Capstone Project - Predicting Peer-to-Peer Loan Payback Amount

Milestone Report

Problem Statement:

Peer-to-peer lenders risk a financial loss if a loan defaults and is not paid back in full. One way to try to minimize the loss is to try try to predict whether a loan will be paid back or not (classification). While predicting if a loan will default or not is valuable, it doesn't take everything into account. For example, there could be a situation where a high-interest loan defaults after being nearly paid off, that is actually pays back the lender more than a low-interest rate loan that is paid in full. Therefore, this project will take a different approach and predict the amount that will be paid back.

My client is an investor that wants to maximize their return on their lending. When an investor obtains a list of available loans they can invest in, they could run the model and generate a prediction of how much will be paid back from their investment, and choose the loans that will maximize their return.

Data

The data for this project comes from https://www.lendingclub.com/info/download-data.action. The data is available for download in csv format. Lending Club loans are for either 3 or 5 years, therefore, the data used was for years 2012-2013 since this is the latest data that contains completed 5-year loans. There are 188,183 loan observations, and 145 columns in the original data set.

Data Dictionary

Variable	Description
bc_open_to_buy	Total open to buy on revolving bankcards.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
disbursement_method	The method by which the borrower receives their loan. Possible values are: CASH, DIRECT_PAY
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
funded_amnt	The total amount committed to that loan at that point in time.

grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
num_actv_rev_tl	Number of currently active revolving trades
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
open_acc	The number of open credit lines in the borrower's credit file.
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
purpose	A category provided by the borrower for the loan request.
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
title	The loan title provided by the borrower
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_pymnt	Payments received to date for total amount funded
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Data Wrangling

The first steps were remove the unnecessary data. Only the variables that contain information known before the loan is funded will be useful for predicting the amount that will be paid, so any column that contains information that is collected after the loan is funded will be removed, except for `total_pymnt`, which is the dependent variable. Incomplete loans were removed from the data, since the dependent variable `total_pymnt` is not yet known for these.

Next, I took a look at the missing values. There were several columns where all observations were null, which were dropped.

There were 20 columns that had the same number of missing observations. I found that there was no data collected for these columns in the first 6 months of the 2-years of data. I removed these observations from the data, leaving me with 1.5 years of complete data, rather than 2-years with a lot of imputed data.

There were a few variables that represent months since a delinquency occurred, such as months since there was a delinquency on revolving credit. In these cases, I suspect that missing values represent loans where the applicant hasn't had the delinquency, which is meaningful but is something that can't be represented numerically like the number of months. Therefore, I converted these variables to category, with categories ‘none’, ‘1-2 years’, etc.

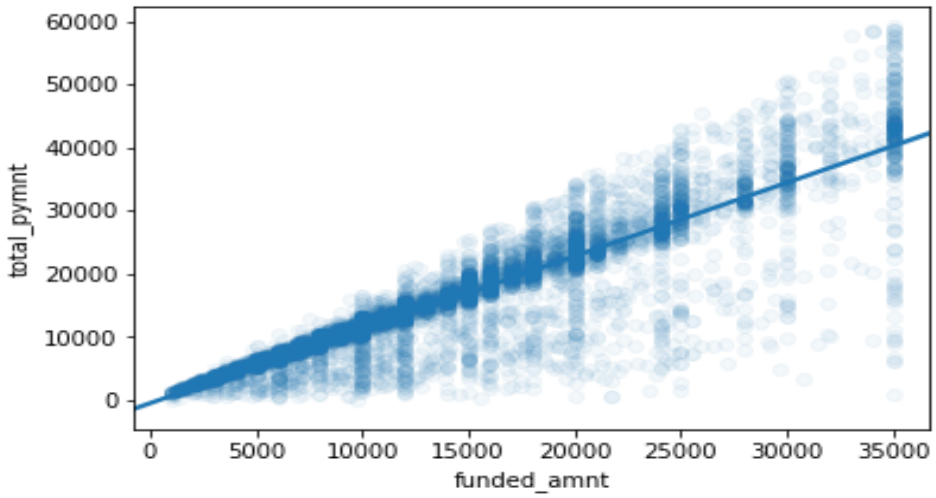
The remaining missing data was replaced with the median.

I also checked for columns that only had one unique value, and removed them since they won’t add anything to the model.

