Loan Default Prediction:

Maximizing Return on Peer-to-Peer Lending Investments

Samuel P. Sears

Bellevue University

Abstract

In this project, loan data for a peer-to-peer lending platform is obtained and analyzed with a goal of maximizing return for investors. Variables from the dataset are evaluated to determine if they can be used for prediction of the target variable, loan status. Each variable is then analyzed to describe its distribution and its association with whether or not a loan was paid in full. Following exploratory data analysis, preprocessing and predictive modeling is used to create a loan selection process that maximizes return on investment. This loan selection process is compared to using FICO scores and loan grade for selecting loans. Finally, ideas for further analysis and key learnings are provided.

Keywords: unbalanced classification, machine learning, binary classification, pycaret, loan prediction

Loan Default Prediction:

Maximizing Return on Peer-to-Peer Lending Investments

Traditionally if someone needs a loan, whether it be for a small business, educational expenses, wedding, or anything else, they would go their bank or some other financial institution. That financial institution has the ability to loan money to borrowers because of pooled funds from people who utilize the bank. In a way, financial institutions take other people's money and lend it to borrowers. Peer-to-peer lending takes the financial institution out the equation and provides the opportunity for individuals to fund loans for borrowers (Kagan, 2020). The peer-to-peer lenders are individual investors looking to make a better return than savings accounts offer. According to Bankrate, the average interest rate for savings accounts as of June 30, 2021 was 0.06 percent in the United States (Goldberg, 2021).

One of the pioneering companies of the peer-to-peer lending industry was LendingClub. However, in 2020, the company announced it would be ending its peer-to-peer lending model (Frankel, 2020). Nevertheless, several platforms still exist today including Peerform, Upstart, Prosper, Funding Circle, and Payoff to name a few (Black, 2021). Before LendingClub exited the industry, it left a mass of publicly available loan data which countless individuals had scraped and compiled into Kaggle datasets. This provided a perfect opportunity for analysis and predictive modeling.

Business Problem and Hypothesis

In this data, various pieces of information were gathered on each loan including whether the loan was paid in full or not. As an investor, the outcome of the loan is of the utmost importance. If a loan is not paid in full, there would likely be a loss on their investment. Therefore, if the outcome of a loan could be accurately predicted prior to selecting the loan for investment, the investor would be able to ensure a return on their investment. This isn't the only aspect to consider when maximizing returns, however. Loans that are less likely to be paid in full also are typically loans with higher interest rates and higher returns rates. Therefore, being able to predict the outcome of not just any loan, but higher

interest loans in particular, would equate to the highest return. Given the data made available on Kaggle, could a predictive model be utilized to create a profitable loan selection strategy?

Data

The data for this project came from a dataset made available on Kaggle which provided information on accepted and rejected loans from 2007 to 2018 (George, 2019). In total, there was over 2.2 million accepted loans available to be analyzed. For each loan, 151 data points were provided. Due to time constraints, only the first 60 variables were able to be fully researched and considered. The goal of researching the variables was to determine if there was duplicate, not useful, or future information. Only information that would be available at the origination of the loan could be considered. After reviewing an article from DataQuest (Osei, 2019) and reviewing the data dictionary from LendingClub (LendingClub, n.d.), several variables were eliminated. In total, 14 numeric and 6 categorical variables were used in predictive modeling. The target variable used was loan_status. For this analysis, only loans with final outcomes were considered. This meant that only loans with the status of Charged Off and Fully Paid were analyzed. Refer to Appendix A for a full list of variables used and their descriptions.

Exploratory Data Analysis

The first step in the project was to explore the data to become familiar with the distributions of values for each variable and to look for key insightful information for investors. In practice, this meant segmenting the data by ranges of values within variables and observing the fully paid vs charged off rates within the segments. Further, a comprehensive profile report was generated which provided summary and correlation information for all variables.

