Fraud Detection for PPP Loans

Ajanya Sharma, Aayush Bakre
Instructor: Steve Taylor

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

- Explore loan data from the Paycheck Protection Program (PPP), which provided relief to small and medium-sized businesses during the COVID-19 pandemic and identify probable fraudulent loans. The primary objective is to reduce frauds in the future by applying anomaly detection methods to identify outliers and building machine learning models to potentially detect possible frauds.

Technology

- Python was used for data pre-processing and cleaning
- Python libraries matplotlib, plotly, seaborn utilized for exploratory analysis and dashboarding
- Python to identify anomalies using pandas, matplotlib, IsolationForest in slkearn

Current & Future Work

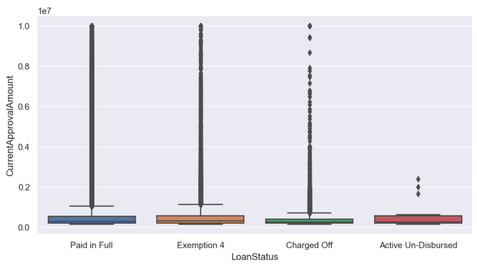
- 22% of the $510 billion sanctioned was identified as possible fraud and Maine accounted for highest fraud to loan ratio
- Government agencies and policymakers may use the data to evaluate the effectiveness of the program and make necessary adjustments to future relief efforts
- Researchers may use data to study the economic impact of the pandemic on small businesses and to identify patterns and trends in the distribution of PPP loans
- Incorporate possible frauds with portal like Datamerch, Experian, Iapps to improve due diligence for debt-based Venture Capitalists

Exploratory Analysis

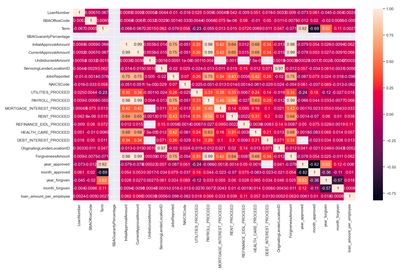
Visualizations for this project are intended to aid the user get an essential understanding of the context of the subject and get a feel of what the dataset is trying to convey on the high level.

Supplementary intentions with the visualizations are to be informative for the user to know where to start looking for anomalies first. Ultimately, we want the the visualizations to aid the process of identifying anomalous phenomena by guiding the user's attention to some extreme unruly observations as a starting place for the outlier analysis.

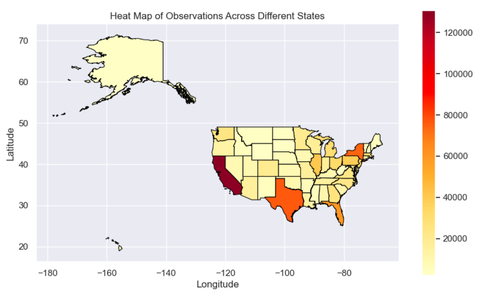
Some examples of this effort can be seen below.



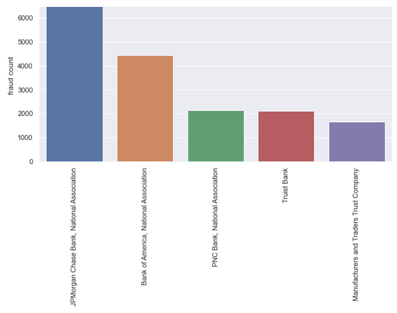
Approval Amount Outlier Analysis



Correlation heat-matrix to identify key dependencies



Heatmap of number of loans approved across all the states



Top 5 lenders vs Probable Fraud count

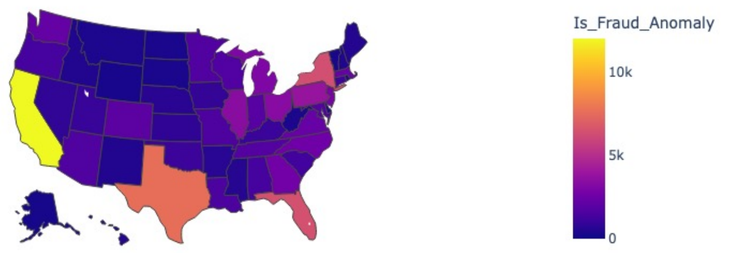
Anomaly Detection

Standardization is done using standard scalar
Higher the anomaly score, higher the probability of loan being fraud

Isolation forest is being used to calculate anomaly scores
Anomaly score is calculated for features:
Initial Approval Amount, Payroll proceed and Jobs Reported

We then identify a loan being fraud if it lies in the top 95% of anomaly score calculated

State-wise Probable Frauds



State-wise Loans vs Frauds