Small Business Loans

Decision Making on Loan Approval

U.S. Small Business Administration

Ed Lee
Classification, METIS, July 2022

Background

- SBA audits and approves loans for small businesses
- Banks lend funds based on SBA aproval

https://www.sba.gov/

Background

- But some businesses end up failing to pay the loan back

→ "Charge-off"

https://www.sba.gov/

Which businesses are likely to default?

https://wwd.com/

The Dataset

Records of loans approved by SBA

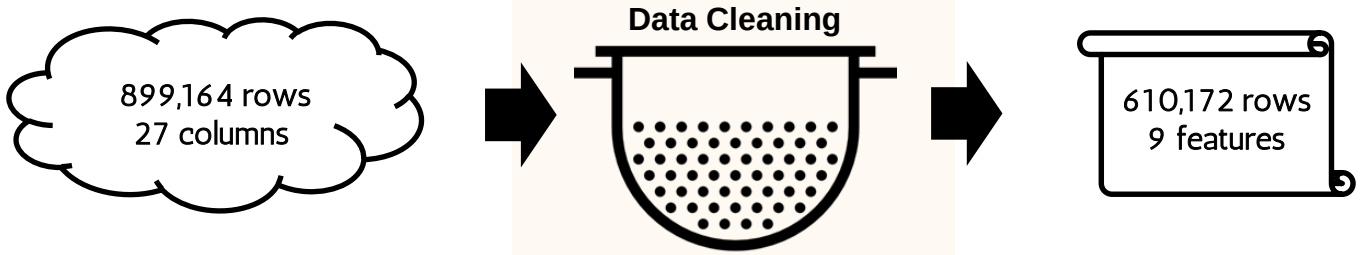
- Raw dataset contains 899,164 entries and 27 columns
- Includes traits of each business, amount approved, loan status (paid in full / charged off)

Tools

All the tools I have utilized were run through Python 3.8

Pandas	EDA, data cleaning, data manipulation, feature engineering
Matplotlib, seaborn	Plots, model visualizations
Scikit-learn	Building classification models (LogReg, random forest, naïve Bayes, AdaBoost), preprocessing data, evaluating model metrics
XGBoost	Building classification models (Xgboost classifier)
imblearn	Testing for class imbalance

EDA & Data Clening



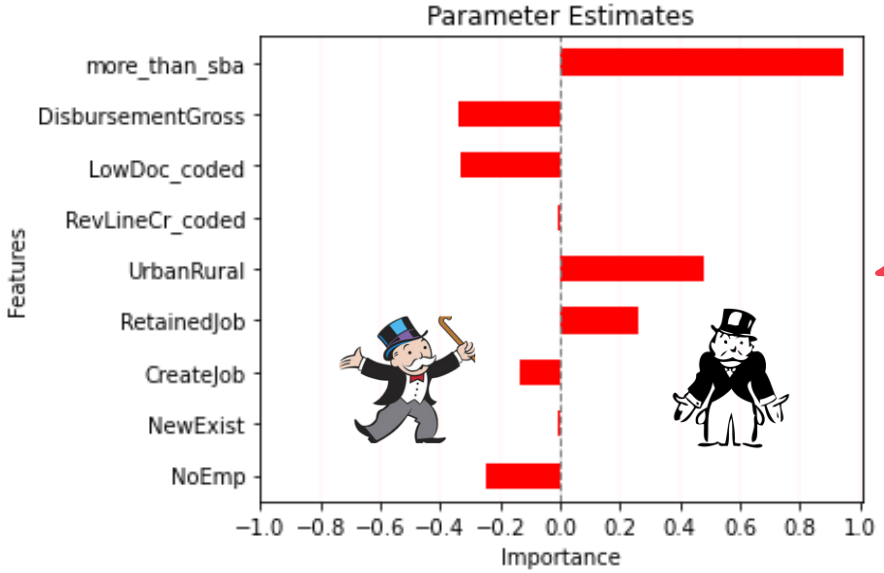
- Raw dataset

- Removing rows with invalid values
- Preprocessing
- Feature selection

- Target: Loan Status
- Imbalance: ~20% of all target values were positives (charged-off)

Feature Importance

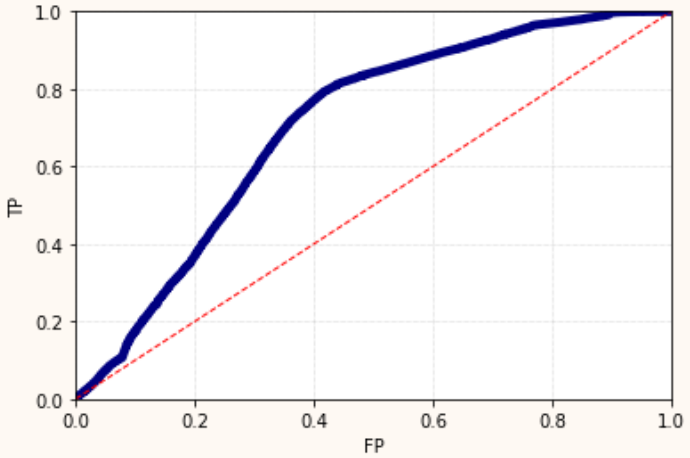
Before jumping into model comparisons, let's get a rough idea of each feature's impact



Provides clients insights on which factors to scrutinize upon

Result – Models 🎯

‘Starting Point’ - ROC (LogReg)