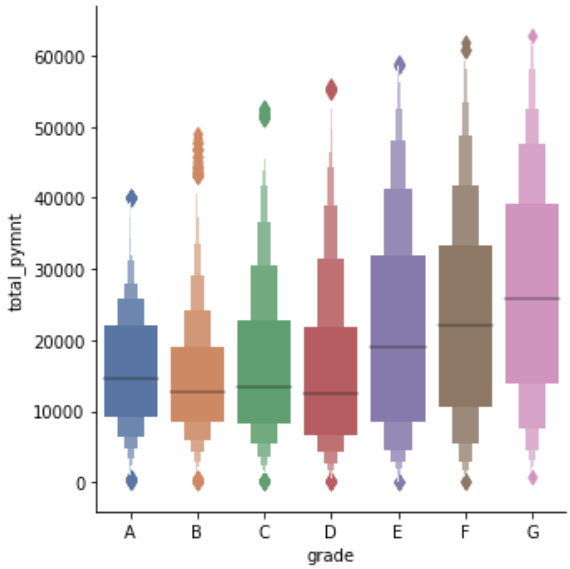
Once the number of columns was narrowed down, I checked to make sure they were all of the right type. Dates were converted to date variables, and objects were converted to categorical. Two variables were percentages, but read in as strings, so I removed the ‘%’ symbol, and converted them to floats.

Data Storytelling

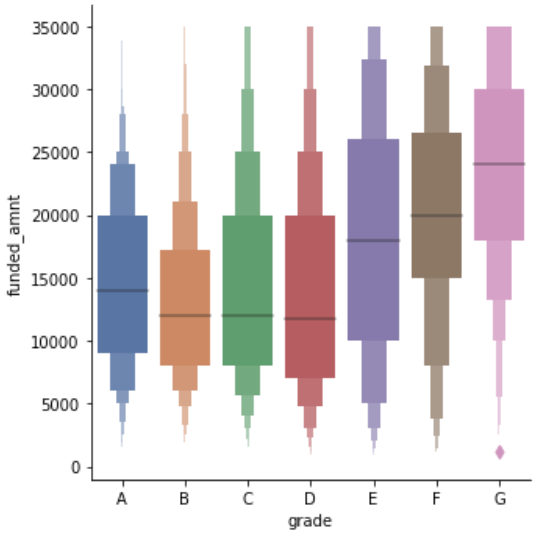
The original dependent variable was intended to be the amount paid back on the loan, however, it is so heavily influenced by the original loan amount, it is hard to interpret the effects of the independent variables.



Looking at the total payments vs the amount funded shows that there is a strong linear relationship. While this is expected, it causes some difficulties. For example, loans with worse grades have higher amounts paid back than loans with better grades.



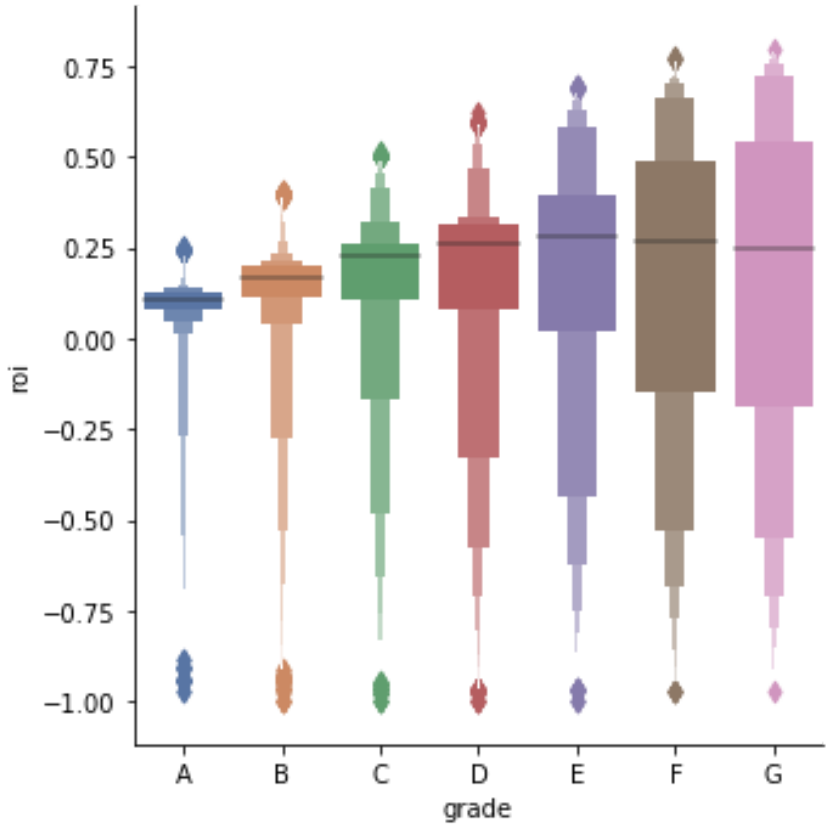
Looking at the amount repaid by grade, it looks like the G-grade loans pay back the most. Since G is the lowest grade, I expect that the only thing that could account for the high repayment amount is a high initial loan amount.