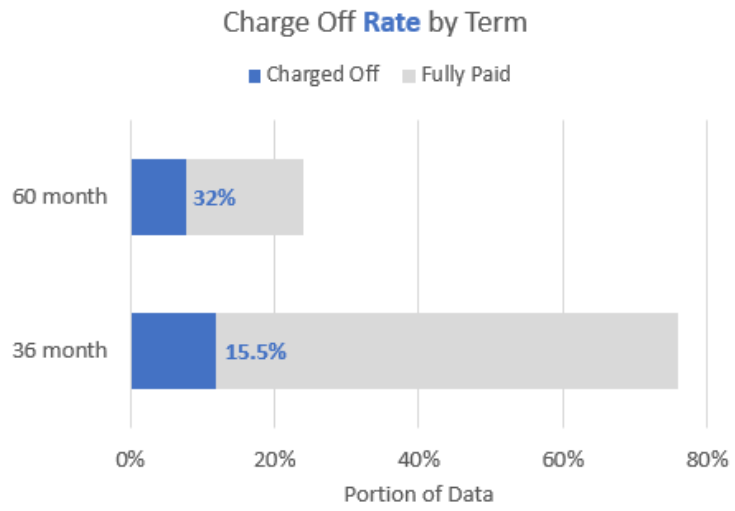
The first finding in exploratory data analysis was that only 19% of the loans were charged off. The imbalance in the target variable was an important issue to address when fitting a model later in the project since models typically have difficulty learning the underrepresented class. Another finding applicable to modeling was that eight explanatory variables had outliers and 10 had skew. Since some

models are sensitive to outliers and skew, transformations and normalization would be considered in preprocessing the data for modeling.

Beyond that, segments were also analyzed in hopes of finding segments of loans with abnormally large or small charge off rates. For the two term lengths for the loans in the data, 36 and 60 months, the later had a charge off rate of over double that of the prior at 32%. However, there were about three times as many 36-month loans as there were 60-month loans as shown in Figure 1.

Figure 1

Charge Off Rates by Term

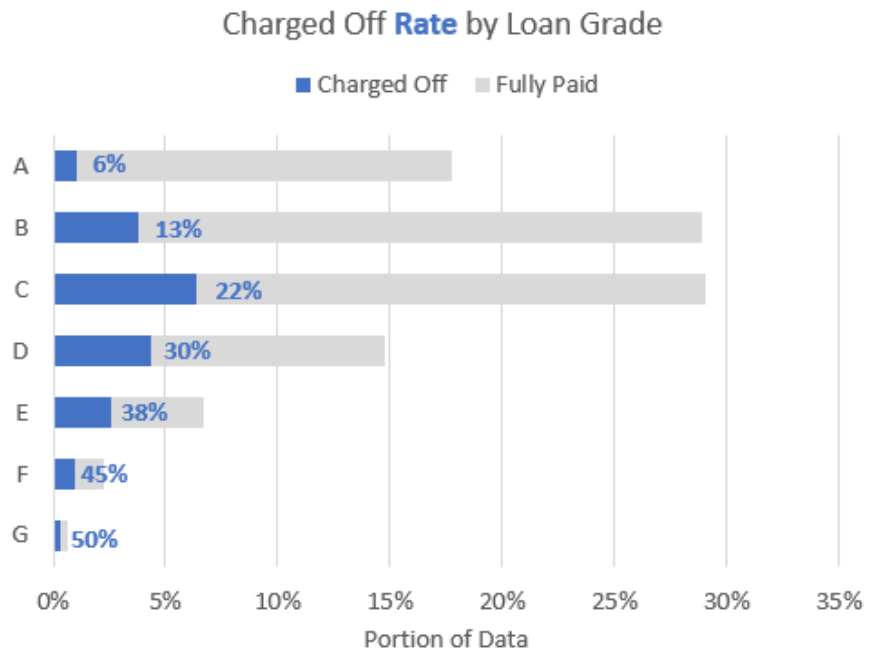


When breaking out the data by monthly installments, a positive association was seen between the amount of monthly installments and the charge off rates. In other words, the higher the monthly installment, the more likely it was that a borrower would have their loan charged off. However, these higher monthly installments were less prevalent. Similarly, higher debt-to-income ratios were also associated with higher rates of charge off. To contrast, there were negative associations between FICO scores, annual income, and loan grade and charge off rate. This was expected as lower values in each of those variables would typically indicate a riskier loan. It's normal for a borrower with a high FICO score and high annual income to get a higher-grade loan which is deemed less risky. This dataset supported

the notion that loan grades reflect risk. Charge off rates increased as the loan grade decreased as shown in Figure 2.

Figure 2

Charge Off Rate by Loan Grade



The general theme from the findings in exploratory data analysis was that there were no surprises in terms of charge off rates. As an investor, finding a segment of loans that were traditionally deemed riskier, but had lower rates of charge off would have been the best-case scenario. For example, if we found a segment of customers with low FICO scores but low charge off rates, investing in those borrowers would prove to be profitable. Their loans would have higher returns. This was not observed, though. Perhaps a predictive model would be able to help in this space.

