DATACREW

FRAUD DETECTION FOR ONLINE PAYMENT PLATFORM

Sunday, 17th September 2023

Team DataCrew

- Funsho-Akande Ifeoluwa
- Maryam Habib
- Ridwan Abdurahman

PROBLEM STATEMENT

- An online payment platform processes millions of transactions daily, making it vulnerable to various types of fraudulent activities. These activities pose a significant threat to both the business and it’s customers. To safeguard the platform and enhance user experience, we aim to leverage the power of data science and machine learning to proactively detect and prevent fraudulent transactions.

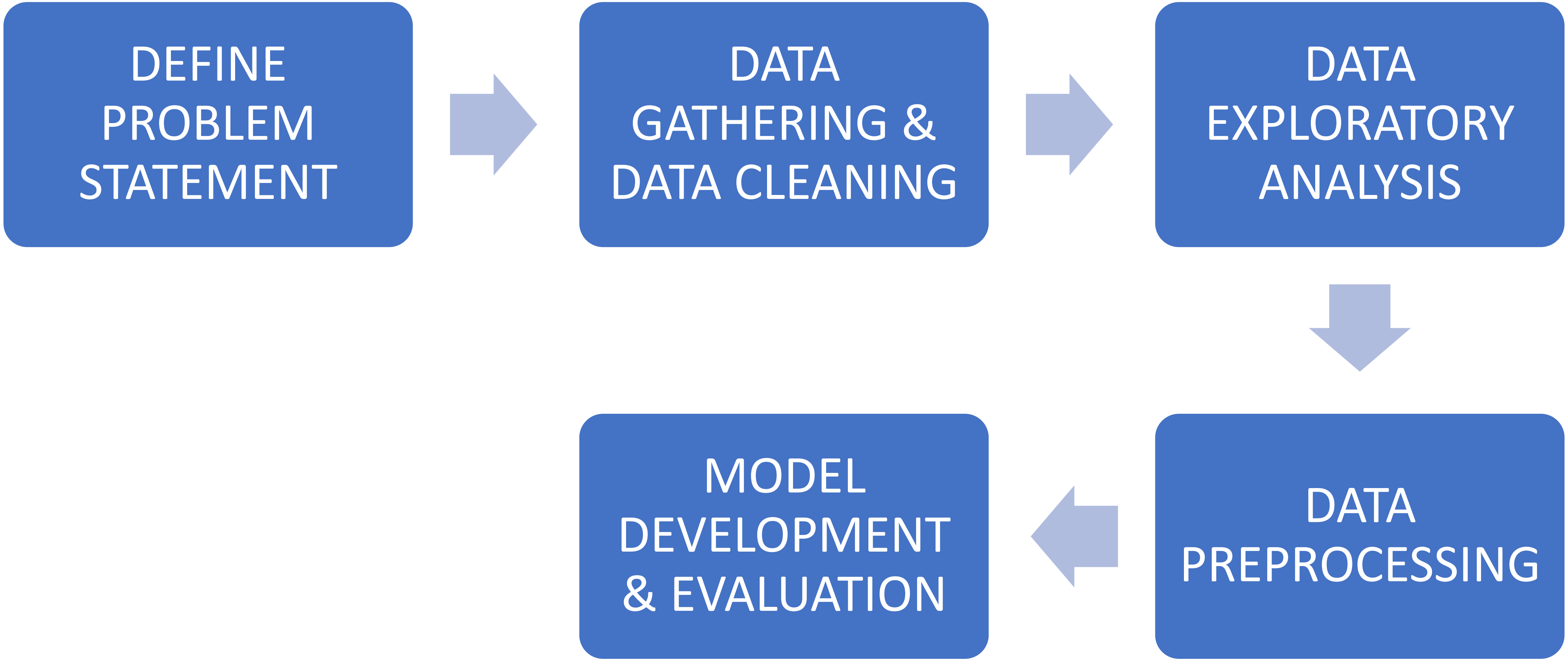
OBJECTIVES

- The primary aim is to enhance the security of the platform. By identifying fraudulent transactions early

- To build an advanced machine learning model to predict whether a given transaction is potentially fraudulent or not

- Provide accurate and efficient detection of fraudulent activities to assure users that their transactions are safe and secure.

FLOW PROCESS



DEFINE PROBLEM STATEMENT

DATA GATHERING & DATA CLEANING

DATA EXPLORATORY ANALYSIS

DATA PREPROCESSING

MODEL DEVELOPMENT & EVALUATION

THE DATASET

The Fraud Detection Dataset contains the following data

- Transaction Data: Transaction ID, User ID, Transaction Amount, Transaction Date and Time, Merchant ID, Payment Method, Country Code, and Transaction Type.

- User Data: User Age, User Gender, User Account Status, User's Transaction History, User's Credit Score, and User's Email Domain.

- Merchant Data: Merchant Category and Merchant's Reputation Score.

- Transaction Details: Transaction Status, Location Distance, Time Taken for Transaction, and Transaction Currency.

- Device Information: Device Type, IP Address, Browser Type, and Operating System.

