a automated method of fitting a regression model to the select the predictor variables. We use R package to do Stepwise selection. And we first use 12 features as input. The output of final model are 8 features: $loan\_amnt + funded\_amnt\_inv + int\_rate + installment + revol\_util + last\_fico\_range\_high + last\_fico\_range\_low + total\_bc\_limit$. We eliminate the feature total_bc_limit because its p-value in logistic regression is bigger than 0.05. Thus, we have 7 features in the model. But when we actually predict the result of default we found the performance of model is awful for every evaluation, for example the accuracy is below 50%. Thus, we need to further select the features.

The final step for the feature selection is correlation matrix. We use correlation matrix to do final selection for left 7 features. From figure 5(a) we can see some features have strong correlation with each other. For example, loan_amnt has strong positive correlation with installment and funded_amnt_inv. And last_fico_range_high has strong positive correlation with last_fico_range_low. Thus, the reason of bad performance of model is it contains highly correlated features.Thus we decide to drop the feature funded_amnt_inv, loan_amnt, last_fico_range_high. Then, we finally have four features. Figure 5(b) shows the final correlation matrix, and we can see there are no strong correlation features left.



(a) Correlation matrix before feature selection

(b) Correlation matrix after feature selection

Figure 5: Final feature selection

Table1 shows the final features in the model. Grade is a special features which give the rank of customer based on his/her past history, which A-7,B-6, . . . , G-1. And finally we can use thes six features to fit logistics regression model.

Table 1: Final features in model

Feature	Description	Data type
installment	The monthly payment owed by the borrower if the loan originates.	Float
last_fico_range_low	The blow boundary range the borrower's FICO at loan origination belongs to.	Float
grade	LC assigned loan grade(0-7)	Object
revol_util	The amount of credit the borrower is using relative to all available revolving credit	Float

5 Fitting,Stepwise selection and Prediction

Then we use a final dataset with 4 features and 20000 observations to fit the logistic regression. We split the dataset into 70% training data and 30% testing data. Then we first use cross validation to tune the parameter C, which is inverse regularization parameter. And we find that the best parameter is C=1. Then we use training data to fit the logistic regression and use testing data to get the prediction of the model.

6 Evaluation and diagnostics

6.1 Odds ratio

First we can see that from table 2 the p-value of all coefficients are smaller than 0.05, which means that all the coefficients are valid and significantly different from the 0.

Secondly, we can see the coefficients of grade, revol_util , and last_fico_range_low are negative, which means if all these three feature increase one units the probability of default will decrease. While for installment it is opposite of other three feature. The probability of default will increase when installment(log) increase one unit.

Finally we can interpret the odd ratios. For the installment features, if installment feature increases one unit(log) the odds of defaulting will increase 35%. While other three features are the opposite to installment. For example when grade increases one unit the odds will decrease 21%. This is intuitive because we set low grade as 0. When grade increases the probability of defaulting will decrease. While the change of

Revol_util and last_fico_range_low features has small influence on odd, which means the odds change 1% and 2% respectively when the two features increase one unit.

Table 2: Performance of the model

Variables	Coefficients	Standard error	Z	P-values	Odd Ratios
Intercept	11.25	0.261	43.145	0.000	
log(installment)	0.3030	0.030	10.045	0.000	1.35
grade	-0.2395	0.015	-16.151	0.000	0.79
revol_util	-0.0052	0.001	-6.151	0.000	0.99
last_fico_range_low	-0.0181	0.000	-63.762	0.000	0.98

6.2 Accuracy and confusion matrix

First, we use accuracy and confusion matrix to evaluate the model. Accuracy is the percentage of true prediction out of total sample. We can see the accuracy of prediction is high which is 81%, which indicates the features and model are suitable to predict the loan default.

To further investigate performance of model, we use confusion matrix(figure 6) to calculate precision, recall and F1 score. We can see all three scores are near 81%, which are high. And from figure 6 we can see the main disagreement is from wrong prediction of Will-Default to Will-Pay , which we may increase sample size or select more features to improve.

Table 3: Score of model

Attributes	Accuracy	Precision	Recall	F1 score
Score	81%	81%	81%	81%

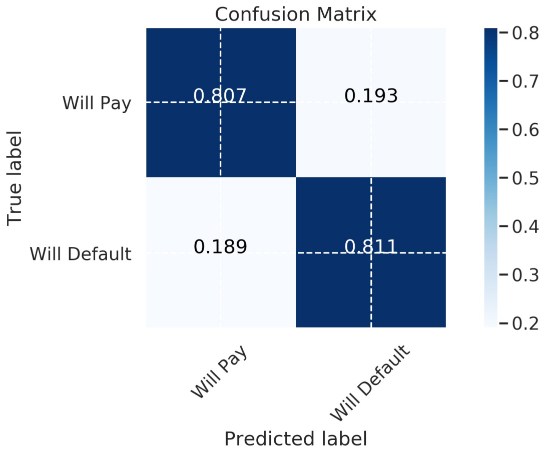


Figure 6: Confusion matrix

6.3 ROC curve

The model also show excellent performance on ROC curve. We can see the area under curve(AUC) is 0.88 which is close to 1. It means our model has the good diagnostic capability of the binary classifier system when the discrimination threshold changes.

