Default Predictions on Peer-to-Peer Loans

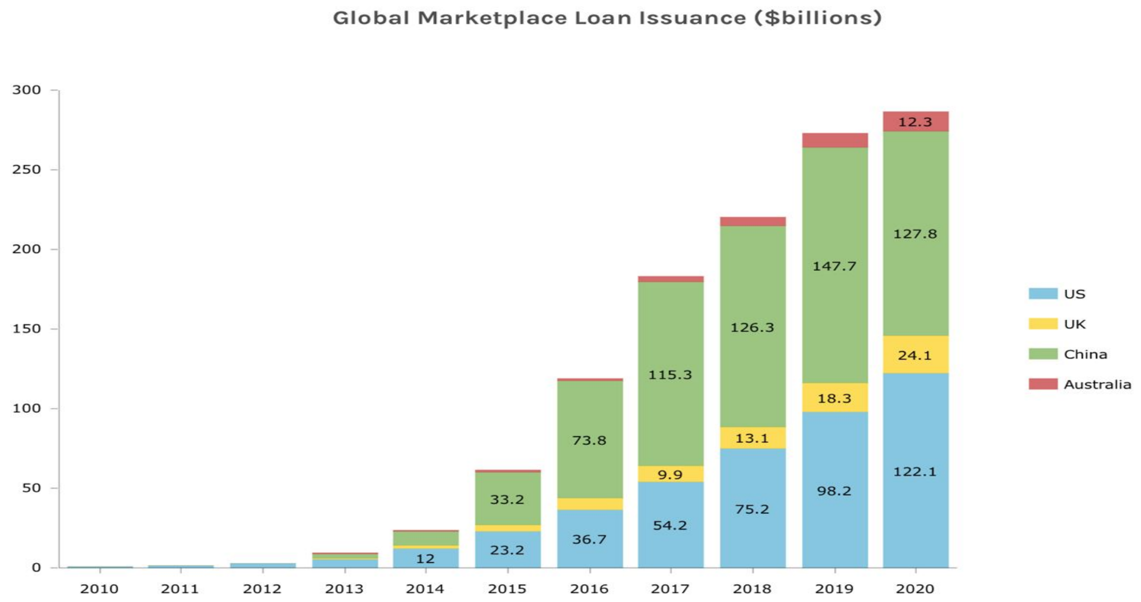
Data Mining for Business Analytics
Stern School of Business
New York University

Soumya Bandyopadhyay
Lizhou Chen
Xing Jin
David McLaughlin
Henika Sahajwani
Chirag Singhal

Introduction

The explosion of peer-to-peer marketplaces is one of the biggest stories of today's technology industry. The online platforms facilitate transactions between individuals so that one person can transfer an unused resource to another who needs it and is willing to pay. While ride-sharing company Uber and lodging site Airbnb may be the most prominent examples, the platforms have spread to numerous other businesses from pet-sitting to dining out. Peer-to-peer startups are also moving into personal lending traditionally done by banks and credit cards. The model is based on the idea of uniting borrowers who are unwilling or unable to gain access to traditional bank loans with investors who are willing to lend to them in order to earn comparatively higher returns, particularly in today's low-yield environment. While banks act as intermediaries, channeling funds from savers to borrowers, the peer-to-peer sites enable lenders and borrowers to transact directly with one another. In return for screening borrowers and evaluating their risks, they earn transaction fees for funded loans.

The growth of peer-to-peer lending has led to speculation that the platforms could upend traditional retail banking. Loan origination has doubled every year since 2010 to $12 billion in 2014, according to Morgan Stanley, which predicts that global loan issuance will reach $290 billion by 2020 with the biggest marketplaces in the U.S. and China.[1]



The projected growth stems from operating advantages peer-to-peer lenders have over brick-and-mortar banks. Their costs are significantly lower because they don't have to maintain retail branches and don't have capital requirements imposed by regulators. Their technology also gives them the scale to reach borrowers and lenders around the world. As a result, these marketplaces are positioned to offer loans faster than traditional banks and at lower rates.

