Data Analysis on Lending Club Loan Data ECE/CS 498 DSG Final Project Spring 2020

1st Jiashuo Tong
jtong8
Dept. of Mechanical Science and Engineering in UIUC

2rd Rongqi Gao
rongqig2
Dept. of Statistics in UIUC

***Abstract*—Lending Club is the largest peer-to-peer lending company. It has released the Lending Club Loan Data which contains more than 2 million loan records. We have developed a classification model for this data set. A feature selection strategy involving weight of evidence, information value and Lasso regression is designed to reduce the number of features from 144 to 17. For imbalanced response, down-sampling was used to improve the model performance. For multinormal classification, decision tree and random forest were used, and the test accuracy is about 0.9. For binary classification, random forest, gradient boosting with decision tree, ridge logistic regression and SVM were used. After tuning parameters, the highest AUC is 0.88 with gradient boosting with decision tree method and the average AUC is around 0.8, which indicated the effectiveness of the features selection and model selection. The recommended features are the financial background and credit records,e.g., income, debt ratio (debt amount/total income), credit score (grade and sub-grade), whether payment is made on time and any settlement negotiation records. Spark was used to dealt with the huge data set.**

For NPL ratio analyzing, we first find out some important variables through visualizing, calculating the condition probability and the work in prediction loan status. We leveraged Random Forest model to have a better understanding about how each classification node generated. Finally, we find out the model results is reasonable and consistent with the domain knowledge.

Key Words: Financial risk, weight of evidence, information value, machine learning, AUC, Spark, NPL

I. Introduction

Lending Club is one of the world's earliest and largest peer-to-peer lenders which allow the borrowers to create unsecured personal loans. Because the history of peer-to-peer lending is relatively short [1], the industry lacks a complete picture of the customer groups, and a mechanism that reduces the risk of bad debts. The problems have motivated us to seek solutions from the Lending Club Loan Data, a data set that is released by Lending Club and contains information of all its loans issued from 2007 to 2015. Specifically, we have created a classification model that predicts the borrowers' loan status using their payment information.

There are several Github projects using the same data set [2, 3, 4]. Husted [2] does extensive exploration on the data set. However, the model, with all the features involved, only achieves 0.77 precision and 0.64 recall on the balanced data set (with equal amount of default and ́fully paid́loans), and 0.76 precision and 0.80 recall on the original unbalanced data set. Thus, this model needs improving as it computes with a massive volume of redundant data while achieving a moderate performance. Corliss [3] directly drops the features with more than 30% missing rate, and the features that are unlikely to be available to potential investors. Such decisions are risky because the dropped features can include important information for predicting the loan status. In addition, he takes accuracy as the single metrics for evaluating the classification model, thus failing to demonstrate its performance based on other essential metrics, e.g., precision and recall. Mandge [4] achieves impressive accuracy, precision, and recall (close to 100%) for both the random forest classifier and the gradient boosting classifier which are implemented with PySpark. Unfortunately, the consideration for feature selection is missing from the data modeling process, resulting in lengthy training time.

We try to overcome the limitations of existing models as described above. One characteristics of our work is the integration of an elaborate feature selection scheme into the model. The scheme has ensured reduced computational cost while keeping the predictive information at its finest. Another advantage of our work is the use of multiple evaluation metrics, including accuracy, precision, and recall. It guarantees the usefulness of our model. Besides, we have done a brief EDA in the beginning of the project to understand how the issued loans are related to the borrowers' income, purposes, etc.

II. Methods

A. Data set

Our dataset is released by Lending Club, which is an American peer-to-peer lending company. The data set has a total of 2260668 records, and as many as 145 features. Looking into the excel file, it contains complete loan data for all loans issued through Lending Club from 2007 to 2015. including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the ”present” contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others.

B. Features

For this project, our goal is to detect the default and find the possible reason of abnormal behavior, that is we are pursuing a model with high prediction accuracy as well as interpretable. Therefore, feature engineering addresses the most of our concern of the project. For the 144 variables in the raw data, we first removed features with 99% of missing values and we got 99 numerical features and 28 categorical features left. Then the features selection of weight of evidence (WOE) and information value (IV) is used to find which features is significant to the response [5] .

