Loan Default Prediction

Dominic Maranta

Abstract

- Loan Default Prediction Dataset
- Sourced, cleaned, and modeled the data using Python.
- Classification Problem using Logistic Regression, KNN, Decision Trees, Random Forests, Boosting, and SVM
- Random Forest satisfaction classifier with 74% accuracy was the strongest validated model.

Problem Statement

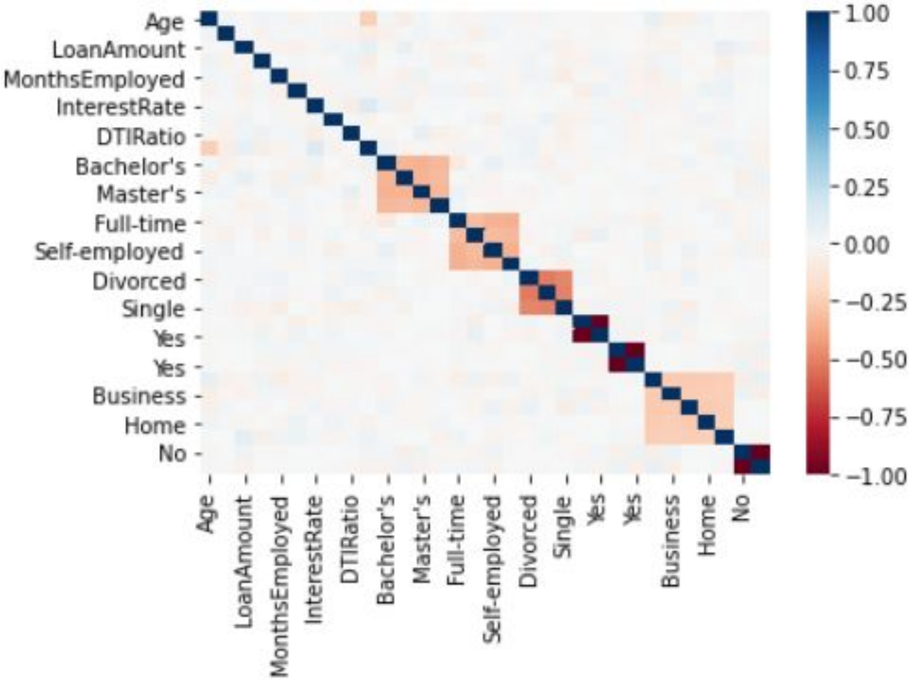
- Construct a classifier to predict borrower loan default.
- How can our predictions help borrowers and lenders?
- Use EDA to explore and prepare the data.
- Compare and Evaluate machine learning classification models.

Data Preprocessing

- Data: 18 variables with 255,000 instances
- 16 Predictors and 1 binary outcome: Loan Default or Not
- Select 2000 observations to allow for model runtime efficiency
- No missing values in the 2000 observations
- Remove one unnecessary Loan ID variable.
- Clean variables and create factors or numeric variables
- Create 80/20 Test Training Split

Data Understanding

- Loan amount, Debt to income ratio, loan purpose, and employment type all important predictors.
- Past research shows unemployment a leading driver of loan default



Evaluation

- Cross validation and test metric comparison: Accuracy/ROC Curve, Recall, Precision, and Confusion Matrix
- Test Decision Trees, Logistic Regression, K-Nearest Neighbors (KNN) models, AdaBoost, Gradient Boosting, Support Vector Machines (SVM), and Random Forests.
- Modeling using the SKlearn package in Python
- Cross Validation Comparison allows the tuning of various hyperparameters

Evaluation

- After testing and tuning all models Random Forests had the best classification accuracy of 74%.
- Gradient Boosting followed with a classification accuracy of 73.2%.
- SVM produced the lowest final accuracy of 47%

Discussion

- Random Forest model has validation accuracy of 74.25% and an ROC AUC value of 0.7415.
- Highest specificity, precision, sensitivity, and recall as well.
- Most Important Predictors: Loan amount, DTI Ratio, employment status, and loan purpose.

Outcomes

- Borrowers and Lenders can use this prediction model to help predict and identify loans in danger of default.
- Both parties can take action for default mitigation early on in the loan life.

Future Work

- Expand dataset to include all 255,000 records.
- Use this data to better tune SVM models and improve performance.
- Expand the model selection by including deep learning models such as Neural Networks.

Thank you!