The growth of the platforms depends on attracting new borrowers who are creditworthy and new investors who are willing to fund loans. Without a sufficient amount of qualified loan requests from borrowers, then investors

[1] https://www.morganstanley.com/ideas/p2p-marketplace-lending

may be forced to turn elsewhere for other opportunities, and without sufficient funding from investors, then borrowers may be unable to secure loans at attractive rates and may search for other sources of capital. And not only must these sites attract new borrowers and lenders, but the existing ones have to be repeat users.

Attracting both borrowers and lenders depends significantly whether both parties trust the platform, which in turn depends on the site accurately evaluating the credit risk of borrowers and the likelihood they will default on their loans. Borrowers will only use a site if they are fairly evaluated and offered interest rates that are competitive with what would be offered by banks and credit-card issuers, while lenders want investments that provide an attractive return relative to the risk. They will tend to shift their capital to competitor sites or other asset classes if a platform mispriced loans by underestimating or overestimating their risk. The growth of peer-to-peer sites provides investors with additional choices of where to direct funding, putting added pressure on the platforms to ensure that their predictions of borrower risk are accurate.[2] Otherwise they risk losing investors to competitors with more precise models.

Methodology

We model individual loan data from Lending Club Corp., the largest peer-to-peer lender, to predict defaults on a given loan.[3] Investors can select loans to individuals and fund a fraction of the loan in $25 increments. They can also purchase whole loans. Loans range from $1,000 to $35,000 for up to 60 months and are typically used for purchases or refinancing credit-card balances.

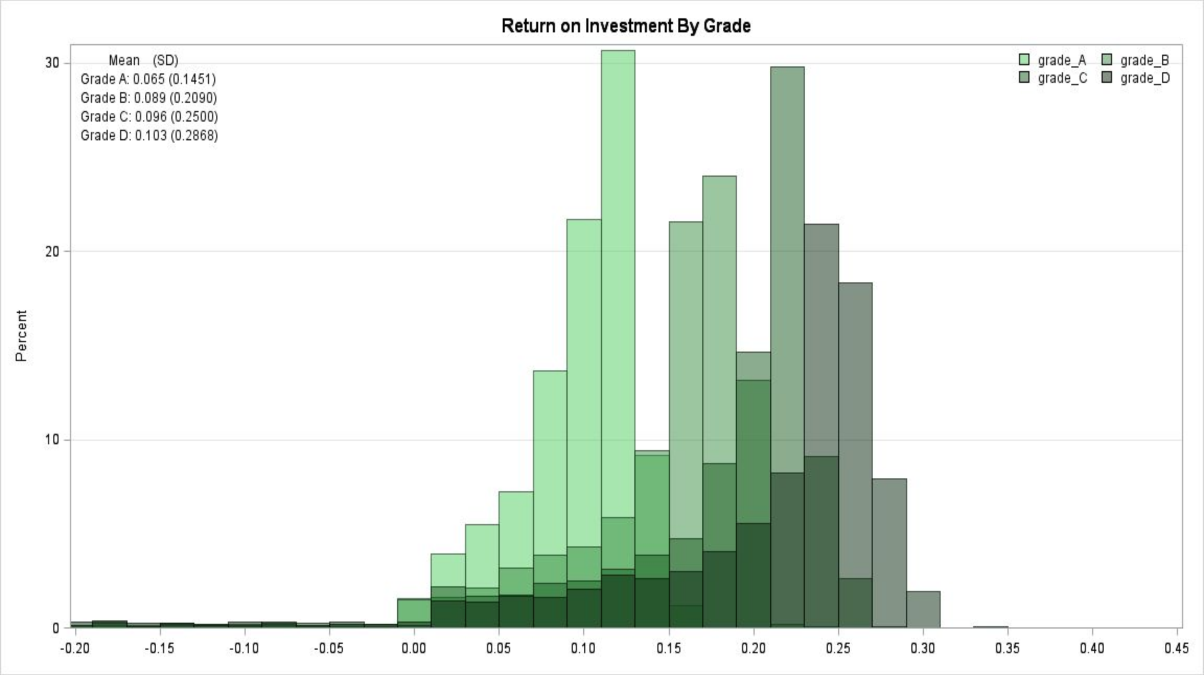
Default predictions are valuable for investors in choosing which loans to fund because they face several significant risks. First, when a borrower becomes delinquent Lending Club is not obligated to make payments to the investor. In addition, investors can't take legal action against borrowers to collect late payments. They must instead rely on Lending Club to pursue borrowers and are required to pay a collection fee of up to 35 percent, which reduces any recovery on the loan. Investors also don't have access to detailed financial information about borrowers. Lending Club doesn't always verify a borrower's income and doesn't verify other information such as employment, instead using third parties such as credit-reporting companies to provide basic data.

