Team Number: IC22004

Paycheck Protection Program (PPP), a Small Business Administration (SBA)-backed loan, helps businesses keep employed during the COVID-19. The SBA has periodically released data on the more than 11.5 million approved applications, but it also has removed applications that had been previously present in the dataset for hidden reason. This project focused on loans to businesses in Georgia with 550k non-removed and 22k removed loans from the PPP database. The project goals were:

1. Find out characteristics of the removed loans.
2. Compare characteristics of removed and non-removed loans.
3. Build a predictive model with the characteristics we found to predict whether a loan would be removed from the dataset.

In this project, we performed data processing, exploratory data analysis (EDA), and predictive modeling using Python. In the data processing, we organized data such as extracting the first two digits from the NAICS code to simplify categorization, classifying multifarious business types into four major classes, and many more. During EDA, we identified 16 variables that had different characteristics between non-removed and removed loans. For instance, among all the business types, the sole company had the highest application removal rate. Also, applications approved by lenders in April and May 2021 had a high removal rate at the SBA level. In the predictive modeling, we fed these 16 input variables into the XGBoost model, resulting in an Area Under Curve (AUC) score of 0.949 the true positive rate was 89.9%. These results showed our model achieved our accuracy objectives.