Loan Status Prediction

P12: Yuanming Song, Yuyang Jiang
Department of Computer Science
North Carolina State University
{ysong33, yjiang35}@ncsu.edu

1 Background & Introduction

1.1 Research Background

Peer-to-peer (P2P) lending is a new, online-based financial intermediary connecting people willing to borrow (borrowers) with people willing to lend their money (lenders). Borrowers and lenders are connected through online P2P lending platforms. P2P lending platforms can provide loans with lower middle costs than traditional platforms, such as banks, because of their online presence and the elimination of costs associated with "bricks and mortar" locations with staff. This fact supports the offer of more competitive terms for both borrowers and lenders.

In the days before peer-to-peer (P2P) lending, if you needed money for personal uses or purposes, you had a few standard options: you could rack up your credit card debt, apply for a loan from a bank, or borrow from your family or friends. Each of these approaches carried its risks, difficulties, and complexities.

Back in the year 2007, the LendingClub company, now becomes the world's largest peer-to-peer lending platform, saw an opportunity to disrupt these traditional options by creating a P2P lending platform to directly connect individual borrowers and lenders. Thus, they removed the traditional role of the big bank or corporation (in the case of bank loans and credit cards) while greatly expanding an individual borrower's reach (in the case of family and friends).

A key aspect of LendingClub's model was the frictionless and simple experience for the borrower and lender. The application process for borrowers was straightforward - after providing information such as the purpose of the loan, income information, and information required to retrieve credit report data, the borrower would either be approved or denied. If approved, LendingClub would assign a credit rating ranging from A to G, with A representing borrowers with the highest credit quality and G representing the lowest quality. (In the public data set of all loans issued provided by LendingClub, F and G rated loans were discontinued in 2017 due to high default rates.)

An interest rate would also be assigned to the loan commensurate with the loan grade and credit quality. The process was friendly for the lender as well. After creating and funding an account, one could easily browse and invest in any of the hundreds of thousands of loans seeking funding. The lender could search and filter through loans based on aspects such as loan grade, income, debt-to-income ratios, and loan purpose, to find specific loans that fit the lender's investment criteria.

1.2 Research Goal

While LendingClub presented investors with an exciting and novel investment opportunity, one of the key concerns was the risk of default. LendingClub loans were unsecured personal loans, meaning that there was no collateral backing the loan. If the borrower defaulted, the lender would generally lose all remaining interest and principal.

While the investment losses incurred in a loan default could be severe and daunting, it is also well understood that predicting defaults is a task that is well suited for a classification model. Therefore,

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

our primary goal was to train multiple models to accurately predict loan defaults. With a high-performing model in hand, a LendingClub lender would be able to navigate the thousands of loans with more confidence to achieve higher returns with less risk.

Another important goal of our project was to determine how these features help to classify the future status of the loan (charged-off or default). And reach high classification results using appropriate models. Findings from this process would also be critical in further enhancing investment returns while reducing risk by addressing the concerns a discerning investor would ask, such as which loan attributes were most correlated with defaults, what the average interest rates were for each loan grade, and how the quality of loans may have changed over time.

Lending Club specializes in extending different types of loans to urban customers, which is decided on the basis of the applicant's profile. Accordingly, the data considered in this work, contains information about past loan applications and whether they were "defaulted" or "not".

2 Proposed Method

2.1 Data Description

The data set we use is a unique real-world data set comes from LendingClub P2P platform, which includes all accepted loan data with original 151 features and 2,260,701 customers data from 2007 to 2018.

The features includes The number of payments on the loan, Employment length in years, The total number of credit lines currently in the borrower's credit file, Number of mortgage accounts, etc. All our records are matured, i.e. we know their final loan resolution status.

2.2 Data Preprocessing

After going through the data set, we noticed that excluding the id entry, each of the variables in the original data contains missing values, ranging from 33 to 2260701, we remove the variables with more than 30% missing values. After this step, there are 92 variables left (150 in total).

Then, we calculate the correlation coefficient matrix and drew the heat map of the data set (shown in fig 1). To further reduce the dimension of the data set, we removed highly correlated variables — those with correlation coefficients greater than 0.98. They could hardly contribute much too to our model. In this step, we are dropping 7 columns (85 variables left). What's more, we set the correlation coefficients' threshold as a hyperparameter in our models, we will further compare the performance of the model with correlation coefficients > 0.98 and the model with correlation coefficients > 0.95.