Investors can choose among loans with different risk profiles based on letter grades assigned by Lending Club using a proprietary scoring model. The grades range from A to G, with A the being safest and G the riskiest. Investors can earn yields of as much as 29 percent for the riskiest loans and 5 percent for the least risky. Approximately 84 percent of Lending Club loans are fully paid while the remaining are charged off, meaning the debt is deemed uncollectible, which results in a loss to the investor. Loans are considered defaulted 120 days after the last due date of payment and are charged off after 150 days.

The following chart shows the distribution of returns for each grade. As would be expected, the mean returns and volatility as measured by standard deviation rises with increasing risk (i.e. lower grades).

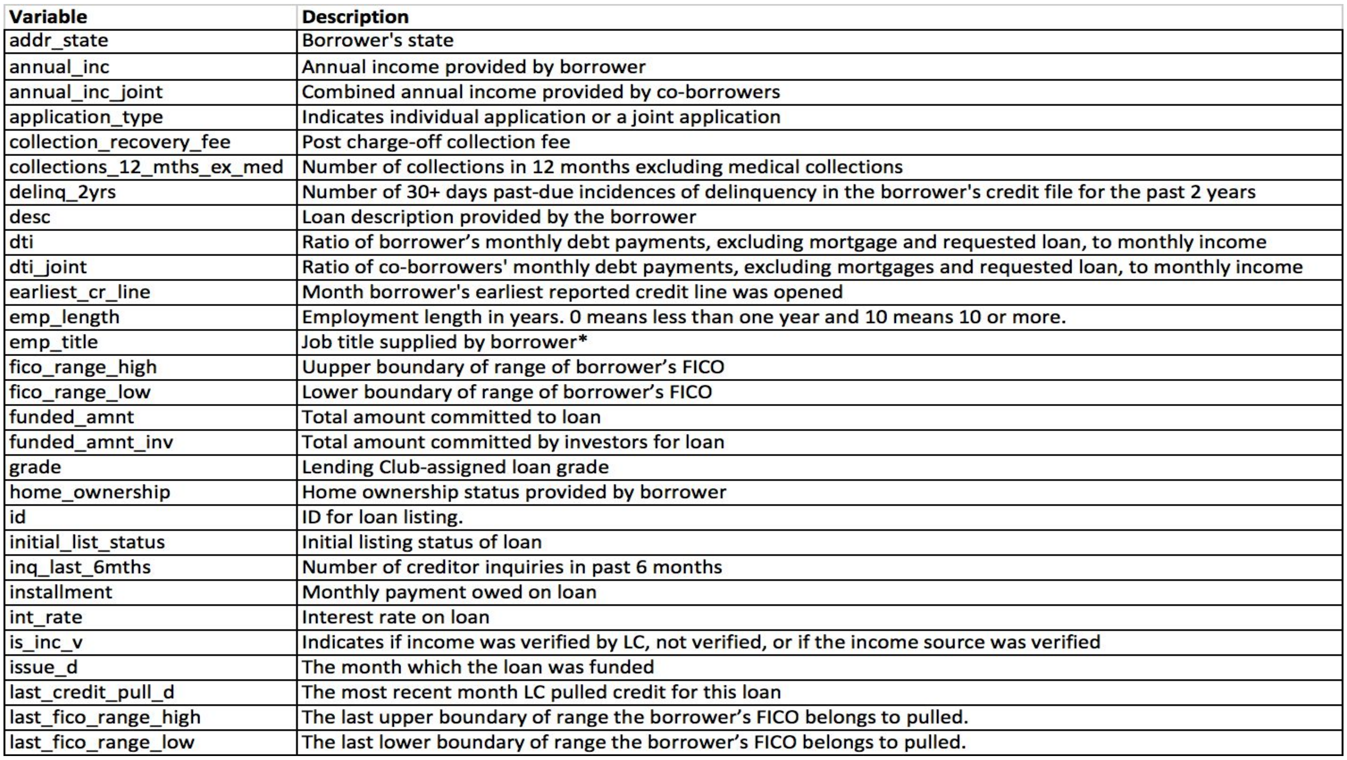
[2] http://www.bloomberg.com/news/articles/2015-05-14/wall-street-loves-peer-to-peer-loans-despite-concerns-of-a-bubble

