Default and Investment Return Prediction in Online Peer-to-peer Lending

Chiyuan Cheng

JUNE 2020

Data Science Career Track, Springboard

1. Introduction

Peer-to-peer (P2P) lending is the new practice of lending money to individuals or small businesses via online service that matches lenders with borrowers. Lending Club (LC) is the world's largest P2P lender according to their issued loan volume and revenue.[1] With a much higher rate of returns compared to the traditional investment tools[2], P2P lending has become one of the fastest-growing investment platforms.

The process of LC starts with borrowers submitting their loan applications to the system, and the loan applications will be either approved or denied based on the credit review. The LC platform uses proprietary models to determine the interest rate of approved loans based on the information and credit history of borrowers. Investors can diversify their portfolios by investing loans they pick up with the suggested risk grade in each loan. However, in contract with the traditional investment, P2P lending presents a higher credit risk, because the borrower has a higher chance to not pay off his/her loan, leading to the loan default. Therefore, it is desirable for investors to be able to independently evaluate and screen the credit risk of available loans quickly, and invest in the selected loans with lower risks and higher returns.

This motivates us to build machine-learning classification and regression models to predict the credit risk and optimal investment return with the LC historical loan dataset. The goal of this project is to answer the following answers:

1. What are the key attributes for a successful P2P loan investment?
2. How can we use data to help investors to reduce their investment risk and increase their return?
3. What is the optimal return can investors expect from the P2P loan investment?

Specifically, we built and evaluated classifiers to predict whether a given loan will be fully paid by the borrower, as well as regressors to predict the annualized return of investment in a given loan. Finally, we recommended a simple loan selection strategy to maximize investment return based on the predictive models.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

LC has made its historical data publicly available for their registered customers.[3] The data we used contain comprehensive information on all loans issued between 2012 and 2019, which have 20 zip files with a data size of 2.6 GB. The data has various attributes, containing roughly 150+ columns that have categorical, numerical, text, and date fields. Additionally, we removed all loans which had not been terminated, and left only "Fully Paid" and "Charged Off" in the "loan status" column (Figure 1). The term "Defaulted" and "Charged Off" will be used interchangeably throughout this report. They have the same meaning where the borrower is no longer expected to pay off his/her loan.

2.2 Data Wrangling and Cleaning

The columns with more than 20% missing (NaN) values were dropped. Because most of the numerical features are not a normal distribution, we used the interquartile range (IQR) to remove the outliers. Briefly, IQR is the difference between the third quartile and the first quartile (IQR=Q3-Q1). The outliner is defined as the observation that is below or above 1.5 times IQR. We also removed some disqualifying features, which are not relevant to predictive models. The remaining missing values were then removed via the removal of the entire loan instance (row) itself. Overall, the preprocess reduces the dimension of the dataframe from 2,736,278 rows and 150 columns to 1,040,120 rows and 26 columns.

3. Data Exploration and Statistics

3.1 Loan Status

The first question we ask is what is the amount of defaulted loans LC has declared so far? To address this question, we have to understand there are still current loans that have not been paid off or defaulted yet. The status of loans can change as the days pass by, because there are still a large number of current loans (Figure 1). Therefore, to build the predictive models, we only consider the loans that have been matured, which leads to an imbalance dataset: "Fully Paid" (80%) and "Defaulted" (20%).

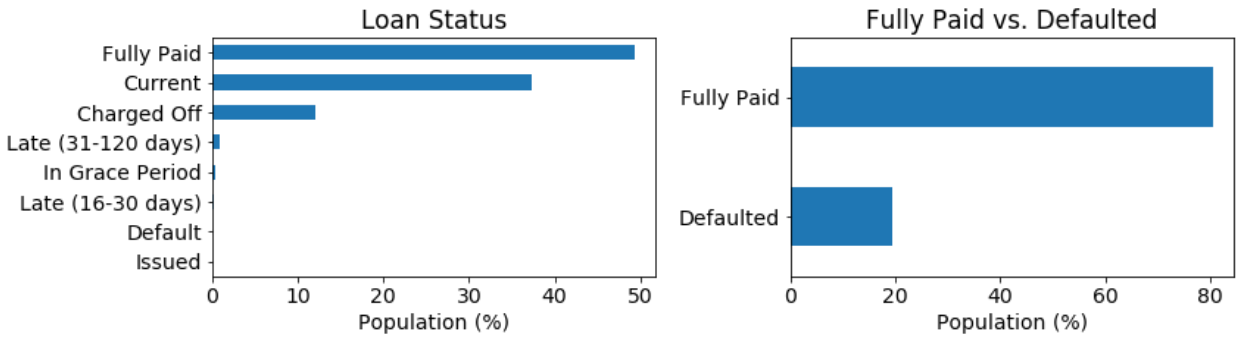

Figure 1. Loan status of LC data (2012-2019)

3.2 Default Rate

Figure 2 demonstrates that 60-month loans have overall higher default rates than the 36-month loans, suggesting that a long-term loan has a higher risk than a short-term loan. Interestingly, the default rate drops significantly after 2018, because many of these loans had not yet reached full maturity and their statuses are still unknown. For this reason, we excluded all the loan data after 2015 into the predictive model.

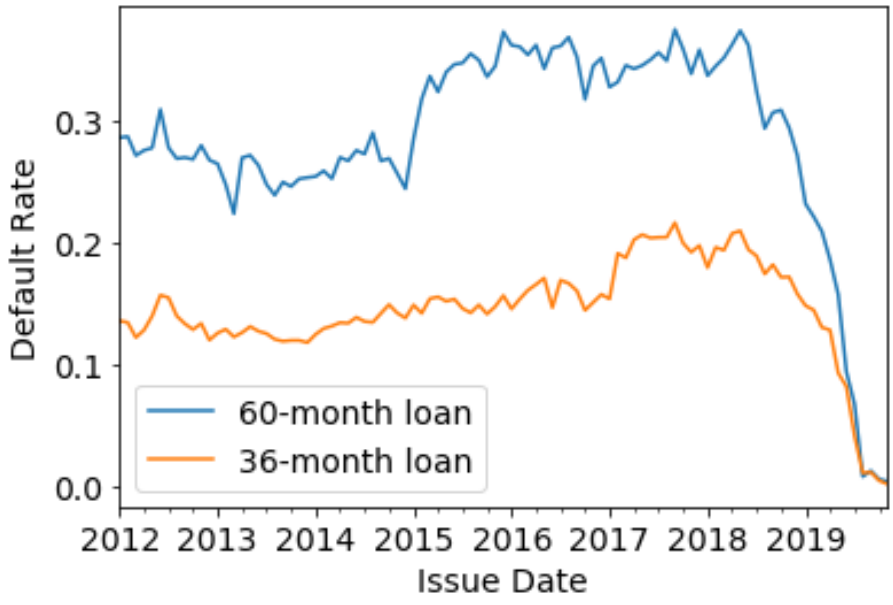

Figure 2. Time series plots of default rate data with different loan terms.

3.2.1 Low-income Borrowers Tend to Default on Loan

We compared the default rate of loan from high-income borrowers (income > median income) and low-income borrowers (income < mean income). Hypothesis **z test** suggests that low-income borrowers are more likely to default on a loan ($p<0.05$, assuming $\alpha=0.05$).