According to the data dictionary (downloaded directly from LendingClub's website) and our preliminary observation, we removed extra useless columns which will not help in prediction like URL, zip code, etc. Now we get our selected data set with 64 attributes.

Furthermore, we still need to deal with the existing missing values and categorical features. For the remaining variables, we replaced float attributes with their median value and replaced other attributes with their mode value. For the categorical features, we converted them into numerical values. (shown in fig 2) like {36 month, 60 month} in term attribute to {30, 60} and {'Source Verified', 'Verified','Not Verified'} in verification_status attribute to {1, 2, 3} (shown in fig 2).

After all steps listed above, we finally got our final data set with 1348059 rows and 64 variables. In our project, we mainly care about two statuses of the loan — 'Fully Paid' and 'Default (Charged Off)'. These two statuses can tell us whether a loan turned out to be a good loan (i.e. fully paid) or a bad loan (i.e. default or charged off). As figure 3 shows, finally in our data set there are 80.02% for the former and 19.98% for the latter.

2.3 Model Training & Selection

First we use part of our cleaned data (26 attributes) to train three different classification models: Logistic Regression, XGBoost and Decision Tree. The best result we got is from LR models with $auc = 0.73$(shown in figure 4). So we will further explore how the LR model would improve after we apply it on the whole data set.

