Lending Club Loans

Predicting Return on Investment for Peer-to-Peer Loans

Introduction

- Peer-to-Peer (P2P) loans are personal loans
- Individuals can borrow up to $40k from Lending Club
- Individuals can invest in consumer credit
- Previously only available to financial institutions

Business Problem

Peer-to-Peer lending platforms provide borrower and investors a place to connect for personal loans. This provides individual investors a new opportunity to invest in consumer credit, but comes with risk in the form of default/charged-off loans and loans that are paid off early. If investors were able to know what loans would have the best return on investment they could have an advantage and see their portfolio increase exponentially.

Background

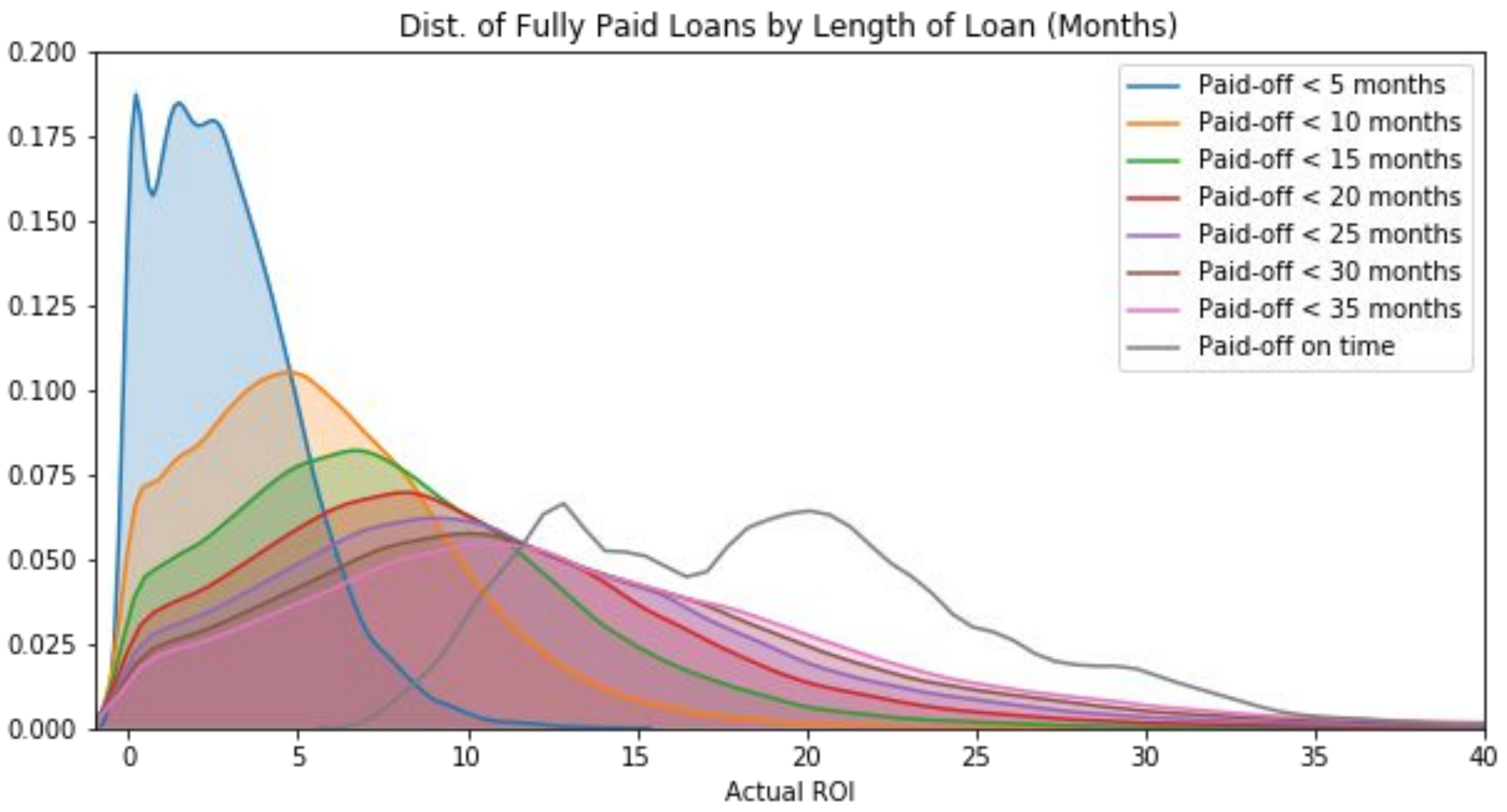
- P2P Market
  - $34 Billion as of 2008
  - $589 Billion by 2025
- Top reasons for P2P Loans
  - Debt Consolidation
  - Credit Card Refinancing
  - Home Improvement
- Past Studies focused on predicting defaults
- Our goal: predicting ROI