[3] Website link to Lending Club data: https://www.lendingclub.com/info/download-data.action

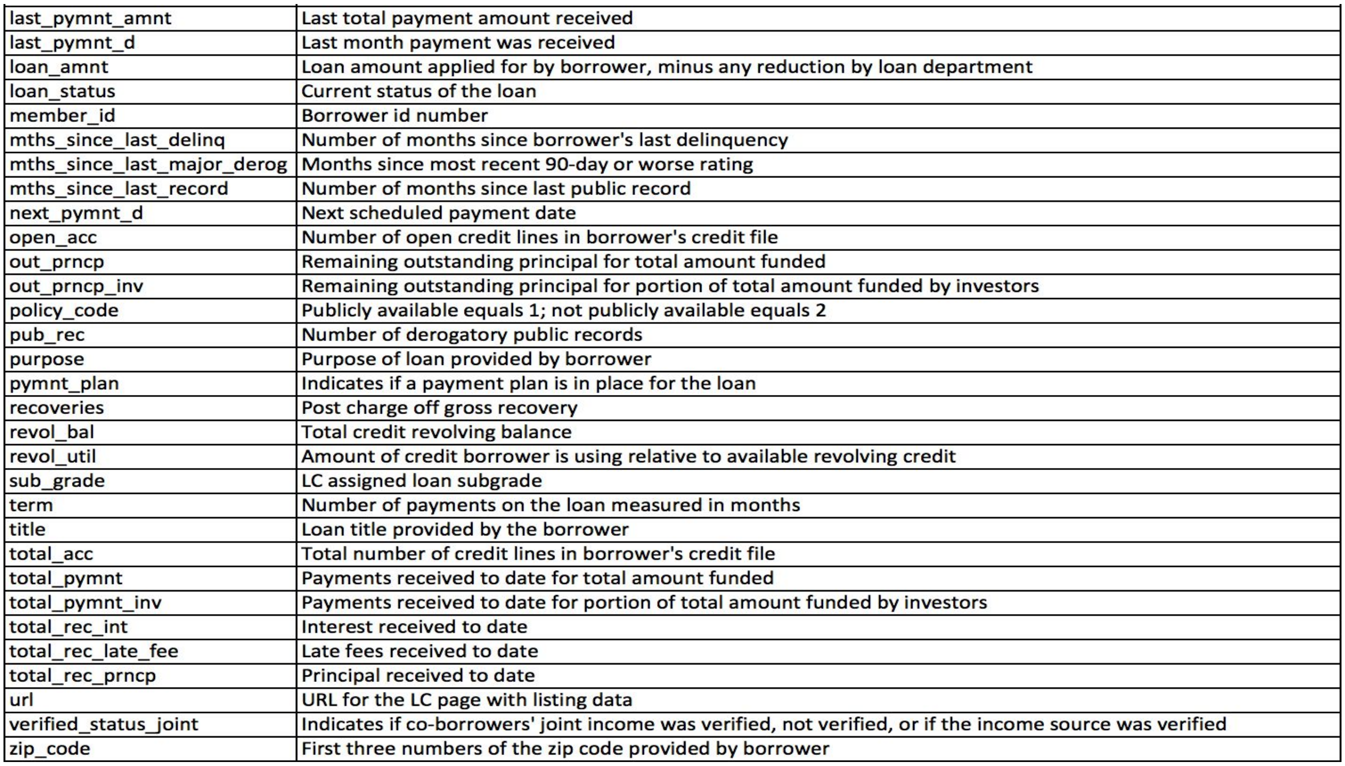


We use financial data about borrowers provided by Lending Club to inform potential investors about which loans to fund. The purpose of the model is to predict the probability of default. This data is then used to produce the group of loan options with the 10 highest expected returns.

