Predicting Defaults for Small-Business Loans

Ray Karpman
March 22, 2022

Lending to Small Businesses

- Advantages:
  - Revenue
  - Public relations
- Disadvantage: risk of default
- **Goal:** data-driven strategy for screening loan applications
  - Avoid higher-risk loans
  - Approve loans likely to be paid

SBA Loan Program

- Small Business Administration (SBA)
  - US Government Agency
  - Provides backing for small-business loans
  - If business defaults, SBA pays lender a portion of the ballance
- Lenders still face serious risk
- **Our model** predicts whether SBA loan will go into default
  - Can help lenders avoid risky loans

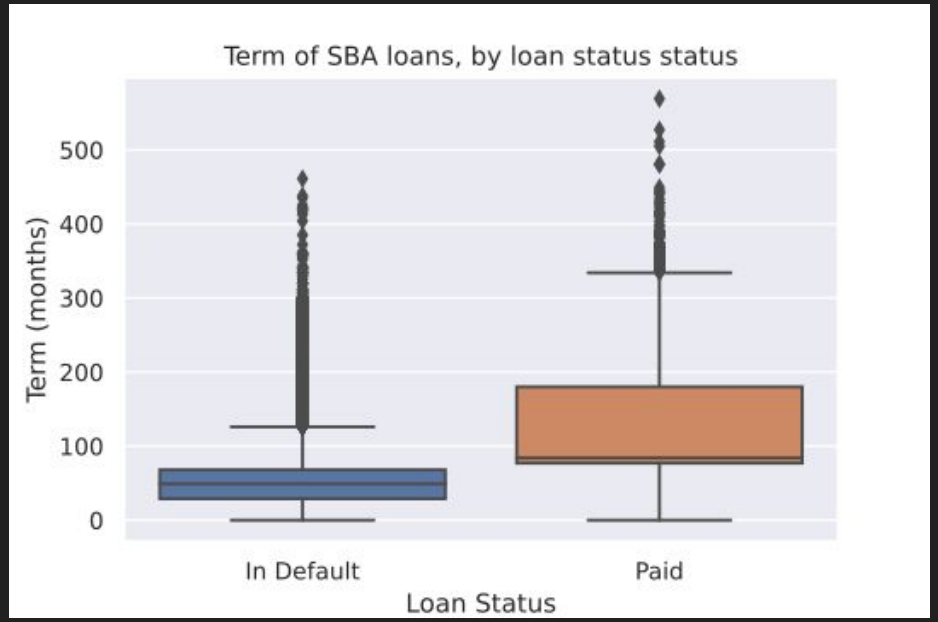
The Data

- From Small Business Administration Database
  - Retrieved by professors Min Li, Amy Mickel and Stanley Taylor.
- Records on over 890,000 loans
- 26 features, loan status (default or paid in full).
  - Characteristics of business:
    - Ex: industry, number employees, new or established business
  - Characteristics of loan:
    - Ex: year approved, term, amount covered by SBA

Data Cleaning

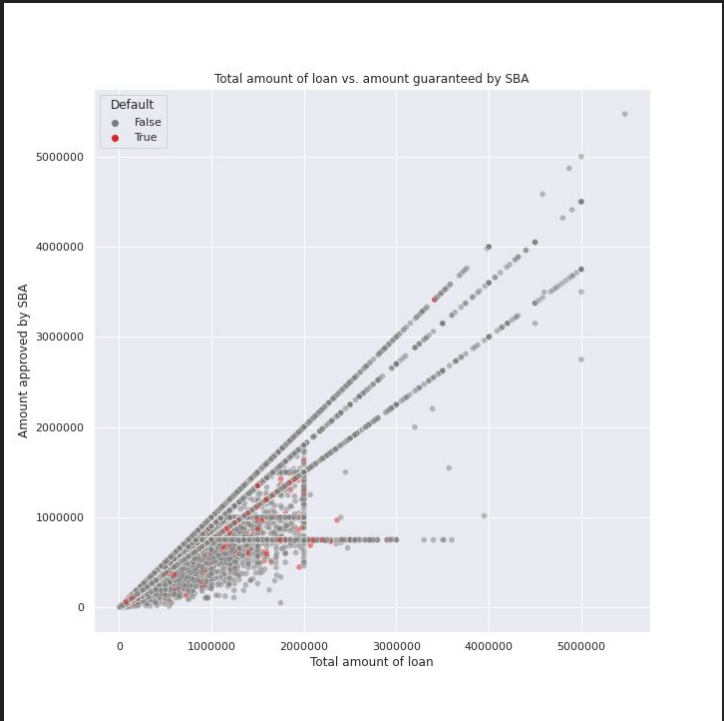
- Discarded loans approved between 1900, after 2010
  - Very little data from outside this range
- Dropped columns that describing outcome of loan
  - Ex: jobs created
- Dropped categorical features with many unique values
  - Ex: bank, zip code
- Retained 847977 records, 11 features

