Predicting P2P Microloan Defaults

Jack Magill and Mike Hallee

Abstract

Goal: To evaluate the relative success of different classification models in predicting loan defaults

Data: 2.3 million micro-loans issued by Lending Club between 2007 and 2018

Methods: Various binary classifiers implemented in scikit-learn

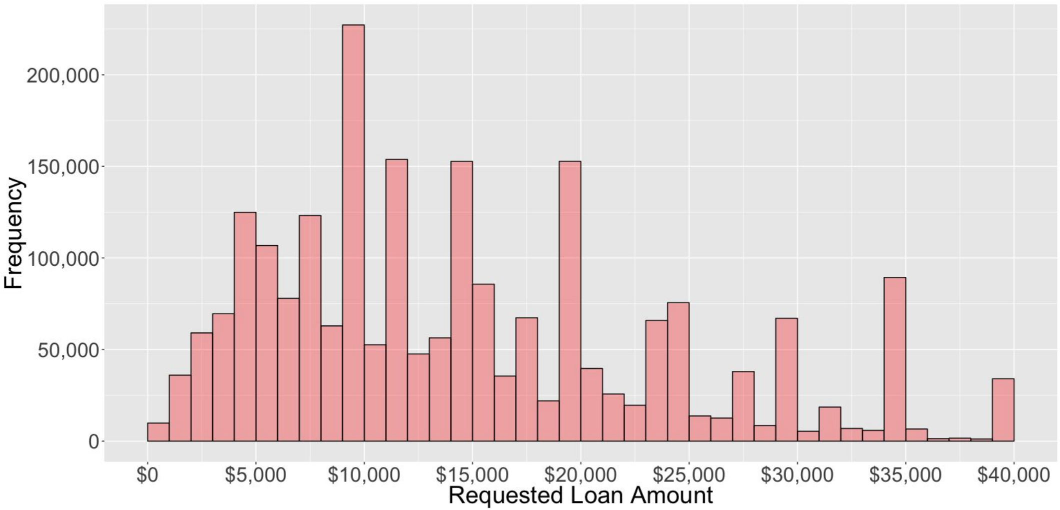
Results:
- Gaussian Naive Bayes is best (23% recall, 25% precision)
- Default probabilities match LC grading system

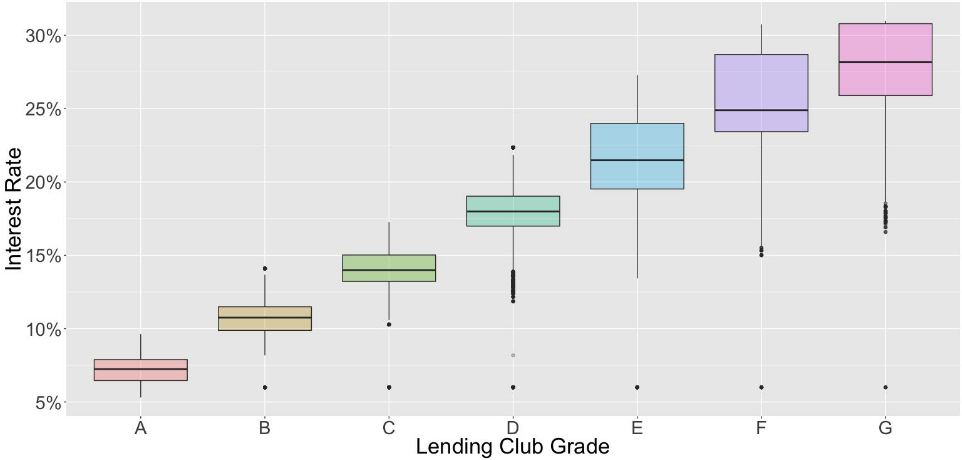
Motivation & Overview

- **Lending Club**
  - Peer-to-peer lending:
    - **Borrowers** apply to Lending Club platform, if approved they are added to the holdings portfolio.
    - **Investors** on the platform choose which borrowers loans, broken up into “notes” that represent a fraction of the loan amount, they want to purchase in exchange for fixed payments with interest.
- **Exploitation of Data for Maximizing returns**
  - Lowest risk category A1 yields 6% interest, whereas highest risk E5 can exceed 30% interest [2]
  - If we can weed out the E5 notes that will default, we can earn highest returns with minimized risk

The Data

- Obtained From Kaggle
- **2.3 million lendings from 2007-2018 [2]**
- 120+ features
  - **Loan Details:** Amount, Interest Rates, Terms
  - **Borrower Details:** Financials, Credit, Employment, Home Ownership, Demographics
  - **Lending Club Measures:** Assigned Loan Grade, Loan Status (payments, loan status: good/late/defaulted, etc)


Frequency
200,000
150,000
100,000
50,000
0
$0
$5,000
$10,000
$15,000
$20,000
$25,000
$30,000
$35,000
$40,000
Requested Loan Amount