Predictive Modeling

The first step in predictive modeling was to split the data. Even though only the loans with final outcomes were used, there were still over 1.15 million rows of data. This would make training a model very time consuming. So, random samples of the data were taken for training and testing. During the

data splitting phase, the class imbalance issue was addressed. Since there was an abundance of data, the majority class, loans that were paid in full, were under sampled so the training data set had an equal number of paid in full and charged off loans. This would ensure the model learned the minority class as well as the majority class. For the test data sets, 5 separate 100-thousand representative samples were taken. In these sets, the proportion of charged off loans to paid in full loans reflected that of the dataset as a whole. These sets were withheld for use in evaluating the model on new data it had not seen before.

For the preprocessing phase, a variety of preprocessing steps were tested to determine their effect on model performance. Model performance was determined by comparing AUC scores. AUC is a standard evaluation metric for unbalanced classification problems and discourages choosing a model that is representational, but does a poor job of separating the target class. Preprocessing steps that were tested included normalization, transformation, removal of outliers, and ignoring low variance. Without getting too technical, these preprocessing steps are methods to deal with skewness, data at different scales, and outliers which can have negative effects on certain models. In the end, it was seen that a CatBoost Classifier with normalization, removal of outliers, and changing loan grade to numeric provided the best AUC at 0.7151. Other models tested were Light Gradient Boosting Machine, Gradient Boosting Classifier, and Extreme Gradient Boosting Classifier. Once the CatBoost model and preprocessing steps were solidified, the model was tuned to optimize AUC using the PyCaret tune_model method.

With the final model complete, it was ready to make predictions on the first test dataset. The AUC on the testing data set was 0.655, a bit lower than the training dataset. Nevertheless, breaking down the model by prediction probabilities showed promise. For each prediction, a binary classification model gives a probability that the observation is a positive. In this case, if a loan was given a high probability, there was a higher chance that it was paid in full. Segmenting out the loans by probabilities

given by the model showed that higher model probabilities indeed correlated with higher rates of loans being paid in full. For example, when the model gave a probability between 95% and 100% that the loan would be paid in full, 98.6% of the time the loan actually was paid in full. When the model gave a probability between 5% and 10% that the loan would be paid in full, it was only paid in full 23.8% of the time.

Application

So how could this model be applied in order to maximize return for investors? One strategy was to invest in any loans that had a model probability greater than a certain threshold. With that threshold set somewhere between 70% and 97%, the return would be about 6% regardless of the threshold as seen in Appendix B. The lower the threshold, however, the more loans that would be included and, thus, the larger the investment would need to be. A similar selection process was used by simply filtering loans by FICO and loan grade as shown in Appendix C. Investing in all loans above a grade B, an investor could expect a 5.1% return on investment. However, they would need to invest over 600 million dollars to ensure that. In short, there's two main issues. The return on the investment isn't incredibly appealing. 6% return over 3 to 5 years barely rivals inflation. Further, to ensure those rates, an extremely large amount of capital would be needed.

Referring back to the original problem statement, predicting whether a loan will be paid in full is not enough. The model would need to be able to predict when a highly profitable loan will be paid in full. Luckily, the return potential was a metric that could be calculated at loan origination using the loan amount, term, and monthly installments. The question then became, if only loans with a potential return above a certain percentage and model probability above a certain percentage were selected, could a high actual return be ensured? A grid search calculating actual returns for different model probability and potential return percentages was conducted to determine which thresholds resulted in the highest returns. The thresholds that resulted in highest mean return were 0.55 model probability

and 45% return potential. In other words, if only loans with a model probability greater than 0.55 and return potential greater than 45% were selected as investments, the average return would be 15.6%. For the 5 different test sets used, the return rate ranged from 11% to 18% as seen in Figure 3. Further, the total investment required ranged from roughly 1.6 million to 2.1 million. Because of this, one of the main issues of having to invest a lot of capital to ensure a higher reward still might not have been achieved. However, this is was expected since diversification is a general theme in investing.

Figure 3

Returns for model probability 0.55 and return potential 45%

```
Model Score 0.55
Return Potential 45%

Test Set A
Investment: $2,094,700.00
Return: $341,931.27 (16%)

Test Set B
Investment: $1,651,150.00
Return: $287,614.09 (17%)

Test Set C
Investment: $2,048,050.00
Return: $324,628.08 (16%)

Test Set D
Investment: $1,763,950.00
Return: $199,015.67 (11%)

Test Set E
Investment: $1,808,325.00
Return: $319,421.78 (18%)
```

Further Analysis

Though this project showed promising results for devising a profitable loan selection strategy for peer-to-peer lending investors, more work is still needed. The method of using probability thresholds and return potential thresholds resulted in somewhat varied results across the different test sets. Therefore, further validation and review would be necessary before moving forward. Another idea for further analysis would be to create a model that doesn't predict whether a loan will default or not, but

to create a model that predicts the return on investment. If the model is able to accurately predict the return, it would be more suited for solving the problem of maximizing returns. This was a key learning in this study. Predicting whether a loan will default or not is not the only aspect to consider for an investor. The investor must be able to predict when a highly profitable, more risky loan will be paid in full or not.

