Loan Approval Screener

Jim Petoskey

Flatiron School

Phase 3 Project

Outline

Business Problem

Summary

Data - Overview, Manipulation, Feature Selection, Feature Importance

Methods

Results

Conclusions

Business Problem

Banks want an easy way
to screen
loan applications.

Make strong inferences with
data from prior applications
about likelihood of approval.

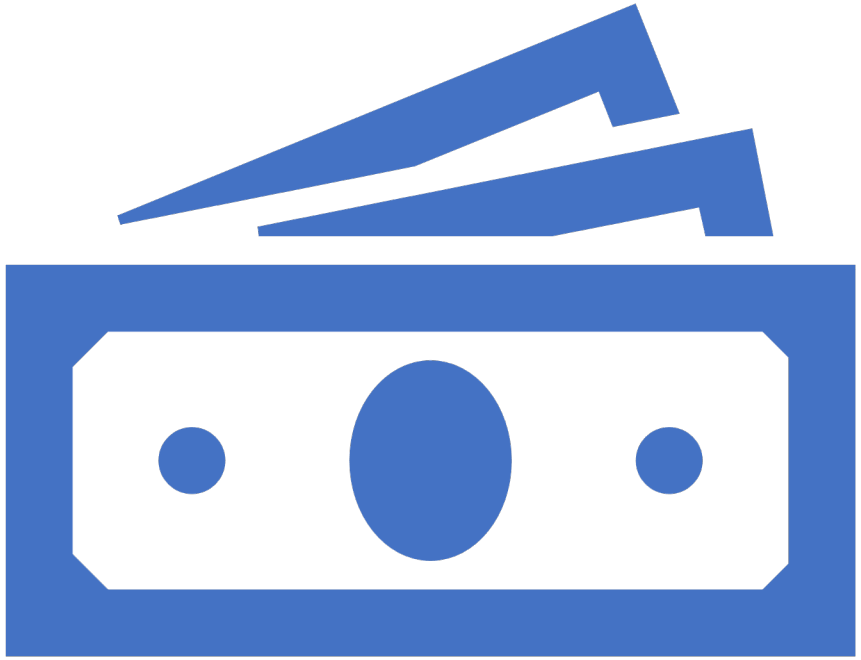
Business Opportunity

- Reduce human-hours for banks
  - Use model to select applications that are highly likely to be approved.
- Improve pipeline for customer experience by automating questions and providing likelihood of approval for customer.

Data - Overview

Loan Application Data

- 307,511 Applications
- 122 Features
  - Type of Loan:
    - 90% Cash Loans
    - 10% Revolving Loans
  - Application Results:
    - 8% Approved
    - 92% Rejected



Data Manipulation

- Class Imbalance Issues:
  - Improve balance of Approval/Refusal ratio
  - Used under-sampling and over-sampling techniques to improve balance
- Replaced missing numerical and categorical data with mean and mode, respectively.
- 100 % of data from 39 features and 1 target feature used for model.

