Lending Club:
Analyzing and Determining Loan Default Probabilities using Lending Platform Data

Grant Cho

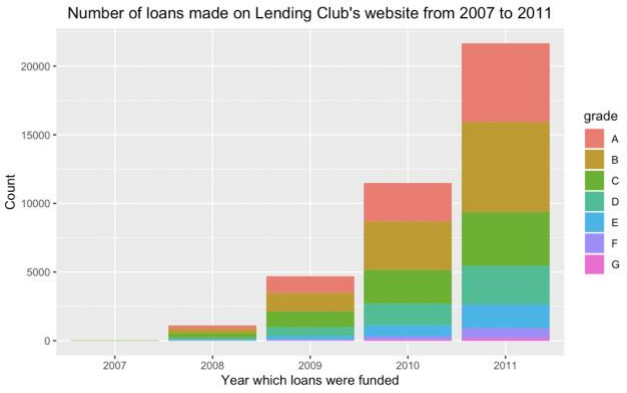
STAT 471
Professor Linda Zhao
November 15, 2019

Executive Summary

Lending Club is one of the foremost peer-to-peer online lending platforms in the world. The company acts as an intermediary and is an attractive substitute to traditional financial institutions by providing easy access for loans and credible return on investments, generating revenue by charging commissions per loan.

The opportunity for Lending Club's growth came around when Treasury rates decreased from 4.68% to 1.97% from 2007 to 2015, increasing demand for services that provide higher rates in an economy so stable and drawing in ambitious investors.

Additionally, Lending Club was able to draw in borrowers since this exponential growth occurred during a time of financial crisis and recovery. Demand for loans increased as yields fell to just over 0% in 2008; however, the supply of creditors likely decreased. In short, Lending Club began at the perfect time because interest rates were low, the online lending market was still not fully saturated given that the Securities Exchange Commission had to step in to establish the fundamental online lending practices. Because of these factors, the number of loans made on Lending Club's website exploded from 7 in 2007 to 21,675 in 2011, growing 3096 times its original value.



That being said, the possibility of debtors defaulting on their loans still exists. Of the 38,971 debtors from Lending Club's dataset, 5468 defaulted on their loans, which amounts to a little over 14% of all verified loans defaulting.

Furthermore, exploratory data analysis (EDA) reveals severe collinearity between several predictor variables, such as loan amount, funded amount, total payment, total payment invested, total received principal, total received interest, total received late fees, etc.

As such, one way to minimize this risk is through LASSO, where the final model would try to predict whether a borrower defaults or fully pays off his/her debt.

LASSO is a tolerant method to deal with multicollinearity. To build the logistic regression model, I removed unique identifiers, such as employer titles and zip codes, since a model with these variables would have no degrees of freedom to make reliable estimates. From there, I used 10-fold cross-validation to find the log(lambda) value that yields the highest area under the curve (AUC) value since a higher value is indicative of a more precise model. Finally, I used backward selection to narrow the model's non-zero coefficients down to only significant predictor variables.

I also used random forest to determine which classification method resulted in a more precise model.

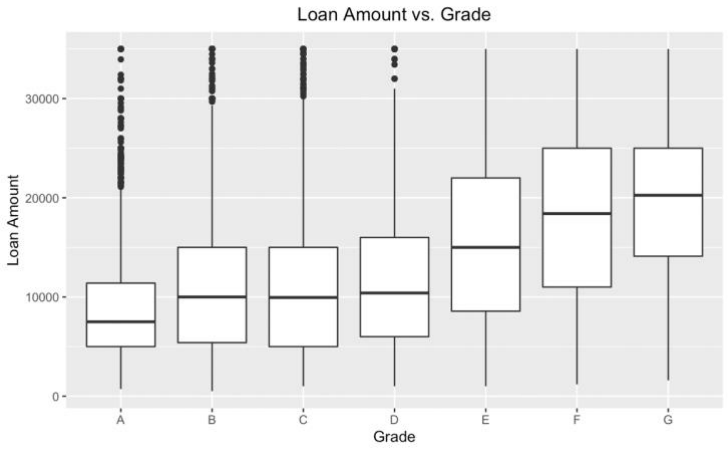
It turns out that random forest generates a slightly less accurate model, outputting an AUC of around 0.67 compared to LASSO's 0.69; however, the lambda penalty for LASSO was set to one standard error greater from the minimum. These AUC values are satisfactory but not excellent by any means, meaning this model isn't an incredibly reliable predictor of whether a borrower defaults or fully pays off his/her debt. Also notable, the misclassification error (MCE) of the LASSO model (0.277) is better than the MCE achieved by the random forest model (0.287).