Interest Rate
30%
25%
20%
15%
10%
5%
A
B
C
D
E
F
G
Lending Club Grade

Preprocessing and Imputation

- Only data known at time of loan issuance is retained for predictive use
- Variables with **<75%** valid values are dropped
- Missing numeric data is **imputed with median** values
- Categorical data is **one-hot encoded**
- Resultant dataset consists of 99 total features

- Defaults/charge-offs/late payments are labeled as **1**

Prior Work

- University of Illinois Paper predicting defaults
    - **Similarity-based model** assigning loans to specific categories and making predictions based off the industry standard risk for that category
- IEEE Conference Paper
    - **Tree-Based Classifiers** (Decision Tree, Random Forest, Bagging) predicting loans that don’t go late/defaulted with high (96%) precision
        - This is an easier task than predicting defaults and charged-off loans which are a minority in the dataset

Classification Methods

1. Stochastic Gradient Descent
   a. w/ Log Loss
   b. w/ Perceptron Loss
2. Support Vector Machine

3. AdaBoost Classifier
4. Random Forest Classifier
5. Decision Tree Classifier

Classifier Evaluation

- The value of this classifier would be to **avoid investing in loans that default**
- Given that there are hundreds of thousands of loans, we aren't really worried about **false positives**
- Good classifiers will have a high true positive rate or **recall**
- ROC and P-R curves can be examined to modify the **decision threshold** to get desired performance

Hyperparameter Tuning

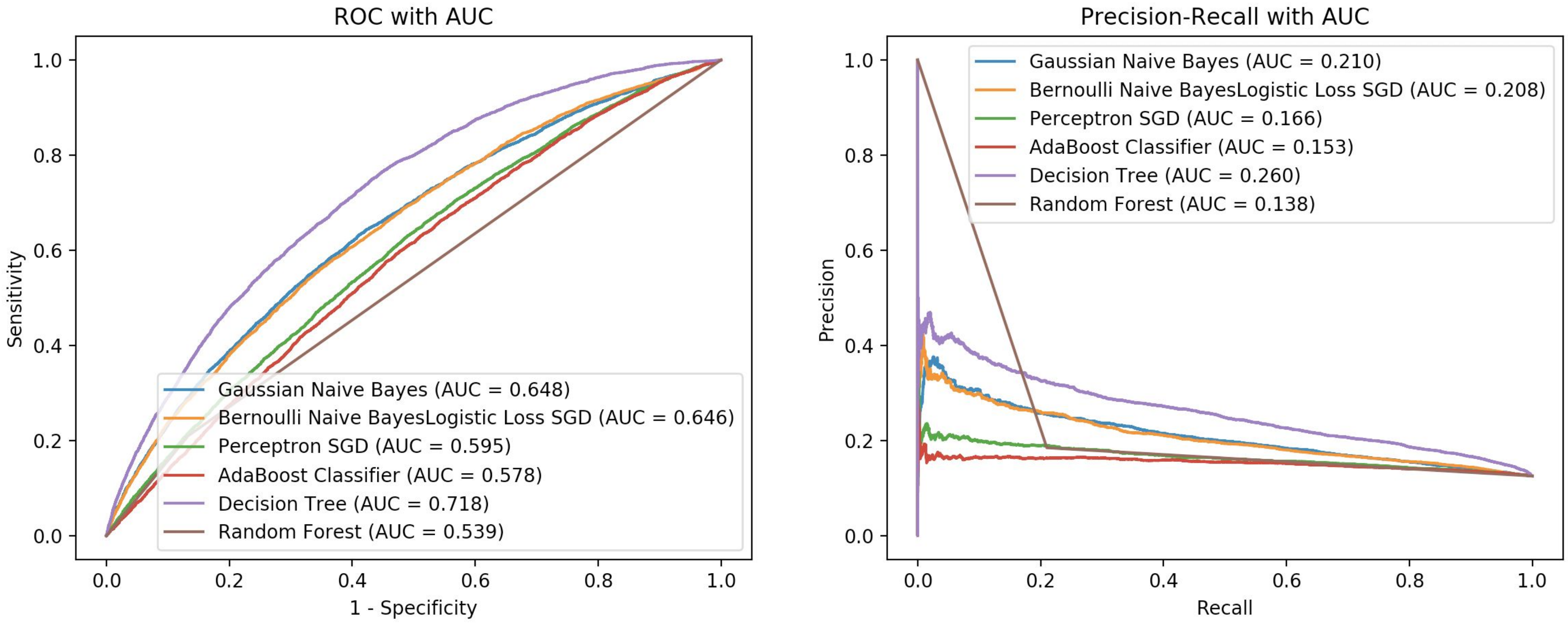
- Classifier **hyperparameters** are tuned using **grid search**
  - Regularization coefficient optimized for descent methods
  - Estimator count, maximum features/depth optimized for other methods