3.2.2 Loan Purpose Affects Interest Rate and Loan Default

The boxplot (Figure 3) shows that the interest rates have large variations among different loan purposes. We used a one-way **ANOVA** test to examine whether these variations are significant. The p value from ANOVA test is 0, which is $p<0.05$ (assume $\alpha=0.05$). It suggests there is at least one loan purpose that has a different interest rate.

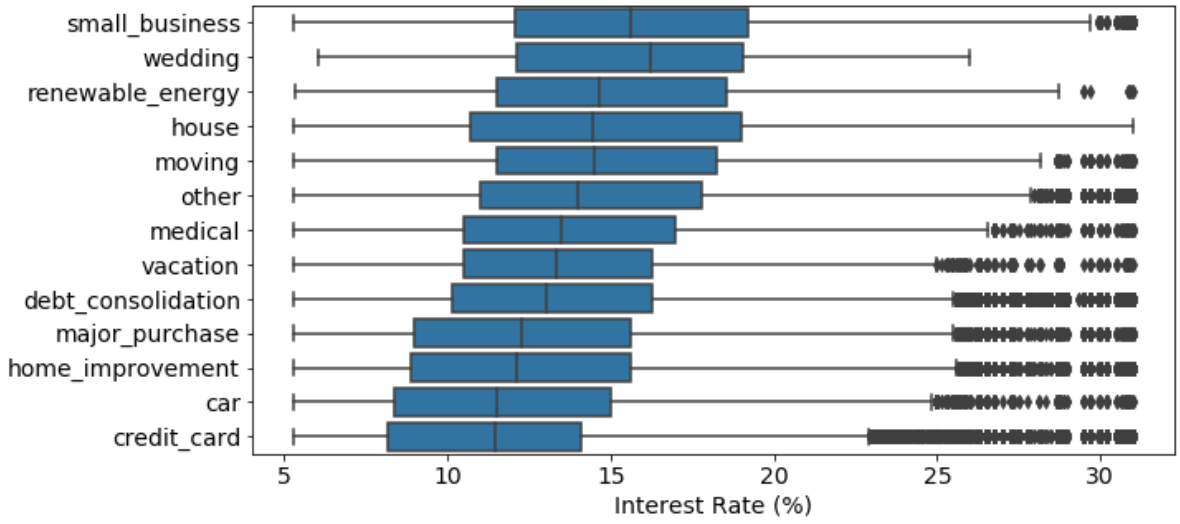

Figure 3. A box plot of Interest rate with the different loan purpose

Next, we used a **chi-square test** to evaluate whether the loan purpose affects the loan default. The result $p<0.05$ ($p=0$, assuming $\alpha=0.05$), suggesting that different loan purposes can affect the default.

3.3 Loan Period

The loan status changes over time, depending on the duration for each loan. For example, some loans can be paid off or defaulted earlier than the given loan term. To understand how soon a given loan will be terminated early, we defined the 'loan period' — the time

duration between the last payment date and the loan issue date. Figure 4 shows the density distribution plot of the loan period with different loan terms. For the non-defaulted, fully-paid loan, the borrowers with a longer loan term tend to pay off their loan earlier before the loan maturity: 93% of borrowers with 60-month usually loans pay off their loans earlier before the loans expire, while 74% of borrowers with 36-month loans pay off their loans early. Interestingly, most of the defaulted loans occurred in the 10th month.

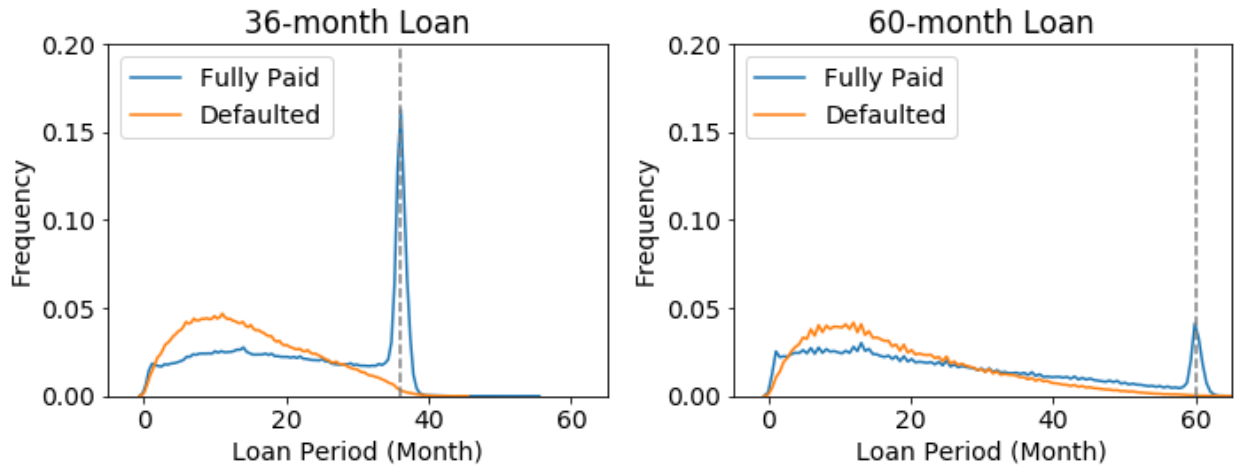

Figure 4. Density distribution plots of the loan period with different loan terms.

3.4 Return of Investment

To ensure the successful investment, besides estimating which loans are less likely to default, we also need to know which loan can yield a higher return. Notably, there are many different ways to define the investment return.[4] In this work, we defined the annualized Return of Investment (Annualized ROI or AROI) in a way similar to how LC calculates Net Annualized Return (NAR) for investors.[5]

$$Annualized\ ROI = \frac{Investment\ Gain}{Investment\ Loss} - 1$$

$$= \left(\frac{x_{TP}}{x_{FA}}\right)^{\frac{12}{m}} - 1 \qquad [1]$$

where x_{FA} is the total amount committed to that loan at that point in time (funded amount), x_{TP} is the total amount of loan received (total payment), and m is the loan period in month. Figure 5 shows that AROI of the non-defaulted and defaulted loans have completely different distributions. Although the median AROI of the non-defaulted loan is higher than that of the defaulted loan, the defaulted loan has a wider distribution of AROI, with many outliers at the higher end. It presents a great opportunity to obtain a higher return from the default loan using anomaly detection techniques.[6]

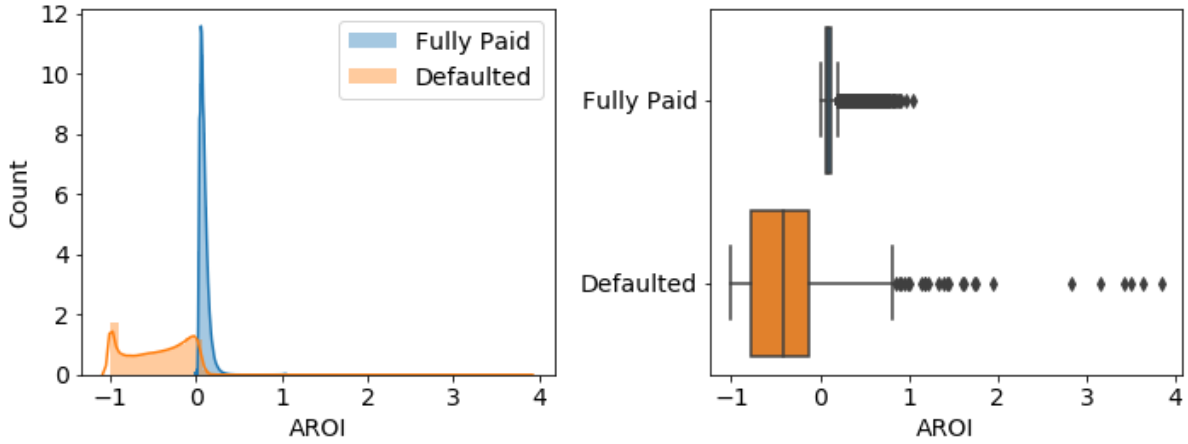

Figure 5. Density distribution plot and box plot of annualized ROI (AROI)