Some risks in this model is that I excluded variables that caused "algorithms" to "not converge" and also those that resulted in a near perfect AUC value of >0.99 since these most likely distorts the overall model.

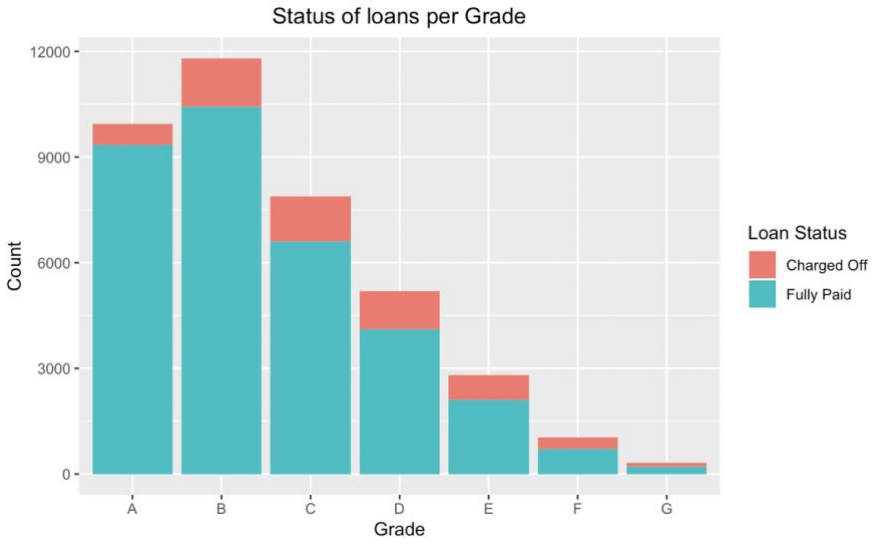
Even so, the final model revealed that the best coefficients to include are term, interest rate, installment, sub-grade, employment length, home ownership, purpose, DTI, inquiry in the last 6 months, public record, revolving line utilization rate, and total accounts. A unit increase of each variable, holding all else constant, is expected to decrease the probability of the debtor fully paying off the loan, with the exception of total accounts, purpose: credit card, purpose: wedding, and employment length of 2, 3, 6, and 9 years.

Exploratory Data Analysis

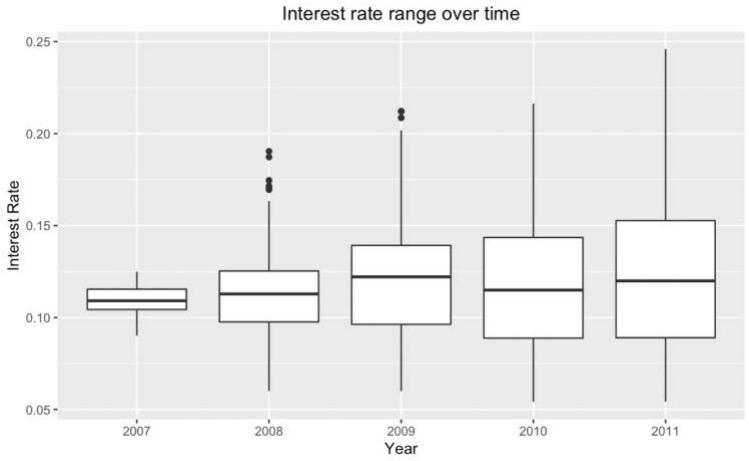
The Lending Club dataset was obtained through the club's website, ranges from 2007 to 2011, and includes 38,971 observations with 38 variables. Irrelevant variables and variables with lots of missing values were dropped.



