Will Borrowers Pay Off Their Loan?

Sylvia Lee

Thinkful Capstone Project

Key Facts about Personal Loans

- Fastest growing form of consumer lending
- Consolidate debt, refinance credit cards, finance major purchases
- Outstanding personal loan debt is $138 billion as of 2018
- About 19 million people have a personal loan
- Average APR in Q1 2019: 33%. Rates range from 7% to 86%
- Higher rates of delinquency than other common loan types

P2P Lending Business Model

- **Investors select loans** based on borrower's info and loan info
- Investors' revenue
  - Loan Interest (5%-36%)
- Business's revenue
  - Origination Fee (0.25%-0.5%)
  - Service Fee (0.5%-1%)

Problem

- Investors lose money when borrowers **stop paying** interest
- Investors lose their investment when borrowers **default** on the principal

Solution

- Investors want to predict whether the borrower will **default**
- Investors benefit from a **predictive model**

Purpose of Study

- Supervised Learning Model that predicts whether borrowers will **default**
- Available data: demographics, credit history, loan characteristics

  Positive = pay off the loan | Negative = default

- False positives: **investors** lose entire principal
- False negatives: **investors** lose interest from borrowers

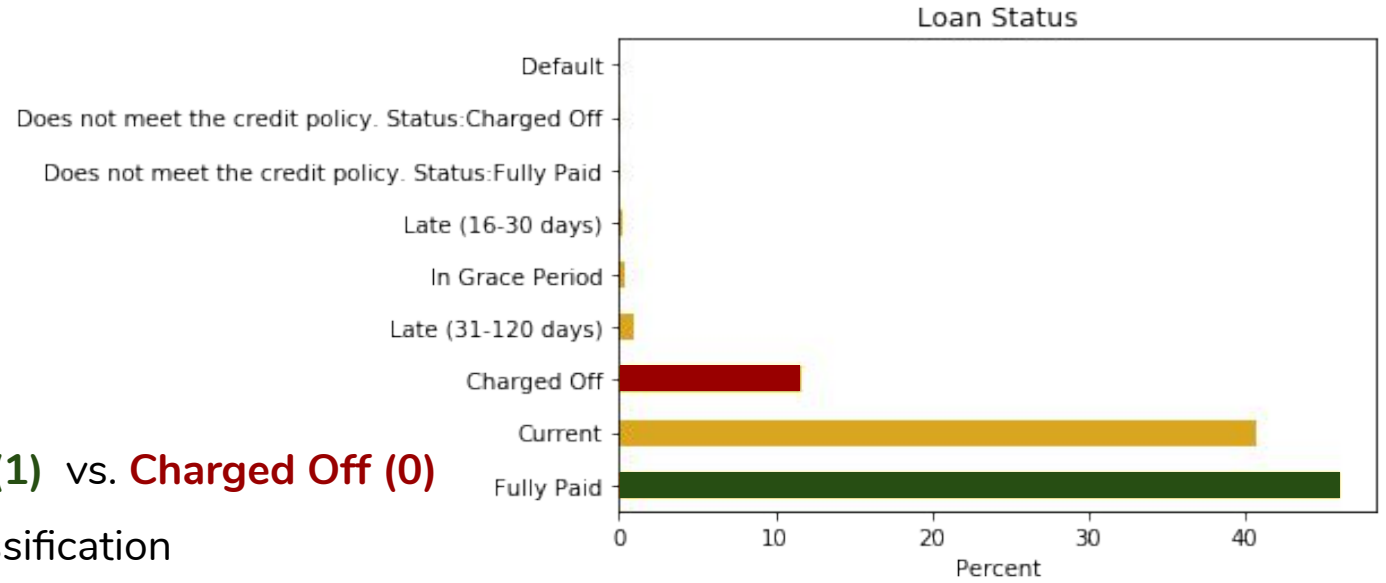
About the Loan Dataset

- LendingClub
- 2007-2015
- 2,260,668 records
- 145 fields
  - 109 numeric variables
  - 36 categorical variables
  - 10 datetime variables
  - 2+ natural language variables

Data Cleaning

- Exclude 58 variables with > 30% missing data
- Exclude 20 variables unavailable at the time of the loan
  - Monthly payment
  - Type of payment
  - Principal received to date
  - Hardship flag
- Include variables with > 90% valid data

Target Variable



- **Fully Paid (1)** vs. **Charged Off (0)**
- Binary Classification

Feature Engineering

- Categorical → Ordinal
  - Grade
  - Subgrade
  - Employment Length
- Cardinality Reduction
  - Zip Code First Digit → Region
  - NLP (Hash): Job Title

Feature Selection

- Logistic Regression
- Add features to baseline model
- Focus on Specificity (True Negative Rate)