3.4.1 Investment Gain and Loss

We separated the AROI into investment gain (AROI >0) and loss (AROI <0), and did the log transform on their absolute values (Figure 6). The distribution of gain is close to a normal distribution, but the distribution of loss is not. We further performed the **Shapiro-Wilk** hypothesis test to check their normality and found both are not a normal distribution.

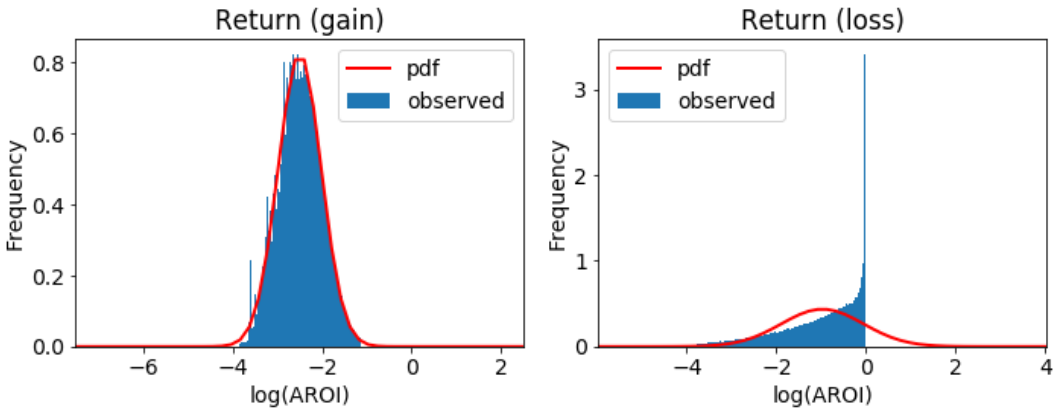

Figure 6. The bar charts of AROI on a log scale.

3.4.2 Outliers in the defaulted loan have much higher returns

Generally, we can expect that investing non-defaulted loans can have a positive return. But what about investing defaulted loans? The notion is that the investment of defaulted loans has a higher risk. On the other hand, it also has a higher interest rate. Therefore, there may be a potential opportunity to have a higher return from the defaulted loan.

Table 1 shows the comparison of the investment gain and loss between non-defaulted and defaulted loans. As expected, all the non-defaulted loans have positive returns. However, only 8.5% of the defaulted loans have positive returns, and they also have much more outliers with higher AROI values (Figure 7). This result tells us that a much higher return can be obtained from the highly risky defaulted loans.

Table 1. Percentage of gain and loss in loans.

Loan Status	Return	Percentage
Defaulted	Gain	8.5%
	Loss	91.5%
Non-defaulted (Fully Paid)	Gain	100%
	Loss	0%

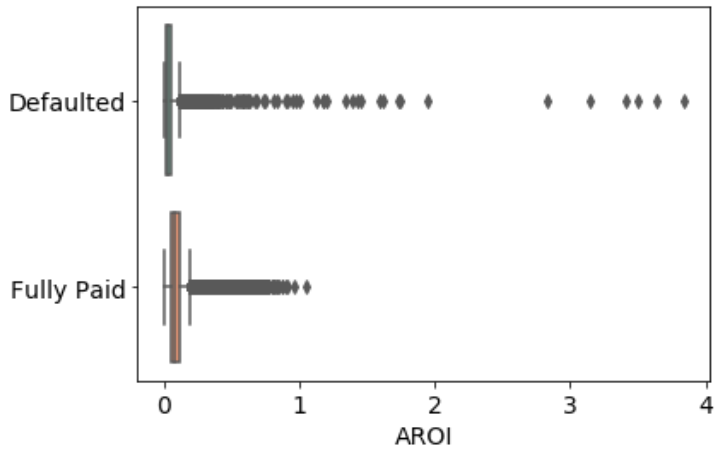

Figure 7. A box plot of AROI in the loan with positive returns (AROI >0)

3.4.3 The AROI of 36-month vs. 60-month default loans are the same

We used a **two-sample bootstrapping hypothesis test** to examine whether 36-month and 60-month defaulted loans have a different return. The result shows p>0.05 (p=0.4882, assume α =0.05), so the null hypothesis can not be rejected. Therefore, average AROI for 36-month and 60-month defaulted lands are the same.

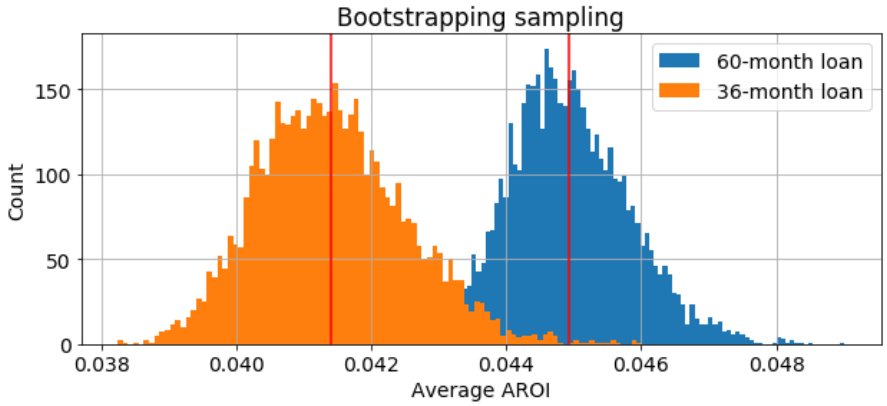

Figure 8. Bootstrapping statistics of average AROI in 36-month and 60-month defaulted loans

3.4.4 AROI in the US States

Next, we calculated the AROI in different USA states (Figure 9). In general, the northwest region has a higher AROI. Idaho (ID), Utah (UT), and Washington (WA) states have the highest AROI. The south and east regions have lower AROI.

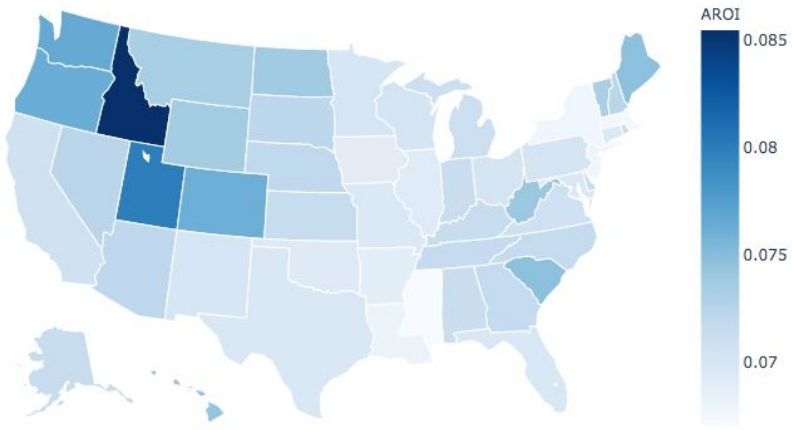

Figure 9. AROI by USA states (2012-2019).

