LendingClub Loan Default and Profitability Prediction

Peiqian Li[1] and Gao Han[2]

[1]peiqian@stanford.edu, Computer Science, Stanford University
[2]gh352@stanford.edu, Stanford University

Abstract & Motivation

Credit risk is the risk of default as a result of borrowers failing to make required payments, leading to loss of principal and interest. In the context of peer-to-peer lending, investors wish to independently evaluate the credit risk of listed loans, and select loans with lower perceived risks to invest in. This motivates us to build machine learning models that can quantify the credit risk with LendingClub historical loan data.

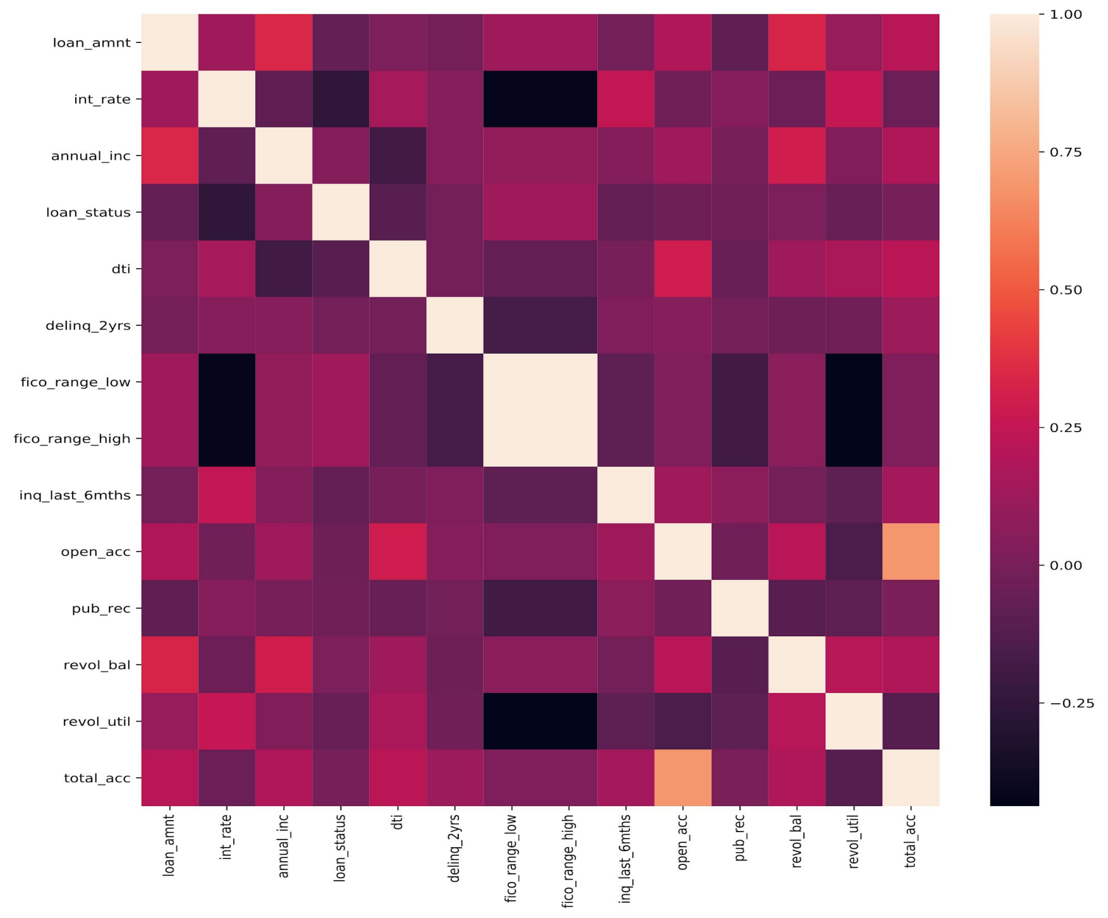
We built classifiers that predict whether a given loan will be paid back in full, using logistic regression, multilayer perceptron neural network, and random forest. All three models achieve an weighted average F1 score of around 0.89.

We then built regression models that predict the net annualized return of a given loan with linear regression, multilayer perceptron neural network, and random forest. The best performing model is the random forest regressor which achieves a coefficient of determination of 0.315. Picking loans with model prediction above 0.132 yields an excellent loan selection strategy. Simulated on the test set, this strategy achieves an average annualized return of 15% with investment in 1.76% of available loans.

Dataset & Features

Our dataset consists of LendingClub historical loans initiated from 2012~2015. Loans are either 36 or 60-month terms, so we filtered out loans whose status are not yet final. We treat "Paid Off" as our positive label, and "Default" or "Charged Off" as negative. The dataset consists of 80% positive and 20% negative examples.