Data Manipulation, continued
Improving Class Balance

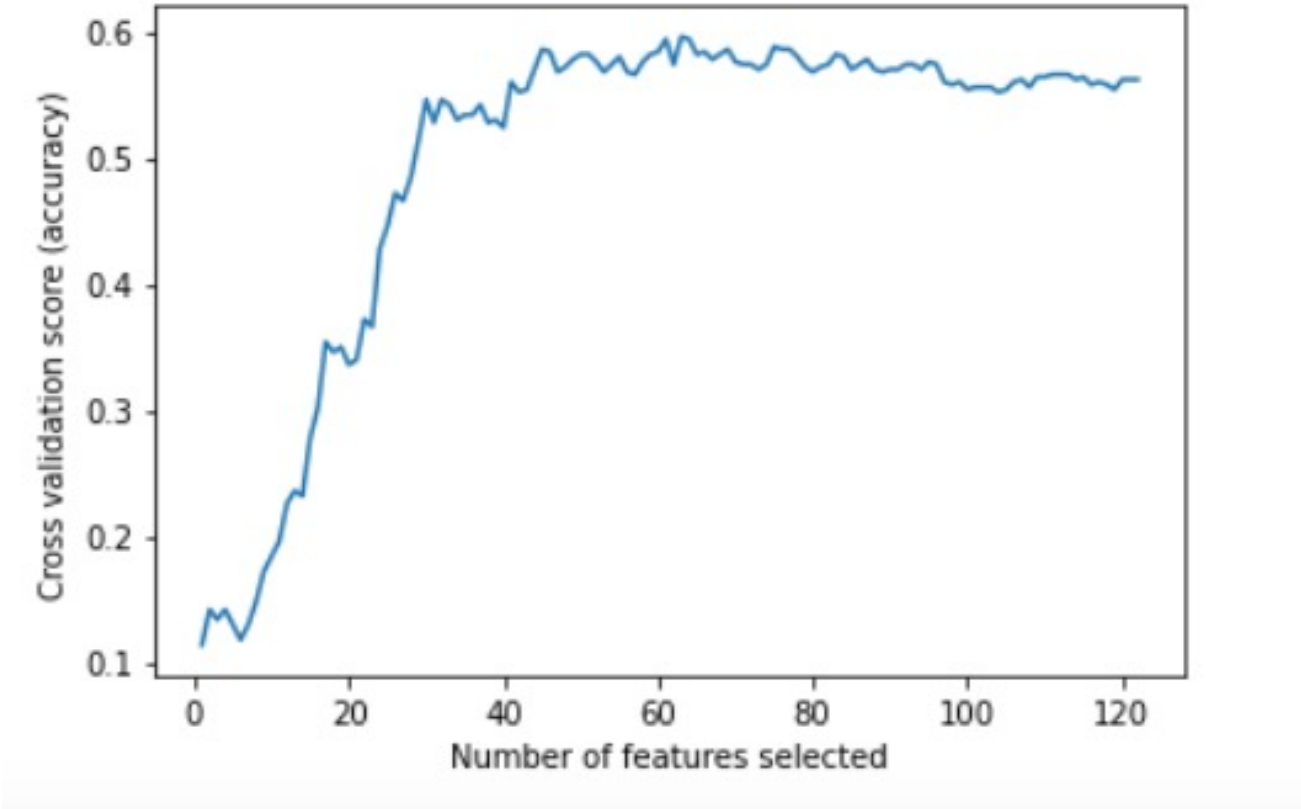
- Initial data set had an 8% prevalence of Loan Approval.
  - Ideally, a classification model will have a 50% prevalence of each class.

Solutions for class imbalance:

- Under-sample Majority Class (Loan Rejection)
  - Randomly selected 40,000 applications that were rejected.
- Over-sample Minority Class (Loan Approval)
  - Used Synthetic Minority Oversampling (SMOTE) to increase loan approvals from 24,825 values to 40,000 values.