3.5 Loan Amount

The loan amount relates to the load demand. If the loan demand of the borrower increases, the loan amount should increase. Understanding the loan demand can allow the LC to have sufficient funds to supply the requested loan upon approval. We first analyzed time-series data of the average loan amount (Figure 10).

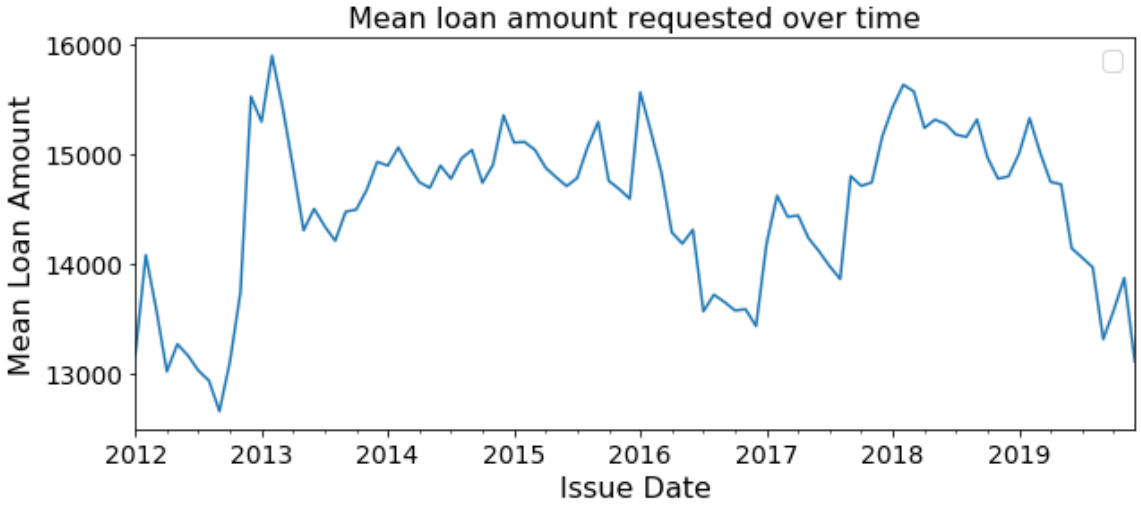

Figure 10. Time-series data for the average loan amount

3.5.1 The time-series data of loan amount is not stationary

When statistical parameters remain constant over time, it means the time-series is stationary, and we should not expect to see any trend in the stationary data. If we assume the past data is stationary, they should remain stationary on future data. Therefore, it is possible to do the forecasting on the time-series data, only if the data is stationary. Here, we used the **Augmented Dickey-Fuller (ADF) test** to check whether the loan amount time-series data is stationary or not. The p value of 0.07804 was estimated from the ADF test. If we assume α is 0.05, we can not reject the null hypothesis because p>0.05. Therefore, it can be concluded that the time-series data of the average loan amount is not stationary. It has a trend and seasonality associated with it.

3.5.2 Decomposition

The non-stationary time-series data of the average loan amount can be decomposed into three different components: trend, seasonality, and noise: The trend clearly shows the loan amount drops between 2016 and 2018, and the seasonal frequency is about 1 year.

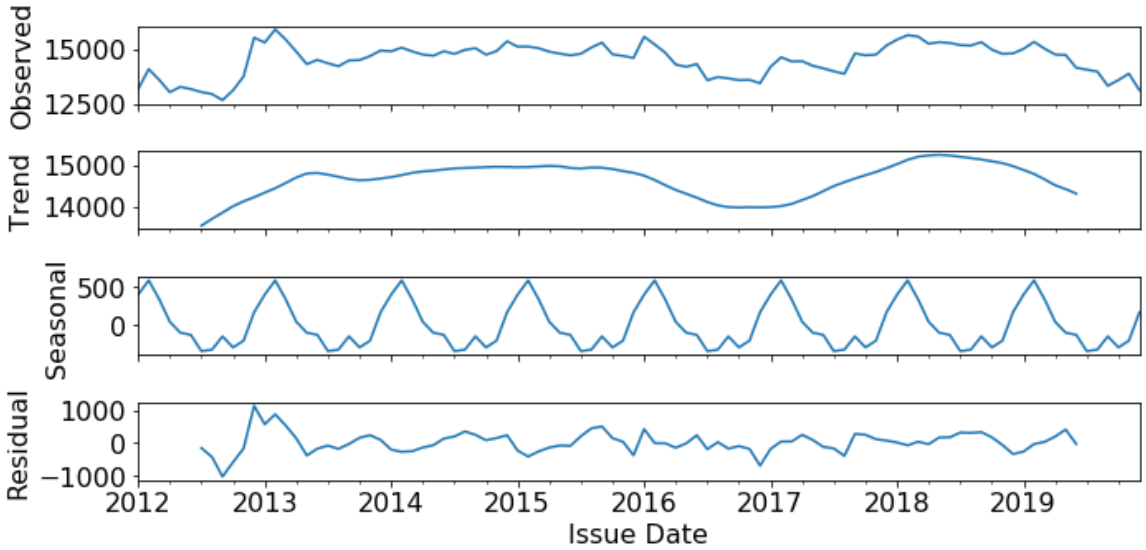

Figure 10. Decompose time-series data of average loan amount

4. Modeling

In this section, we use classification models to predict how likely a loan is to default and use regression models to predict how much return of investment can investors expect to obtain.

4.1 Data Pre-processing

The LC data issued in 2012 -2015 was used to ensure that the status of loans have been finalized. All the datatime variables were transformed into year and month as integral values. The categorical variables were replaced with their one-hot representations. For feature selection, we first eliminated the features whose correlations are less than 1% with target variables. The features that are highly correlated with each other were dropped to avoid multicollinearity.

4.1.1 Label Definition

For the classification model, “Charged Off” is assigned label 1 as the default class and “Fully Paid” is assigned label 0 as the non-default class. For the regression model, we used the annualized return of investment (AROI) calculated from Eq [1].

4.1.2 Imbalanced data

After data pre-processing, the dataset contains 90% negative and 10% positive, which is a common imbalance data problem. In such cases, classifiers tend to be biased towards the majority class, while the minority class is ignored. We used the synthetic minority oversampling technique (SMOTE)[7] to overcome this issue. SMOTE used k-nearest neighbors to oversample new data in the minority dataset based on the distance between the minority data and the randomly selected nearest neighbors. Standardization was then applied to the oversampled dataset.

4.1.3 Train-test split and cross-validation

Since it has been shown that LC dataset is stable with time[8], we can randomly split data into training and test (holdout) sets with 80%: 20% ratio to avoid bias from the time event. Standardization was then performed on all features, so that they have zero mean and one standard deviation. The training dataset was split into 10 folds and cross-validated with

each set to avoid over-fitting. The grid search was used to tune the hyperparameters for each model.

