Fraud Detection in Financial Transactions

By: Biyag “Zig” Dukuray

Introduction to Data Science in Fraud detection

- Fraud Detection in financial transactions refers to the process of preventing and identifying fraudulent activities. These include but aren't limited to unauthorized transactions, identity theft or financial scams.
- This process involves analyzing transactional data, patterns, and customer behavior to detect anomalies and potential instances of fraud.
- Importance of Fraud Protection: Protects businesses in safeguarding assets, reputations and customer trust by minimizing financial losses experienced from fraud.
- When was the last time your bank has sent you a text confirming whether it was you?

Approaches to Fraud Protection

- Supervised Learning: This approach involves training a machine learning model on labeled data, where the labels indicate whether a transaction is fraudulent or not. This is similar to our sentiment analysis assignment we completed.
- Common supervised learning algorithms used for fraud detection include logistic regression, random forests, and gradient boosting machines. We have utilized supervised machine learning especially in Project 4 where I used Random Forest for the Ham/Spam dataset.
- Unsupervised Learning: Such techniques are used when labeled fraud data is scarce. Algorithms such as clustering and anomaly detection identify unusual patterns or outliers in transactional data that may indicate fraud.

Supervised Learning Models

Figure 39: Random Forest - Train Confusion Matrix

Precision: 100.0%
Recall: 99.84%

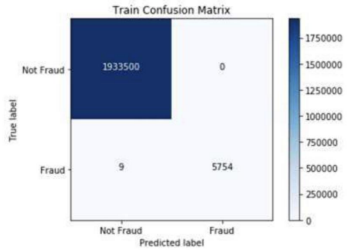
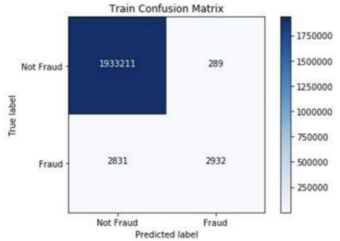


Figure 35: Logistic Regression - Train Confusion Matrix

Precision: 91.03%
Recall: 50.88%



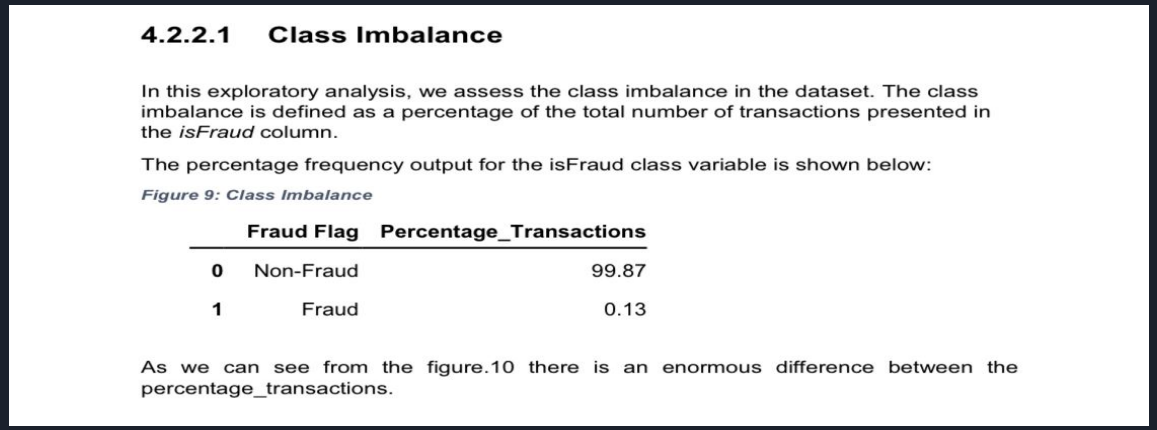
The following table compares the results of the two models:

Table 4: Comparison of Results of Logistic Regression and Random Forest

Model	Train Precision	Train Recall	Test Precision	Test Recall
Logistic Regression	91.03%	50.88%	90.12%	51.7%
Random Forest	100%	99.84%	100%	99.79%

Challenges in Fraud Detection

- Imbalanced Data: Fraudulent transactions are often rare compared to legitimate ones, leading to imbalanced datasets. This imbalance can result in biased models that fail to accurately detect fraudulent transactions while also minimizing false positives.
- This is an example of an imbalance in a dataset that has transaction data:



4.2.2.1 Class Imbalance

In this exploratory analysis, we assess the class imbalance in the dataset. The class imbalance is defined as a percentage of the total number of transactions presented in the *isFraud* column.

The percentage frequency output for the isFraud class variable is shown below:

Figure 9: Class Imbalance

	Fraud Flag	Percentage_Transactions
0	Non-Fraud	99.87
1	Fraud	0.13

As we can see from the figure.10 there is an enormous difference between the percentage_transactions.

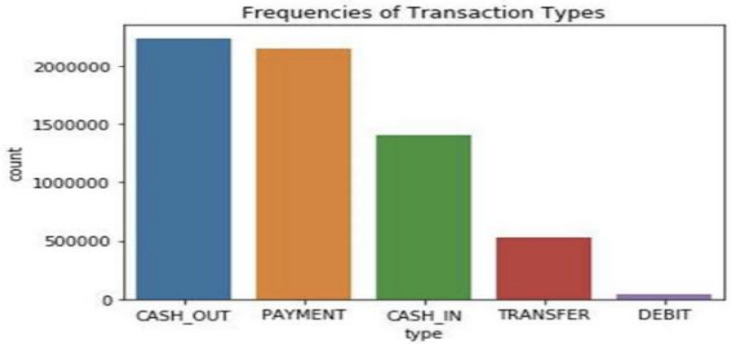
Glimpse at the data in a dataset

4.2.2.2 Types of Transactions

In this section, we are exploring the dataset by examining the 'type' variable. We present what the different 'types' of transactions are and which of these types can be fraudulent.

The following plot shows the frequencies of the different transaction types:

Figure 11: Frequencies of Transaction Types



The most frequent transaction types are **CASH-OUT** and **PAYMENT**.

From the above possible types of transactions, only cash-out and transfer are considered as fraudulent transactions.

Feel free to ask any questions

References:

- https://www.transunion.com/business-needs/fraud-prevention/banking-fraud-detection#:~:text=Banks%20analyze%20historical%20transaction%20data,it%20could%20trigger%20an%20alert.
- https://repository.rit.edu/cgi/viewcontent.cgi?article=11833&context=theses