Key Predictors: Term



Loans that go into default tend to have shorter terms.

Key Predictors: Portion guaranteed by SBA

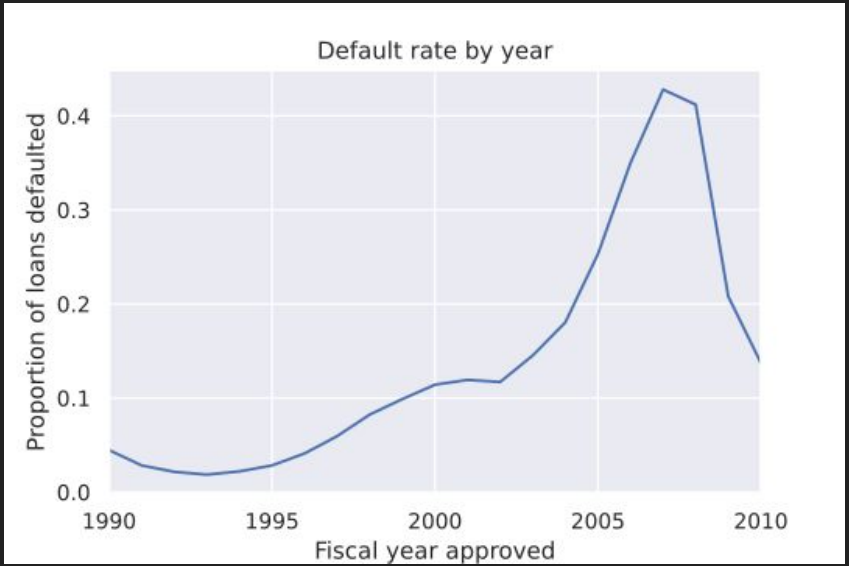


Loans that go into default tend to have:

- Smaller gross amounts approved
- Smaller amounts guaranteed by SBA

For loans that go into default, the amount guaranteed by the SBA tends to be a smaller fraction of total

Key Predictors: Year Approved



- Risk depends on year loan approved
- Reflects overall state of economy
- Loans approved in mid-late 2000's have highest default risk
    - Great recession: 2007-2009

Data preprocessing

- Randomly split data into training and test sets
  - 70% train, 30% test
- Used one-hot encoding for categorical variables
- Standardized all features
  - Scaled to mean 0, standard deviation 1
  - Fit scaler on training data only

Decision-tree based models

- Work well with skewed distributions, correlated features, outliers
- Random forest
  - Fit many decision trees using bootstrap samples, random selection of features.
- Gradient boosting
  - Use small trees, individually weak.
  - Iterative process: at each stage, fit additional tree to account for error remaining after previous stage.