4.2 Classification Problem Overview

4.2.1 Classification Models

The goal is to predict loan default, and it is a binary classification problem. The loan-default prediction of five selected classifiers, **k-Nearest Neighbors (KNN)**, **Logistic Regression**, **Naive Bayes**, **Random Forest**, and **Gradient Boosting**, were tested.

4.2.2 Evaluation Metrics

Although accuracy is the most popular metrics in classification performance, accuracy does not consider the false positive and false negative. It will be a misleading criterion and causes bias results. For this reason, other performance metrics are more appropriate to evaluate the binary classification problems. In this work, we used receiver operating characteristics (ROC), Area Under Curve (AUC), precision, and recall to evaluate the classification model performance. The definition of precision and recall are shown below:

$$Precision = \frac{TP}{TP+FP} \quad [2]$$

$$Recall = \frac{TP}{TP+FN} \quad [3]$$

Where TP is true positive, FP is false positive, and FN is false negative. We used the ROC curves to evaluate and compare the thresholds of different models. The AUC can also provide a summary of each model. In general, a model that can perfectly discriminate default from non-default would have an AUC of 1.0. The ROC curve is generally used in the balanced dataset, whereas the precision-recall ROC curves should be used when there is a moderate to large class imbalance.[9] A dummy classifier was used as our baseline model to predict the most probable outcome for all loan instances. It sets the baseline performance, where other predictive models should be at least able to achieve.

The metrics of all classification models were summarized in Figure 11 and Figure 12. The baseline model had AUC at 0.5 and AUC (precision-recall) at 0.088. The scores of all the models are above baseline scores. Based on the summary metrics for all models, the

ensemble models (Random Forest, Gradient Boosting) achieved the highest AUC: it is able to achieve nearly 100% for both precision and recall, and a high AUC of 98% on both models.

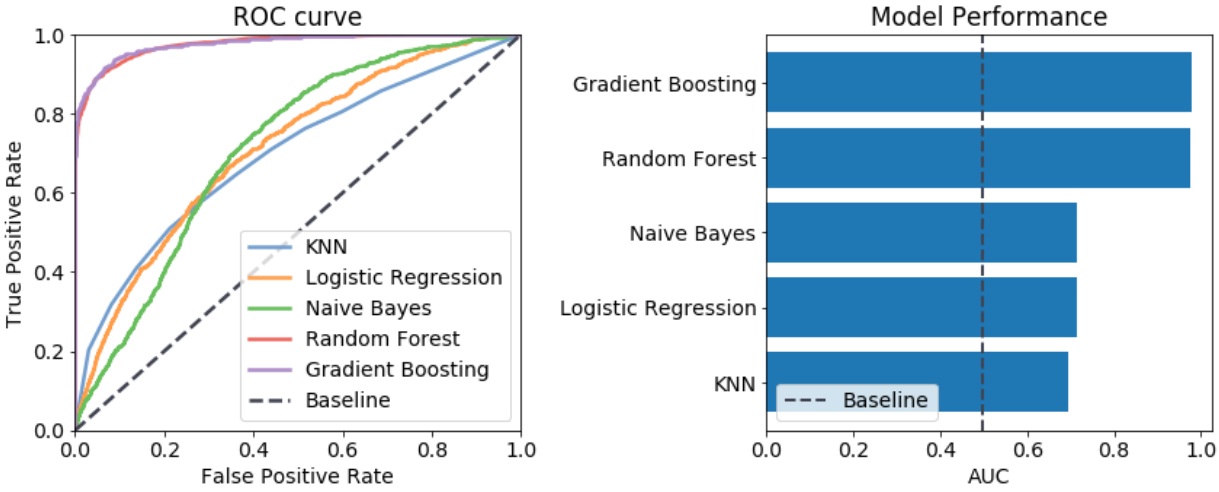

Figure 11. ROC metrics for classification model performance.

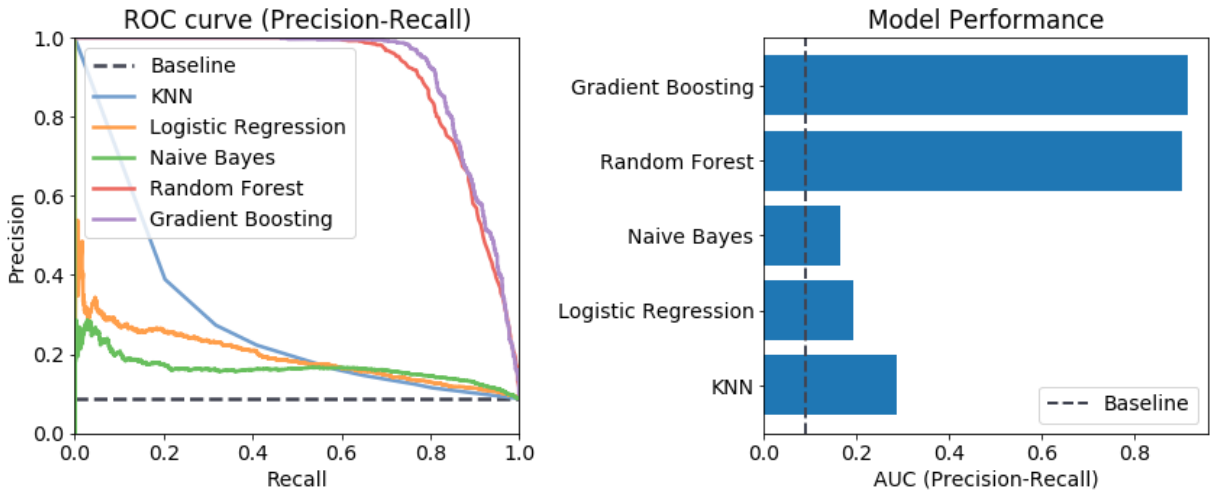

Figure 12. ROC (precision-recall) metrics for classification model performance.

4.2.3 Data Interpretation

To provide interpretation or inference from the model, rather than just to predict whether a loan is defaulted or not, we calculated the feature importance score (Figure 13) based on Random Forest. As expected, the Annual ROI, loan period, and interest rate are three most important features to cause loan default.

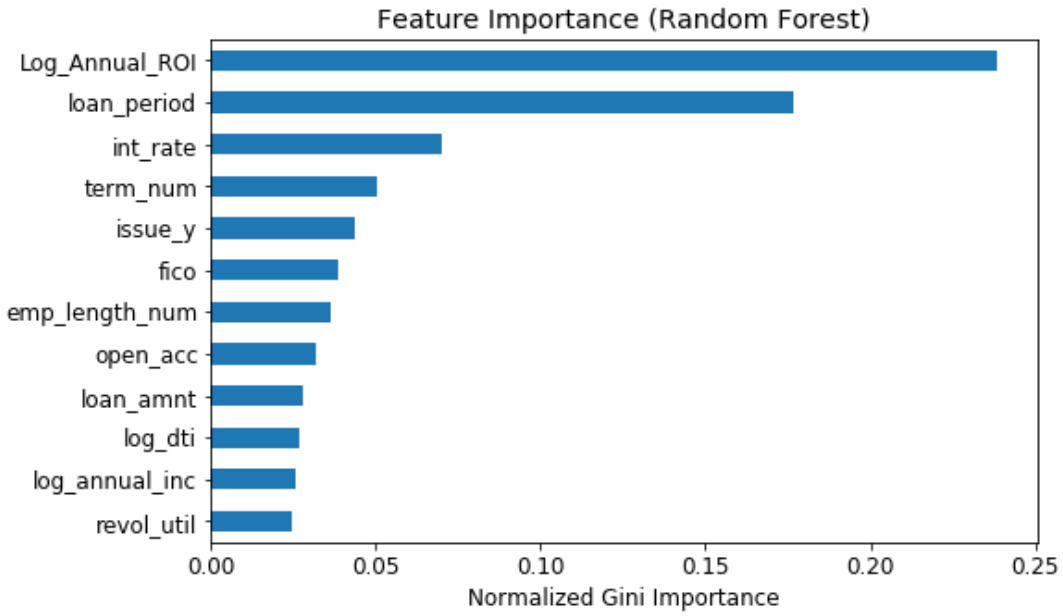

Figure 13. Top feature importance as observed by the random forest model.