To calculate the weight of evidence and information value, the response needs to be transformed into binary form. Since we are interested in the abnormal observations, the 9 level of ”loan status” is divided into two groups with the tag “1” of abnormal observations. Basically, weight of evidence is a quantitative measure of how much information dose the feature provided for the tag “1”. And that measures is estimated by mutual information. Suppose $\boldsymbol{X_i}$ is the $i-th$ level for a feature, the mutual information of $\boldsymbol{X_i}$ and our interested tag ”1” will be as follow.

$$I(Y=1:\boldsymbol{X_i}) = \log \frac{Pr(Y=1 \mid \boldsymbol{X_i})}{Pr(Y=1)}$$

Then calculate the mutual information between $\boldsymbol{X_i}$ and the rest response of y (i.e. tag ”0”), the difference between these two mutual information is intuitively the contribution to tag ”1” provided by $\boldsymbol{X_i}$, which is call weight of evidence.

$$\begin{aligned} WOE(\boldsymbol{X_i}) &= I(Y=1:\boldsymbol{X_i}) - I(Y\neq 1:\boldsymbol{X_i}) \\ &= \log \frac{Pr(\boldsymbol{X_i} \mid Y=1)}{Pr(\boldsymbol{X_i} \mid Y\neq 1)} \end{aligned}$$

Finally, the information value is weighted sum up of WOE of all levels of the feature, the weight here is further considered the amount observations of each level. Notice that the information may take the value of infinity when that level only contains one kind of of response level, which means this level's contribution to out interested response tag is infinity. The cause of this situation is too many levels or classes of a feature, combination is a useful ways to save you from infinity.

After the information value selection, there is 12 numerical features and 8 categorical features, in case of losing any valuable features, we subjectively added 6 features that might be useful. The total amount of feature is 26 for this step.

Further selection is applied to these features to reduce the multicollinearity, Lasso regression is used for this step. The whole feature selection process is provided in Fig. 1.

C. Missing Values

Missing value is one of the major challenges of this project. Given that the majority clients of lending club is normal, features with lots of missing values might also be important to find abnormal clients. So, before WOE IV selection features with too many missing values is turned into missing non-missing variable. After the selection, for features have with more than 50% missing value, we use -999 to impute the missing value, for the rest features mean is used to impute the missing value.

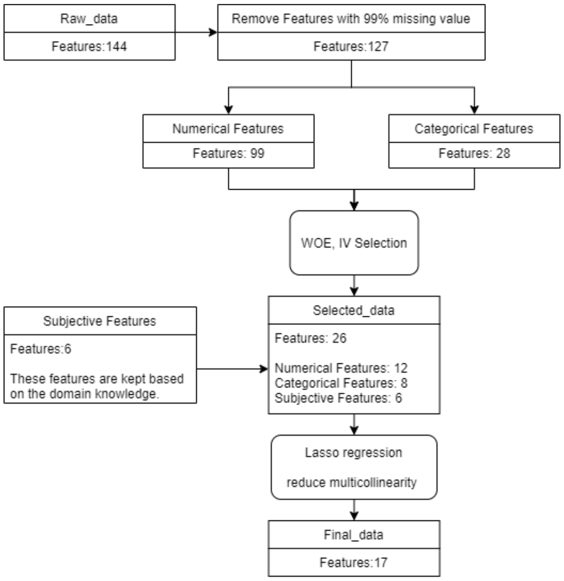


Fig. 1. Flow Chart of Feature selection

D. Imbalanced Data

For binary response, only 13.11% observations is abnormal, which will significantly lower the predicting recall of our model. This is because if we simply assign all the clients are normal, our predicting accuracy will be greater than 85%, yet none of the abnormal clients are detected. To solve this, we introduced the method of down-sampling. The intuition of this method is to reduce the number of normal clients when training the model, as shown in Fig. 2.

