TEAM# IC22030

Team Members: Ling Fang, Yu-Tung Chang, Upasana Mohapatra

To predict the removal of loans from the public dataset, we're working on building a **classification model** to predict whether loans will be removed, across various industries and regions using the given datasets - **removed loans**, the original **public dataset** including all loans, and using intelligence from additional data sets such as the census data, NAICS data on industry classification.

Comparing the average values for loan amount in the full dataset and the combined dataset **(~$44K)** with the average loan amount from the removed dataset**($20K)**, we see a huge difference in loan borrowing. We then estimated the **size of businesses** in the datasets by looking into the jobs retained per service industry(source: NAICS data). Most of the jobs retained in the **removed loan** dataset were from miscellaneous service industries(**mostly B2C**) such as beauty salons and barbershops that are **smaller**. For the **full dataset**, most of the jobs retained were in **bigger service** industries(**B2B, B2C**) like Accommodation and Food Services, Retail Trade, etc.

Furthermore, we looked into the **top lenders and cities** in the removed loans dataset the maximum number of approved loans and removed loans. Some of the more prominent cities from this analysis are – '**Atlanta', 'Decatur', 'Fairburn'** while some of the lenders approving the maximum number of removed loans are **- 'Capital Plus Financial LLC', 'Prestamos CDFI LLC'**. We also used data from the census to understand the average income across congressional districts in the state of Georgia and then compared those values to the size of loans borrowed in those districts. This could be indicative of the fact that districts with a lower average monthly income could more likely be prospects for loan borrowers. All these factors helped us reduce our problem statement into a defined set of hypotheses.

We built a binary classification model by creating a **neural network** so that our model understands best to converse with features that narrate an evident story. Considering the **difference in the size** of the two datasets provided, we believed it would not be fair to train our model based only on patterns observed in a familiar dataset, because the sample size of the dataset comprising of removed loans was close to 4% of the size of the full dataset on which the predictions were to be made. For this reason, we used one-hot encoding to better feed our categorical variables into the model and let the neural network learn patterns as it iterates through the entire dataset. A key tactic used here was to first choose an equally sized and randomly chosen sample from the full dataset and use that for our model first. We used 75% of the removed loans and 25% of the full dataset as our training data. With the help of our model, we were able to minimize the mean average error and we then used the prediction value to be applied to the entire dataset. This exercise returned 542 out of ~550,000 rows that were to be removed based on the values we fed into the training data.

We then looked into patterns in the predicted data that were most evidently similar to our removed sample and recorded similarities and differences.