Maximizing Return on Investment from LendingClub loans

Tom Maryniak 03-11-2020

Introduction

- Lending Club is a peer to peer loan lending site. Lenders can view certain information on the borrower (such as income, purpose of loan, etc.), then lenders can invest in as many or few loans as they want by investing a minimum of $25 per loan request.
- While lending club claims 4-7% Return on Investment (ROI) average across all loans, this projects aims to improve this ROI.

Methodology

- Historical data of Lending Club loans was obtained
- Data was processed and filtered
- Data was fed into a Machine Learning model
- Model was used to predict what ROI was to be expected from any given loan.
- Model can be used to decide whether or not to invest in a loan based on the predicted ROI

Summary of Results

ROI was improved from an average of 4.2% to 8.2%

Hypothetical $10,000 investment	Before	After
Investment	$10,000	$10,000
3 Year return	$420	$820
ROI	4.2%	8.2%

Full data analysis and processing

Data

- Data was acquired from Dataset website Kaggle.com
- This dataset contains 2.2+ million loans with 145 datapoints on each loan
- To optimize the algorithm only certain loans and data points will be collected.

The following columns have been selected to be used in the analysis

- loan_amnt - The listed amount of the loan applied for by the borrower.
- funded_amnt - The total amount committed to that loan at that point in time.
- Term - The number of payments on the loan. Values are in months and can be either 36 or 60.
- Installment - The monthly payment owed by the borrower if the loan originates.
- emp_length - Employment length in years. Possible values are between 0 and 10
- home_ownership - The home ownership status provided by the borrower during registration.
- annual_inc - The self-reported annual income provided by the borrower during registration.
- loan_status - Current status of the loan
- Purpose - A category provided by the borrower for the loan request
- 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.
- delinq_2yrs - The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
- inq_last_6mths - The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- int_rate - Interest Rate on the loan
- total_pymnt - Payments received to date for total amount funded
- Desc - Loan description provided by the borrower

Data Cleaning Methodology (1/2)

- Home ownership was filtered to only include Mortgage, Ownership, and Rent as these made up the majority of the loan applications
- Loan status was filtered to only include Charged Off, Fully Paid, and Default as some of the loans in the data set were still ongoing/late/etc, only the ones that are known good/bad are to be incorporated
- The dataset did not have any 60 month loans that were complete due to the timeframe of the dataset, only 36 month loans were selected.
- If no employment length was given, a value of 0 was given.

Data Cleaning Methodology (2/2)

- Another data point, payment ratio, was made as a ratio of income to installment
- Target variable, ROI, was made that was the total_pymnt/loan_amnt
- Description was quantified as length of the description, and the amount of entries in the description
- Loan purpose and home ownership was quantified with One Hot Encoding
- To help scaling and since these values are self reporting, Income, payment ratio, and description were filtered out to remove outliers

Goal

- After all the filtering the average ROI was 4.2%
- If you invested $10,000 as $25 each across any random 400 of the filtered loans, after 3 years you could expect a total of $10,420 to be returned.
- This is using credit score and by extension interest rate produce a positive ROI.
- The algorithm's success will be based on an increase of this 4.2% ROI

Machine Learning (1/3)

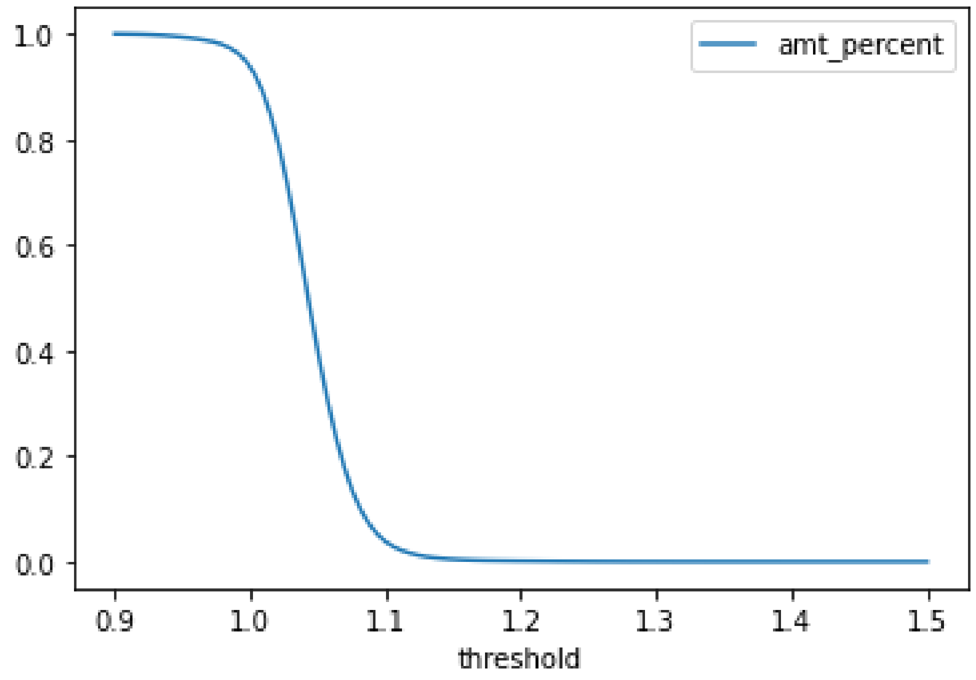
- The clean dataset was split into 75/25 train-test split
- A standard scaler was fit and transformed the train data
- A SGDRegressor model was trained on the training data
- The test data was transformed by the scaler
- The model predicted the ROI for the scaled test data