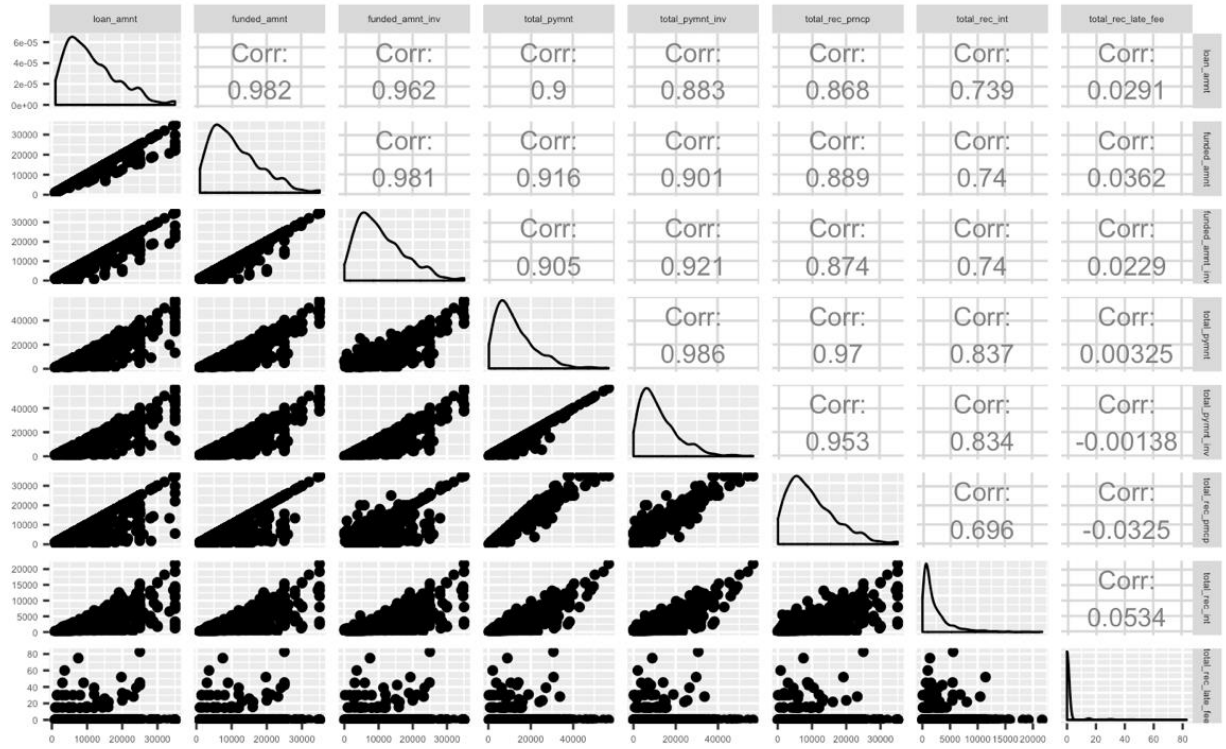
Interestingly, from the boxplot above, it appears that the median loan amount increases as the grade of the credit weakens. However, this data may be misleading as the lower-graded loans may be undercounted.



While the range of interest rates have varied from 2007 to 2011, the median has remained stable for the most part. This stability in interest rate may be attractive to risk-averse lenders who seek stable returns on investments.



There are a few problems with the dataset. The first is that there are unique identifiers that don't add anything to the model, such as zip code and employer. These variables are removed because they act as noise in the data that don't intrinsically produce anything noteworthy prediction-wise. The second is that there's severe collinearity within the dataset, as shown in the pairwise plot below:



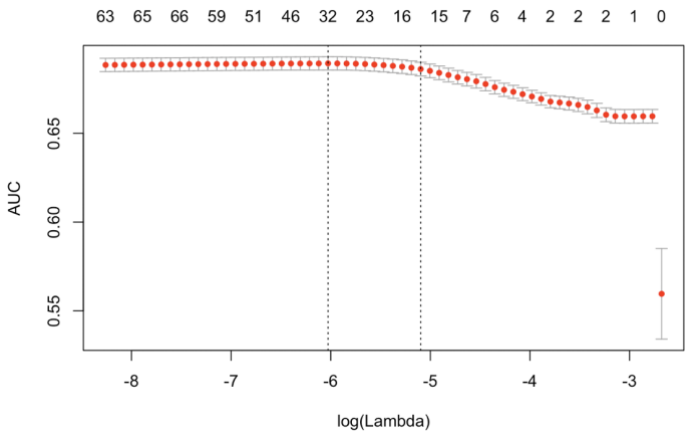
These selected predictor variables from the original dataset not only display redundancies within the data, but also severe multicollinearity - 11 of the above pairs have correlations $\rho > 0.9$. The collinearity is due to most of these variables' classification as post-loan data, which means that investors and borrowers will likely follow a similar course of action after the loan rather than before.

