Matthew Sparacio

Final Project: Loan Default Model Report

12/9/22

Model Report

# File Summary

For this analysis, two data sets were used. A data set containing information about people who took out loans and if they defaulted on them. A holdout set is also used which does not have a column for if they defaulted. Information about these data sets is listed below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **File Name** | **Record Count** | **Column Count** | **Numeric Columns** | **Character Columns** |
| Loan\_train.csv | 29777 | 52 | 29 | 23 |
| Loan\_holdout.csv | 12761 | 51 | 29 | 22 |

# Field Summary

Summary of numeric Variables.

Text

Description automatically generated with medium confidence

Summary of Categorical Variables:

Graphical user interface

Description automatically generated

# Target Summary

The target variables in this analysis is loan status. A loan can either be current (paid off) or default (not paid). We are interested in being able to predict if a person will likely default on their loan. Loan status is current 85% of the time and people default 15% of the time.

Graphical user interface, application

Description automatically generated

A screenshot of a computer

Description automatically generated

# Data Prep and Transformation

To address the date variables, I made issued, last payment, and last credit pull as months since (integer) and earliest credit line as years since. Additionally, I changed the interest rate and revolving line utilization rate from precents to decimals.

Removed many high cardinality variables including emp\_title, issue\_d, url, desc, title, zip\_code, next\_pymnt\_d. Additionally, mths\_since\_last\_delinq and mths\_since\_last\_record was less than 35% complete so both were removed.

The missing values for numeric variables were imputed using the median and all categorical variables were made into dummies. I also down sampled with a ratio of 5 to account for the discrepancy between the amount of current vs default loans. For the neural net model, all variables were normalized.

# Exploratory Data Analysis and Screening

Created a correlation matrix of all the numeric variables to see if any are relationships to be wary of. Looks ok.

Chart

Description automatically generated

Created box plots for each numeric variable comparing the boxplot for current vs default loans. The box plots with the most significant difference when a loan is current vs default are shown below. These can be identified as likely important predictors.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Created column charts comparing the amount of default vs current loans for each level of categorical variables. Below are the variables with the highest variance between categories and likely will be important for predictions.

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

# Anomaly Detection

To account for the anomalies in the data, I created an isolation forest to identify them, then remove them. After fitting the full isolation forest, we got the tree below with rules on if a variable is an anomaly. Using these rules, I removed 12 anomalies.

Diagram

Description automatically generated

# Model Building and Training

The first step in the model building process was to split the data into a 70% training set and a 30% test set to evaluate each model’s performance. Additionally, kfolds splits were set to 5 for tuning.

The first model built was a Lasso logistic regression model using every variable in the data set (besides ones already explained were removed). This is to serve as a baseline model and to potentially help omit some unnecessary variables that will slow down the more complex computations. This model produced many variables that were insignificant but after some analysis, state and sub grade were removed for the heavier machine learning models.

The rest of the models were built on this large subset of data and were also down sampled. Random Forest was the first model tuned and trained. This model was trained using k folds cross validation. The best parameters found through tuning were min n of 6 and 175 trees.

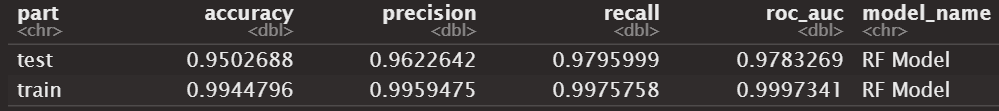
The next model trained was an XG Boost model. This was also tuned using kfolds cross validation. The best parameters the tuning found were 1505 trees, tree depth of 7, and a learning rate of 0.017.

The final model trained was a Neural Network. This model required a new recipe with the same variables but normalized. It was also trained using k folds cross validation and the best parameters were hidden units of 6, penalty of 0.667, and 581 epochs.

# Model Comparison

Below is all the information on how each model performed.

Random Forest:

Chart, line chart

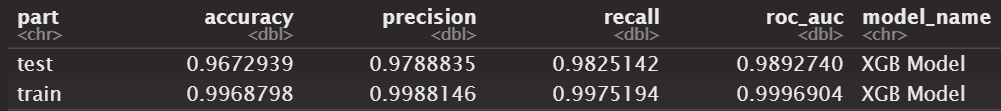
Description automatically generatedChart, histogram

Description automatically generatedChart, treemap chart

Description automatically generatedChart

Description automatically generated with medium confidence

XGBoost Model:

