MACHINE LEARNING AND
PATTERN REGONCTION

CREDIT CARD FRAUD

Machine learning model developed to identify fraudulent credit card transactions, addressing bias through under-sampling techniques.

INDEX

INTRODUCTION

Credit card fraud has emerged as a substantial and expanding concern in the digital era, impacting both consumers and financial institutions globally. With the proliferation of online transactions, the likelihood of fraudulent activities has surged, resulting in substantial financial losses and security breaches.

The primary objective of this project is to construct a comprehensive model capable of accurately identifying fraudulent transactions. By analyzing transaction patterns and identifying key characteristics, we seek to differentiate between legitimate and fraudulent transactions.

This project employs machine learning algorithms, commencing with logistic regression on a balanced dataset. Data preparation entails the utilization of under-sampling techniques, and a thorough model evaluation has been conducted to ensure the attainment of high accuracy and reliability.

Choice of Dependent and Independent Variables & Selection of Algorithms

1

Dependent Variable (Target Variable): Whether a transaction is fraudulent (1) or lawful (0) is our binary target variable. We can spot trends in fraudulent transactions when compared to authentic ones thanks to our binary classification. A clear goal variable is necessary to direct the model's training process because fraud detection is a supervised learning problem.

2

Independent Variables (Features): Important attributes:
Transaction Amount: Unusual or high sums could be a sign of fraud.
Location: Transactions coming from strange places could be a symptom of fraud.
User Behavior: Transaction sequence, frequency, and timing can all show trends; abrupt or odd behavioral shifts could raise suspicions of fraud.
Device Information: Fraud may be indicated by transactions from unfamiliar or unidentified devices.

3

Algorithm Selection: 1. Random Forest .
2. Logistic Regression.
3. Multiple Regression.

Data Preparation

Data Collection and Cleaning:

SMake that the dataset is high-quality and appropriate for model training by fixing any inconsistencies.a clear dataset that eliminates noise while preserving important fraud indications, providing a solid basis for precise model training.

Balancing the Dataset

Fraud detection data is frequently imbalanced, with legitimate transactions significantly outnumbering fraudulent ones. Addressing this imbalance is crucial for the model's performance, ensuring it can effectively identify both fraudulent and legitimate patterns. A balanced dataset, where the model has an equal opportunity to learn from both classes, reduces the risk of overfitting to the majority class.

Splitting the Dataset

Ensure rigorous model evaluation by reserving data for testing purposes. Simulate the performance of the model on unseen real-world data. Maintain a distinct separation between training and testing data to enhance the confidence in the model's generalization capabilities beyond the training data.

Tools Used

We leveraged several Python libraries to streamline and execute our data preparation steps efficiently:

1. Pandas:
2. Sci-kit learn
3. Num-py
4. Matplotlib

Using these libraries provided a structured, repeatable, and efficient process for preparing the dataset, minimizing manual intervention and reducing the chances of errors.

Feature Selection

Feature selection techniques, including correlation analysis and Random Forest feature importance, were used to identify relevant features and improve model performance. By removing redundant and low-importance features, the dataset was streamlined for optimal model performance.

1

Techniques Employed for Feature Selection

To optimize model performance, select only the most pertinent features, reduce noise, and potentially enhance computation time.

2

Selected Features

After evaluating features using the techniques outlined above, we refined a set that strikes a balance between completeness and simplicity. This curated feature set was selected to capture a comprehensive range of behaviors and transactional characteristics associated with both legitimate and fraudulent activities.

3

Rationale Behind Feature Selection

The selected features offer a comprehensive perspective of each transaction, encompassing both behavior-related and transaction-specific attributes. This comprehensive view enhances the model’s capacity to differentiate between normal and anomalous patterns. The refined feature set is anticipated to facilitate the model’s efficient learning, augment its predictive capabilities, and enhance its generalization performance on novel data.

Model Development and Evaluation

In our project, we implemented three distinct models: Random Forest, Logistic Regression, and XGBoost. Each model possesses unique strengths and considerations, and we conducted a rigorous evaluation of their performance to ascertain the most suitable model for our credit card fraud detection application.

1

Model 1: Random Forest

The Random Forest algorithm is an ensemble learning method that generates multiple decision trees during training and outputs the mode of their predictions. The Random Forest model exhibited robust performance across the evaluation metrics, notably excelling in recall, which is paramount in minimizing undetected fraud.