To eliminate redundant variables and to determine an optimal lending/investment strategy, it would be best to build a model from LASSO that would predict whether a borrower defaults or fully pays off his/her loan.

LASSO

I began by excluding unique identifiers, variables that weren't able to algorithmically converge, post-loan data, and variables that returned fitted probabilities numerically of either 0 or 1. Including the latter would result in a near-perfect predictor of the model that wouldn't be helpful in terms of interpretation.

Next, I used cv.glmnet to perform a 10-fold cross-validation on the data with the type of measure calculating the area under the curve (AUC).



According to the plot, the highest AUC is achieved when the number of predictor variables is around 32, with a log(lambda) value of around -5. Instead of using the minimum lambda value, however, I used lambda one standard error since that would create a sparser model with only the most relevant predictor variables, while also not allowing too much leniency in penalizing the coefficients.

```
Analysis of Deviance Table (Type II tests)

Response: loan_status
                    LR Chisq Df Pr(>Chisq)
term                 195.945  1  < 2.2e-16 ***
int_rate              29.812  1  4.760e-08 ***
installment           15.371  1  8.832e-05 ***
sub_grade             65.020 34  0.0010625 **
emp_length            71.210 11  7.189e-11 ***
home_ownership        15.960  3  0.0011557 **
purpose              229.933 13  < 2.2e-16 ***
dti                   21.365  1  3.796e-06 ***
inq_last_6mths        87.037  1  < 2.2e-16 ***
pub_rec                4.289  1  0.0383618 *
revol_util            26.207  1  3.067e-07 ***
total_acc             11.260  1  0.0007919 ***
pub_rec_bankruptcies   1.252  1  0.2631814
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA reveals that public record of bankruptcies is a non-significant variable.

```
Analysis of Deviance Table (Type II tests)

Response: loan_status
                LR Chisq Df Pr(>Chisq)
term             196.482  1  < 2.2e-16 ***
int_rate          29.818  1  4.745e-08 ***
installment       15.353  1  8.920e-05 ***
sub_grade         64.984 34  0.0010728 **
emp_length        71.962 11  5.167e-11 ***
home_ownership    15.929  3  0.0011728 **
purpose          230.033 13  < 2.2e-16 ***
dti               21.404  1  3.719e-06 ***
inq_last_6mths    86.891  1  < 2.2e-16 ***
pub_rec           30.122  1  4.057e-08 ***
revol_util        26.487  1  2.654e-07 ***
total_acc         11.193  1  0.0008208 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

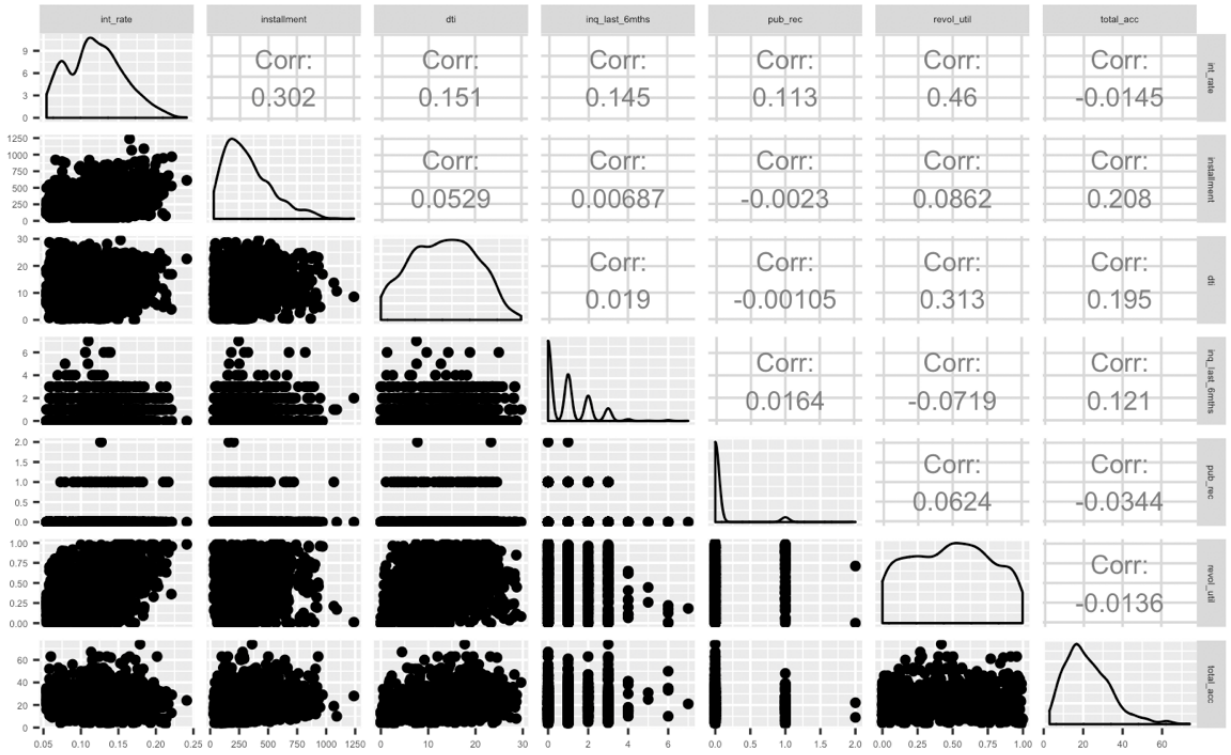
Removing public record of bankruptcies results in model where all variables are significant, so this is used as the final logistic regression model. As mentioned in the executive summary, the summary output reveals that a unit increase of each variable, holding all else constant, is expected to decrease the probability of the