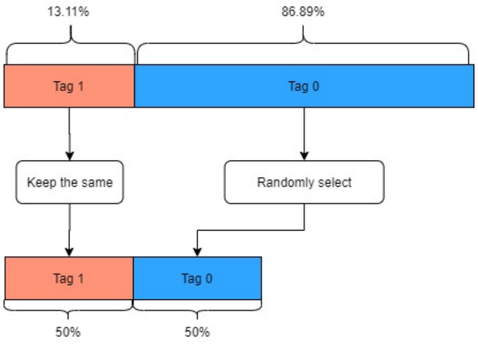


Fig. 2. Down-sampling for Imbalance Data

For verifying the effectiveness of down-sampling method, we compared the test AUC of Random Forest model before and after applying down-sampling. It turns out that before it, our AUC is only 0.56, which basically is random guess.

However, after the process, our predictive AUC is around 0.8, which is much better and acceptable.

E. Model

Predicting the loan status is our first target and it's a classification task. So hereby we will introduce the classification models we applied. Logistic regression is commonly used in binary classification setting because of its efficiency and explicit. Since we transformed our response variable towards binary one, naturally, we first consider the logistic regression method, a type of generalized linear models. For linear regression, we could briefly summarize it as

$$\eta = \beta^T \mathbf{X},$$

where β is the coefficient vector and $\mathbf{X}$ is the predictor vector. Generalized linear regression employs the sigmoid function as an activation function to transform the response variable to the expectation of the response variable. And in binary setting(zero or one), the expectation is actually a probability. We can illustrate it further by the following equation

$$\mu = \frac{1}{1+\exp(-\eta)} = \frac{1}{1+\exp(-\beta^T \mathbf{X})},$$

where μ is the expectation of the response variable.

Random forest is an ensemble method, and could be further categorized as a bagging(bootstrap aggregating) method. Bagging often considers homogeneous weak learners, learns them independently and combines them following some kind of deterministic averaging process. In each iteration, random forest will randomly choose part of the features to fit a decision tree independently. Choosing partial features instead of overall which could decorrelate the decision trees in this random forest.

Different from random forest model, Gradient Boosting Decision Tree(GBDT), our third method, is an ensemble method based on stochastic gradient descent and decision tree. We initialize the model with a constant value. Then in each iteration, we calculate the residuals based on the prior model, followed by fitting a weak learner to the residuals and calculating the gradient descent based on the loss function. In GBDT, the weak learner here is decision tree. Finally, we update the model and loss function according to the equation below:

$$\theta_l = \theta_{l-1} - \alpha_l \nabla_\theta L(\theta_l)$$

where θ_l, α_l are the set of model parameters and the learning rate in this iteration respectively. An appropriate learning rate helps the parameters descend with less oscillation and the loss function approximate the global minimum. One could observe that GBDT learns weak learners sequentially in a very adaptive way. This is a major characteristic of boosting methods, and GBDT is one of them. Stochastic gradient descent is a variant of the gradient descent, which only takes one sample to go through the gradient descent in each iteration. The computational time and cost could be greatly reduced using stochastic gradient descent compared with the regular gradient descent.

Support vector machine(SVM) is a commonly-used method for classification and regression. Constructing a support vector classifier is equivalent to solve the following optimization problem

$$\begin{aligned} &\arg\min_{\omega} \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y_i(\omega \cdot \phi(\mathbf{x_i}) + b) \geq 1 - \xi_i \quad (i = 1, 2, \cdots, m) \\ & \xi_i \geq 0 \quad (i = 1, 2, \cdots, m), \end{aligned}$$

where m is the number of samples, and C is a penalty coefficient. What's more, $\phi(x_i)$ is a function mapping from low dimension to higher dimension. By applying Lagrangian function and dualization, we could have

$$\begin{aligned} &\arg\min_{\alpha} \frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m}\alpha_i \\ s.t. \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \end{aligned}$$

where α is the Lagrangian coefficient vector, and $K(x_i, x_j)$ is the kernel function.