Chart, line chart

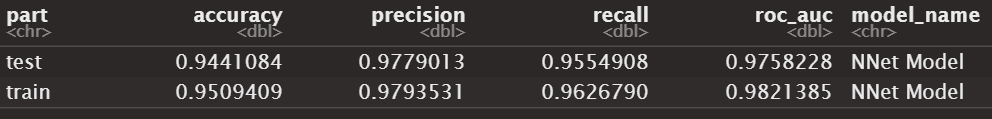
Description automatically generatedChart

Description automatically generatedChart, treemap chart

Description automatically generatedChart

Description automatically generated with medium confidence

Neural Network:

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generatedChart, treemap chart

Description automatically generatedChart

Description automatically generated with medium confidence

Comparing these three models, every model performed quite well. The neural net is the model that is least overfit because of the training and test auc being so similar, but the XGB model performed the best across the board, so that is the model that will be used to make predictions with.

# Best Model

Diving deeper into the XGB model, we looked at the operating ranges to balance TPR to FPR. Using a KS, the optimal operating FPR is 0.06 which is shown on the ROC Curve below as well.

Table

Description automatically generatedChart

Description automatically generated

We also looked at the variable importance of the model. We looked at both the global variable importance and the local variable importance. The global importance shows how important each variable is generally, but the local gives us details on which direction and by how much the variable effects the model. The first chart below shows the global variable importance of the top 20 variables. Next, I created partial dependence plots to analyze these important variables further. The most insightful plots are shown below.

Chart

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generatedChart

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidenceChart, bar chart, histogram

Description automatically generated

Lastly, the SHAP method was used for local explanations. This gives us the most insight into how important a variable was to the model and tells us which direction it influenced the model.

Chart

Description automatically generated

Looking more closely at the predictions of the model, we got the top 10 best true positives (predicted to default with certainty and did), the top 10 worst false positives (predicted to not default with certainty but did), and the top 10 worst false negatives (predicted to default with certainty but did not).

Best True Positives: The best predictions were mostly on 36-month term loans, had a middle of the road grade, and were early in their careers (low employment length). Their last payment amount was fairly low and their last credit pull was the shortest it could be.

Graphical user interface, text

Description automatically generated A picture containing text, monitor, black, screenshot

Description automatically generated A picture containing text, indoor, several

Description automatically generated A screenshot of a computer screen

Description automatically generated with medium confidence Graphical user interface, application

Description automatically generated

Worst False Positives: The worst false positives were often on higher grade loans with low interest rates. They tended to have higher fico ranges and some last payments were well above average.

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

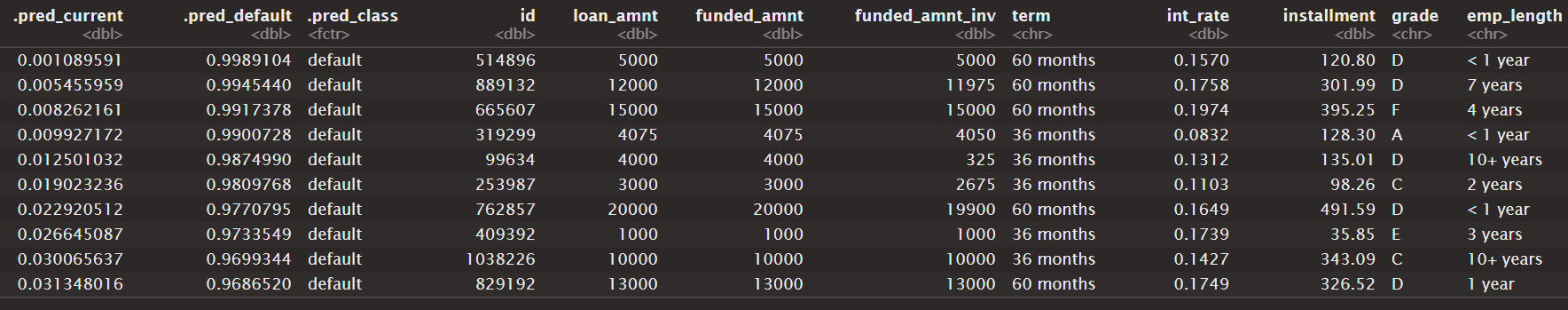
Description automatically generated with medium confidence Graphical user interface

Description automatically generated A screenshot of a computer screen

Description automatically generated with medium confidence Graphical user interface

Description automatically generated

Worst False Negatives: The worst false negatives was fairly diverse and middle of the road. Many of the people with these loans were renters which could be why we predicted they would default.

 A computer screen capture

Description automatically generated with medium confidence Graphical user interface

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence