**Abstract**

In this project, we aim to understand the differences between features associated to loans initially approved but later removed from the payroll protection program(PPP) and predicting which loans are likely to be removed.

The Paycheck protection program was launched by SBA as a stimulus package to aid small businesses hit by COVID, with billions of dollars designated and guaranteed by the U.S Federal Government. Multiple irregularities were noted throughout the program with ineligible large businesses being approved for these relief loans, newly created businesses being sanctioned loans and “phished” loans to individuals belonging to criminal networks. Lenders facilitating these suspicious loans were “fintech” and there seems to have been the possibility of corruption, which is also of particular interest for our project.

We approached this problem from a socio-economic perspective by incorporating the American census data as an overlay on the PPP dataset. We translated NAICS codes by merging the data with an NAICS dataset. By doing so, we were able to identify the industry for approved loans and removed loans.

To understand variable interactions better, we used Tableau to visualize the data and gain insights. Using these insights and by ranking feature importance, we built our models using various classification techniques such as Logistic Regression, Random Forest Classification, Decision Tree classification, etc. We evaluate the models based on accuracy, precision, recall, confusion matrix and AUC. Our final step is to fine tune the selected model using hyper-parameter tuning to get the best possible results.