To successfully implement the support vector classifier algorithm, we need to choose an appropriate kernel function, a penalty coefficient, as well as other hyperparameters. We hereby introduce some prevalent kernel functions.
(i) Linear kernel: $K(\mathbf{x}, \mathbf{z}) = \mathbf{x} \cdot \mathbf{z}$, to apply linear kernel, our samples should be linearly separable.
(ii) Polynomial kernel: $K(\mathbf{x}, \mathbf{z}) = (\gamma \mathbf{x} \cdot \mathbf{z} + \mathbf{r})^d$, where $\gamma, \mathbf{r}, d$ are hyperparameters.
(iii) Gaussian kernel: $K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma\|\mathbf{x} - \mathbf{z}\|^2)$, where $\gamma > 0$ is a hyperparameters.
Currently, Gaussian kernel is the most widely used. Because SVM is intrinsically a linear separator when the classes are nor linearly separable we could project the sample data into a high dimensional space and with a high probability find a linear separation. Gaussian kernel does exactly the same thing, projecting the data into infinite dimensions and then finding a linear separation. Polynomial kernel is not the prior choice in most scenarios because of its numerous parameters. Tuning and computing process could be highly complicated. Furthermore, some works comment that when the dataset is very large, it becomes approximately linearly separable, so that linear kernel might be useful on a large dataset.

III. Results

A. Model results and tuned parameters

Model results on the original dataset

Classifier	Cross-Validation Accuracy	Test Accuracy
Decision Tree	88.89%	88.90%
Random Forest	89.01%	88.98%

Model results on the down-sampled dataset

Classifier	AUC
Random Forest	0.7903
Gradient Boosting with Decision Tree	0.8874
Ridge Logistic Regression	0.7625
SVM with Linear Kernel	0.8004
SVM with Gaussian Kernel	0.7890

Tuned parameters of random forest model on the down-sampled dataset

Parameters	Value
Number of Estimators	500
Maximum Depth	11
Maximum Features per Iteration	9
Minimum Sample Split	100
Minimum Sample Leaf	50

Tuned parameters of gradient boosting decision tree model on the down-sampled dataset

Parameters	Value
Learning Rate	0.07
Number of Estimators	500
Maximum Depth	13
Minimum Sample Split	1200
Minimum Sample Leaf	80
Subsample proportion	0.9

Tuned parameters of ridge logistic regression model on the down-sampled dataset

Parameters	Value
Penalty Coefficient	0.0100

Tuned parameters of (linear kernel) model on the down-sampled dataset

Parameters	Value
Penalty Coefficient	0.0160

Tuned parameters of (Gaussian kernel) model on the down-sampled dataset

Parameters	Value
Penalty Coefficient	0.0695
Gamma	0.0001

B. Customer Analysis

From the graphs below, we could get some general insights about the lending club clients.