References

Black, M. (2021, April 29). *Best Peer-to-Peer Lending.* Investopedia. https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp

Frankel, M. (2020, October 8). *LendingClub Is Ending Its P2P Lending Platform – Now What?*. The Ascent. https://www.fool.com/the-ascent/personal-loans/articles/lendingclub-ending-its-p2p-lending-platform-now-what/

George, N. (2019). *All Lending Club loan data*. Kaggle. https://www.kaggle.com/wordsforthewise/lending-club

Goldberg, M. (2021, July 2). *What is the average interest rate for savings accounts?*. Bankrate. https://www.bankrate.com/banking/savings/average-savings-interest-rates/

Kagan, J. (2020, May 11). *Peer-to-Peer (P2P) Lending*. Investopedia. https://www.investopedia.com/terms/p/peer-to-peer-lending.asp

LendingClub. (n.d.) *LCDataDictionary.* [Data Set]. https://resources.lendingclub.com/LCDataDictionary.xlsx

Osei, D. (2019, June 19). *Data Cleaning and Preparation for Machine Learning*. DataQuest. https://www.dataquest.io/blog/machine-learning-preparing-data/

Appendix A

Variable	Type	Description
loan_amnt	Numeric	Loan amount
term	Categorical	Term length of the loan, either 36 or 60 months
installment	Numeric	Monthly repayment amount
grade	Categorical	Loan rating, A-G rating where A is least risk
emp_length	Numeric	Employment length in years, 0-10+
home_ownership	Categorical	Home ownership status, Rent, Own, Mortgage, or Other
annual_income	Numeric	Self-reported annual income of borrower
verification_status	Boolean	Indicates if income was verified
purpose	Categorical	Category provided by the borrower for loan request
dti	Numeric	Debt-to-income ratio of the borrower
delinq_2yrs	Numeric	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
earliest_cr_line	Numeric	Number of months between loan issue date and earliest reported credit line for the borrower
inq_last_6mnths	Numeric	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
open_acc	Numeric	The number of open credit lines in the borrower's credit file
pub_rec	Numeric	Number of derogatory public records
revol_bal	Numeric	Total credit revolving balance
revol_util	Numeric	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit

total_acc	Numeric	The total number of credit lines currently in the borrower's credit file
application_type	Categorical	Indicates whether the loan is an individual application or a joint application with two co-borrowers
fico	Numeric	Average of fico_range_low and fico_range_high
loan_status	Categorical	Current status of the loan, target variable

Appendix B

Using Model Probabilities to Select Loans for Investment

	Score	Investment	Return	Return Percent
0	> 0.70	$301,186,150.00	$18,806,361.82	6.2%
1	> 0.71	$284,085,750.00	$17,627,334.02	6.2%
2	> 0.72	$267,696,900.00	$16,639,044.08	6.2%
3	> 0.73	$251,656,250.00	$15,596,469.67	6.2%
4	> 0.74	$235,879,850.00	$14,460,279.06	6.1%
5	> 0.75	$220,747,475.00	$13,524,766.95	6.1%
6	> 0.76	$206,089,025.00	$12,570,202.89	6.1%
7	> 0.77	$190,849,525.00	$11,492,270.98	6.0%
8	> 0.78	$176,867,425.00	$10,662,798.26	6.0%
9	> 0.79	$161,892,175.00	$9,817,142.03	6.1%
10	> 0.80	$147,862,725.00	$9,067,650.84	6.1%
11	> 0.81	$133,191,275.00	$8,145,937.33	6.1%
12	> 0.82	$118,258,325.00	$7,188,015.74	6.1%
13	> 0.83	$103,720,675.00	$6,319,160.97	6.1%
14	> 0.84	$89,823,425.00	$5,499,568.62	6.1%
15	> 0.85	$76,467,425.00	$4,622,547.34	6.0%
16	> 0.86	$63,164,875.00	$3,792,146.20	6.0%
17	> 0.87	$50,746,500.00	$2,955,886.49	5.8%
18	> 0.88	$38,845,550.00	$2,226,391.95	5.7%
19	> 0.89	$29,185,475.00	$1,748,669.84	6.0%
20	> 0.90	$20,412,200.00	$1,234,790.02	6.0%
21	> 0.91	$13,677,925.00	$831,879.07	6.1%
22	> 0.92	$8,269,075.00	$559,382.64	6.8%
23	> 0.93	$4,165,625.00	$277,211.14	6.7%
24	> 0.94	$1,412,125.00	$97,570.22	6.9%
25	> 0.95	$449,550.00	$28,157.46	6.3%
26	> 0.96	$124,300.00	$5,190.48	4.2%
27	> 0.97	$12,600.00	$813.90	6.5%

