Fraud Data Guardian

Helping Banks Uncover Financial Frauds.

A presentation on Leadership in Analytics.

- Leadership in Analytics (ALY6120)
- **Guided by:** Sasan Mehrabian
- **Submitted by**: Group – 1
- **Date of submission :** 16 May’2023

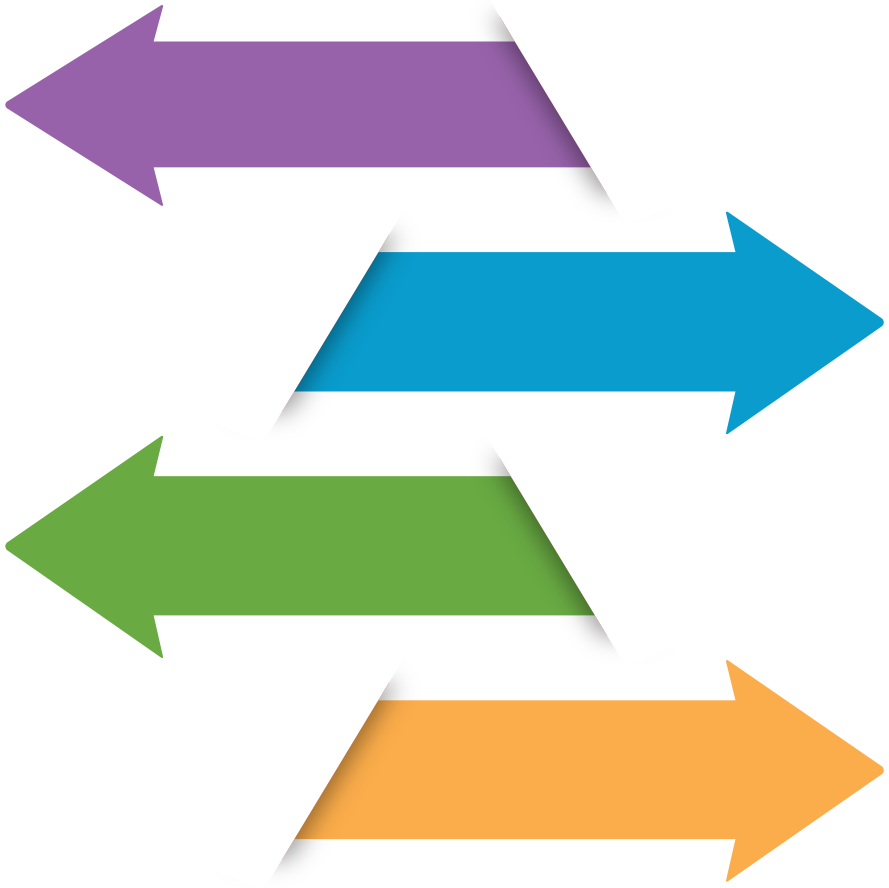
BUSINESS BACKGROUND AND OBJECTIVE

Business Background

Addressing **fraud detection** within the organization **through predictive analytics models**, transitioning from hindsight to foresight and enabling informed decision-making.

Model Refinement

Iterative process to fine-tune models and address data quality issues, **aiming for improved accuracy and effectiveness**.



Methodology

Utilizing **XGBoost and Random Forest models** for **analysing customer's transaction data**, including data pre-processing, model training, and performance evaluation.