Categorical features are transformed into one-hot representations. For partially missing values, they are replaced with either mean, zero or the max of their respective columns to avoid unnecessary penalty or reward when training the model. For columns with same values across all examples (including empty value), they are dropped. Finally, all features are standardized so they have 0 mean and 1 variance. The size of our final feature set is 1097. We then ran PCA on the dataset with the hope to further reduce feature set size. With 95% variance threshold, about 900 features were returned, which is close to the total number of features and hence no significant reduction in the feature space for correlation between features). Therefore, we decided to keep all of the features. Dataset is split using 0.7 training and 0.3 test.



Heatmap showing correlation among select numerical features.

For classification model, both Default and Charged Off are assigned label 0 and Fully Paid is assigned label 1. For regression model, we use annualized return rate calculated from loan amount, total payment made by the borrower, and the time interval between loan initiation and the date of last payment.

Loan Default Classifier

Random Forest achieves the best result with 0.70 negative f1-score and 0.94 positive f1-score. Random Forest model is trained with 200 trees and with at most 50 features to select randomly at each split. Gini loss is chosen as the objective to optimize.

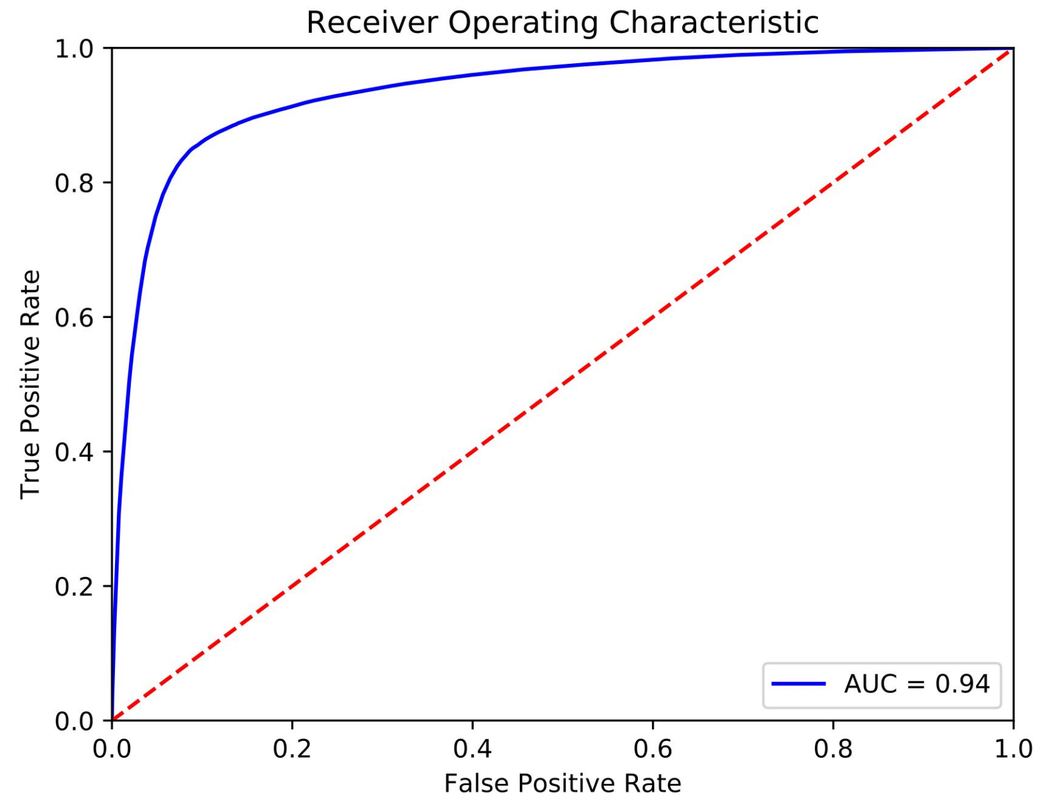
$$Gini = 1 - \sum_{j=1}^{2} p_j^2$$

After training, we arrived at the following result for both training and test set.

Test	Predicted 0	Predicted 1	F1-score
True 0	27760	15007	0.7
True 1	8750	172142	0.94

Train	Predicted 0	Predicted 1	F1-score
True 0	99163	0	1.0
True 1	0	422707	1.0

The ROC curve for Random Forest shows that as threshold is varied, false positive rate starts to increase sharply once true positive rate crosses 0.8. Any attempt to increase true positive rate will be at the expense of increasing false positive rate.



In additional to Random Forest, we also explored Logistic Regression and Neural Network, which did not achieve results as good as Random Forest.

	Train 0 f1	Train 1 f1	Test 0 f1	Test 1 f1
LogReg	0.74	0.92	0.73	0.92
NN	0.75	0.94	0.72	0.93

Annualized Return Regressor

In order to make better informed credit risk evaluations with more fine-grained prediction than the probability of loan default provided by classification models, we build regression models to predict the annualized investment return if we were to invest in a specific loan. Our label is the net annualized return (NAR) defined as

$$y = \left(\frac{x_{TP}}{x_{LA}}\right)^{\frac{1}{365/D}} - 1$$

where x_{TP} is the total payment made by the borrower, x_{LA} is the loan amount, and D is the number of days between loan initiation and the date of last payment.