2

Model 2: Logistic Regression

To address class imbalance, we employed under-sampling techniques to ensure that our training dataset was more representative of both fraudulent and non-fraudulent cases. The logistic regression model demonstrated satisfactory performance, with interpretable coefficients that facilitate the understanding of the influence of various features on fraud prediction.

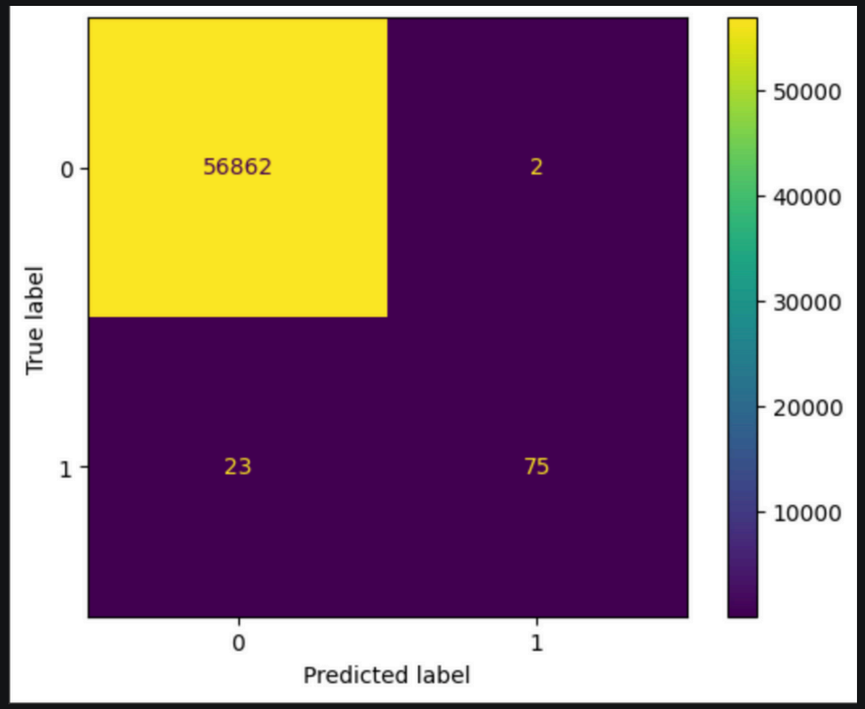
3

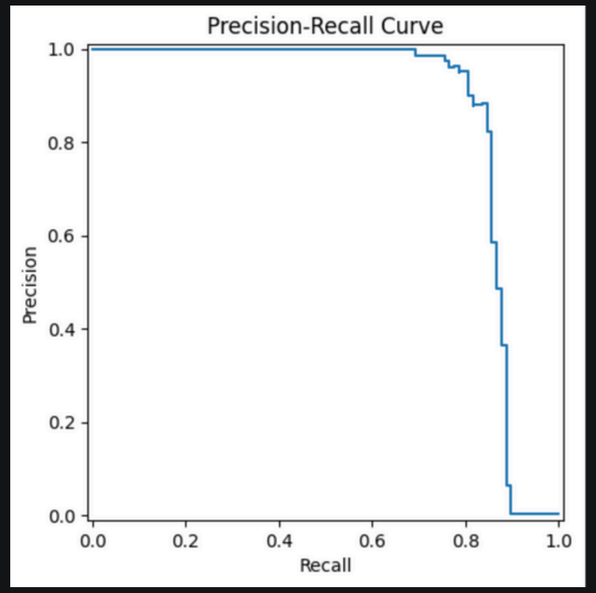
Model 3: XGBoost

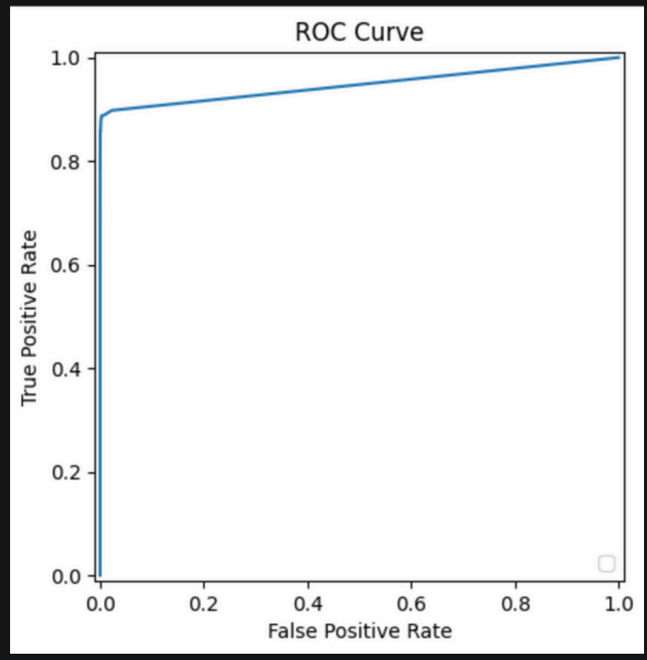
XGBoost (Extreme Gradient Boosting) is a highly efficient implementation of gradient boosting that has gained popularity for its performance and speed in machine learning competitions. XGBoost demonstrated superior accuracy and F1-score compared to Random Forest and Logistic Regression, particularly after parameter tuning. Consequently, it emerges as a strong contender for our final model selection.

Random Forest:

```
Accuracy: 0.9995611109160493
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.97      0.77      0.86        98

    accuracy                           1.00     56962
   macro avg       0.99      0.88      0.93     56962
weighted avg       1.00      1.00      1.00     56962

Confusion Matrix:
 [[56862     2]
 [   23    75]]
```