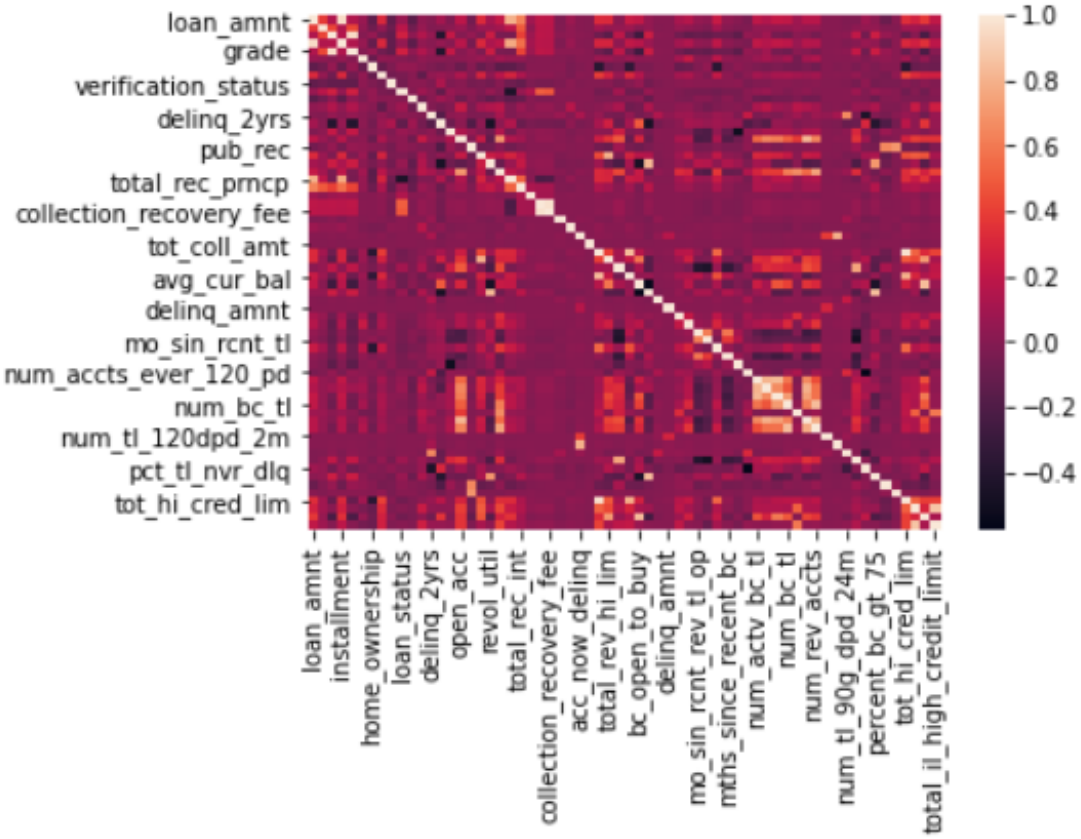

Figure 1: Correlation coefficient matrix heat map

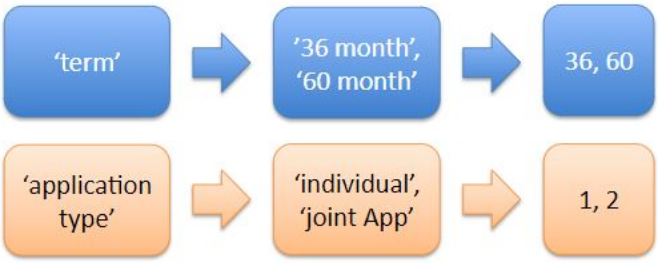

Figure 2: Change object data to numerical ones

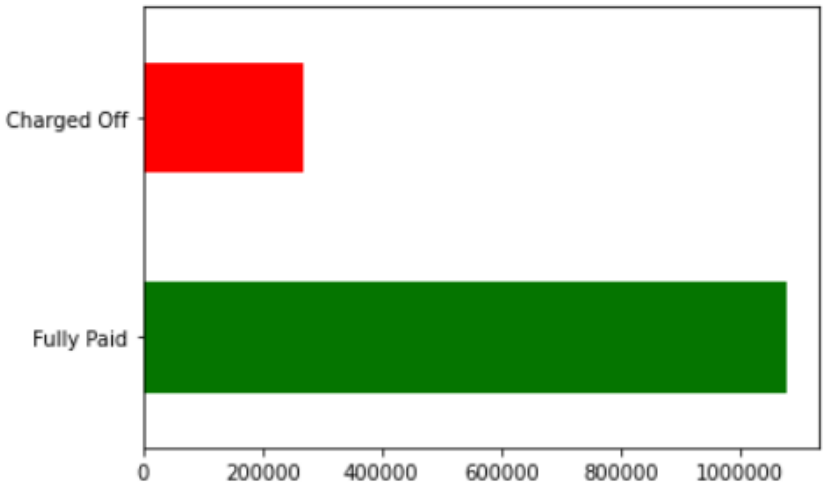

Figure 3: Correlation coefficient matrix heat map

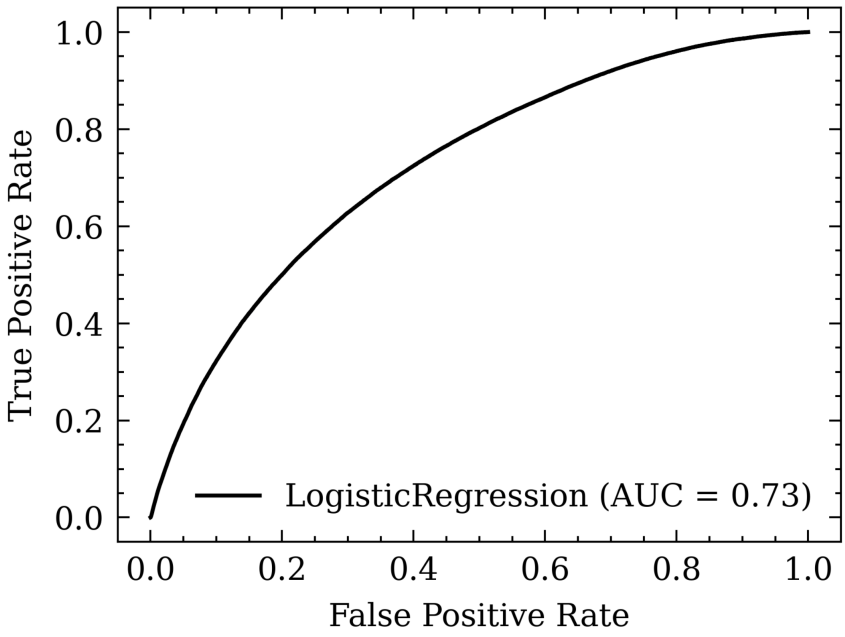

Figure 4: LR on partial data

After simply applying the Logistic Regression model to our whole-cleaned data set, we get an Accuracy of around 99%. But noticed that our data set is really imbalanced. So Accuracy means nothing here. To evaluate the performance of the model, we need to use AUC as the metric. Besides, we should deal with imbalanced data.

To handle the imbalanced data, we applied the up-sampling and down-sampling to the data set. A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

One way to solve this problem is to over-sample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training set prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

For the down-sampling, it can be defined as removing some observations of the majority class. down-sampling can be a good choice when we have a ton of data, millions of rows, etc. But a drawback to down-sampling is that we are removing information that may be valuable. Down-sampling can be defined as removing some observations of the majority class. This is done until the majority and minority class is balanced out.