Looking at the funded amount by grade confirms that the G-grade loans indeed have the highest median loan amount of about \$25,000.

This suggests that the amount requested affects the assigned grade, rather than a pre-determined grade affecting how much an applicant is allowed to borrow.

This still doesn't do a very good job of showing how well a loan performs compared to how it is expected to perform based on the original amount funded. To achieve this, I created an ROI variable, which is the difference between the amount repaid and the amount borrowed divided by the amount borrowed. The ROI would be zero if the borrower paid back exactly what they borrowed, and negative if they repay less than they originally borrowed. The ROI is expected to be above zero to reflect the interest paid.

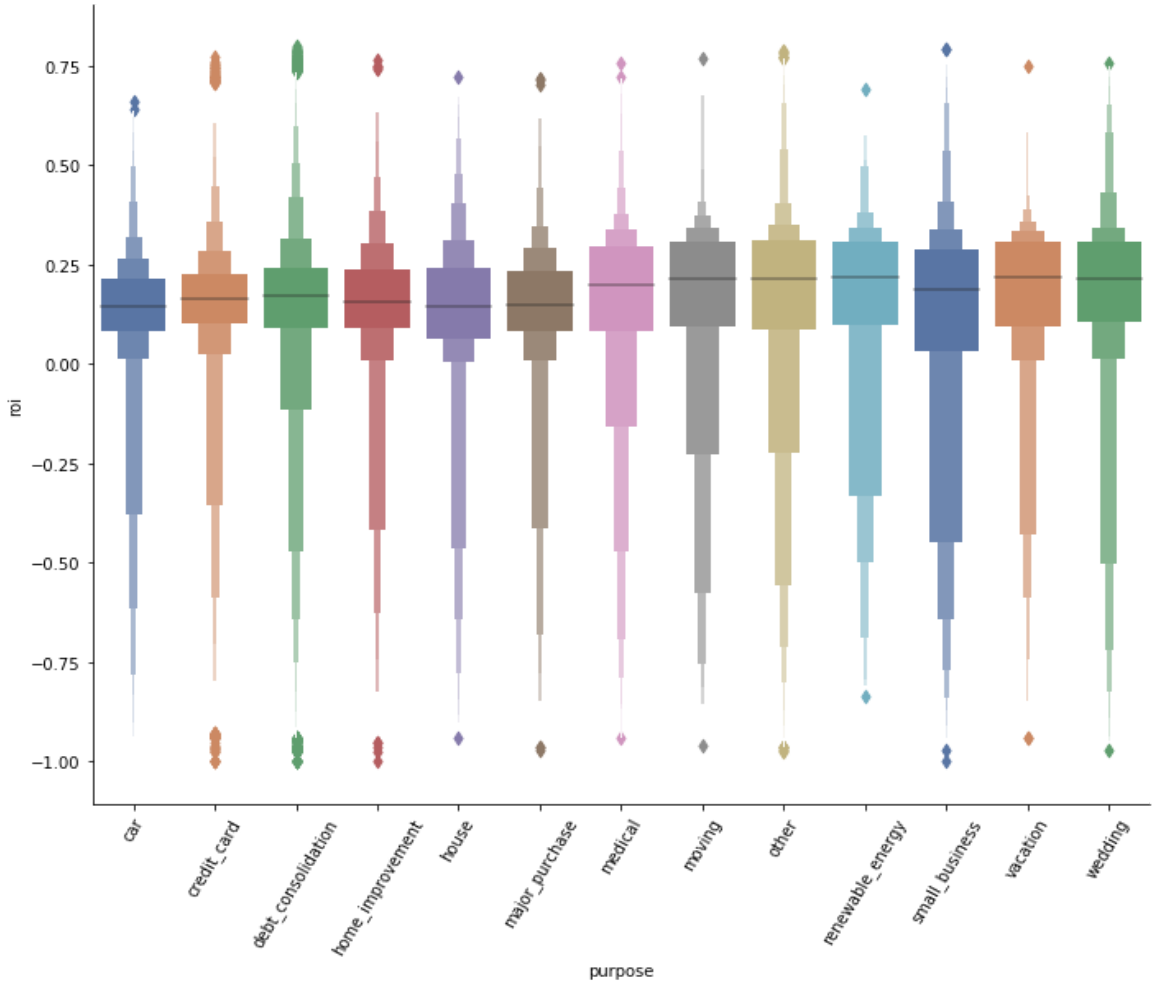


Looking at the ROI by grade is much more interesting than just the amount repaid by grade since it removes the effects of loan size.