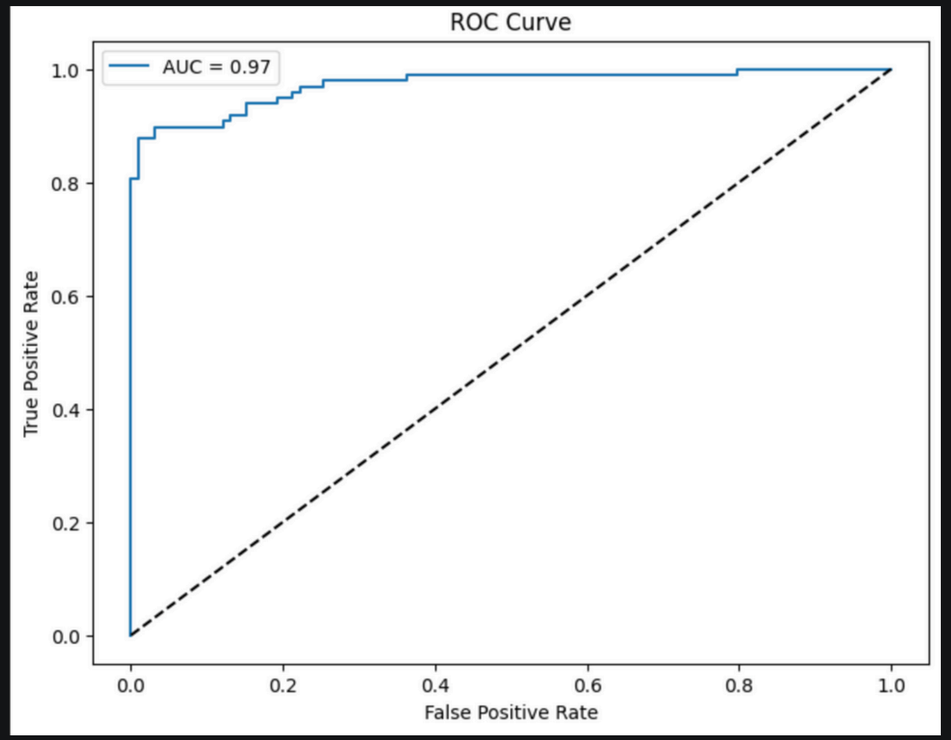


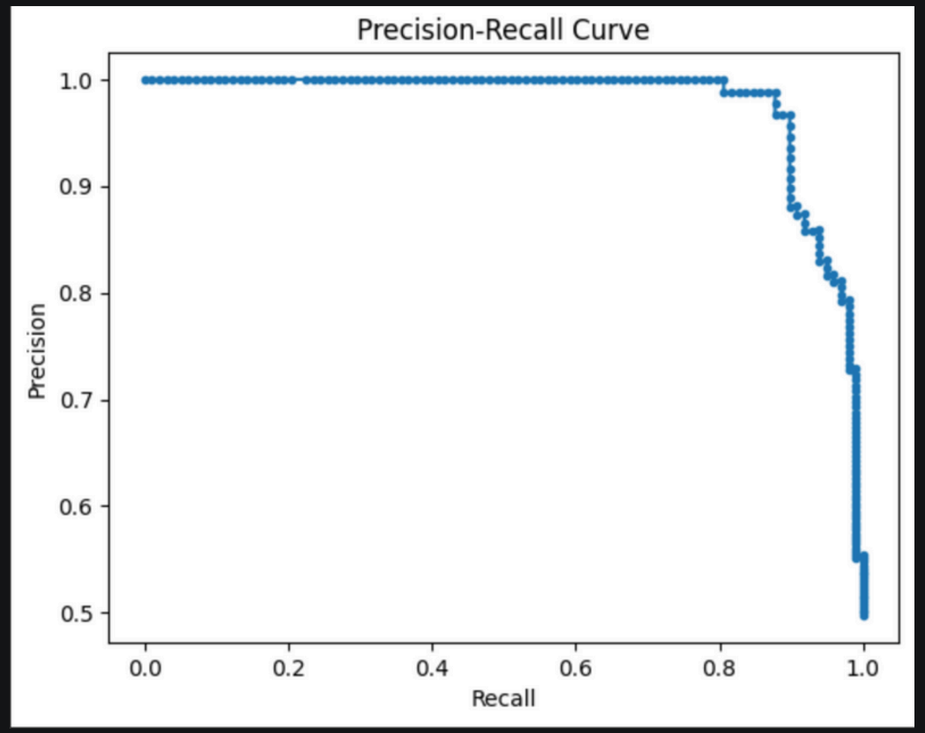


Logistic Regression:

```
Accuracy on Training data :  0.9542566709021602
```

```
Accuracy score on Test Data :  0.9289340101522843
```