What's more, we also apply grid-search on all our models to tune the hyper-parameter, which should further improve the performance of all models.

3 Plan and Experiment

All our code is written in Python. Some packages we are using: **Pandas** for data cleaning, **scikit-learn** for building models, **matplotlib** for data & results visualization.

We implement tree-based models XGBoost and Random Forest and also combine them with the two sampling strategies and compare their performance with the existing Logistic Regression models. Besides, we used grid-search on each model to make sure that they are giving their best performance.

XGboost as a machine learning algorithm, can implement default forecast by automatic iteration without manual intervention supervision and have profound theoretical and practical significance in the context of P2P industry default prediction is pursuing automation gradually.

Implementation of the Random Forest algorithm involves the training stage construction of several decision trees, and predictions emanating from these trees are averaged to arrive at a final prediction. Since the algorithm uses an average of results to make the final prediction, the Random Forest algorithm is referred to as an ensemble technique. Decision Trees are designed to optimally split the considered dataset into smaller and smaller subsets, in order to predict the value being targeted.

4 Results

Figure 5 shows the results we get from the Logistic Regression model with different sampling strategies.

After applying grid search, the result is shown in figure 6. We also compared different thresholds of the correlation coefficient's influence on our LR models. The LR model with oversampling and grid search gives us the best result with more than 0.99 ACC and 0.99 AUC. What's more, oversampling works a little bit better than downsampling in our data. The reason could be that we might remove some information when downsampling our training set.

After all, we have also explored which attribute contributes most to our model. As figure 7 shows, the attribute 'total_rec_prncp' which means the principal received to date contribute most to our model. The analysis of variable importance also shows the importance of removing attribute with high correlation coefficient.

5 Conclusion

We achieve an extremely high AUC with a simple logistic regression model, by using the full data set after pre-processing, simple sampling methods help us reach the best results.