Data

- From Kaggle, original Lending Club data
- Created new target variables
  - Annualized ROI (continuous)
  - Positive Annualized ROI (binary)
    - Annualized ROI > 0 = 1
  - Target Annualized ROI (binary)
    - Median Annualized ROI of dataset = 7.56%
    - Annualized ROI > 7.56 = 1

Data

	Mean	Median
Expected $ ROI	$4,425.84	$2,656.20
Actual $ ROI	$144.05	$1,144.90
Expected ROI %	27.58	22.33
Actual ROI %	1.92	11.56
Annualized ROI %	-1.34	7.56


Dist. of Fully Paid Loans by Length of Loan (Months)
Paid-off < 5 months
Paid-off < 10 months
Paid-off < 15 months
Paid-off < 20 months
Paid-off < 25 months
Paid-off < 30 months
Paid-off < 35 months
Paid-off on time
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25
30
35
40
Actual ROI

Data Prep for Modeling

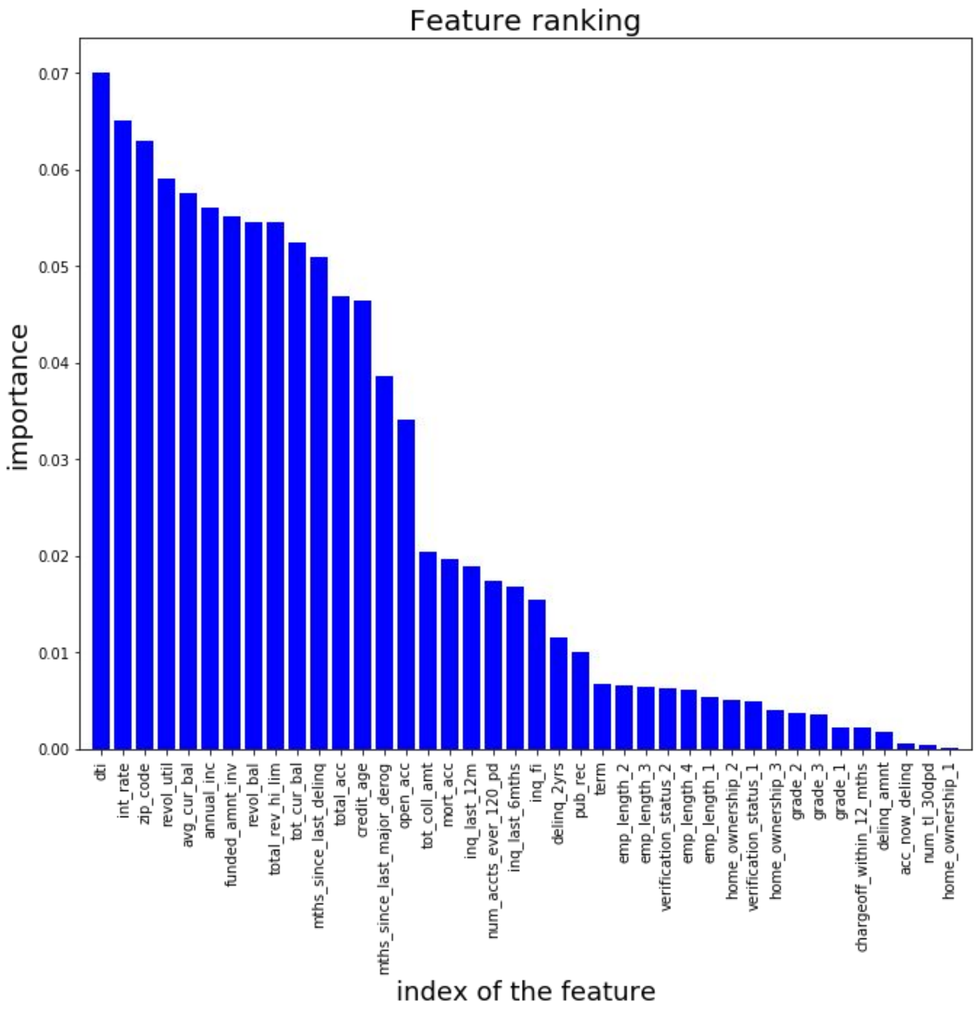
- Drop redundant variables (Installment)
- Drop variables unknown to Investor (Total Received Interest)
- Binary encode categorical variables
- 80/20 Train - Test Split
- Separate labels from datasets
- Standardize on training data
    - Apply to test data