Evaluating performance: The f1-score

- Data is imbalanced–most loans are paid off
- Traditional accuracy score may be misleading.
- Use f1-score instead.
  - Harmonic mean of precision of recall.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

Optimizing the model

- With scikit-learn defaults, random forest and gradient boosting gave similar f1-scores using cross-validation on training data.
  - Random forest: 0.78, gradient boosting: 0.79
- Tuned hyperparameters for both, using grid search with three-fold cross validation on training set.
  - Random forest: n_estimators, max_depth, criterion
  - Gradient boosting: n_estimators, max_depth, learning_rate

Hyperparameter tuning results

- Random forest: no significant improvement.
- Gradient boosting: tuning max_depth improved performance
  - Compared top two gradient boosting models from grid search
  - Chose model with n_estimators=100
    - Similar performance, faster training

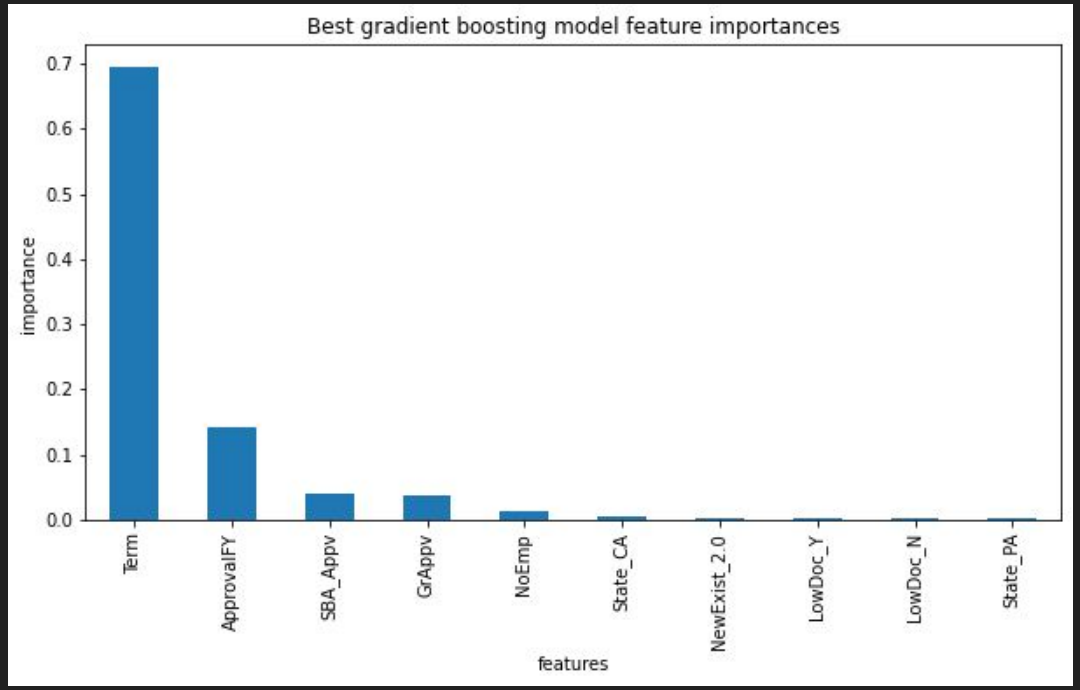
learning_rate	max_depth	n_estimators	f1 -score	Mean fit time
0.1	11	200	0.846	25.6 minutes
0.1	11	100	0.845	12.5 minutes

Final Model: Performance metrics

Accuracy	Precision	Recall	f1-score
0.95	0.83	0.86	0.85

- Fit time on training set: 14 minutes 2 seconds
- Prediction time on test set: 2 seconds

Final Model: Important features



Conclusions

- Final model gives very reliable predictions
  - Can make data-driven decisions about whether to approve a loan
- All key predictors are features of loan itself (e.g term)
  - Not based on characteristics of business
- **Conclusion:** SBA vetting borrowers very effectively

Future Improvements

- Better encoding for categorical variables
  - One-hot encoding not appropriate for variables with many values
  - Use numerical encoding, or sort into low/medium/high risk levels
- Accounting for economic conditions
  - Year approved is proxy for state of economy over life of loan
  - Create indicator feature for economic conditions
    - Could run multiple scenarios for a given loan
  - Incorporate economic forecast into model