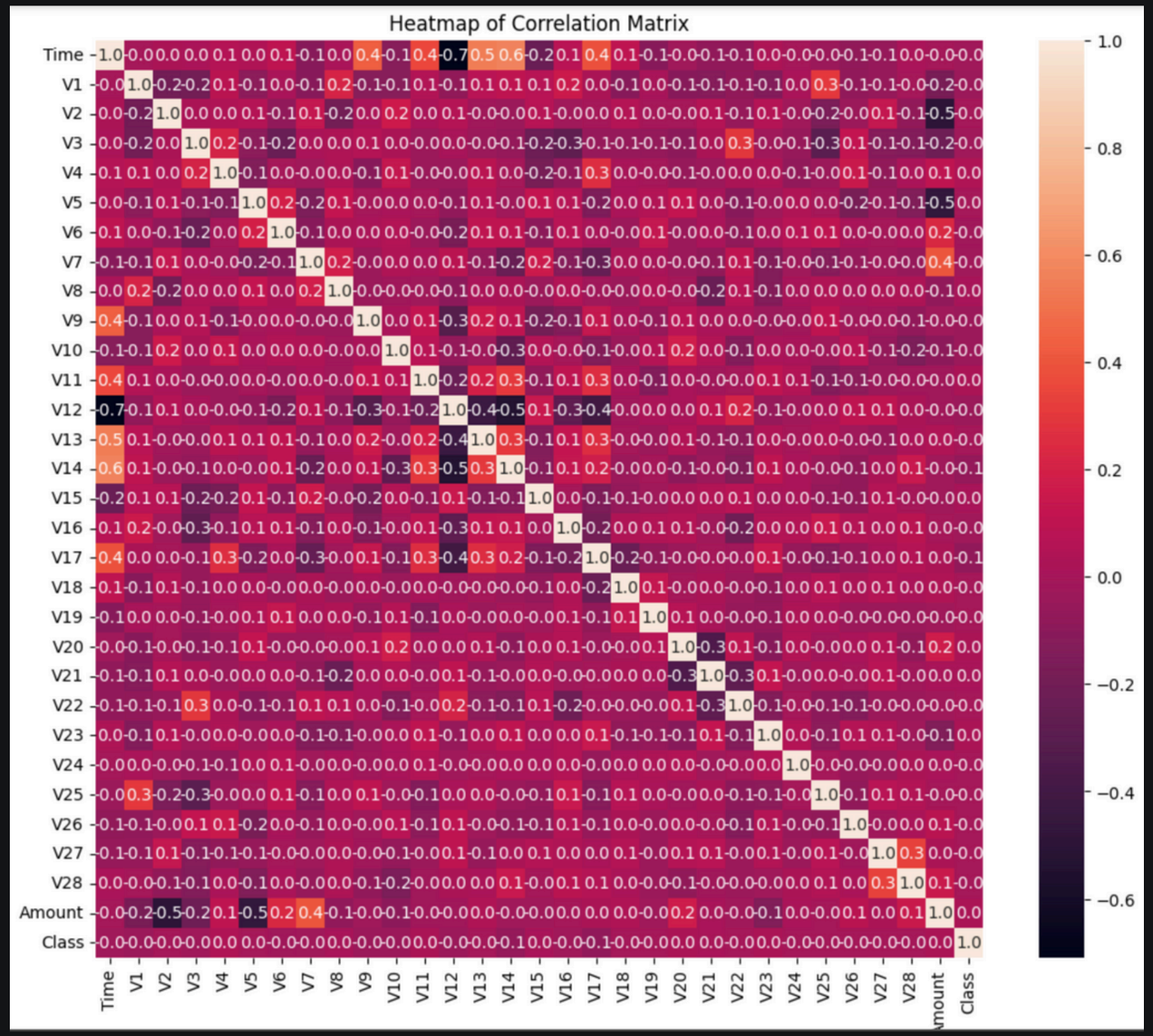




Mutliple Regression:

```
Mean Absolute Error: 24.382480739537627
Mean Squared Error: 4332.337602906516
R-squared: 0.9180139759725466
Root Mean Squared Error (RMSE): 65.82049531040097
```

```
const     0.000000e+00
Time      2.774643e-06
V1        0.000000e+00
V2        0.000000e+00
V3        0.000000e+00
V4        0.000000e+00
V5        0.000000e+00
V6        0.000000e+00
V7        0.000000e+00
V8        0.000000e+00
V9        0.000000e+00
V10       0.000000e+00
V11       1.063632e-03
V12       6.421959e-35
V13       3.929611e-21
V14       0.000000e+00
V15       1.005303e-10
V16       1.647603e-05
V17       1.407728e-56
V18       0.000000e+00
V19       0.000000e+00
V20       0.000000e+00
V21       0.000000e+00
V22       0.000000e+00
V23       0.000000e+00
V24       1.262358e-21
V25       0.000000e+00
V26       3.449328e-10
V27       0.000000e+00
V28       2.563712e-79
Class     1.573316e-21
dtype: float64
```