We further used Shapley values, which is a tree-based model, to get more insight from the models. The Shapley Additive explanations (SHAP) value plot (Figure 14) can show the positive and negative relationships of the predictors with the target variable. The SHAP plot shows the top three strongest predictors for the loan default are interest rate, AROI, and loan amount. By plotting the impact of a feature on every sample, we can also see important outlier effects from the data distribution. For example, while AROI is not the strongest predictor, the color code shows the patterns, such as lower AROI (blue) increases the chance to default, while a shorter loan period (blue) has a lower chance to default.

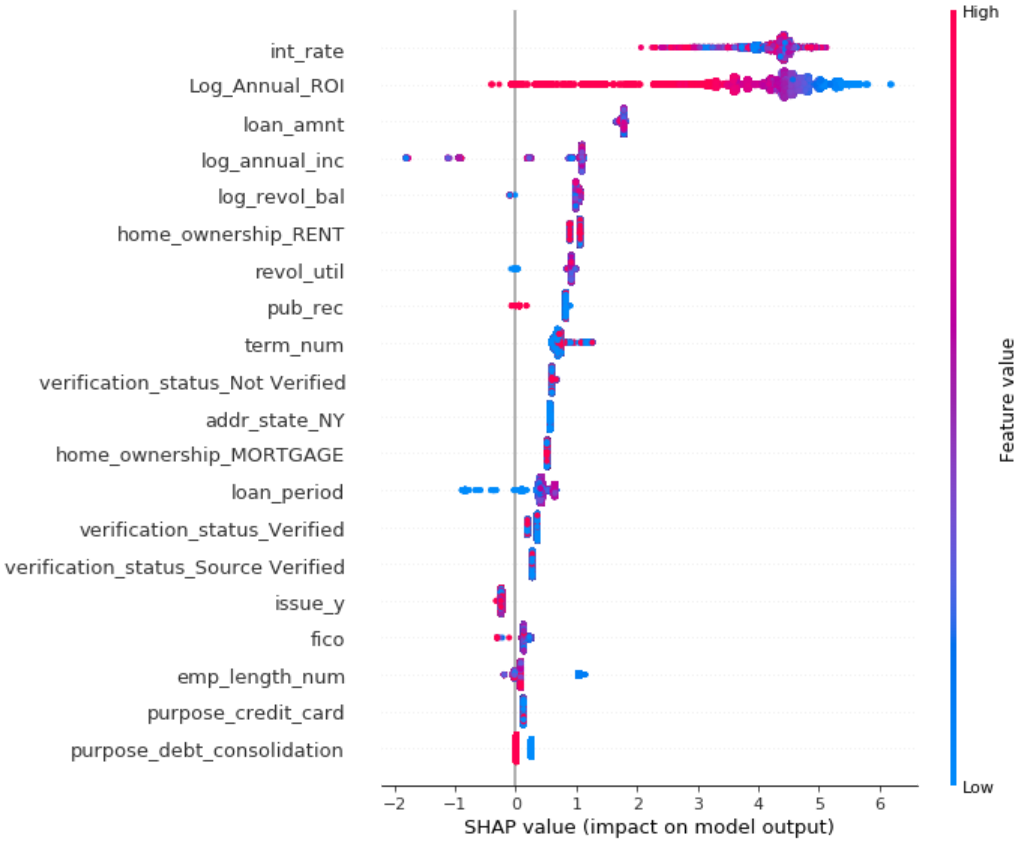

Figure 14. A SHAP value plot of the classification problem.

4.3 Regression problem overview

4.3.1 Regression Models

The goal is to predict the annualized return of investment. We used **k-Nearest Neighbors (KNN)**, **Lasso**, **Ridge**, **Random Forest**, and **Gradient Boosting** to predict the investment return.

4.3.2 Evaluation Metrics

We intend to predict the investment return if the investors were to invest in a given loan. We used the coefficient of determination R^2, and root mean squared error (RMSE) to evaluate the performance of the regression models.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}\right)^2} \qquad [4]$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2} \qquad [5]$$

Where $\hat{y}^{(i)}$ is the model prediction on $x^{(i)}$, and $\hat{y}$ is the mean of the true labels. The coefficient of determination tells us how much variability of the true AROI can be explained by the model.

From the results shown in Figure 15, we can see that the performance of ensemble models (Random Forest and Gradient Boosting) beats both linear regression (Lasso, Ridge) and KNN: R^2 of both Random Forest and Gradient Boosting models are >0.9. It is likely due to the fact that the combination of models is able to capture more patterns and non-linear relationships. However, for the Random Forest, the R^2 of the training set (0.99) is slightly higher than the test set (0.92), indicating the overfitting. The best performing Gradient Boosting regressor achieves RMSE of 0.06 on both training and test sets, which implies that the predicted AROI is estimated to differ from the true AROI by 0.06.

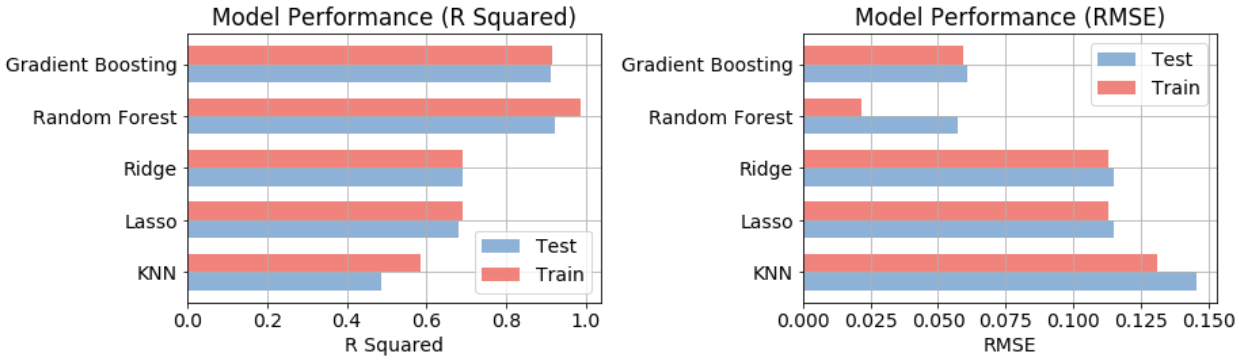

Figure 15. Metrics for the regression model performance

4.3.3 Data Interpretation

To better understand which features influence the return, we first calculated the coefficient from the Lasso regression (Figure 16). It shows that higher interest rate, longer

loan term, loan default, lower credit score, and shorter loan period have a positive impact on the investment return.