The least-risky A-grade loans have the lowest median ROI of around 10%, with a mean of 8%.

Grade E loans have the highest mean ROI of 16.0% and highest median ROI of 28.4%, meaning that on average, E loans will give investors the highest returns. This demonstrates that the higher interest rates on the lower-graded loans more than makes up for the losses due to default through grade E. Past grade E, the defaults start eating more into the additional profits made from the higher interest rates.

Loan Purpose



The ROI by purpose can be broadly put into two groups. Loans for cars, credit cards, debt consolidation, home improvement, and house have lower ROIs, but are less risky as they have lower standard deviations.

Loans for medical, moving, other, renewable energy, vacation, and weddings tend to have better ROIs than the first group, bit also more risk of default with higher standard deviations.

Small business loans are the one category that is hard to put into these two groups, as it has the lower median ROI of the first group, and the higher risk associated with the second group. It also has both the lowest mean ROI and the highest standard deviation, making it, on average, the worst loan purpose to invest in.

I set up a hypothesis test to determine whether the ROI between the first and second groups is statistically significant:

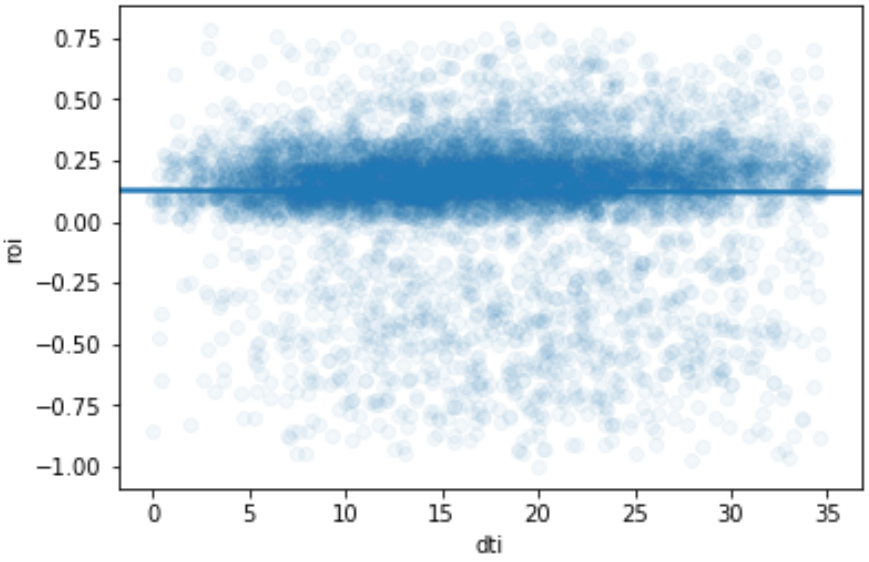
Null hypothesis: There is not a difference in mean ROI between group_1 and group 2 borrowers.
Alternate Hypothesis: There is a difference in mean ROI between group_1 and group 2 borrowers.

Test Statistic: Difference of mean ROI.
Significance level: 0.01

The observed difference in mean ROI was 0.6 %. I took 10,000 bootstrap permutation samples, and calculated the test statistic. None of the 10,000 permutation samples produced a difference of means as great as 0.6%. Therefore, the null hypothesis was rejected.

ROI by DTI

Looking at ROI by DTI (Debt to Income), there is a very slight negative correlation that is hard to detect visually.



I created two groups, a low DTI group with DTI below 10, and a high DTI group with DTI above 25, and set up a hypothesis test to determine if the mean ROI between the groups is statistically significant:

Null hypothesis: There is not a difference in mean ROI between low DTI borrowers(DTI < 10) and high DTI borrowers (DTI >= 25).

Alternate Hypothesis: There is a difference in mean ROI between low DTI borrowers(DTI < 10) and high DTI borrowers (DTI >= 25).

Test Statistic: Difference of mean ROI.

Significance level: 0.01

The observed difference in mean ROI was 0.8 %. I took 10,000 bootstrap permutation samples, and calculated the test statistic. Only one of the 10,000 permutation samples produced a difference of means as great as 0.6%, resulting in a p-value of 0.0001, which is below the 0.01 significance level. Therefore, the null hypothesis was rejected.

Next Steps

The next steps for the project will be to perform the machine learning to make the predictions. This will involve creating dummy variables for the categorical data, then running and tuning linear regression, random forest, and XGBoost models. Then I can compare the results with various metrics including R-squared, residual analysis, and a graph of the expected vs predicted results.