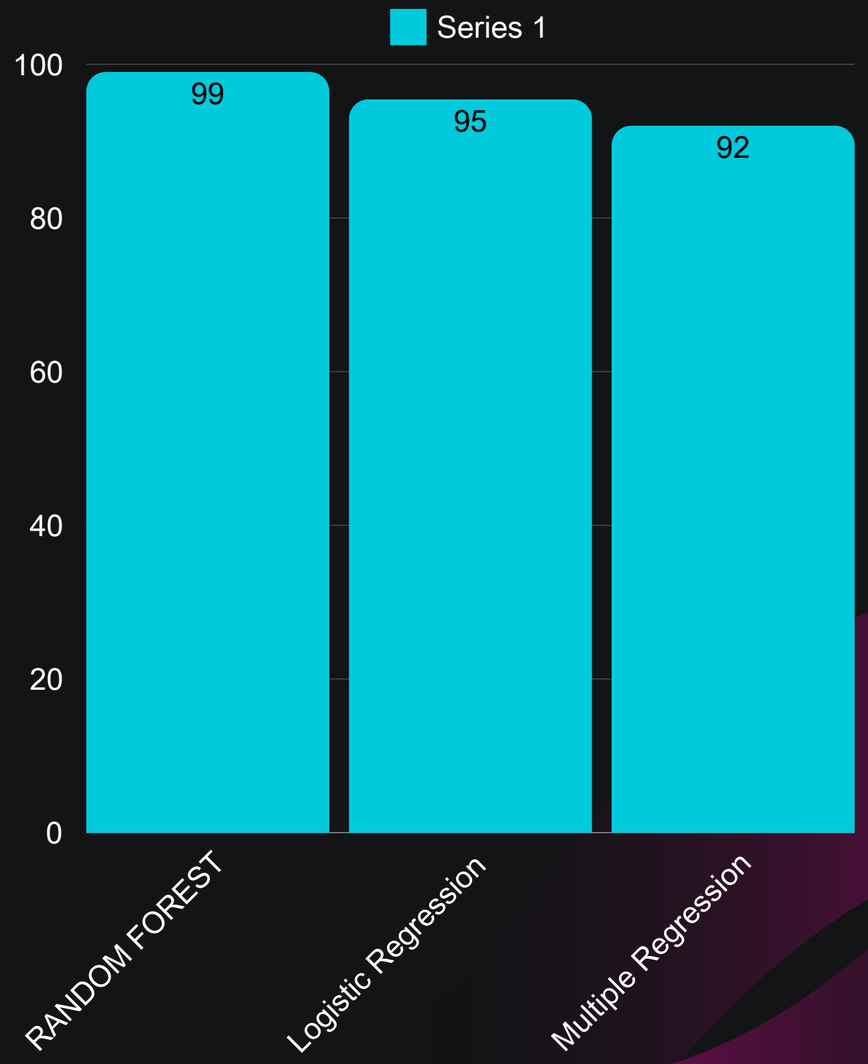


Optimization

In our pursuit of the most effective fraud detection model, we employed various optimization techniques to enhance model performance. This phase was critical for improving accuracy and ensuring the robustness of our selected models.

Hyperparameter tuning involves adjusting pre-training parameters to enhance model performance. GridSearchCV was used to systematically explore parameter combinations for Random Forest and Logistic models.