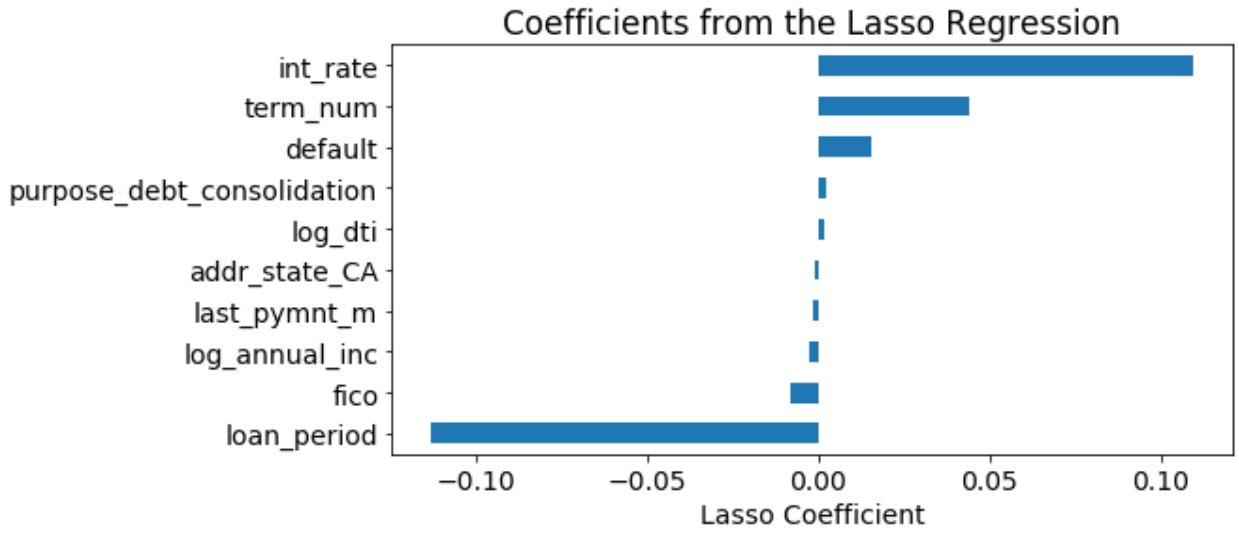

Figure 16. Coefficients from Lasso regression

The feature important scores calculated from Random Forest (Figure 17) obtained the consistent result with the Lasso regression, suggesting the loan period, interest rate, and default are the most important features for the investment return.

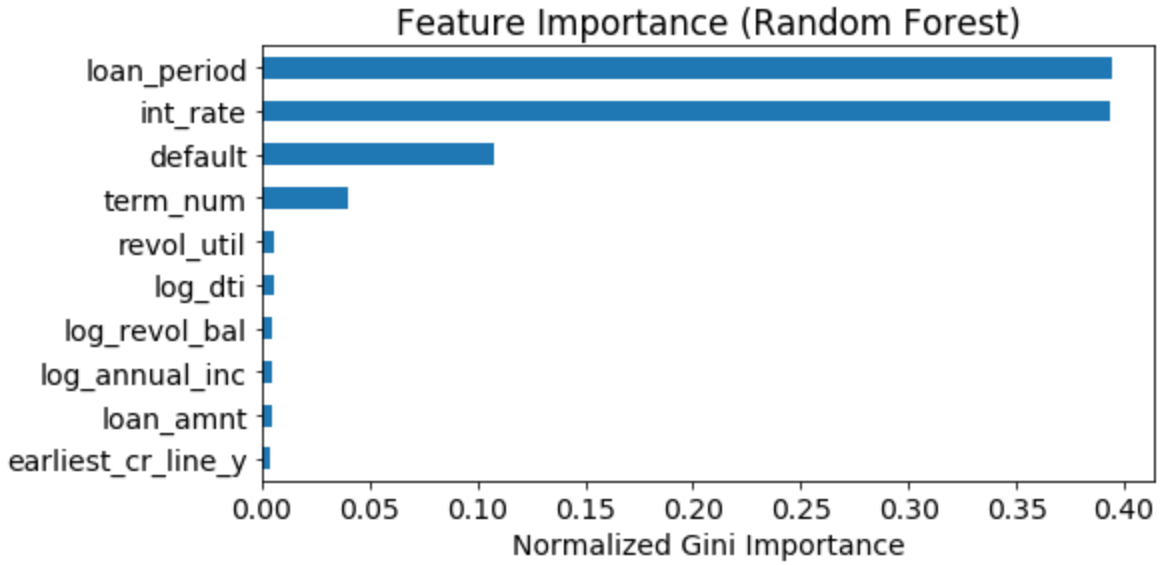

Figure 17. Feature Importance from Random Forest Regressor

To gain more insight to better interpret our model, we used the SHAP value plot to analyze the features with respect to the target variable (AROI), shown in Figure 18. The

result shows that a shorter loan period, a higher interest rate and a shorter loan term can increase the AROI. Overall, the loan period and loan amount have a negative impact on the return, while the interest rate has a positive impact on the return.