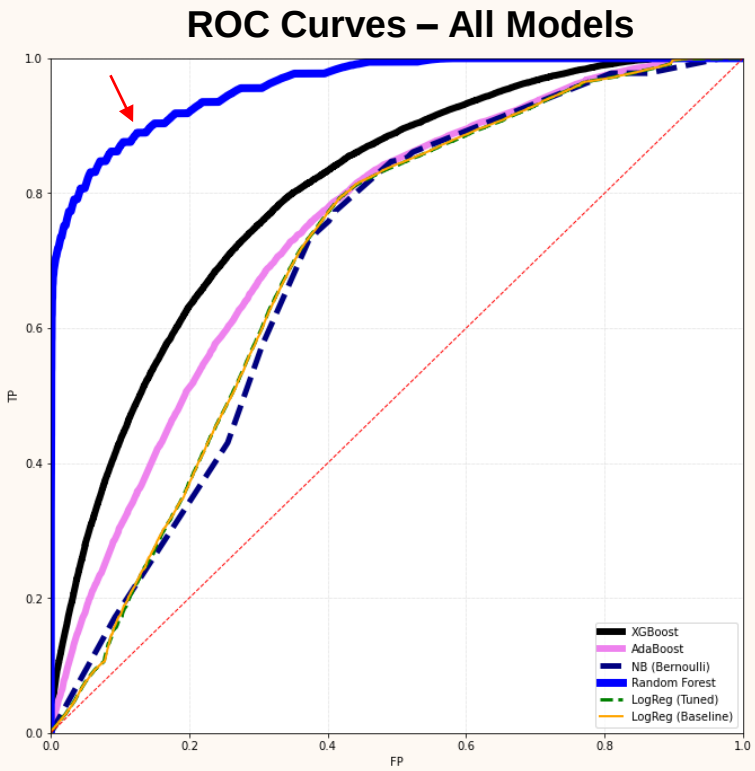


Accuracy: 0.820
Precision: 0.979
Recall: 0.002
AUC: 0.703

Total six models were built:

1. 'Baseline' Logistic Regression (default parameter)
2. Tuned LogReg (with GridSearchCV)
3. Random Forest (tuned with GridSearchCV)
4. Naïve Bayes
5. AdaBoost
6. XGBoost

- Class imbalance was ultimately addressed via class weights



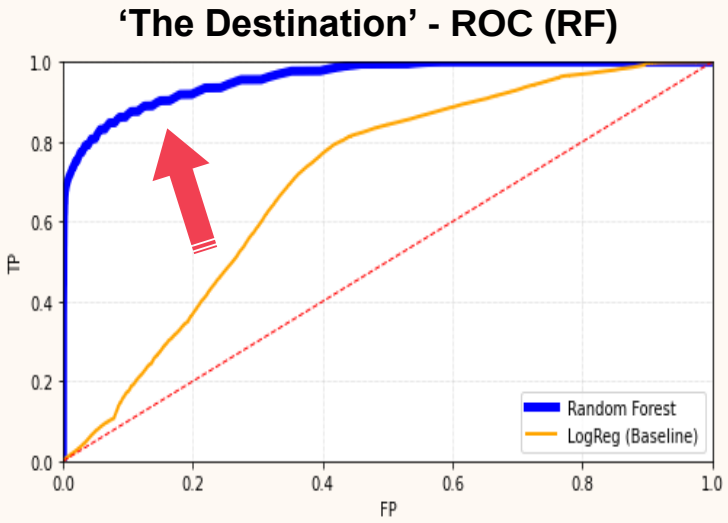
Result – Models 🎯

Choosing the best model:

- Model performances were compared via plotting ROC curve and evaluating its area (ROC AUC)
- Random Forest model yielded the best model

The Champion:

Random Forest Model



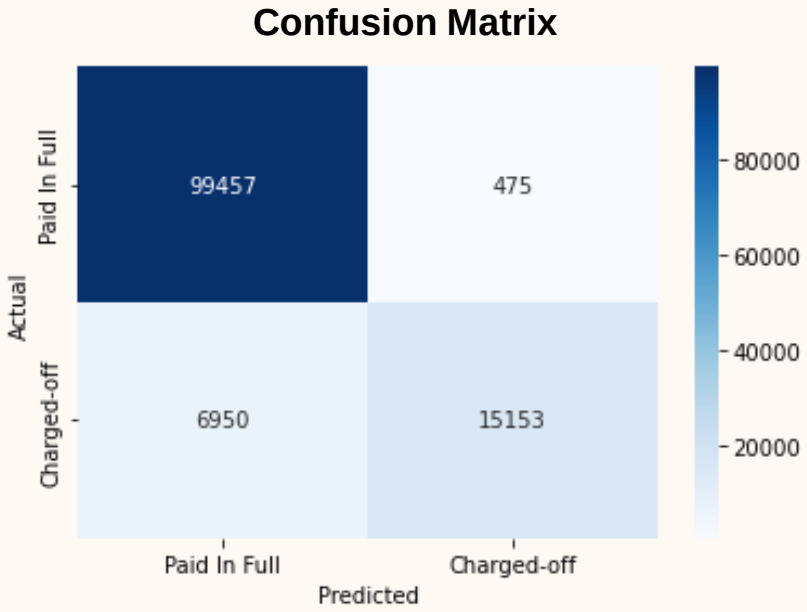
Metrics improvement over the baseline:

Accuracy: 0.820 → 0.938

Precision: 0.979 → 0.970

Recall: 0.002 → 0.681

AUC: 0.703 → 0.961



The Champion:

Random Forest Model

Metrics improvement over the baseline:

Accuracy: 0.820 → 0.938

Precision: 0.979 → 0.970

Recall: 0.002 → 0.681

AUC: 0.703 → 0.961

Further Works

https://www.bamboohr.com/small-business/

Further optimizations and visualizations

- Employing more features from the dataset
  ex) Business type, franchise, location…
- Visualizing predicted values, possibly via geocoding with tableau
- More thorough feature engineering
- Further fine-tuning (hyperparameters, etc)

Thank you!

Any questions?

U.S. Small Business Administration