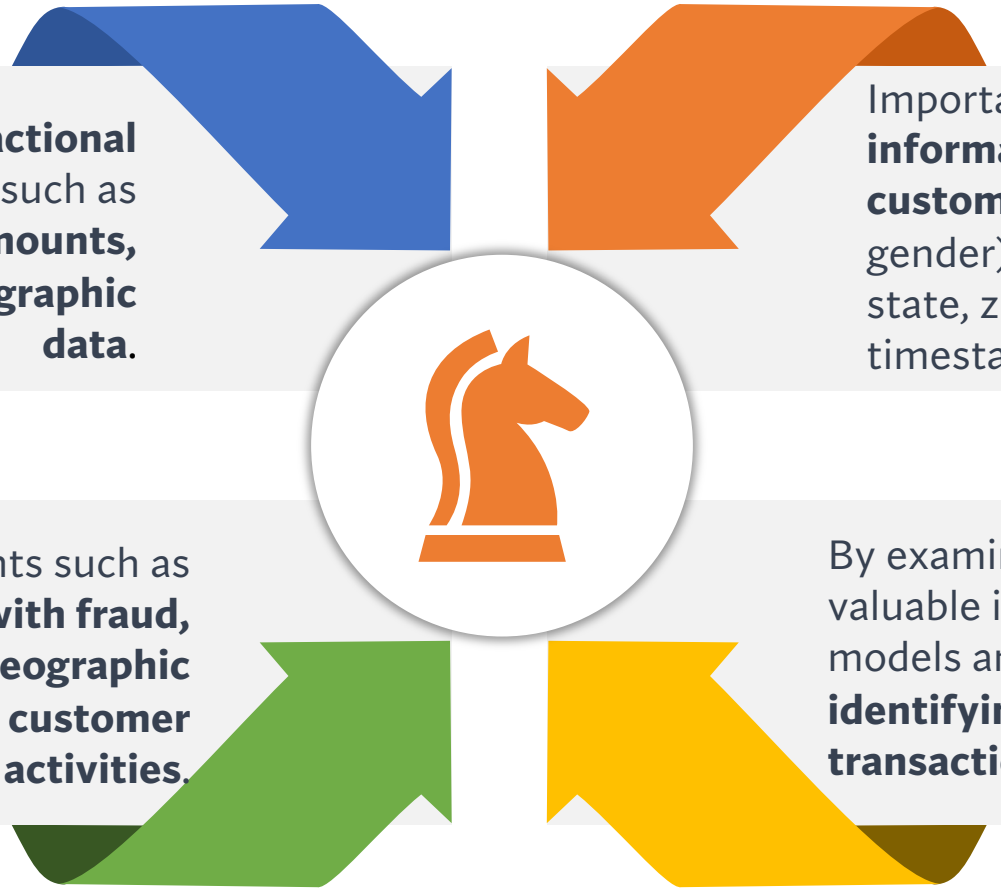
Key Outcomes

Accurate identification of fraudulent activities, proactive fraud prevention, and **enhanced fraud detection capabilities** for informed decision-making.

DATA UNDERSTANDING

The dataset consists of **transactional records** with various attributes such as **merchant details, transaction amounts, customer information, and geographic data**.

Important columns include **merchant information, transaction amounts, customer details** (e.g., name, age, gender), geographic indicators (e.g., city, state, zip), job/occupation, and timestamps.



Analysing the data reveals insights such as **merchant categories associated with fraud, unusual transaction amounts, geographic patterns**, and **correlations** between **customer characteristics and fraudulent activities**.

By examining the data and its elements, valuable insights can be derived to develop models and **techniques for effectively identifying and combating fraudulent transactions.**