tag = 0(F́ully Paid)́
tag = 1(D́efault)́

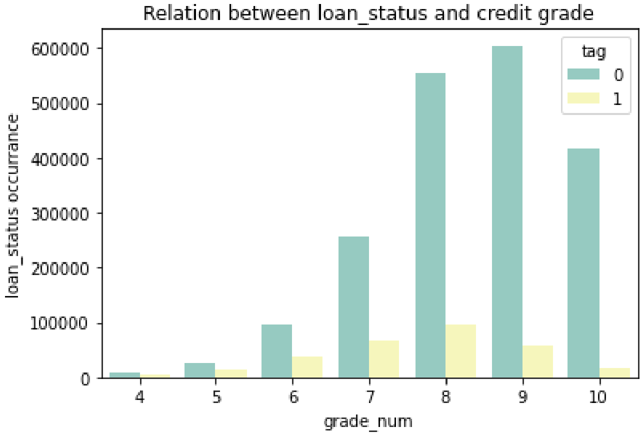


Fig3: Credit grade VS loan status

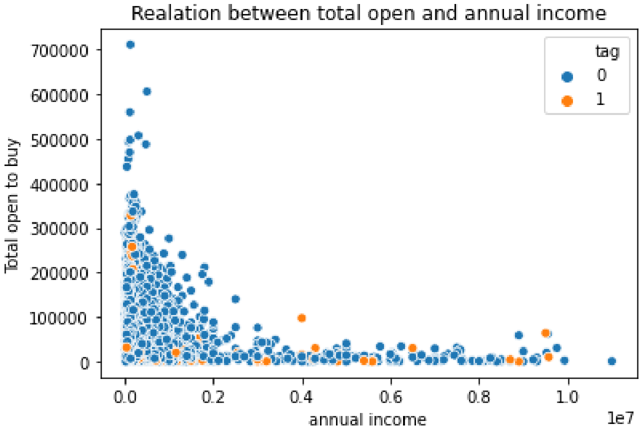
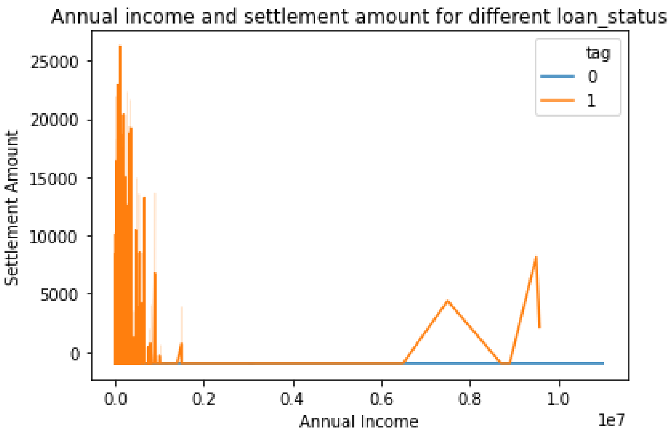


Fig4:Annual inc VS Total open(under different loan status)

Fig5:Annual inc VS Settlement amnt(under different loan status)

From the fig3 we could know, the credit grade for clients who default are more likely to be grade B,C or D. We could infer that few applicant with grade below D will be approved unless they have sufficient evidence to proof their ability to [ay fully]. Client with grade A have benign records and tends to



keep it. However, for clients with moderate credit grade, there might be some reasons to prevent them from paying on time, like unemployment, soar expenditure and etc which become potential default risk. These risk will not influence their credit score instantly and could not be easily detected by the lending club.

From the fig4, the range for total open to buy is decreasing as annual income increase. Default loan status usually have small open to buy for each income level. From the fig 5, we could know that fully paid client do not have settlement need to be negotiated. For those could not fully paid their debt, finally they might undergoing a settlement negotiation stage. With the increase of annual income, the settlement amount decrease first and increase. For those who have higher income, they are able to borrow larger amount of money from lending club. For those who have lower annual income, they might not be capable to pay back the money with their salary on time.

These insights about lending club clients help us know more about the relationship among variables which is beneficial for further research.

C. NPL Ratio Analysis

Here we focus on the severity the default and delinquency might be. NPL, non-performing loan, is an important risk measurement for a financial company. It represent the amount of loan which is related to default. .We would also explore the NPL ratio(proportion of bad debt) to help Lending Club control their risk. We used delinquency loan amount divided by loan amount to represent the NPL ratio.

Exploratory Data Analysis Through exploratory data analysis we could find out the following relationship between NPL ratio and other key dependent variables.Next, we applied the Random Forest Model and finally find out the model we used to do the classification leverage the relationship feature to classify the records.