Feature Reduction

- Random Forest Regressor
  - Used to extract feature importance
  - Top 15 features put into separate dataset
- Principal Component Analysis
  - 95% variance explained by 3 components...initially
  - After standardizing data, 95% variance explained by 32 components
  - Not much better than original 40 variables

Top 5 Features

- Debt-to-Income Ratio
- Interest Rate
- Zip Code (3 digit)
- Utilization of Revolving Credit
- Avg. Current Credit Balance



Regression Modeling

- Trying to Minimize Root Mean Squared Error
- RMSE: the average error between predicted ROI percentage and actual
- R-Squared: amount of variability explained by a model
- Goal: Minimize RMSE

Annualized ROI

- Continuous variable showing percentage return
- Linear Regression
  - Root Mean Squared Error
  - On full training dataset: 27.701
  - 10-fold Cross Validation: **27.704**
  - R-squared: <u>0.033</u>
- Random Forest Regression
  - RMSE on full training dataset: 12.428 (Overfitting?)
  - 10-fold Cross Validation: **29.295**
  - R-squared: <u>0.805</u>

Classification Modeling

- Trying to Maximize Precision Score
- True Positives: loans predicted to have positive/target ROI
- Precision: ratio of True Positives to all predicted positives
- Goal: Minimize negative/low ROI loans that are predicted as positive

Positive ROI

- Binary: ROI > 0 = 1
- Random Forest Classifier
- Random Forest Classifier w/ Top 15 Features
- Logistic Regression

Positive ROI - Random Forest

- Accuracy: 99.22
- Precision: 99.2
- Recall: 99.84
- AUC: 0.982
- **CV Precision: 82.79**

	Predicted Negative	Predicted Positive
Actual Negative	84,754	2,983
Actual Positive	663	372,623

Positive ROI - Random Forest w/ Top 15 Features

- Accuracy: 99.22
- Precision: 99.19
- Recall: 99.85
- AUC: 0.982
- **CV Precision: 82.42**

	Predicted Negative	Predicted Positive
Actual Negative	84,711	3,026
Actual Positive	578	372,708

Positive ROI - Logistic Regression

- Accuracy: 81.08
- Precision: 81.58
- Recall: 98.99
- AUC: 0.519
- **CV Precision: 81.58**

	Predicted Negative	Predicted Positive
Actual Negative	4,290	83,447
Actual Positive	3,763	369,523

Target ROI

- Median Annualized ROI for dataset = 7.56%
- Binary: ROI > 7.56 = 1
- Random Forest Classifier
- Logistic Regression
- Logistic Regression at different thresholds

Target ROI - Random Forest

- Accuracy: 98.74
- Precision: 99.35
- Recall: 98.13
- AUC: 0.987
- **CV Precision: 66.43**

	Predicted Negative	Predicted Positive
Actual Negative	229,280	1,476
Actual Positive	4,312	225,955

Target ROI - Logistic Regression

- Accuracy: 66.77
- Precision: 64.32
- Recall: 75.16
- AUC: 0.668
- **CV Precision: 64.31**

	Predicted Negative	Predicted Positive
Actual Negative	134,737	96,019
Actual Positive	57,189	173,078

Target ROI - Logistic Regression at 0.75 Threshold

- Accuracy: 52.72
- Precision: 71.79
- Recall: 8.79
- AUC: 0.527

	Predicted Negative	Predicted Positive
Actual Negative	222,795	7,961
Actual Positive	210,011	20,256

Tuning Positive ROI Random Forest

- Grid Search
- First set of parameters:
  - Number of Trees: 3, 10, 30
  - Max Features at each branch: 10, 20, 30
- Second set:
  - Bootstrap = False
  - Number of Trees: 3, 10
  - Max Features: 2, 3, 4
- Best Estimator
  - Max Features: 30
  - Number of Trees: 10

Final Model: Positive ROI Random Forest

- Accuracy: 78.22
- Precision: 83.11
- Recall: 91.81
- AUC: 0.557
- **CV Precision: 83.11**

	Predicted Negative	Predicted Positive
Actual Negative	4,266	17,448
Actual Positive	7,658	85,884

Conclusion

- Investors look for any advantage
- 83% Success when predicting Positive ROI
    - Advantage over random investment selection
- Further steps
    - Neural Networks
    - Boosting
    - Ensemble of simple models