DATA PREPARATION

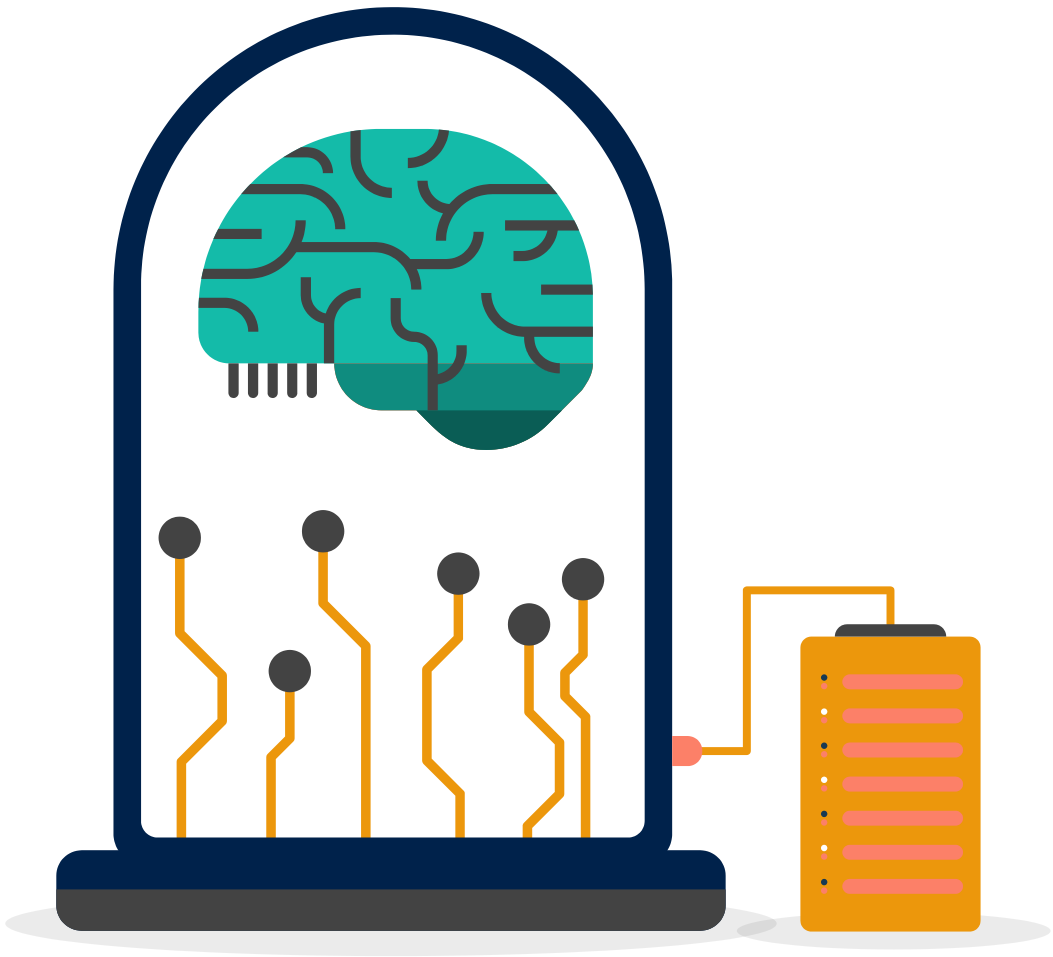
Data Preparation Steps:
Several steps were taken to prepare the data for modeling, including transforming categorical features and addressing missing data and outliers.

Transformation Techniques:
Categorical features were encoded using Ordinal Encoder to make them compatible with the chosen models.

Handling Missing Data:
Missing data was addressed through strategies such as imputation or excluding incomplete records. The specific method used should be clarified.

Outlier Handling:
Outliers in the transactional data were identified and dropped them using the z-score method. This ensures that outliers do not significantly impact model performance and prediction accuracy.

