Lending Club Data Analysis

Katherine Zingerman, Mohini Agarwal, Shane Kok, Hannah Ho, Edward Eustachon

Overview of Lending Club

The Lending Club is a peer-to-peer lending company that connects crowdsourced lenders with borrowers from across the world. The loans are rated from "A-E" based on a proprietary Lending Club algorithm, that takes into account a number of different factors and outputs a rating. Each loan is then assigned an interest rate that corresponds with its Lending Club rating. Our dataset was available on the Lending Club website. All of the information that the company collects, when a person applies for a loan, was included in the dataset, regardless of whether the data was actually used to determine the loan's interest rate. Out of these predictors, we ran our models with 5 main predictors: "Loan Amount", 'FICO Score", "Loan Term", "Employment Length", "Annual Income", and "Public Record Bankruptcies" The data also included the interest rate that Lending Club had assigned the loan, which was our response variable.

Goal of Analysis

The report below aims to manipulate the Lending Club data of various transactions to determine the most important factors that affect loan interest rates. The main goal of the analysis is to:

1. Find the predictors that are the most statistically significant in determining loan interest rates in the Lending Club
2. Create a model that accurately predicts the most profitable loan interest rate for Lending Club users

Explanation of Models

Linear and Polynomial Regression

Our group ran Simple & Multiple Linear and Polynomial Regressions for the train data set of 3750 data points.

Linear Regression and Multiple Linear Regression

Individual linear models were made by regressing Interest rate on FICO average, Annual Income, Loan Amount, and Employment Length. The outcomes showed all the variables as significant except Employment Length. The variable that had the largest effect on output, is FICO average, giving an out-of-sample RMSE of 0.386.

Multiple Linear Model was made to regress Interest rate with all the variables taken together. From the model, Employment Length and Public Record Bankruptcies column seemed insignificant. The model gave an out-of-sample RMSE of 0.28.

However, another multiple linear regression model without the above two insignificant variables gave a higher out-of-sample RMSE. This can be attributed to the assumption that the significant variables function effectively only in the presence of the insignificant variables. Out-of-sample RMSE for the 3-variable model was 0.419, which was much higher than the model with all 5 variables.

Also, no interaction variable seemed significant enough.

Polynomial Regression

From the above Linear models, FICO score was the most important predictor. A polynomial regression model was built on higher powers of FICO score. The out-of-sample RMSE for a cubic power of FICO score was 0.386. Even with increasing the power of the FICO Score, the RMSE stayed stagnant. From the above linear and polynomial regression models, the multiple linear model with all the variables was the best model so far with FICO score as the most significant predictor.

K-Nearest Neighbors

The K-Nearest Neighbors model is used to identify a predicted output for a desired x-value, by choosing the "K" number of nearest neighbors, the most similar x-values to the chosen one, and averaging their y-values to determine the predicted output. The model found that when K=50, it was able to most accurately predict interest rate. KNN chose k=50 to achieve optimal efficiency without overfitting. Given the nature of the data, the KNN model appeared to be the least efficient model, as it had an overall RMSE of 0.368, which is the highest RMSE out of all the models.

Ridge Regression

The group calculated the level of correlation per variable under the assumption that all variables were relevant to the loan interest rate. The optimal lambda value was 0.097 with the lowest coefficients on the Public Record Bankruptcies and Employment Length. The top two most relevant variables, best explaining the variation in interest rate, were Loan Term and FICO score.

Lasso

The Lasso Model had an RMSE of 0.2875 based on a lambda value of 0.0005679. The model ended with 3 predictors that helped us to predict interest rate. These predictors were "Loan Term", "Annual Income", and "FICO Average" . Out of these 3 predictors, The Loan Term's coefficient was the largest, which means that the payback period of the loan has the largest effect on a loan's interest rate, according to the Lasso Model. Furthermore, the model ended up zeroing out 3 predictors as well. "Loan Amount", "Employment Length", and "Public Record Bankruptcies" were all predictors that got zeroed out by the model due to their low influence on the final loan interest rate.

Trees

Our initial tree had 136 leaves or terminal nodes. We used the rpart library to cross-validate the alpha parameter of our function. Based on an out-of-sample RMSE, the best pruned tree had 33 leaves. The resulting tree had an RMSE of 0.269, thus outperforming our linear and KNN models.

Random Forest

We used B = 100 and B = 500 to bootstrap samples of our dataset. This corresponds to the number of trees we will aggregate to create our random forest. We considered $m = \sqrt{6}$ *and* $m = 6$ for variables available at every split. Of the 4 models we created, the best had B = 500 and $m = \sqrt{6}$ with a minimum RMSE of 0.267.

Boosting

We considered different values of tree depth (D), lambda (λ), and number of trees (B) to aggregate. We considered D = 4 or D = 10; λ = .001 or λ = .2, and B = 1000 or B = 5000, resulting in 8 different boosting models. Our best boosting fit had D = 10, λ = .001, and B = 5000 as parameters and an out-of-sample RMSE of .261.

Conclusion

	Multiple Regrn	**KNN**	**RR**	**LASSO**	**Tree**	**RF**	**Boosting**
RMSE	0.281	0.368	0.276	0.275	0.278	0.272	0.267

The boosting model is the best as it has the lowest RMSE. The LASSO, RF, and tree methods all agreed that the most important features are FICO scores and Loan Term Length.