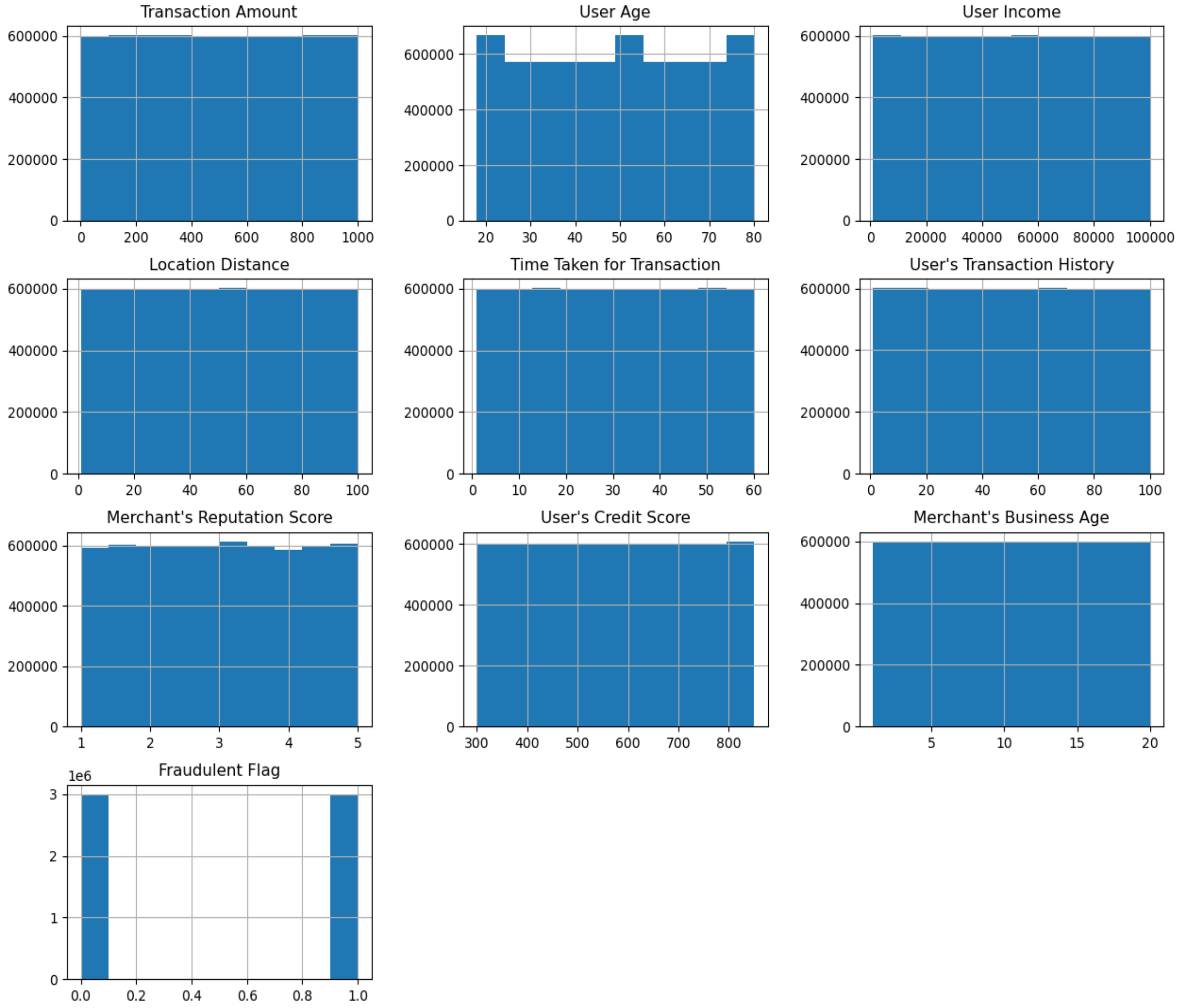
- Additional Context: Transaction Purpose and User's Device Location.

DATA CLEANING & DATA EXPLORATORY ANALYSIS

- Datatypes were changed to suit each column

- Certain columns were dropped to avoid data leakages

- Checked for class imbalance



DATA PREPROCESSING

- Data Transformation:
- Data Splitting: The dataset was divided into training and testing sets using the train-test split method. This separation ensures the evaluation of models on unseen data to gauge their generalization capabilities.
- Label Encoding: Label encoding was applied to the categorical columns, converting them into numerical representations

MODEL DEVELOPMENT

- A Decision Tree Classifier model achieved on almost all metrics a 49%
- A Gradient Boosting Classifier was used to find the most important features and best parameters to train the data on
- The model was evaluated using both the test and validation set
- Arriving at a model of that passed the 50% mark by a little on some of the evaluation metrics

```
              precision    recall  f1-score   support

           0       0.50      0.59      0.54    599406
           1       0.50      0.41      0.45    600594

    accuracy                           0.50   1200000
   macro avg       0.50      0.50      0.50   1200000
weighted avg       0.50      0.50      0.50   1200000
```

LIMITATION

- With such a large dataset, it proved difficult for the model to learn and limitations due to computational resources prevented the trial of other models

CONCLUSION

- For the purpose of predicting the if a transaction is fraudulent or not , the gradient boosting classifier produced a better result than the decision tree classifier and might show further improvement with some hyperparameter tuning

THANK YOU