IBM COURSERA ADVANCED DATA SCIENCE CAPSTONE

Fraud Detection : Exploring methods to classify accounts as fraud or not
-Stakeholder

Soham Mukherjee

Outline

- Data Set
- Use case
- Solution

Data Set

The Bank Account Fraud (BAF) suite of datasets has been published at **NeurIPS 2022** and it comprises a total of 6 different synthetic bank account fraud tabular datasets.

This suite of datasets is:

- Realistic, based on a present-day real-world dataset for fraud detection;
- Biased, each dataset has distinct controlled types of bias;
- Imbalanced, this setting presents a extremely low prevalence of positive class;
- Dynamic, with temporal data and observed distribution shifts;
- Privacy preserving, to protect the identity of potential applicants we have applied differential privacy techniques (noise addition), feature encoding and trained a generative model (CTGAN).

I have used 1 of the 6 datasets - Base.csv for this project

Source - https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022/data

Data Set - Snapshot

Snapshot of 1 row and all the columns.

Data Dictionary - The full data dictionary can be found in this link - https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf

Total columns - 32	Numeric Columns - 20
Total records - 1000000	Categorical - 5
Target Field - fraud_bool	Binary - 6

fraud_bool	0
income	0.3
name_email_similarity	0.986506310633034
prev_address_months_count	-1
current_address_months_count	25
customer_age	40
days_since_request	0.0067353870811739
intended_balcon_amount	102.45371092469456
payment_type	AA
zip_count_4w	1059
velocity_6h	13096.035018400871
velocity_24h	7850.955007125409
velocity_4w	6742.080561007602
bank_branch_count_8w	5
date_of_birth_distinct_emails_4w	5
employment_status	CB
credit_risk_score	163
email_is_free	1
housing_status	BC
phone_home_valid	0
phone_mobile_valid	1
bank_months_count	9
has_other_cards	0
proposed_credit_limit	1500.0
foreign_request	0
source	INTERNET
session_length_in_minutes	16.224843433978073
device_os	linux
keep_alive_session	1
device_distinct_emails_8w	1
device_fraud_count	0
month	0

Use Case

Objective : The objective of the project is to build an automated solution to classifying an account as a fraudulent account or not.

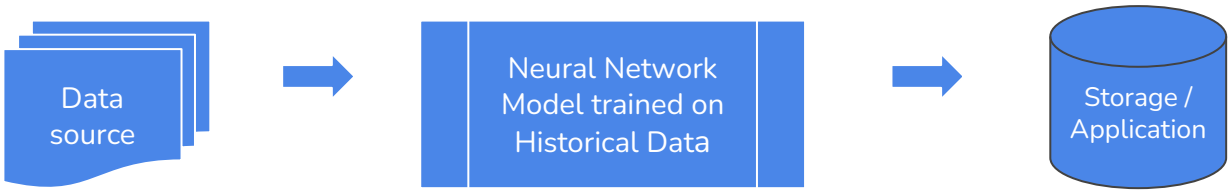
Procedure : The process followed here is to build a data driven machine learning or deep learning model that can learn from historical data and which can be used to classify new records as fraudulent or not.

Solution

The solution is to use a neural network based solution.

Comparison was made between LogisticRegression and Neural Network. Both performed well, however, given a neural network' ability to undergo training with backpropagation, adjusting the weights and biases in the interconnected neurons to minimize errors and optimize performance.

The neural network model has been trained on almost 1 million records comprising of both Fraudulent and non-Fraudulent records.



The trained model has an high accuracy of 96.7 % with a training loss of 8%