last pymnt	Average NPL Ratio
Received	0.305276
Not Received	0.426260

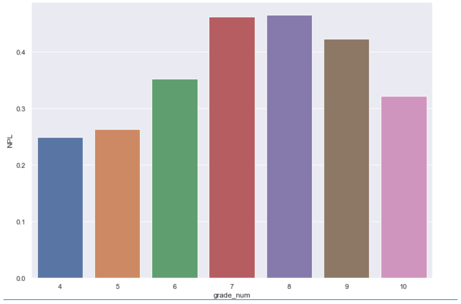


Fig. 6. NPL Ratio Grade number

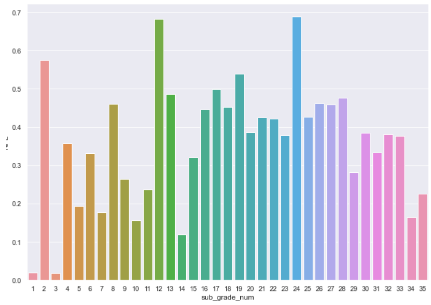


Fig. 7. NPL ratio Sub grade

debt settlement	Average NPL Ratio
Yes	0.423574
Not	0.543953

Intuitively, we assume the NPL ratio is closely related to each borrower's credit grade, whether paid the last due, whether being in a settlement negotiation stage and etc.It is reasonable to include these variables in our classification model. In addition, from the exploratory data analysis results, these intuition could also be verified. From the table we could find the if a borrower did not pay the last due, the NPL ratio is much higher than that of those who pay the due on time. If borrower came across a settlement negotiation process which means he/her is unable to afford the debt payment. Usually, it indicates the default amount accumulated and make borrower hard to afford.

Dividing the not performing loan borrower into three classes as following:

Cluster	Center	Count	Std
0	0.195	6994	0.512
1	7.575	16	11.457
2	28.022	164	3.178

It represent three kinds of customers. First client cluster represents those who forget to pay by incident or just began to default the contract. Second client cluster represents those who could not afford the debt payment gradually, but it is possible for them to pay back all the debt in the future. The third cluster client represents those who will not able to pay in the near future and most possible will come into a debt settlement stage.

Lending club could bear the risk brought from first cluster client to some extent. However, lending club should avoid approving the application of second and third cluster applicant for fear of severe default risk.

Classification Model(Random Forest) With the extraction technique, We ran the Random Forest Model (tree depth = 5, number of tree = 10) on the processed data.

	Pred Class:0	Pred Class:1	Pred Class:2
TrueClass:0	6994	0	0
TrueClass:1	2	14	0
TrueClass:2	37	0	127

Precision is 0.9945. Recall is 1. Accuracy is 0.9946. Although the accuracy, precision and recall are all higher than 0.95, only 87.5 % of cluster 2 and 77.42 % of cluster 3 was detected. What we could do further is to come out with better data division method and collect more data of second and third cluster client to improve our models. For the model part, we are supposed increase the penalty on misclassifying the cluster 2 and cluster 3.

Tree Model Visualization

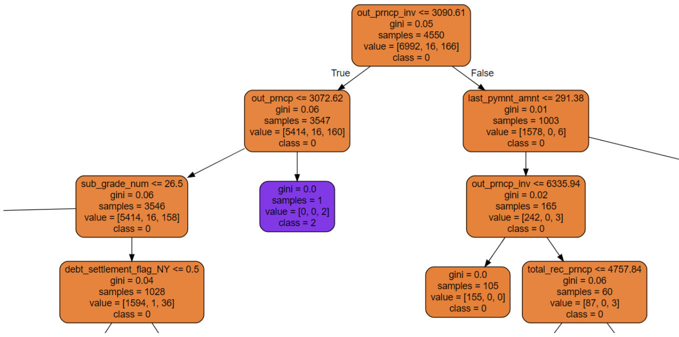


Fig. 8. Visualization of 6th subtree

From the classification results we could find that the first node is related to the remaining debt balance. In addition, the second node contain the division criteria about last due payment which is also consistent with our initial intuition. Credit subgrade and settlement stage are also two top classification criteria in our Random Forest model.