The following predictor variables provided by Lending Club were used to build the model:



Variable	Description
addr_state	Borrower's state
annual_inc	Annual income provided by borrower
annual_inc_joint	Combined annual income provided by co-borrowers
application_type	Indicates individual application or a joint application
collection_recovery_fee	Post charge-off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	Number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
desc	Loan description provided by the borrower
dti	Ratio of borrower's monthly debt payments, excluding mortgage and requested loan, to monthly income
dti_joint	Ratio of co-borrowers' monthly debt payments, excluding mortgages and requested loan, to monthly income
earliest_cr_line	Month borrower's earliest reported credit line was opened
emp_length	Employment length in years. 0 means less than one year and 10 means 10 or more.
emp_title	Job title supplied by borrower*
fico_range_high	Uupper boundary of range of borrower's FICO
fico_range_low	Lower boundary of range of borrower's FICO
funded_amnt	Total amount committed to loan
funded_amnt_inv	Total amount committed by investors for loan
grade	Lending Club-assigned loan grade
home_ownership	Home ownership status provided by borrower
id	ID for loan listing.
initial_list_status	Initial listing status of loan
inq_last_6mths	Number of creditor inquiries in past 6 months
installment	Monthly payment owed on loan
int_rate	Interest rate on loan
is_inc_v	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The last upper boundary of range the borrower's FICO belongs to pulled.
last_fico_range_low	The last lower boundary of range the borrower's FICO belongs to pulled.



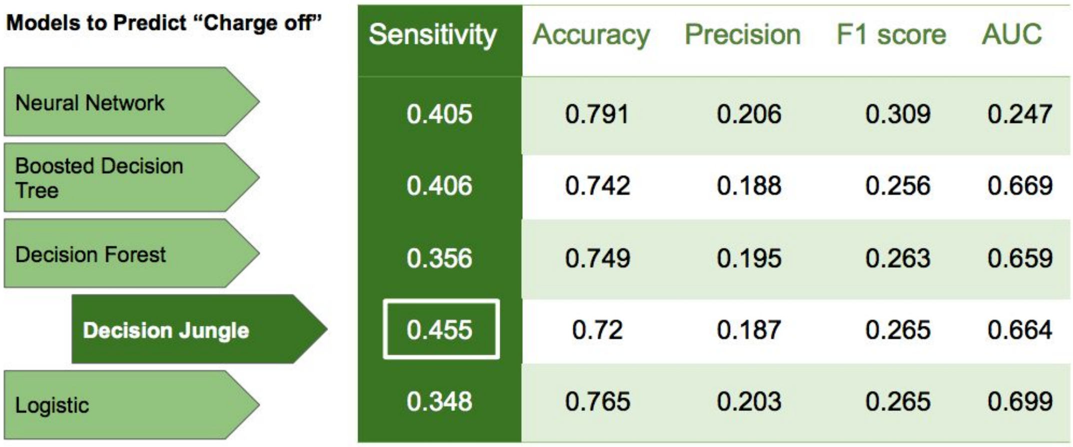
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	Loan amount applied for by borrower, minus any reduction by loan department
loan_status	Current status of the loan
member_id	Borrower id number
mths_since_last_delinq	Number of months since borrower's last delinquency
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	Number of months since last public record
next_pymnt_d	Next scheduled payment date
open_acc	Number of open credit lines in borrower's credit file
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
policy_code	Publicly available equals 1; not publicly available equals 2
pub_rec	Number of derogatory public records
purpose	Purpose of loan provided by borrower
pymnt_plan	Indicates if a payment plan is in place for the loan
recoveries	Post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Amount of credit borrower is using relative to available revolving credit
sub_grade	LC assigned loan subgrade
term	Number of payments on the loan measured in months
title	Loan title provided by the borrower
total_acc	Total number of credit lines in borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
url	URL for the LC page with listing data
verified_status_joint	Indicates if co-borrowers' joint income was verified, not verified, or if the income source was verified
zip_code	First three numbers of the zip code provided by borrower