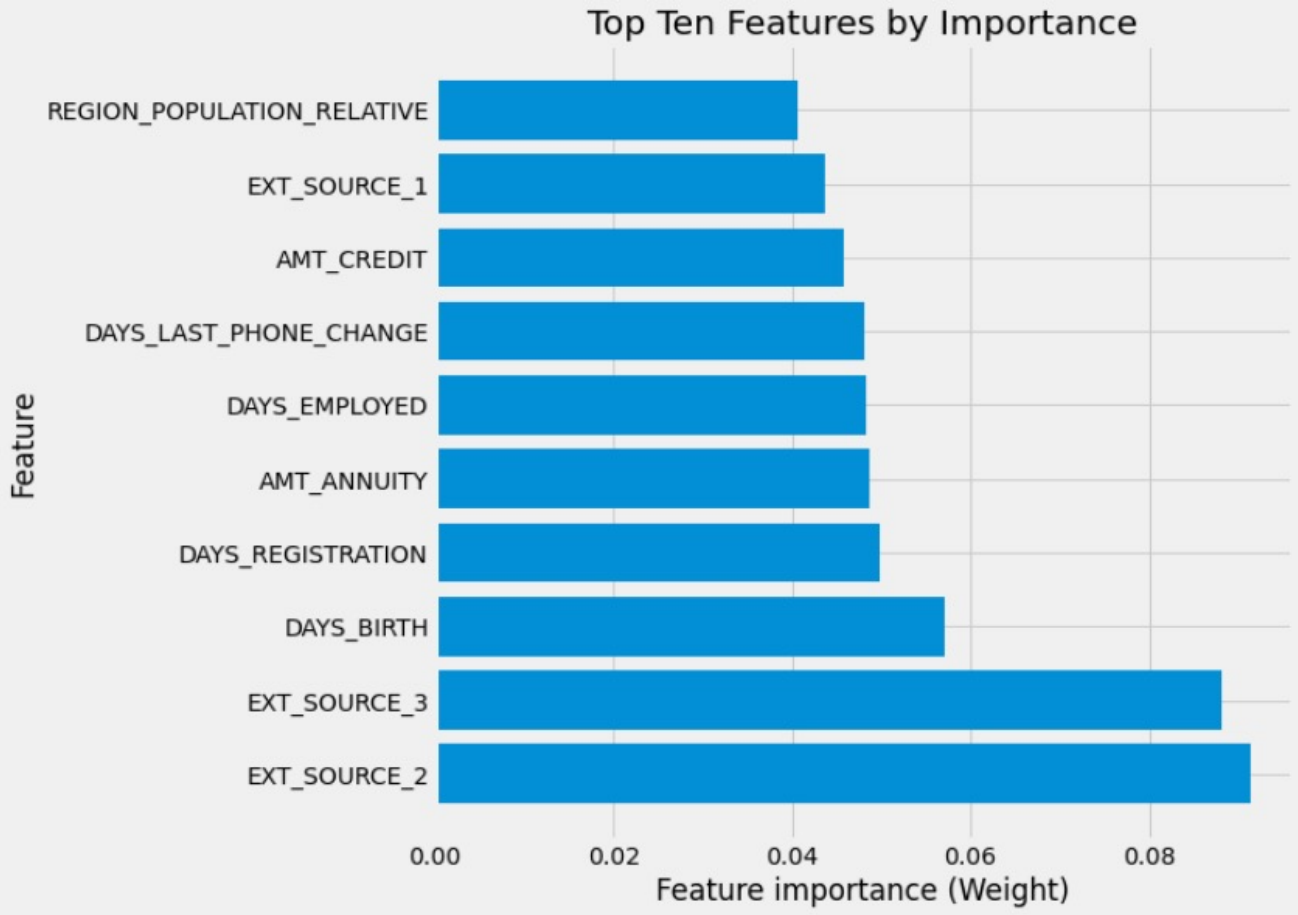
Feature Selection

```
Optimal number of features : 63
```



- Optimal number of features to include was 63.
- I chose 39 features, and after encoding the categorical features, the final models included 61 columns.
- I would have added more features, but I was limited in computational power on my machine.

Feature Importance



- Top Ten Features by importance in the 6th Random Forest Model.
- Top singular model produced
- EXT_SOURCE not defined, but could be a form of credit score.
- Important to note these features so they are included on future loan applications:
  - Birth Date
  - Employment start date
  - Amount of Credit Available
  - Common Annuities

Methods

- Classification Modeling
  - Main Models:
    - Random Forest
    - Gradient Boost
    - Voting Classifier
- Optimize for Loan Approval, or True Positive Rate
  - Value high True Negative Rate, as well.
  - Best scores for True Positive Rate valued because that means the model will have a high likelihood of flagging a
- Identify features with most influence on Loan Approval

Final Model Methods

- Voting Classifier
  - Used 3 models:
    - Top Random Forest optimized for True Positive Rate
    - Top Random Forest optimized for True Negative Rate
    - Top Gradient Boost Model optimized for True Positive Rate
  - Combination of these models yielded better results than any one of the models alone.
    - This is called an ensemble model and optimizes the best parts of each model.
    - 2x heavier weight was applied to the top Random Forest model for True Positive Rate

Results of Best Model

- 84% True Positive Rate
  - 84/100 applications that were approved, were flagged for approval by model
    - Maintained a 40% True Negative Rate, which means many applications can be successfully filtered out.

- Model as a Screener:
  - 84% of applications that are approved, will be recommended for viewing by a person.
  - 31% of applications can be filtered out, with an 84% chance that they would have been rejected.

Conclusions

- This model could save banks time and money by accurately filtering loan applications.
  - Could reduce applications read by people by over 30%
  - 84% likely that approved loans will be recommended for review by a person
    - Makes loan officer’s job simpler and faster.

- Top features for banks to follow in this model:
  - Birth Date
  - Employment start date
  - Amount of Credit Available
  - Common Annuities

Further Studies Recommended

- Gather data on more relevant features
  - Credit Score
  - Cash and Invested Savings
  - Retirement Savings
  - Debt to Income Ratio
- Test model on other samples of market data
  - Iterate and improve model for these samples or market segments
- Improve pipeline for loans by developing digital questionnaire
  - Attain data inexpensively – directly from consumer
  - Tailor questions to garner data from most relevant features