IV. Discussion

A. *Substantiated Conclusions*

We have provided the model results in the previous section. The first table shows the model results on the entire dataset, without down-sampling. We could observe that firstly, the results of decision tree and random forest are very close to each other, which contradicts our knowledge that the ensemble method should perform better than a individual learner. Take a close look at the entire dataset. One could find that it is highly imbalanced that the number of observations with bad loan status(positive one) is much greater than that with benign loan status. Under this specific scenario, the model results are unreliable even if the cross-validation and test accuracy are high. Therefore, we should consider down-sampling the entire dataset to keep the two numbers approximately the same.
The second table illustrates the model results on the down-sampled dataset. Regarding the down-sampled dataset, we instead consider ”AUC” as our metric. AUC is a more general and useful metric compared with accuracy, because it takes both true positive and true negative into account. We should be able to conclude following things from the second table. Gradient Boosting Decision Tree performs best among all models we tried. As an ensemble method, which trains hundreds of weak learners and updates itslef in each iteration, it’s not surprised that GBDT leads a better result than logistic regression and support vector machine. While gradient boosing decision tree took more time to converge and fit, it is actually faster than the other ensemble method - random forest, in our experiment, and we believe the ”sequential training” characteristic should be one of the reasons. GBDT fits the residuals from the previous model instead of the overall dataset to the new learner in each iteration, which might decrease the computation cost. With stochastic gradient descent or mini-batch stochastic gradient descent, GBDT could even converge faster and be closer to global minimum, meaning a better result. It’s a little bit unexpected that the SVM with Gaussian kernel not performs as well as SVM with linear kernel. We have proposed two possible reasons. Firsly, we did not investigate the parameter gamma of Gaussian kernel a lot. We might not tuned the gamma to the best value. Secondly, the entire dataset might be approximately linearly separable, so that the linear kernel outperformed the Gaussian kernel.

Furthermore, from the final 17 selected predictors, we could primarily summarize them as three categories. The first one is the attributes of the loan itself, such as loan amount, current settlement that has been paid. The second one is the outside credit, such as the credit from banks and credit scores given by some financial institutions. The last category is the personal background, like the annual income, percentage of trades that never be delinquent, and total collection amount ever owed. Based on this, we recommended lending club to consider the applicants based on their personal financial background, like income, debt ratio(debt amount/total income) , credit records, like credit score(Grade and sub-grade), whether pay the due on time and any settlement negotiation records.

B. *Limitation and Improvement*

Data Set limitation The number of default especially severe default is much fewer than that of benign records, which is reasonable in real life because the apparently unqualified applicants will not be approved by lending club. As a consequence, it is hard to divide a balanced data set for each classes. It is a problem we should always consider during model training.

Model Limitation Given that the Lending Club Loan data is larger than that in our mini-project or homework, spark is used to save us from endless memory error. However, it turns out that spark might not be a powerful tools for machine learning. The results of this is we also provided lots of our results with Pandas and Sklearn.

V. Member Contributions

Generally, the workload of this task is divided equally for all of us. Specifically, Jiashou contributed more on the environment setting up for pyspark, model selection and writing the report. Yilin focused more on the model results, feature engineering as well as preparing for the presentation. The NPL part is finished by Chengzhuang, also she provided nice charts for our analysis. Rongqi's contribution is also the feature engineering, tuning models and writing the report.

References

[1] A. Bachmann, "Peer-to-peer lending – a literature review," *Journal of Internet Banking and Commerce*, vol. 16, no. 2, pp. 425–466, 2011.

[2] A. Husted, *LendingClub Loan Data*, 2019 (accessed May 1, 2020). [Online]. Available: https://github.com/jalexander03/100119-Lending-Club-Loan-Data/blob/master/ProjectBook.ipynb

[3] J. Corliss, *Predicting Loan Defaults on LendingClub.com*. [Online]. Available: https://github.com/jgcorliss/lending-club/blob/master/predicting_chargeoff.ipynb,year=

[4] R. Mandge, *Lending-Club-Loan-Analysis*, 2018 (accessed May 1, 2020). [Online]. Available: https://github.com/Rohini2505/Lending-Club-Loan-Analysis/blob/master/Code_lending_club/CISC-5950%20Project%20Notebook.ipynb

[5] Y. Wang and A. K. C. Wong, "From association to classification: inference using weight of evidence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 764–767, 2003.