Things Learned

- DTI, Public Bankruptcy Records, Revolving Credits → higher specificity
- Adding job title doesn't improve performance and it's expensive to run
- 1.16M X 40 dataset gives 0.19 Specificity and 0.80 Accuracy
- 1.07M X 60 dataset gives 0.13 Specificity and 0.80 Accuracy

Final Features

Credit Status/History (28)	Credit Status Tax Lien Public Record Bankruptcies Revolving Credit DTI Delinquent Status
Demographics (4)	Residency and Home Ownership Employment Status
Loan Features (8)	Term, Grade, and Application Type Loan Amount and Purpose Interest Rate Verification Status

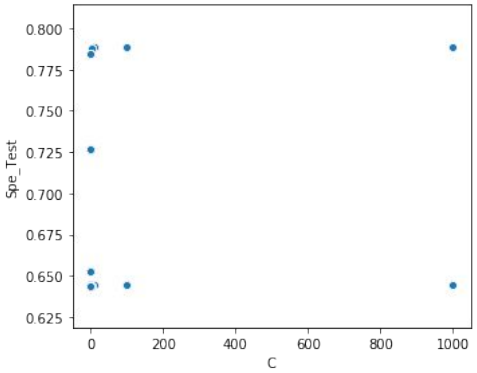
Building Models

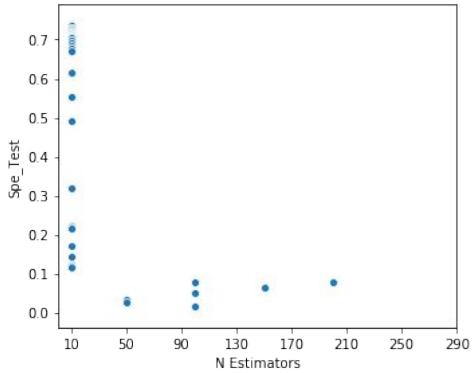
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting Model

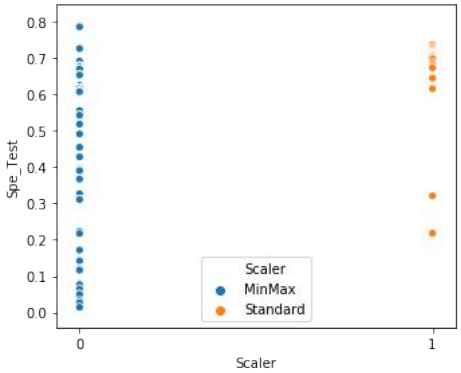
Tuning Hyperparameters

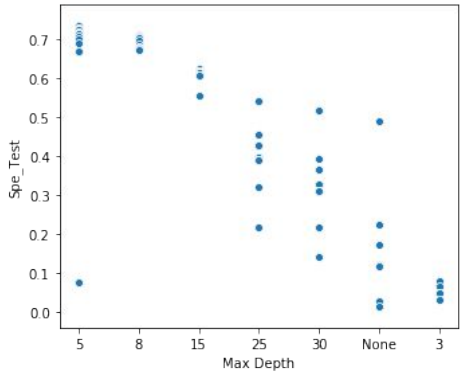
Scaler	Min Max, Standard
Class Weight	Default (None), Balanced
C	0.001, 0.01, 0.1, 1, 10, 100, 1000
# Estimators	10, 50, 100, 150, 200
Max Depth	5, 8, 15, 25, 30
Min Sample Splits	2, 5, 10, 15, 100