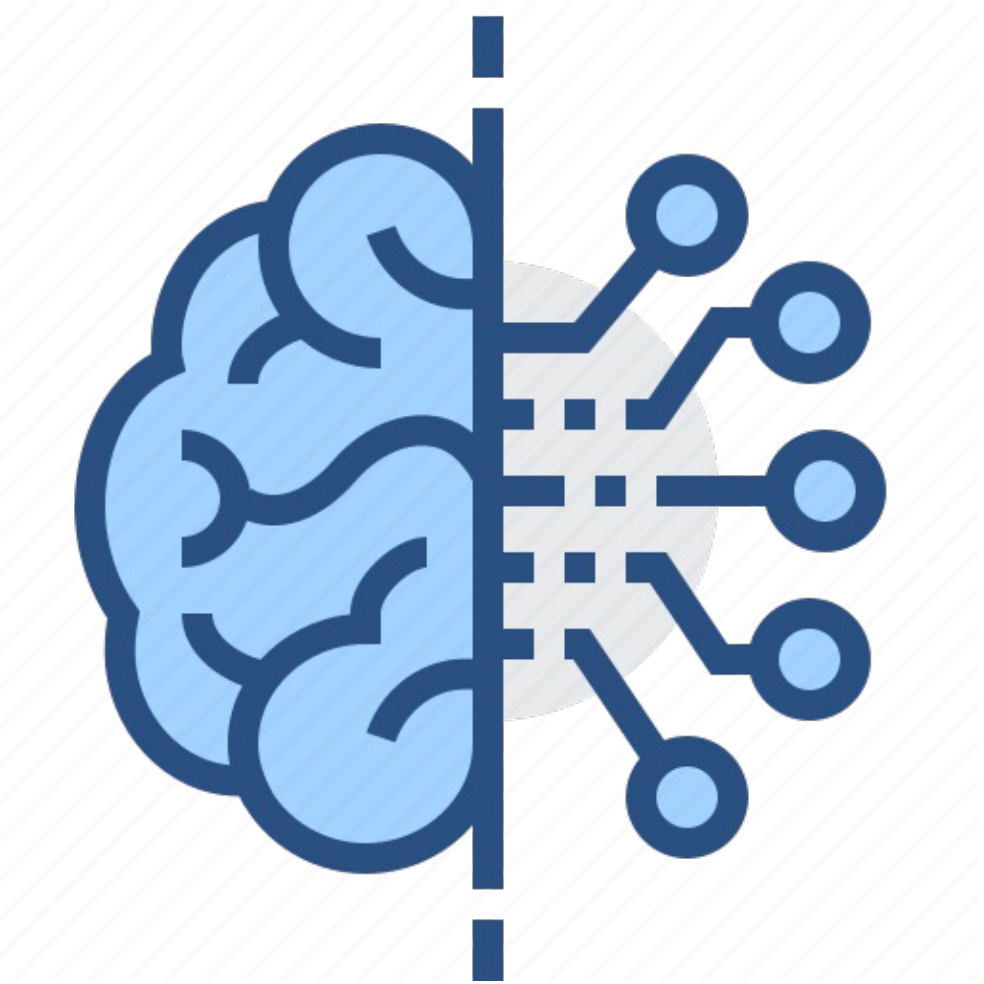
SMOTE:
The SMOTE (Synthetic Minority Over-Sampling Technique) were used to handle the imbalanced data.



MODELING

Random Forest
Random Forest is a type of machine learning algorithm that creates multiple decision trees and combines their predictions to make a more accurate and stable prediction.

XGBoost
XGBoost is a machine learning algorithm that combines multiple decision trees to make accurate predictions for classification and regression tasks, known for its speed and scalability.



EVALUATION

Model	Selected Features	Accuracy	Precision	Recall	F1-Score	Execution Time
RandomForest Classifier	15	0.97	0 = 1.00 1= 0.09	0 = 0.97 1 = 0.70	0 = 0.99 1 = 0.17	28.31s
XGBClassifier	17	1.00	0 = 1.00 1 = 0.73	0 = 1.00 1 = 0.90	0 = 1.00 1 = 0.81	194.94s

CONCLUSION

Random Forest Classifier

- Achieved an accuracy of 97%, indicating that it correctly classified 97% of the samples.
- The precision of 99.54% implies that when the model predicted a sample as fraud, it was correct 99.54% of the time.
- The recall of 70% means that the model identified 70% of the actual fraud cases.
- However, the F1-Score of 17% suggests a lower overall performance in terms of balancing precision and recall.

XG-Boost Classifier

- Achieved a perfect accuracy of 100%, indicating that it correctly classified all the samples in the dataset.
- The precision of 99.86% implies that when the model predicted a sample as fraud, it was correct 99.86% of the time.
- The recall of 90% indicates that the model identified 90% of the actual fraud cases.
- The F1-Score of 81% suggests a good overall performance in terms of balancing precision and recall, although it is lower than the accuracy due to the imbalanced nature of the dataset.

PROJECT FLOW

- In the given flow, the data was initially **divided into a 70% train and 30% test** ratio for model training and evaluation.

- Next, the train dataset was filtered to identify customers who had more **than 3 fraud transactions**. These customers were flagged as potentially suspicious and a separate dataframe called Red Flagged was created to store their information.

- Following that, the trained **model was used to make prediction**. Based on the predictions made by the model, **customers who were predicted to have more than 3 fraud transactions were identified**.

- The names of these customers were then filtered and added to the Red Flagged dataframe.

Thank you!