Credit Risk via Lending Club
STAT 571: Modern Data Mining
Liming (Martin) Ning, Kevin Sun, Hanson Wang, William Walsh

Executive Summary

Lending Club, a leading peer-to-peer lending platform that matches individuals seeking credit with investors seeking to earn returns above dwindling treasury yields. Low rates helped create demand for new financial products that can produce yield. Lending Club has designed a platform that takes advantage of this credit supply to produce a lower cost of borrowing for individuals by cutting out highly-regulated middlemen like credit card companies, consumer banks, and credit unions. Lending Club, through providing greater access to cheaper capital, has helped millions of borrowers escape crippling credit card debt, and fund small businesses. For borrowers, Lending Club offered the following features to deliver value: 1) lower interest rates compared to traditional banks, and 2) an entirely online loan application process. On the other side of the market, Lending Club added value to investors through the following features: 1) attractive returns and lower volatility portfolios, and 2) a simplistic way to build a portfolio with dependable loans.

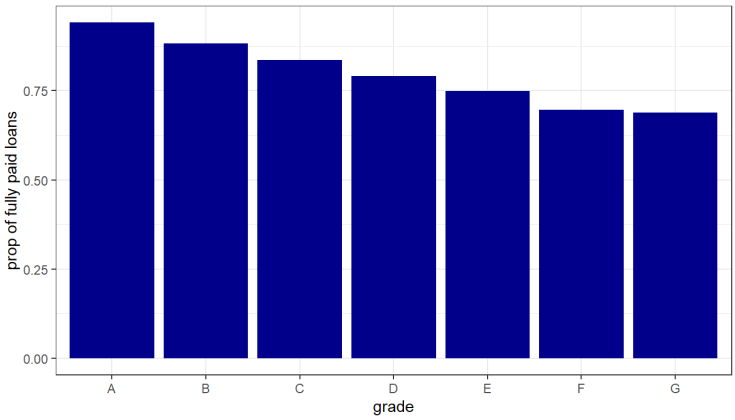
This paper's goal is to identify an optimal investing strategy in this emerging asset class for an external investor seeking to lend on Lending Club. The investing strategy will seek to create a portfolio of loans on Lending Club's peer-to-peer platform that has above-market returns. By developing a statistical model to classify loans as "loans to invest in" and "loans to pass in," with the goal of building the best portfolio that collectly has the greatest default-adjusted return, the investor can use the developed model to select loans to extend credit for. The developed model considers numerous attributes of a loan, such as the loan's amount, stated use, sponsor's personal income and home ownership status, the loan's interest rate and monthly payment, geographic information, and Lending Club's ascribed grade to the loan amui

The paper's data set includes data on around 38,000 approved Lending Club loans from 2007 to 2011, each with 38 attributes. Loans pre-rejected by Lending Club are excluded from the dataset, so our analysis is limited by the missing data. When considering future loans outside of this dataset, the trained model could underestimate the influence of certain variables that were used in pre-screening. This risk seems low if the designed model is used on future lending club loans as these loans would face the same rejection filter as the data set the model trained on.

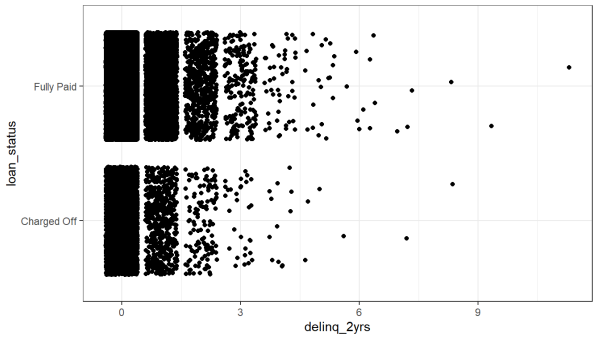
Ultimately, a logistic regression model was chosen to classify lending tree loans as "to be invested in" or "to be passed on." The resulting model showed that larger loans to people with higher annual income of shorter term, with lower interest rates, and less derogatory public records have lower defaults.

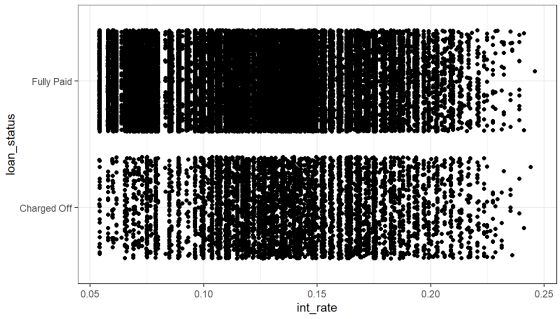
Part 1: Important Risk Factors for Default

As shown in the barplot below, the percentage of fully-paid loans decreases steadily from Grade A to Grade G. Approximately 6% of Grade A loans defaulted whereas 31% of Grade G loans defaulted.



For simplicity we start with a simple question: How does loan status relate to the annual income of the individual? We observe that defaulted loans were held by people with an average income of 62,638, while non-defaulted loans were held by people with an average income of 70,082. Therefore we can see that defaults are negatively correlated with low income.





We try to explore the relationship between the loan status and delinquencies in the past two years. The plot seems to be counter-intuitive: past defaults do not seem to predict current defaults. It is surprising that the relationship between interest rates and defaults is not as clear as what is predicted in theory: defaulted loans do not seem to have a significantly higher interest rate. Calculation reveals that the mean interest rate for the default group is 0.139 while that for the non-default group is 0.117. This difference is much smaller than the within-group difference, indicating that there are more risk factors to identify. Next we will go directly into multiple logistic regression and build a model.