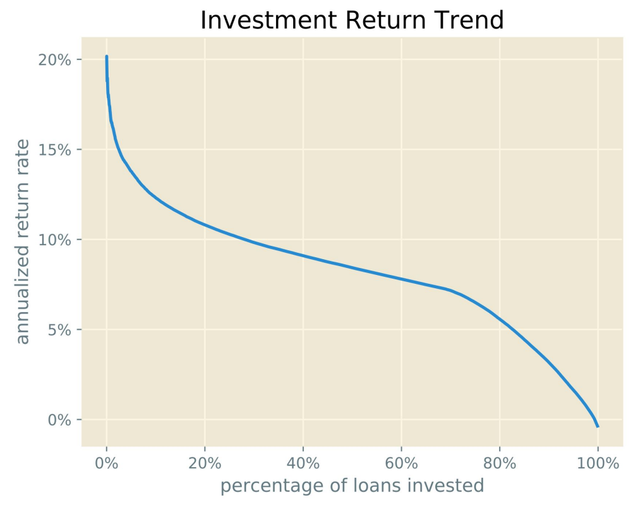
Linear Regression with no regularizations overfits the training data significantly. Ridge regression alleviates this problem with L2 regularization. Fully-connected neural network (20, 10, 5, 3) with ReLU activation function minimizing the squared loss achieves a better mean squared error (MSE) than the linear models.

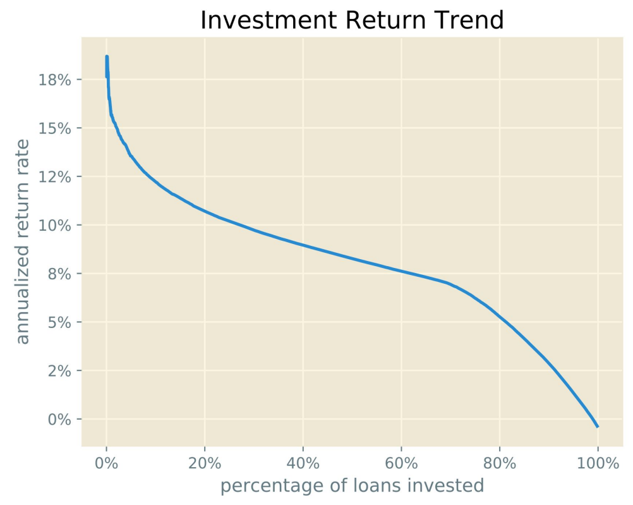
Model	Train MSE	Test MSE	Train R^2	Test R^2
LinearReg	0.040	5.014	0.243	-10^{23}
RidgeReg	0.040	0.040	0.243	0.238
NeuralNet	0.036	0.037	0.324	0.306
RF 4	0.037	0.037	0.298	0.295
RF 8	0.035	0.036	0.329	0.312
RF 10	0.034	0.036	0.356	0.315

Random Forest regression model achieves even better metrics (smaller MSE and larger coefficient of determination R^2). To speed up training and reduce overfitting, we limit the depth of the decision trees to 4, 8, and 10. As the depth limit increases, we see that bias decreases and variance increases.

Our best model, RF 10, gives rise to a simple yet very effective loan selection strategy: invest in loans with predicted NAR greater than 15%.

Loan Selection Strategy





The above two figures represent the relationship of how NAR varies as more investment is made on training (left) and test (right) set, investing in at least 100 loans, $25 each. On the training set, a prediction threshold of 13.2% NAR achieves 15% actual NAR. The same prediction threshold when simulated on the test set also delivers 15% actual NAR with 1.76% of loans selected.

Discussion & Future Works

Both classification and regression models perform better than naive models that randomly predicts with 50-50 chance or blindly predict the majority class. Comparing our models with those from related work, ours have better precision / recall and are more practical in terms of enabling implementable investment strategies in the sense that the return rate is higher than S&P 500's 10% annualized return for the past 90 years.

If more time is permitted, we would test models on finalized loan data between 2016 - 2018 to see if they generalize well. Furthermore, we could inject declined loan data from LendingClub to combat class imbalance problem better. The idea of performing freeform text analysis to utilize features like loan descriptions entered by applicants is also worth entertaining.

References

[1] S. Chang, S. D.-o. Kim, and G. Kondo, “Predicting default risk of lending club loans,” 2015
[2] K. Tsai, S. Ramiah, and S. Singh, “Peer lending risk predictor,” CS229 Autumn, 2014.
[3] A. Gutierrez and D. Mathieson, “Optimizing investment strategy in peer to peer lending,”, 2017
[4] B. Pujun, C. Nick, and L. Max, “Demystifying the workings of lending club,”
[5] “How we measure net annualized return — lendingclub.” https://www.lendingclub.com/public/lendersPerformanceHelpPop.action.