Executive Summary - Kyle Phillips

Business Problem

The framework for this report stems from a situation in which I was hired by a major financial institution to build and compare several machine learning models to predict whether or not a loan is likely to default. The goal in building these models is to find the most accurate predictor of a loan’s status but also be able to explain the logic and methodology behind it. The data being used in this case is based on previous instances of loans being defaulted or not which is what the models will be trained on, and then a separate dataset to attempt and predict loan status on. In addition to building the models and determining significant variable factors in the data, I will also display my top 10 best predictions of a defaulted loan, top 10 predictions of a current loan, and top 10 incorrect predictions on both instances. This report will go through the entire modeling process as well as offer final insights into recommendations as to how the financial institution can tailor the results to improve their business practices and understanding of customers and loans.

Key Findings

1. In the exploratory analysis, it was determined that some variables have instances that are so rare that they are either 99% to one side or only have a single value through the entire dataset. This is evident with the variables “application\_type” and “pymnt\_plan.
2. The variable last\_credit\_pull\_d is the most significant variable and predictor in the dataset, as it is usually an outlier in any of the variable importance graphics, specifically select months which display the highest number of defaulted loans.
3. There are a few select variables that do most of the work in predicting loan\_status that stand out among the rest. Other than couple, there are many variables that aren’t significant predictors.

Model Performance & Interpretation

After formulating a logistic regression model, random forest model, and XGBoost model, the XGBoost turned out to be the best predictor of loan\_status, as defined by the following results:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| XGBoost Model Metrics | | | | | | | |
| # Variables | Partition | Accuracy | ROC AUC | LogLoss | Precision | Recall | F1 |
| 22 | Training | 0.99832 | 0.99998 | 0.02523 | 0.99585 | 0.99299 | 0.99442 |
| Testing | 0.95635 | 0.98291 | 0.11701 | 0.87031 | 0.83259 | 0.85103 |

This model displays a great ability of separability as displayed in the 0.98291 ROC AUC on the test data set. This model also has a great ability to predict correct instances, backed by the 0.95635 accuracy score.

Recommendations

1. Focus attention on why certain months such as September 2016, and why it is such a high predictor of defaulted loans. This can be done for other months and years as well. If there is some insight into why certain months see more defaults, then action can be taken in the company’s practices to highlight those and determine causes of that.
2. Most of the important variables included in the model are numeric variables relating to information dealing with payments or amounts of the loan. Given this information, the company can try to tie their efforts of reducing defaulted loans into more numeric variables or predictors.