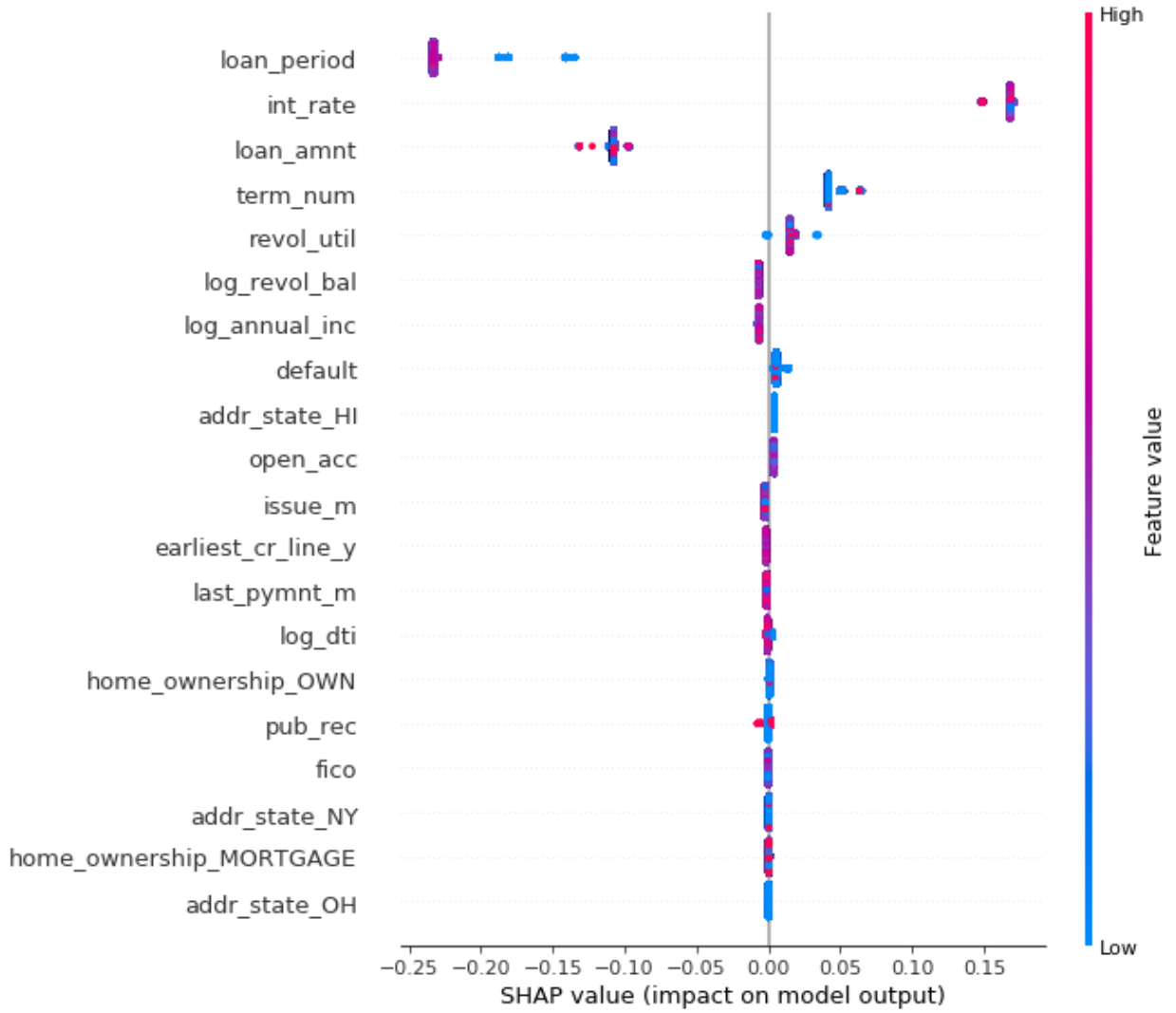

Figure 18. A SHAP value plot of the regression problem.

5. Using Model and Recommendation

In this study, we found Gradient Boosting and Random Forest obtain the best model performance for the classification and regression problems. Gradient Boosting combines the outputs of many simple and weak predictors to build a powerful predictor with

improved performance over the base learner tree, so this technique is known as "boosting", where a new tree boosts the performance of the original version.[11]

In terms of training speed, previous researchers revealed that the Gradient Boosting performs faster compared to other tree-based learning like Random Forest.[11] During the training of Random Forest, this model utilizes fully grown trees and combines all trees to make a prediction through the voting method (i.e. bagging.) However, Gradient Boosting leverages the combination of weak learners to make a better prediction (i.e. boosting)

Throughout the process of model development, fine-tuning and optimization procedures play important roles to find the best fit of hyperparameters to build a better model with lower errors. After all the models are trained, it is a necessary task to perform validation to evaluate a trained model on the test dataset with proper metrics. This validation process verifies whether the model is performed as expected. Moreover, this process also ensures the model can generalize well to the test datasets.

6. Assumption and Limitation

A combination of the best classification and regression models can be very useful in formulating a loan selection strategy. For the current return from LC, the median value of return received (including interest and late fee earned and principal recovered) by all investors divided by their initial principal is about 5 %.[12] If we apply our predictive models (Gradient Boosting) to randomly pick up 100 loans from the non-defaulted loan and calculate their returns, we can increase the mean AROI of the test set from 5% to 9.5%. Moreover, if we only use our best regression model to predict the AROI on the test set and choose the top 100 loans, we will obtain the mean AROI of 35%. A combination of the best classification and regression models will further improve the performance of our loan selection strategy. If we predict the default loan by the best classification model, use the best regression model to predict the AROI in the predicted defaulted loan and choose the top 100 loans with the highest predicted AROI, we will get a mean AROI of 40%. This value is a factor of 8 higher than the return from LC.

Table 2. Select investment strategy

Investment Strategy	Mean Return (AROI)
LC (benchmark)	5%
Randomly choose 100 loans from the predicted non-defaulted loans	9.5%
Apply regression model on the predicted non-defaulted loan, and choose 100 loans with the highest AROI	24%
Choose 100 loans with the highest AROI from predicted non-defaulted loans	35%
Apply regression model on the predicted defaulted loans, and choose 100 loans with the highest AROI	40%

The limitation of this investment strategy is that the classification model uses the matured loans (i.e. the loan status has been finalized) as the training set. However, to predict the loan return and provide suggestions to investors, most of the loans are still ongoing and not finalized yet. To improve the model performance, we should include the current loan (not expired yet) into the test set and test the performance of the classification model.

7. Future Work

In this work, we obtained a prediction threshold based on the classification and regression model on the training set, and simulated the strategy on the test set. Both sets comprise loans initiated within the same periods (2012-2015). We can check to see if this strategy generalizes to future loans by testing it on the loan after 2015 where most of the loans have not finalized yet. Practically speaking, this would be a much more useful metric for investors.

There are definitely some factors that contribute to default but not captured by features in our dataset. We can add external features, such as the demographic data for each state from the Census dataset, or from macroeconomic metrics that have been historically

correlated to the default rate. In addition, we can use deep learning or dimensionality reduction techniques (e.g. PCA) to predict the loan return since they require less feature engineering process. Because we have shown the time-series loan amount data is not-stationary, we can build a model to predict the average loan amount to forecast the future loan demand.

Finally, we can also make better use of existing features in the LC dataset. One example is "loan description", where the borrower enters at the time of loan application. Instead of dropping this feature, we can apply natural language processing techniques, such as TF-IDF, to put this feature into the models.

8. Conclusion

From the comparison of each model, we found that Random Forest and Gradient Boosting perform the best with default prediction and investment return prediction. We conclude Gradient Boosting is slightly better than the Random Forest based on their performance metrics. P2P lenders can take advantage of the predictive modeling discussed in this work to help investors make smart decisions when evaluating loan applications. Although identifying defaulted borrowers in advance can help investors lower the investment risk, developing investment strategies to accurately assess and predict the return can help investors choose the right loans with optimal returns.

9. References

1. https://en.wikipedia.org/wiki/LendingClub
2. https://towardsdatascience.com/turning-lending-clubs-worst-loans-into-investment-gold-475ec97f58ee
3. https://www.lendingclub.com/statistics/additional-statistics?
4. D. Ruppert, D.S. Matteson, "Statistics and Data Analysis for Financial Engineering", 2nd edition, 2015, Springer.
5. "How we measure net annualized return — lending-club." https://www.lendingclub.com/public/lendersPerformanceHelpPop.action. (Accessed on 12/08/2018).

6. Y. Zhao, Z. Narullah, Z Li, “Pyod: A python toolbox for scalable outlier detection”, arXiv:1901.01588, 2019
7. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, arXiv:1106.1813, 2002
8. M.C. Cohen, C.D. Guetta, K, Jiao, F. Provost, “Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science”, Big Data, 2018, 6(3), 191-213.
9. T. Saito and M. Rehmsmeier, ”The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”, PLoS ONE, 2015; 10(3): e0118432.
10. A Ghorbani, J Zou, “Data Shapley: Equitable Valuation of Data for Machine Learning”, arXiv:1904.02868, 2019
11. G. Biau, B. Cadre, L. Rouvière, “Accelerated gradient boosting”, Machine Learning, 2019, 108, 971–992.
12. https://www.lendingclub.com/investing/investor-education/benefits-of-diversification