- How do we determine which parameters are ‘best’?
  - **Maximize Recall** since our goal is to maximize true positives (identifying and avoiding loans that would default) and minimize false negatives (investing in a loan that would default)

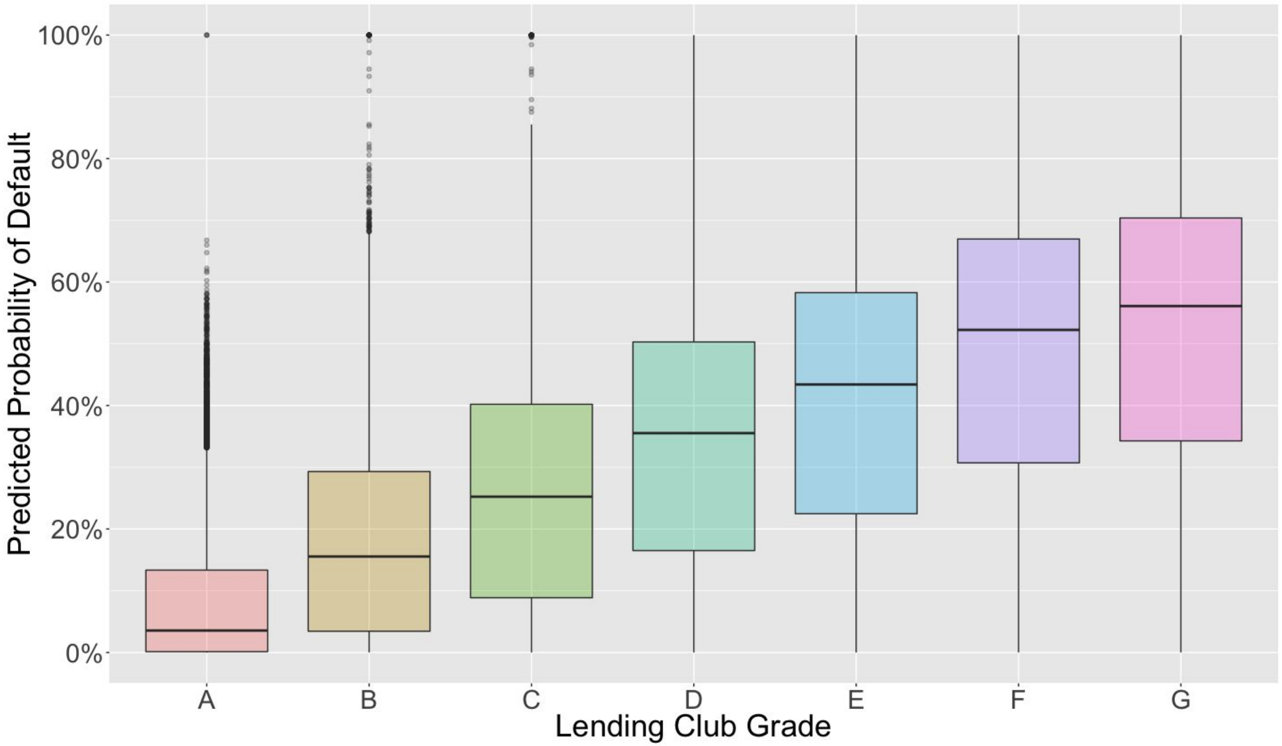
Results

Classifier	ROC AUC	PR AUC	Accuracy	Precision	Recall	F1-Score
Gaussian NB	0.648	0.210	0.816	0.251	0.233	0.242
Bernoulli NB	0.539	0.138	0.784	0.185	0.210	0.197
Perceptron SGD	0.646	0.208	0.869	0.331	0.045	0.079
Decision Tree	0.718	0.260	0.874	0.433	0.010	0.020
Random Forest	0.595	0.166	0.873	0.205	0.005	0.010
AdaBoost	0.578	0.153	0.874	0.224	0.003	0.005

Results: ROC and P-R Curves



How are Lending Clubs grades?



Discussion

- Predicting default is a **difficult problem.** If it were easy, then there would be no risk.
- Lending Club's grading system does follow an **increasing probability of default**
- Decision threshold of classifier could be decreased (increase) if you are risk averse (seeking)
  - Higher Risk → Higher Expected Returns

References

Wendy Kan. Lending club loan data. https://www.kaggle.com/wendykan/lending-club-loan-data, 2019.

Lending Club. Note interest rates and fees. https://www.lendingclub.com/investing/investor-education/interest-rates-and-fees. Accessed: 2019-04-23.

Michael J. Shaw and James A. Gentry. Using an expert system with inductive learning to evaluate business loans. Financial Management, 17(3):45–56, 1988.

Vinod Kumar L, Natarajan S, Keerthana S, Chinmayi K M, and Lakshmi N. Credit risk analysis in peer-to-peer lending system. In 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), pages 193–196, Sep. 2016.