borrower fully paying off the loan, with the exceptions of total accounts, purpose: credit card, purpose: wedding, and employment length of 2, 3, 6, and 9 years.

The loss ratio of picking up a bad loan to that of missing a good loan is around 2 to 1. This means that the threshold should be $\frac{2/1}{1+2/1} = \frac{2}{3}$.

fit.logit.pred	Charged Off	Fully Paid
0	501	852
1	4967	32651

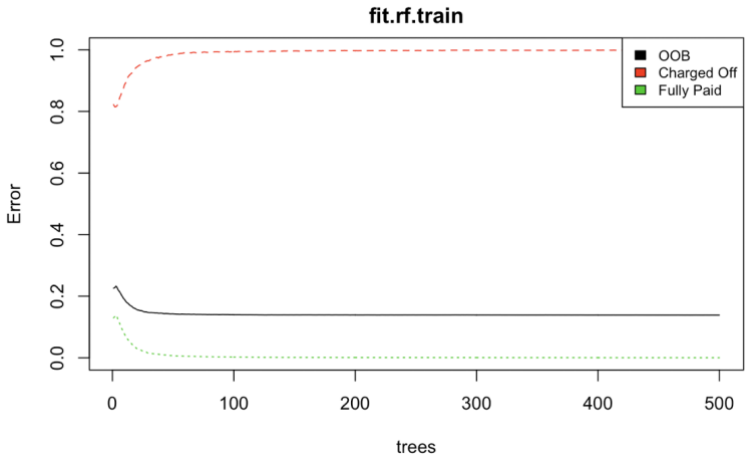
Using this threshold, the misclassification error for the logistic regression model is calculated as the sum of incorrect predictions over all observations with a penalty imposed on false positives, resulting in an MCE of $\frac{2*4967+852}{38{,}971} = 0.277$. Finally, the resulting AUC using this penalty is 0.6942, indicating a model that is satisfactory in its predictions, although not outstanding by any means. This reveals that the outcome of loans is dependent on factors likely not considered within Lending Club's existing dataset.



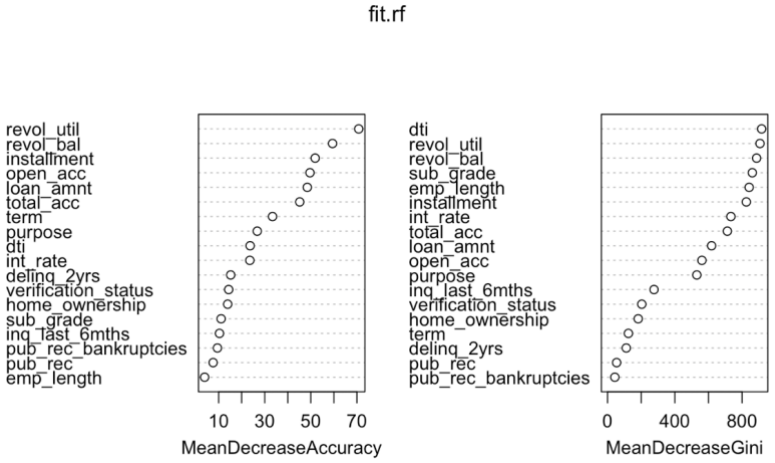
There also doesn't appear to be any severe collinearity among the numeric variables within the final model, which is excellent when it comes to identifying key predictors that affect the overall prediction.

