CREDIT CARD FRAUD DETECTION

Machine Learning Project

By Aryan Verma & Jappan Sharma

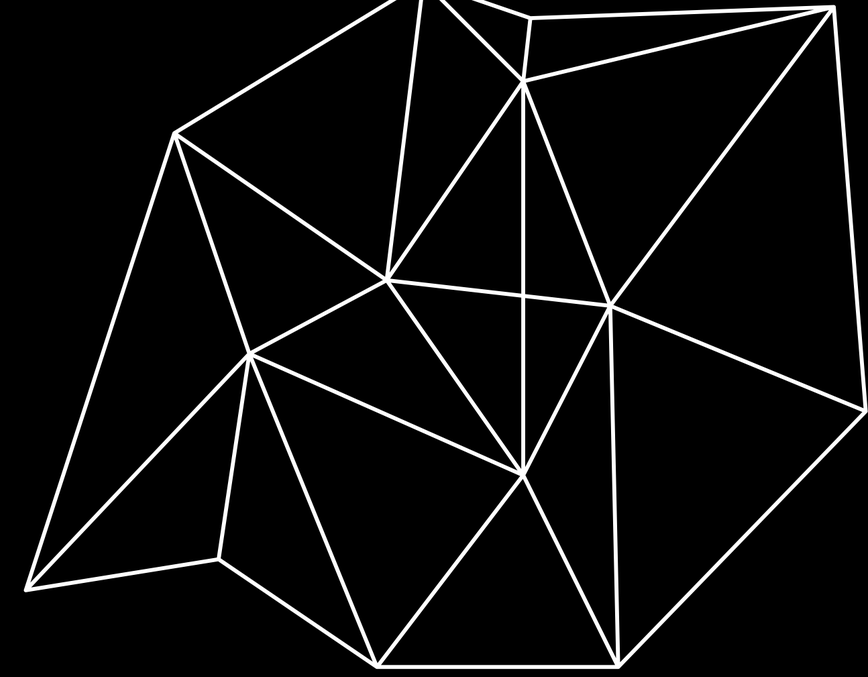
AIM OF THE PROJECT

The challenge is to recognize fraudulent credit card transactions so that the customers of credit card companies are not charged for items that they did not purchase.

MAIN CHALLENGES INVOLVED IN CREDIT CARD FRAUD DETECTION

1. Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.
2. Imbalanced Data i.e most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones
3. Data availability as the data is mostly private.
4. Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.
5. Adaptive techniques used against the model by the scammers.

HOW TO TACKLE THESE CHALLENGES?



Challenge 1

The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible.

Challenge 2

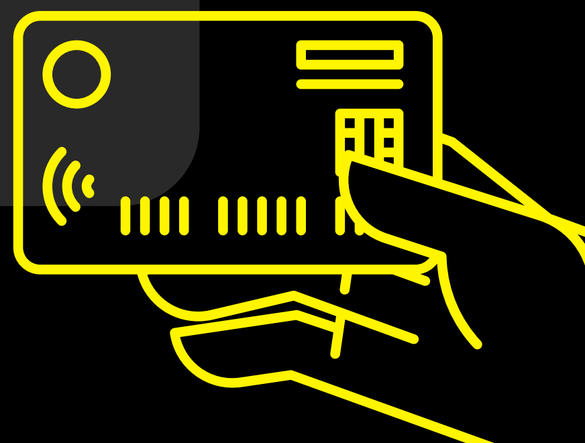
A more trustworthy source must be taken which double-check the data

Challenge 3

We can make the model simple and interpretable so that when the scammer adapts to it with just some tweaks we can have a new model up and running to deploy.

CODE OVERVIEW

- Importing necessary libraries.
- Loading and understanding the data.
- Checking imbalance in the data.
- Printing details of fraudulent and normal transactions.
- Plotting the correlation matrix.