Before building the model, any data observed after the loan was issued such as payment information, recoveries, and remaining outstanding principal were excluded. These variables cannot be observed before investors decide to fund a loan. They are also target leaks for default. Individual credit scores were not used because they weren't provided, though the grade system adopted by Lending Club positively reflect individual credit scores. Missing data were imputed by grade clustering. Exogenous variables such as unemployment and business cycles were not merged into the analysis of the dataset, as observations were treated as cross-sectional data.

Analysis

We used a decision jungle to predict defaults because it has the highest sensitivity, which measures how effective a model is at limiting the number of false-negative results in predictions. False negatives are loans that the model predicts will stay current but in fact default, whereas false-positive results are loans that are predicted to default but actually don't.

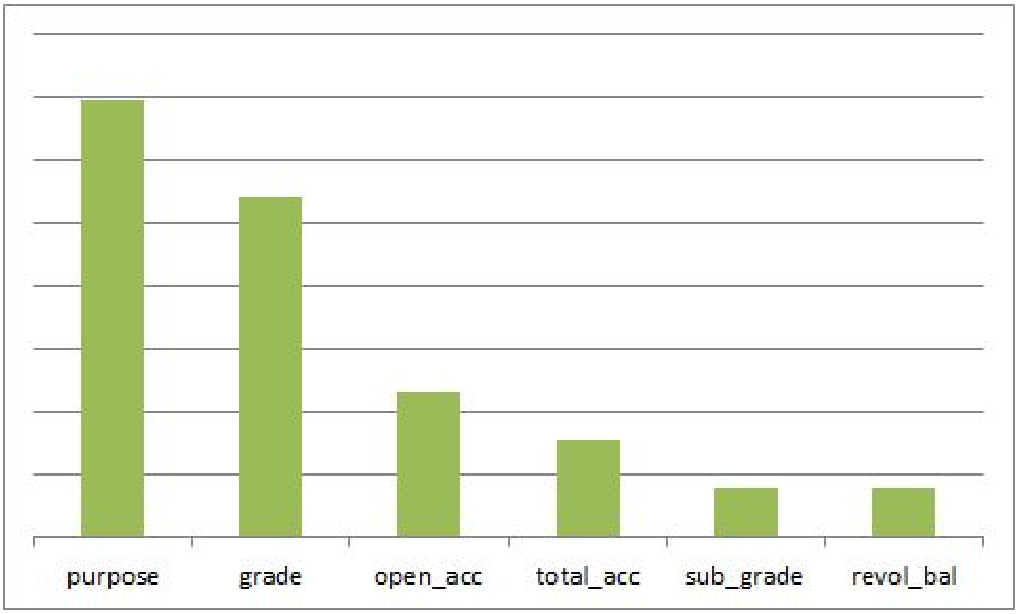
False negative carry the much higher costs to investors compared to the opportunity cost that comes with a false positive. We calculate the opportunity cost as the average return on loans, which is 8.5 percent. The average loss on a defaulted loan was 47.4 percent. Therefore, we focus on sensitivity. Decision jungle had the highest among the five tested at 0.455.



Models to Predict "Charge off"	Sensitivity	Accuracy	Precision	F1 score	AUC
Neural Network	0.405	0.791	0.206	0.309	0.247
Boosted Decision Tree	0.406	0.742	0.188	0.256	0.669
Decision Forest	0.356	0.749	0.195	0.263	0.659
Decision Jungle	0.455	0.72	0.187	0.265	0.664
Logistic	0.348	0.765	0.203	0.265	0.699