Appendix C

Loan Selection Based on FICO and Loan Grade Thresholds

	FICO	Investment	Return	Return Percent
14	> 840	$639,125.00	($19,325.81)	-3.0%
13	> 830	$1,825,800.00	$17,235.66	0.9%
12	> 820	$4,722,875.00	$117,580.25	2.5%
11	> 810	$9,590,725.00	$330,300.34	3.4%
10	> 800	$18,035,550.00	$680,673.76	3.8%
9	> 790	$28,987,950.00	$891,151.02	3.1%
8	> 780	$42,523,125.00	$1,363,990.21	3.2%
7	> 770	$61,093,100.00	$1,845,681.05	3.0%
6	> 760	$84,770,100.00	$2,279,095.04	2.7%
5	> 750	$116,039,075.00	$3,527,805.02	3.0%
4	> 740	$159,835,250.00	$4,793,856.81	3.0%
3	> 730	$222,367,525.00	$7,128,162.05	3.2%
2	> 720	$313,465,825.00	$10,546,842.22	3.4%
1	> 710	$439,177,375.00	$15,958,308.19	3.6%
0	> 700	$594,961,275.00	$21,532,477.57	3.6%

Figure C1. Loan selection by investing in all loans with borrower's FICO scores above a certain threshold.

	Grade	Investment	Return	Return Percent
0	A	$245,454,075.00	$12,622,301.91	5.1%
1	B+	$622,146,300.00	$31,666,138.74	5.1%
2	C+	$1,035,893,950.00	$41,948,090.44	4.0%
3	D+	$1,258,940,825.00	$42,853,184.45	3.4%
4	E+	$1,376,237,325.00	$40,603,722.92	3.0%
5	F+	$1,416,926,275.00	$38,697,269.68	2.7%
6	G+	$1,429,701,975.00	$37,753,355.10	2.6%

Figure C2. Loan selection by investing in all loans with grades above a certain threshold.

	FICO	Grade	Investment	Return	Return Percent
0	> 840	A	$436,200.00	$6,159.57	1.4%
1	> 800	A	$11,807,100.00	$455,398.07	3.9%
2	> 760	A	$49,446,675.00	$2,171,819.22	4.4%
3	> 720	A	$135,319,775.00	$6,510,864.90	4.8%
4	> 840	B	$566,050.00	($134.35)	-0.0%
5	> 800	B	$15,907,025.00	$573,350.46	3.6%
6	> 760	B	$70,185,925.00	$2,564,155.24	3.7%
7	> 720	B	$225,860,150.00	$10,241,684.42	4.5%
8	> 840	C	$639,125.00	($19,325.81)	-3.0%
9	> 800	C	$17,512,925.00	$653,429.16	3.7%
10	> 760	C	$80,224,850.00	$2,664,969.18	3.3%
11	> 720	C	$282,183,275.00	$11,639,889.65	4.1%
12	> 840	D	$639,125.00	($19,325.81)	-3.0%
13	> 800	D	$17,898,000.00	$700,645.42	3.9%
14	> 760	D	$83,221,150.00	$2,456,632.45	3.0%
15	> 720	D	$302,634,900.00	$11,410,011.01	3.8%
16	> 840	E	$639,125.00	($19,325.81)	-3.0%
17	> 800	E	$18,002,325.00	$679,649.28	3.8%
18	> 760	E	$84,289,525.00	$2,369,303.48	2.8%
19	> 720	E	$310,319,300.00	$10,929,565.07	3.5%
20	> 840	F	$639,125.00	($19,325.81)	-3.0%
21	> 800	F	$18,002,325.00	$679,649.28	3.8%
22	> 760	F	$84,644,125.00	$2,321,314.05	2.7%
23	> 720	F	$312,625,500.00	$10,760,524.95	3.4%
24	> 840	G	$639,125.00	($19,325.81)	-3.0%
25	> 800	G	$18,035,550.00	$680,673.76	3.8%
26	> 760	G	$84,770,100.00	$2,279,095.04	2.7%
27	> 720	G	$313,465,825.00	$10,546,842.22	3.4%

Figure C3. Loan selection by investing in all loans with grades and FICO scores above a certain threshold.