Project Proposal
Identifying Early Signs of Bank Account Fraud

Anthony Coffin-Schmitt (awc93), Jackie Lasseter (jl2638), Elana Pocress (erp49)

Course Code: ORIE 5741, Cornell University (Spring 2024)
Github Link: https://github.com/AnthonyCS/orie_5741_project

March 17, 2024

Problem Addressed and Importance

Fraudulent applications in the financial sector carry significant risk to businesses and the public. Identity theft, fraudulent credit and loan transactions, lost time and revenue are just a few of the risks from bank fraud. The Federal Trade Commission reported over $10 Billion in financial fraud loss in 2023. Ideally to minimize losses, the fraud should be detected as early as possible and at the source of a new account application. Waiting to catch fraudulent transactions as they occur is inefficient and carries higher risk. This project proposes leveraging big data analytics techniques on a newly available dataset, the Bank Account Fraud (BAF) suite, to identify warning signs of potential fraud in bank account applications.

Questions to Answer

- Can a trained model accurately predict whether an application is fraudulent?
- What factors have the highest predictive power of fraudulent applications?
  - To what extent is each feature contribute to the model's ability to distinguish between fraudulent and non-fraudulent applications.
- The accuracy and efficiency of different models in detecting fraudulent applications.

Benefit to Stakeholders / Financial Sector

The above questions result naturally from a desire to understand, predict, and mitigate fraud. From a banking perspective, predicting which consumers may be risky can inform loan offerings, credit card openings, and can prevent lawsuits that arise from fraud claims. It is in the best interest of the banking industry to leverage machine learning models to maximize their profits and protect their shareholders. However, this may come at the expense of consumers who are predicted to be fraudulent but actually are not. The topic of "fairness" in a dataset is particularly important to consider when certain predictors may be sensitive to certain characteristics of a person. For example, zip code can often be used as a proxy for race. In this case, it is important to create a model that does not overfit to any particular feature. In our proposed approach we will discuss how feature engineering will play a role in our analysis.

Dataset Description

Released in 2023 at NeurIPS Datasets and Benchmarks, BAF is large-scale and realistic tabular dataset that maintains privacy. It contains a large sample of 2.5 million bank application instances generated from a generative model (CTGAN) trained on anonymized original banking data. Within the dataset suite, there are six instances, five of which represent a specific bias type and lastly a 'base' dataset. Each instance contains a 1 million sub-samples from the large sample. We will primarily be working with the 'base' dataset as that most closely resembles the original banking data.

The labeled tabular datasets contain 30 features and also temporal information. There is also information on the protected attributes of different groups, e.g. age, employment status and income.

Links to the dataset:

- https://arxiv.org/abs/2211.13358
- https://github.com/feedzai/bank-account-fraud
- https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf

How Does the Dataset Help Answer the Question

Since banking application data is difficult to find due to privacy issues, this data allows us to dig deeper into detecting which banking applications are indeed fraudulent. Our data set contains information regarding application materials and whether the application ended up being fraudulent as a binary output. Given this information, we will be able to construct a logistic regression model to estimate the probability of a transaction being fraudulent based on a variety of input features.

Proposed Approach

Initially, we will conduct exploratory data analysis to better understand the data and the included features. Better understanding of the data could provide opportunities for feature engineering to make a more rich dataset. Clustering the data could also be helpful to visualize and see if any distinct groups form. Performing a Principle Component Analysis would help determine the most important features. Sampling different features of interest could also give cluster insights with k-NN or k-means. We will perform feature engineering, as necessary, to transform data for our use case.

Performing classification either with kernel SVM and/or logistic regression to predict if an application is fraudulent or not will be a major focus. We will likely try out several different vanilla models to see how they perform and then tune hyper-parameters and other parameters to decrease loss and improve efficacy. If time and computational resources permit, we may investigate more modern models and possabily neural nets.

Conclusion

This project will provide a valuable opportunity to improve early fraud detection and prevent financial loss. We will be analyzing a bank account fraud dataset to look at factors that are relevant for predicting fraud. In doing so, we aim to answer the questions of: what characteristics are correlated with fraudulent behavior (and to what degree) and how fairness is evaluated in a model for predicting fraud. This will be done using logistic regression and SVM to find a model that will be most effective and efficient.