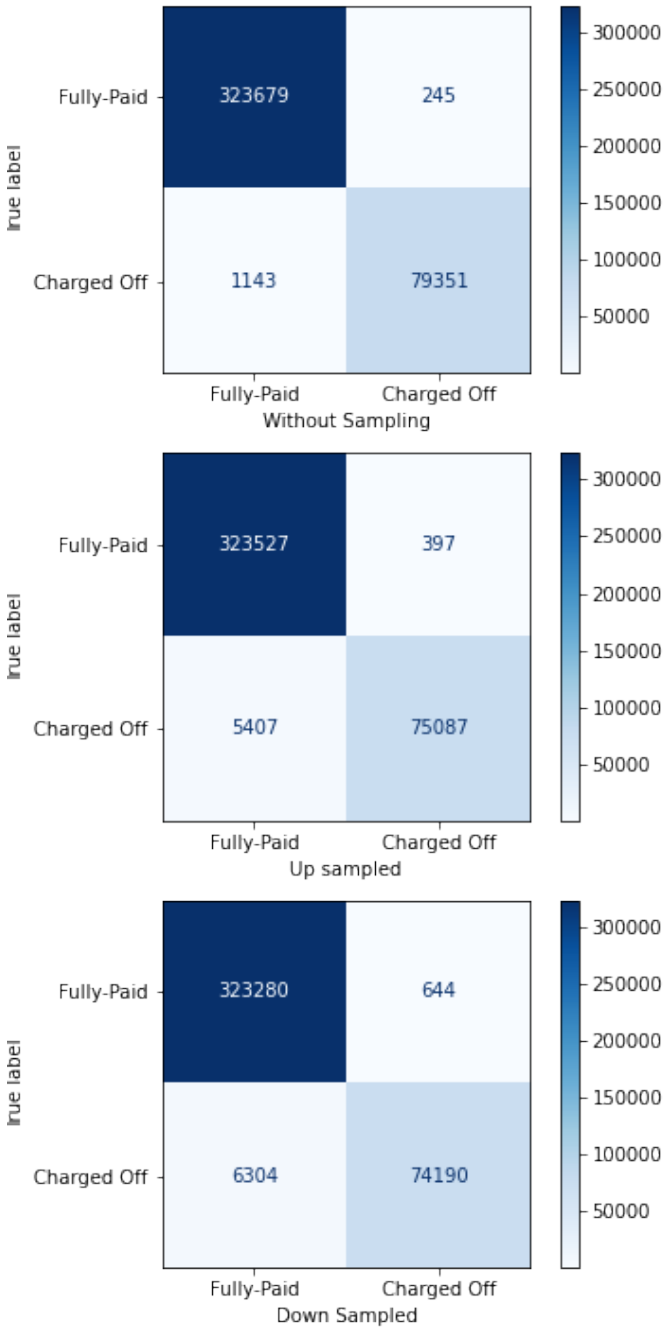

Figure 5: Confusion matrix of three different LR models

Model	Correlation coeff > 0.95	Correlation coeff > 0.98
Sklearn Dummy Classifier	accuracy_socre = 0.8 roc_auc_score = 0.5	accuracy_socre = 0.8 roc_auc_score = 0.5
Logistic Regression (LR)	accuracy_socre = 0.81 roc_auc_score = 0.523	accuracy_socre = 0.95 roc_auc_score = 0.880
LR + Oversampling Minority Class	accuracy_socre = 0.78 roc_auc_score = 0.729	accuracy_socre = 0.9856 roc_auc_score = 0.9658
LR + Downsampling Majority Class	accuracy_socre = 0.65 roc_auc_score = 0.652	accuracy_socre = 0.9828 roc_auc_score = 0.9598
LR + Oversampling Minority Class + GridSearch	{'C': 10, 'penalty': 'l1'} cv=3, scoring='roc_auc' accuracy_socre = 0.78 roc_auc_score = 0.729	{'C': 0.01, 'penalty': l2} cv=3, scoring='roc_auc' accuracy_socre = 0.9969 roc_auc_score = 0.9950
LR + Downsampling Majority Class + Gridsearch	{'C': 100, 'penalty': 'l1'} cv=3, scoring='roc_auc' accuracy_socre = 0.65 roc_auc_score = 0.652	{'C': 0.01, 'penalty': l2} cv=3, scoring='roc_auc' accuracy_socre = 0.9968 roc_auc_score = 0.9947

Figure 6: Performance of LR models with different methods

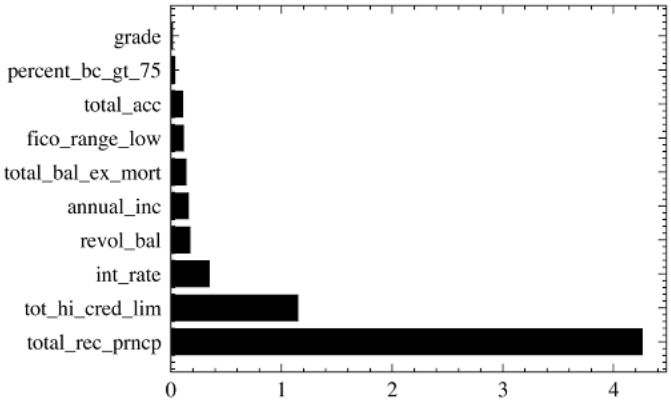

Figure 7: Variable Importance

If we use only 26 variables, we can only reach an 0.73 AUC by the logistic regression model.

Using models like SMOTE to handle imbalanced data, XGBoost(one of the most widely used tree-based models), Cross Validation, and GridSearch, can not help us reach an 0.8 AUC if only using 26 variables.

So in this project, we find the Data outweighs the model in our models.

6 References

[1] https://github.ncsu.edu/yjiang35/engr-ALDA-Fall2022-P12

[2] George, N. (2019, April 10). *All lending club loan data.* Kaggle. Retrieved December 5, 2022, from https://www.kaggle.com/datasets/wordsforthewise/lending-club

[3] Teply, P., & Polena, M. (2020) *Best classification algorithms in peer-to-peer lending.* The North American Journal of Economics and Finance, 51, 100904.

[4] Li, P., & Han, G. (2020). *LendingClub Loan Default and Profitability Prediction.* URL: http://cs229. stanford. edu/proj2018/report/69. pdf, 12.

[5] Zhang, W., & Wang, C., & Zhang, Y., & Wang, J. (2020). *Credit risk evaluation model with textual features from loan descriptions for P2P lending.* Electronic Commerce Research and Applications, 42, 100989.

[6] Gu, Y., & Guo, L., & Ma, C., & Wang, H., & Wei, X. (2022, November). Peer to Peer Lending Risk Analysis: Predictions from Lending Club. In 2022 3rd International Conference on E-commerce and Internet Technology (ECIT 2022) (pp. 750-759). Atlantis Press.

[7] Anon. Online personal loans + full-service banking at LendingClub. Retrieved December 5, 2022 from https://www.lendingclub.com/