Matt Siriani

July 24, 2023

Rutgers Data Analytics Bootcamp

Credit Risk Classification Analysis

This week, we focused on using various techniques to train and evaluate a model based on loan risk. We did this by using a dataset of lending activity from a peer-to-peer lending service company and build a model based on the credit of buyers.

The overall factors that contribute to this analysis include the size of the loan, the interest rate of the loan, the borrower’s income, debt-to-income ratio, the number of accounts held by the borrower, negative marks against the barber, and the total debt.

The dataset contains over 77,000 datapoints. I split the into training and test sets. Afterwards, I used the training set to build the first logistic regression model. The first model was applied to the testing dataset. The purpose of the initial model was to see whether a loan to the borrower in the testing set would be low or high-risk results.

The breakdown of low and high-risk data points based on what I was able to observe featured 75,000 low-risk and 2,500 high-risk data points. To resample the training data and ensure that the logistic regression model had an equal number of data points to draw from, the training set data was resampled with the “RandomOverSampler” module from imbalanced-learn. By doing this, it generated 56,277 data points for both low-risk (0) and high-risk (1) loans, based on the original dataset.

The resampled data was used to build a new logistic regression model (Logistic Regression Model 2). The purpose of Logistic Regression Model 2 was to determine whether a loan to the borrower in the testing set would be low- or high-risk.

Logistic Regression model one showed that there was a 93% precision rate, a 94% accuracy score, and a 95% recall rate. The second model had the same precision rate of 93%. However, the Accuracy Score and recall rate for the second model were both 100%.

In conclusion, logistic regression model 2 is less likely to predict false negative results. However, based on the confusion matrices for each model, logistic regression model 2 predicted slightly more false positives (low-risk when the actual was high-risk).

If the goal of the model is to determine the likelihood of high-risk loans, neither model scores above 90% precision. Logistic Regression Model 2 had fewer false predictions of the testing data overall and would be the best model to use based on the high accuracy and recall of this model.