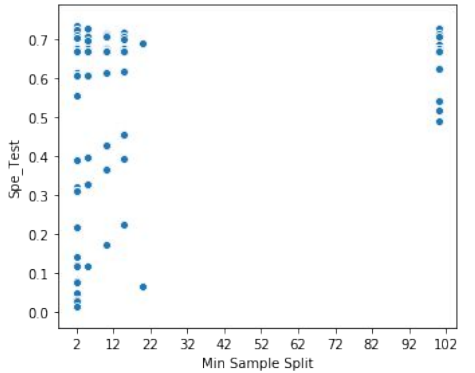
Hyperparameters and Testset Specificity







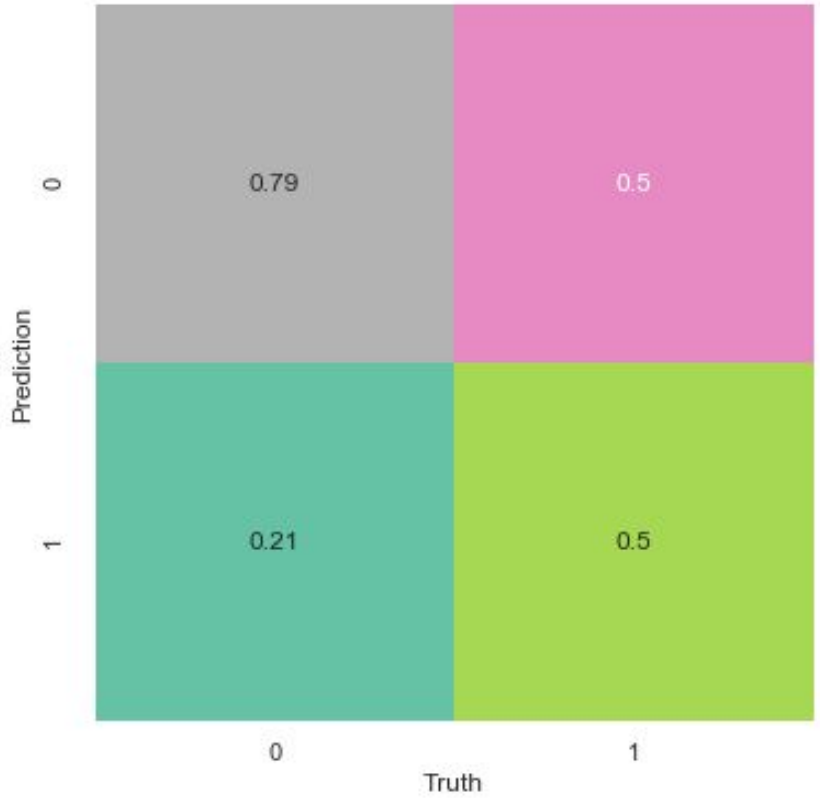




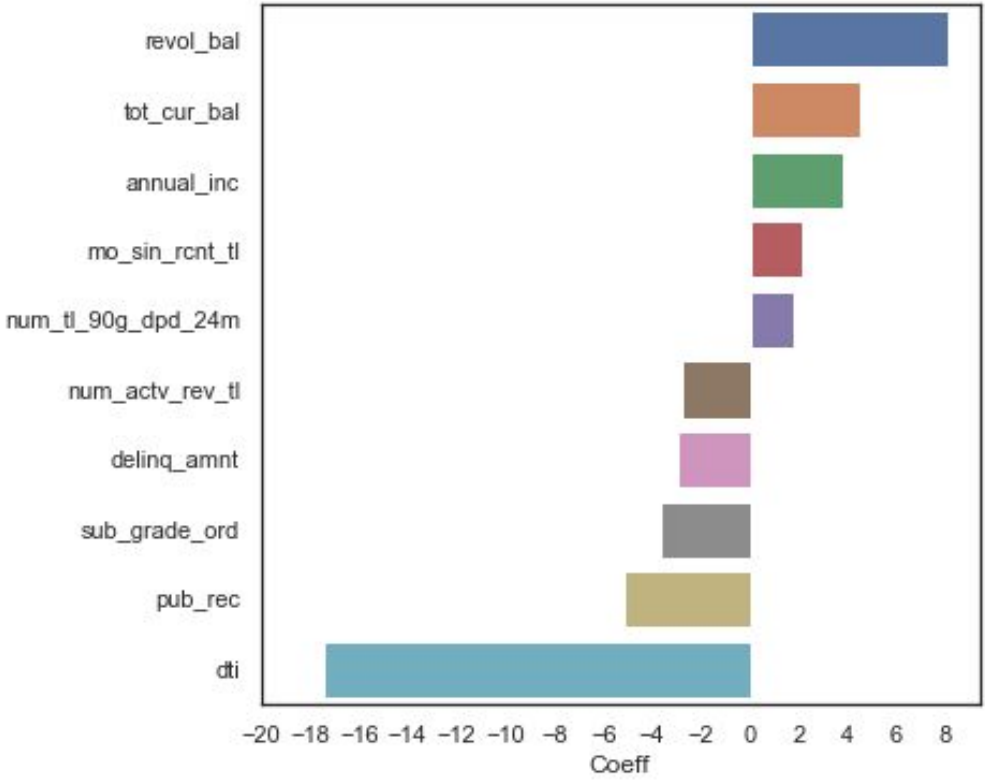
Final Model

Model	Accu_Train	Accu_Test	Sen_Train	Sen_Test	Spe_Train	Spe_Test	F1_Train	F1_Test	Scaler	C	Class_Weight
Logistic Regression	0.665665	0.557294	0.670871	0.499984	0.644746	0.788825	0.762663	0.644203	MinMax	1000	Balanced

Confusion Matrix



Feature Importance



Summary

- Logistic regression model
- 79% accuracy (4 x better than random guess)
- This model will help investors make more money

Future Work

- Evaluate model performance on more recent data when available
- Improve current true positive rate of 0.5 to make the model more valuable
- NLP job titles with correct model
- Improve missing data imputation