Team Number: IC22004

Paycheck Protection Program (PPP), a Small Business Administration (SBA)-backed loan, helps businesses keep employed during the COVID-19. SBA released millions of approved applications, but it removed some previously present applications for untold reasons. In the IC22, we focused on PPP data in Georgia with 550k non-removed and 22k removed loans and our goals were to 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 analysis, we performed data processing, exploratory data analysis (EDA) and predictive model building by using Python. In the data processing, we extracted the first two digits from the NAICS code to simplify categorization, and classified multifarious business types into four major classes. In the step of EDA, we identified 16 variables that had different characteristics between non-removed and removed loans. For instance, among business types, the sole company had an unreal-high removal rate, and applications approved in April and May 2021 had a high removal rate. We also found that compared to non-removed applications, removed applications had a lower percentage of initial approval amount with decimals. In the model building step, we fed these 16 impactful variables to predictive models, and XGBoost model had the best Area Under Curve (AUC) of 0.949.