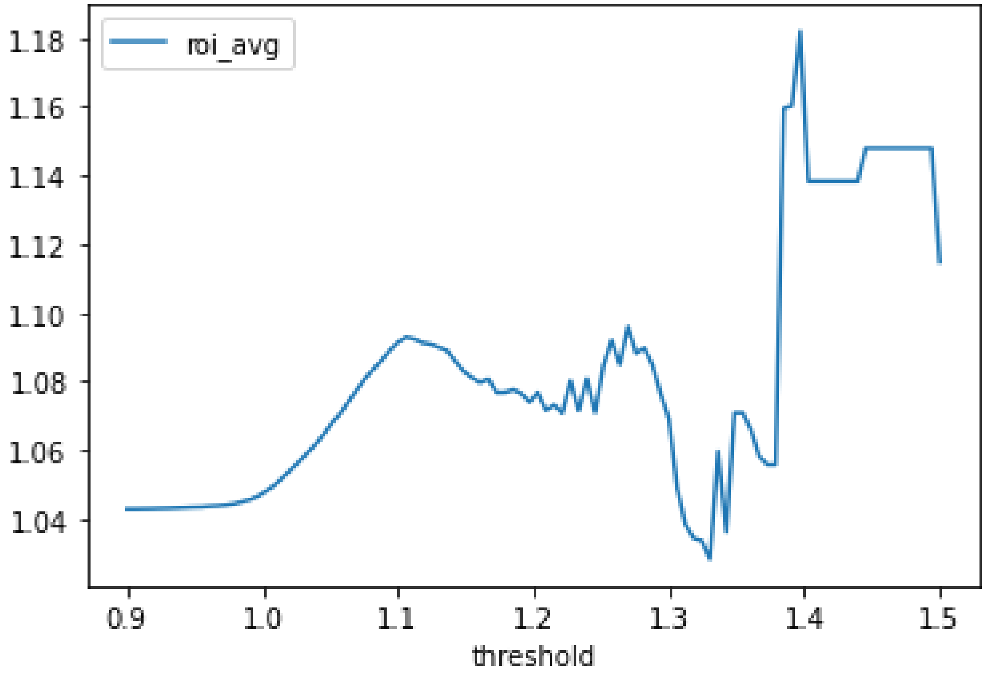
Machine Learning (2/3)

- Due to the complexity of the problem, the conventional score is quite low; but using a threshold to only approve loans that the algorithm will predict to have an ROI above the threshold will yield the best results. This is used as a pseudo-confidence score

Machine Learning (3/3)

- These 2 charts can be used together to find a recommended threshold.
- While increasing the threshold can theoretically increase the ROI, due to the smaller amount of loans that algorithm will approve the data gets sparse and sporadic. A threshold of about 1.077 will only approve about 12% of the loans but the average ROI of those loans will be 8.2%





Results

If the same data that was used to train the model is provided of future requested loans, the model will approve about 12% of the requested loans, but the average ROI will be 8.2% nearly double the goal of 4.2%.

Resources

- https://www.kaggle.com/wendykan/lending-club-loan-data
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?highlight=standard%20scaler#sklearn.preprocessing.StandardScaler
- https://github.com/tommybox3377/LendingClubROI