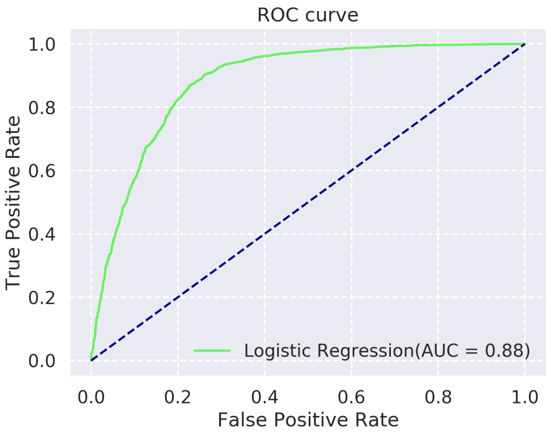


Figure 7: ROC Curve

6.4 Residual plot

In this part, we test the hypothesis of model. Figure 8 shows the deviance residuals plot and QQ plot. We can see most of residuals are in the scale of [-3,3]. But it is clear to see that there are many big negative outliers in the deviance residuals. And also because of these outliers the residual become deviate from normal distribution from QQ plot. Thus, we need to remove the outliers from the data. The basic rule of remove is remove the data which has the residuals are in 95% confidence interval.

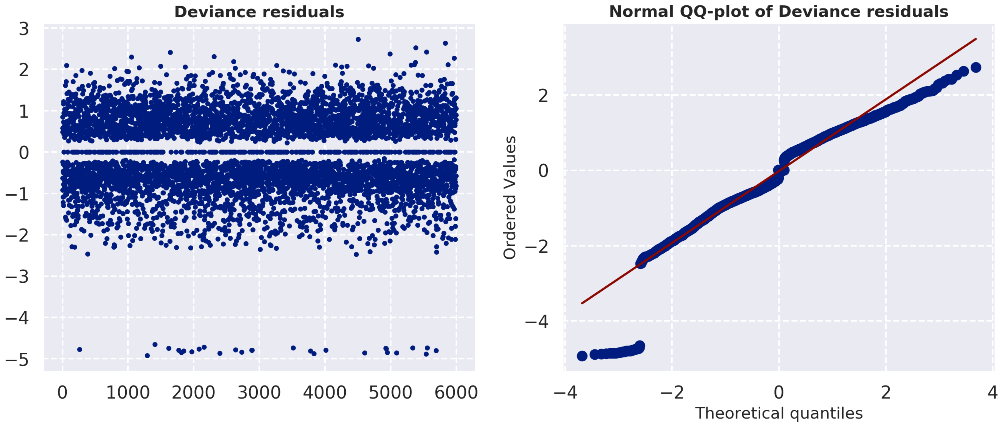


Figure 8: Deviance residuals plot and QQ plot

7 Final model

We remove the outliers of test dataset and fit the model again and we found there are many improvement in the model.

First we recheck the residuals and QQ plot. We found that there is no outliers in the deviance residuals. And the residuals still have small derivation from normal distribution at head and tail part, which may because we only use log function in installment feature. Values of revol_util and last_fico_range_low are too large for other features.

Secondly, we check out the other diagnostics and result shows in table 4. Accuracy, precision,recall and F1 score increase 2%. And AUC from 88% increase 92%. This indicates that the outliers significantly influence the model and prediction. And we can get the experience that next time we fit the model we should check the outliers and residual plot first.

Finally, for the coefficients we see positive coefficients becomes bigger and negative coefficients become smaller when compared with table 2. Especially for installment feature the coefficient changes from 0.3 to 0.6. And odd ratio of installment feature also have great increase, which increase to 1.8. The possible explanation is that the outliers are caused by installment feature and after removing outliers the installment has significatly change.

Table 4: Score of final model

Attributes	Accuracy	Precision	Recall	F1 score	AUC
Score	83%	83%	83%	83%	0.92
Variables	Coefficients	Standard error	Z	P-values	Odd Ratios
Intercept	17.26	0.620	27.834	0.000	
log(installment)	0.6127	0.066	9.295	0.000	1.8
grade	-0.2546	0.032	-7.914	0.000	0.78
revol_util	-0.0119	0.002	-6.224	0.000	0.99
last_fico_range_low	-0.0294	0.001	-38.485	0.000	0.97

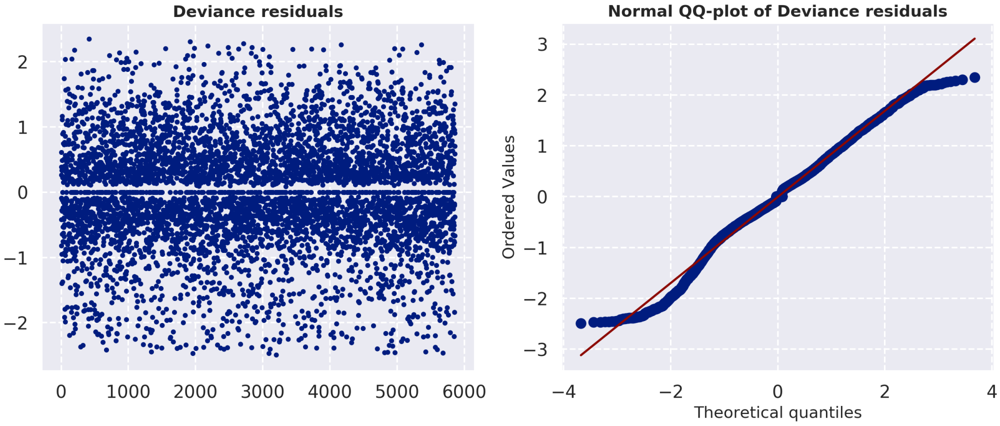


Figure 9: Deviance residuals plot and QQ plot after removing outliers.

8 Conclusion

At the beginning, usually there are many features; univariate model, information value and correlation map are useful tools to select features. And the accuracy of fit model based on stepwise algorithm, ROC, confusion matrix and even p-value are all helpful. However, one improvement in this task could be the expansion of dataset, especially for some highly concentrated features, which may have large effect on the prediction.

References

[1] Lending data description. http://rstudio-pubs-static.s3.amazonaws.com/290261_676d9bb194ae4c9882f599e7c0a808f2.html.