Random Forest

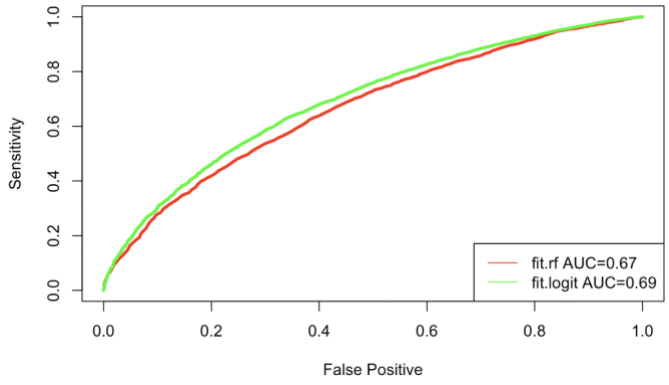
Next, I used random forest to not only determine the accuracy of the LASSO model, but also see if trees are a better statistical method to predict for loan status. I started by splitting the data into training and testing data. I then used the training dataset to obtain a prediction using random forest.



The error for out-of-bag, fully paid, and charged off stabilize as the number of trees grows. Evaluating the model reveals that the random forest model isn't as accurate as the one LASSO created, with the testing error and AUC values equaling 0.1439 and 0.668 respectively.



The left graph above ranks the importance of 18 variables from most important to least. Randomly permuting the revolving line utilization rate will result in an average mean squared error (MSE) increase of 70%, while doing the same for employment length only increases the MSE by <10%. This makes sense since an individual who is more leveraged will find it harder to pay back the debt and will be at a greater risk of defaulting on his/her loan.



The graph above reveals that the final model generated by LASSO is slightly more accurate, with an AUC value of 0.69. While both AUCs are satisfactory, it is far from stunning, indicating that there are possibly other factors outside the dataset that affect a borrower's ability to pay off debt.

predict.rf.y	Charged Off	Fully Paid
Charged Off	5	6
Fully Paid	1863	11117

Moreover, the misclassification error achieved through random forest is 0.287, which is actually worse than the one LASSO produced.

Conclusion

Regardless of the method used, both outcomes produce satisfactory - but not fantastic - final models with LASSO and random forest yielding AUC values of 0.69 and 0.67 respectively. The misclassification error for LASSO is 0.277, whereas the one obtained through random forest is 0.287, implying that the LASSO model is more accurate given the loss ratio penalties. Investors and Lending Club must also be cognizant that there are important confounding variables that the dataset may not have taken into account.

The final model obtained via LASSO resulted in the inclusion of 12 predictor variables absent of any severe collinearity: term, interest rate, installment, sub-grade, employment length, home ownership, purpose, DTI, inquiries in the past 6 months, public record, revolving line utilization rate, and total accounts opened. A unit increase of each variable, holding all else constant, is expected to decrease the probability of the debtor fully paying off the loan, with the exceptions of total accounts, purpose: credit card, purpose: wedding, and employment length of 2, 3, 6, and 9 years.

As for random forest, revolving line utilization rate is the most important factor in the model as it's not only highly statistically significant, but also a random permutation of the variable will result in a 70% average increase in the mean squared error.

I would recommend investors and Lending Club alike to focus their attention on revolving line utilization rate given its discovered importance. A borrower with a consistent history of heavy leverage is more likely to default than one who has a lower debt to limit ratio. Lending Club could establish a stricter debt to limit ratio requirement, which would increase the overall expected return on investments but would also likely decrease returns for the company as it depends upon commissions from each loan to generate revenue. To compensate, Lending Club could increase the interest rate it charges lenders – not borrowers – in the form of a risk premium.

Lending Club could also prioritize borrowers with more active credit lines as the predictor variable is also shown to not only be statistically significant, but also crucial in the overall model in terms of average increase in MSE. Those with more accounts could be given more opportunities to borrow, increasing the number of loans and strengthening expected return on investments.