We observe on average, loans in default tend to be held by those with lower income.

Return-Maximizing Classifier

We first manually identify 13 covariates which have a big possibility to be correlated with defaults. We will only include these variables into our model for a better interpretation.

Before building the model, we first put aside 9,000 observations as the testing set. We will build our model with the rest of the observations and select the best one by comparing the real testing error.

We first fit the full model and then conduct backward selection. We proceed until all variables are significant at 10% in the regression. The result of our final model from multiple logistic regression is presented in the appendix (figure 1).

By setting the loss ratio to 1/2, the misclassification error we get from the training data: 0.2758. The threshold for the logit we choose: 0.3333. The mce of the training set and the test set is similar, indicating that our model is not overfitted. Let's interpret our model results. loan amount, annual income are negatively correlated with defaults, while term length, interest rate, dti and derogatory public records are positively correlated with defaults. It indicates that we should value more on the former covariates and consider more about the latter covariates when making investments.

A number of different models were used in the model building process, including Random Forest, Decision Trees, and Logistic regression. By comparing fitted models with the area under the ROC curve, logistic regression was found to be the best of the three models.

We also tried out Random Forests, which are extensions of Decision Trees. We got the following confusion matrix and AUC ROC (Area under the curve: 0.5834):

	FALSE	TRUE
FALSE	6386	853
TRUE	451	105

False = No Default, True = Defaulted

The model accurately classified the vast majority of non-defaulting loans.

Random Forests offered a slight improvement over Decision Trees, which had an AUC ROC of 0.5.

Part 2: Why has LendingClub been successful and grown so fast?

Lending Club grew rapidly from a combination of a fundamentally better lending product combined with an asset-light marketplace business model capable of scaling rapidly.

Revolutions in capital markets that make lending to riskier individual borrowers have fueled innovation in modern economies. Micheal Milken's discovery that a diversified portfolio of individually risky but weakly correlated "junk bonds" could yield a higher risk-adjusted return than "safer" bonds greatly reduced the cost of capital for smaller companies. Fixed Income funds, such as Oaktree, were able to pick and choose which junk bonds to invest in, building individual portfolios that both generated superior returns for their investors and led to lower costs of borrowing for future junk bond issuers, since multiple parties bid up their loans. Lending Club has brought these capital market innovations to personal and small business loans by both a) greatly increasing the number of investors capable of investing in consumer credit, a field traditionally limited to highly-regulated financial institutions, and b) let investors, not regulators, decide what borrowers are most credit worthy, lowering the cost of borrowing for the market as a whole.

Furthermore, Lending Club's business model, a marketplace that takes on *no credit risk of its own*, allowed the company to scale rapidly, not limited by the need for working capital while loans wait to be sold. Lending Club monetized it's platform through fees, instead of interest rate spreads, letting the free market decide the cost of borrowing instead of a centralized financial institution leads to better price discovery and a lower cost of risk Financial institutions are highly risk averse, whereas, collectively, a diversified investor group can tolerate defaults more easily by diversifying their portfolio across many loans. By widening the consumer credit market supply to include less risk-averse lenders, Lending Club is able to lower the cost of borrowing for a wider range of borrowers.

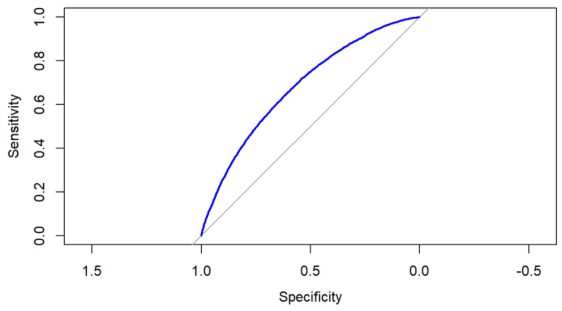
Part Three: Recommendations: *How should lending club modify their selection rules to increase returns for investors?*
A part of increasing returns for investors comes from minimizing downside risk– for the loans that do enter default, either from insufficient periodic cash flows to meet burdensome interest payments or a lack of money to pay the loan at maturity, lending club should seek to design systems aimed at maximizing recoveries for their lender partners while maintaining favorable borrower terms to encourage new loan origination. For corporate credit markets, the United States has developed a robust corporate reorganization process aimed at preserving enterprise value and maximizing creditor recoveries. This process affords powers to debtors that help them negotiate with creditors to reach a mutually beneficial arrangement. Lending club should work to develop a more robust loan "workout" procedure for the loans that do enter default. Possible elements to a "workout" process include deferring interest payments to later periods (when the borrower may have the ability to repay), lowering the principal amount of the loan, and / or pushing back the maturity of the loan. While creditors may initially resist such changes, they may support the development of a standardized workout process if it leads to *higher recoveries for loans that become delinquent*. This suggestion should not affect the return brought by paid-off loans while lessening the losses from failed loans, raising overall returns for lending club investors.

Appendix

Figure 1

	Dependent variable:
	loan_status
loan_amnt	-0.00000**
	(0.00000)
term60	0.073***
	(0.005)
int_rate	1.461***
	(0.060)
annual_inc	-0.00000***
	(0.00000)
dti	0.001**
	(0.0003)
pub_rec	0.053***
	(0.008)
Constant	-0.040***
	(0.008)
Observations	29,971
Log Likelihood	-9,919.000
Akaike Inf. Crit.	19,853.000
Note:	*p<0.1;* ***p<0.05;*** p<0.01

Figure 2



The ROC curve of our model is presented below. The value of the AUC is 0.681.