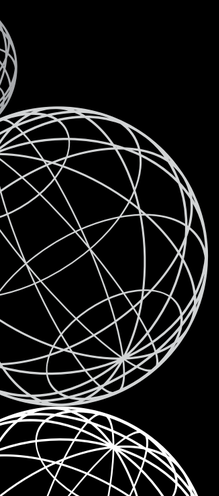
ALGO USED

1. **Isolation Forest Algorithm**
2. **Local Outlier Factor(LOF) Algorithm**

ISOLATION FOREST ALGORITHM

This method is an algorithm with a low linear time complexity and a small memory requirement. It builds a good performing model with a small number of trees using small sub-samples of fixed size, regardless of the size of a data set.

The Isolation Forest algorithm isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The logic argument goes: isolating anomaly observations is easier because only a few conditions are needed to separate those cases from the normal observations.



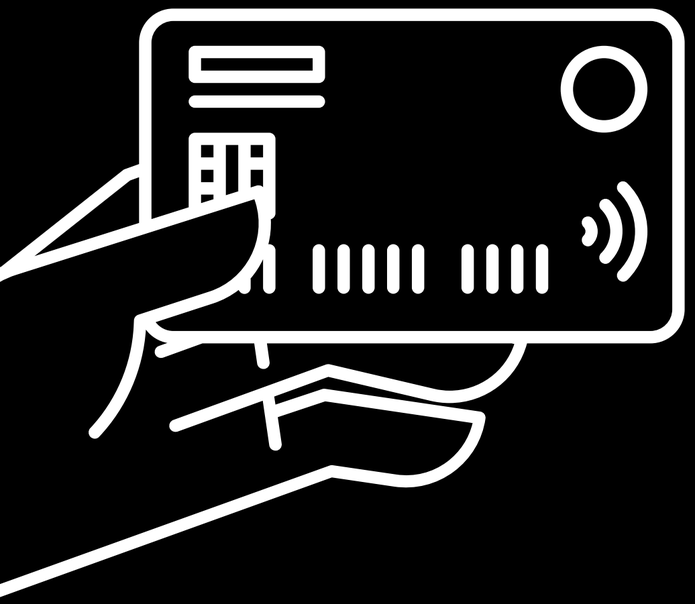
LOCAL OUTLIER FACTOR ALGORITHM

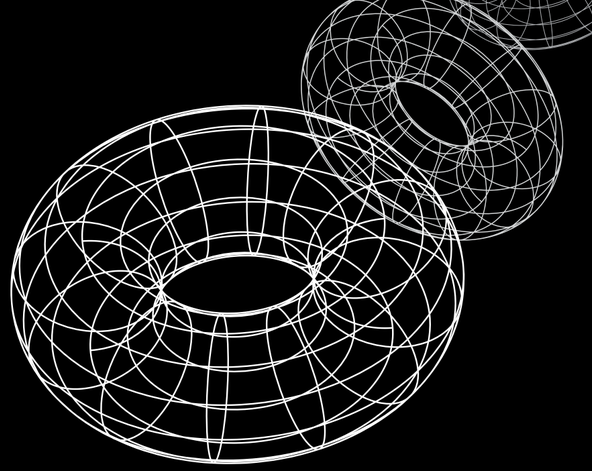
The LOF algorithm is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outlier samples that have a substantially lower density than their neighbors.

The number of neighbors considered, (parameter n_neighbors) is typically chosen 1) greater than the minimum number of objects a cluster has to contain, so that other objects can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by objects that can potentially be local outliers. In practice, such informations are generally not available, and taking n_neighbors=20 appears to work well in general.

OBSERVATIONS :

- Isolation Forest detected 73 errors versus Local Outlier Factor detecting 97 errors vs. SVM detecting 8516 errors
- Isolation Forest has a 99.74% more accurate than LOF of 99.65% and SVM of 70.09
- When comparing error precision & recall for 3 models , the Isolation Forest performed much better than the LOF as we can see that the detection of fraud cases is around 27 % versus LOF detection rate of just 2 % and SVM of 0%.
- So overall Isolation Forest Method performed much better in determining the fraud cases which is around 30%.
- We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense.We can also use complex anomaly detection models to get better accuracy in determining more fraudulent cases





THANK YOU

Made By Aryan Verma (102103222) & Jappan Sharma (102103228)