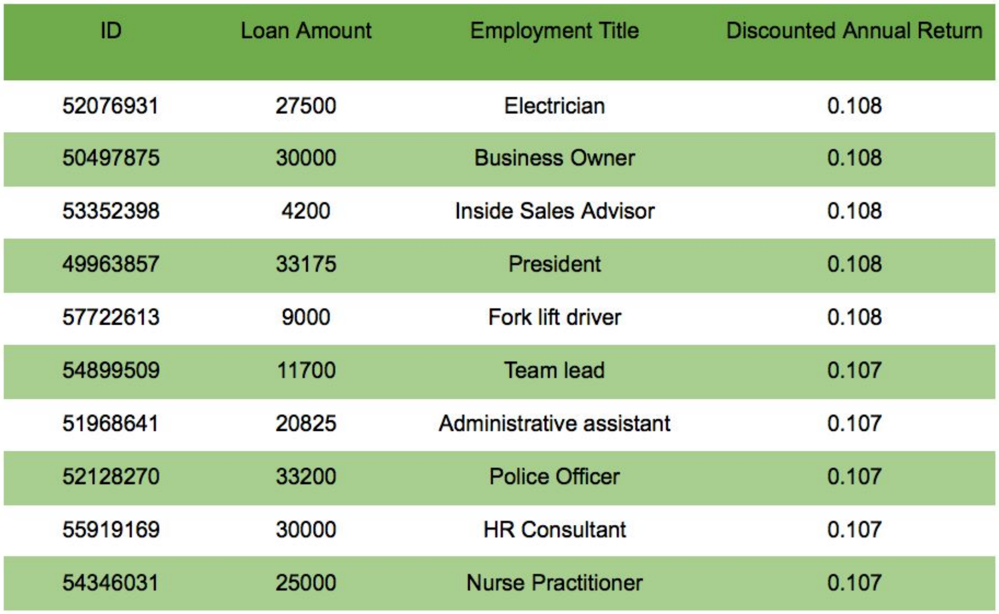
*As we have low target samples [Prob(charge off=1)= 0.15], the actual threshold when comparing models were set at 0.15

The most significant variables in the model are loan purpose (most of the defaulted loans were for debt consolidation), grade, and the number of open credit lines held by the borrower. The fact that grade is only the second most significant variable in our model suggests that the grading system adopted by Lending Club has potential flaws. Ideally, the grade system will correctly predict the default probability, and by correcting for the risk, Lending Club will then assign appropriate interest rates with each loan.



*Based on selected best model: Decision jungle

We can use the default model to generate a list of loans from 2015 with their predicted discounted annual returns. The table below shows the top 10 expected returns.



ID	Loan Amount	Employment Title	Discounted Annual Return
52076931	27500	Electrician	0.108
50497875	30000	Business Owner	0.108
53352398	4200	Inside Sales Advisor	0.108
49963857	33175	President	0.108
57722613	9000	Fork lift driver	0.108
54899509	11700	Team lead	0.107
51968641	20825	Administrative assistant	0.107
52128270	33200	Police Officer	0.107
55919169	30000	HR Consultant	0.107
54346031	25000	Nurse Practitioner	0.107

The loans all provide returns above 10 percent. Though not shown in this table, these loans carry high risks, as evaluated by Lending Club. An annual return of 10.8 percent compares favorably to the historical annual return of the S&P 500 stock index, which is about 9.6 percent, and is almost double the return on 10-year Treasury bonds.[4] As a result, it appears that peer-to-peer loans offer investors returns that are high enough to encourage investor participation in the platforms in the future. However, earning this level of return will probably require funding the highest-risk borrowers, which may turn off some lenders, particularly retail investors. In addition, the ranking of the best investment opportunities doesn't take into account the volatility of returns to determine how the risk-return profiles compare to other investment options. Calculating the volatility of returns would be an important step for future research to help determine the viability of peer-to-peer lending.

Our model identified potential flaws in Lending Club's grading system. Although the grade system provides a straightforward method for investors to evaluate the risk without sophisticated modeling techniques, it's reliability is questionable. With the reevaluation of the grade model, we can use post-observational adjustments to build a better model that more accurately represents the risks that investors face. This would serve as a correction method that helps both Lending Club and investors better realize the risks associated with individual borrowers. With our model, and given more data, we would be able to provide a better prediction of the probability of default, creating greater opportunity for investors to maximize their returns.

[4] http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html