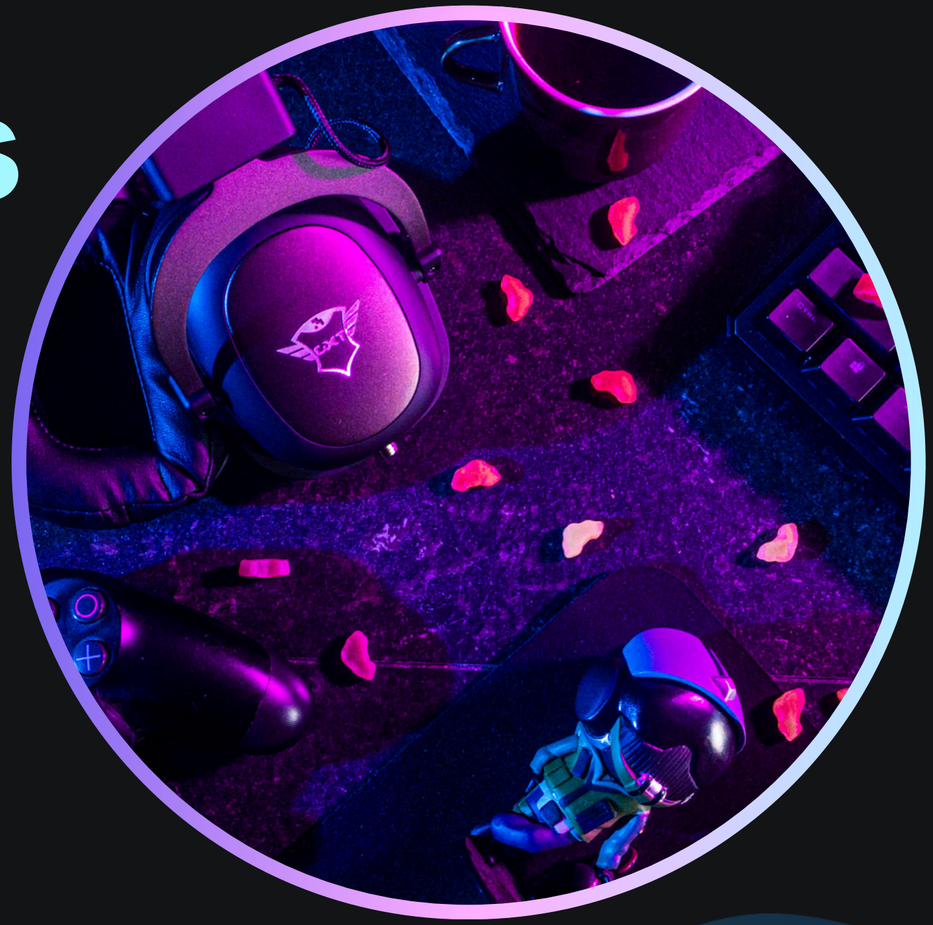
Moreover, with regards to the Multiple Linear Regression, this was done to detect to what extent other variables impacts the Amount and the MSE and R2 were the parameters used to ascertain the how suitable the model is and the quality of the prediction itself.



Model Comparison

Random Forest is the most effective model for fraud detection due to its high accuracy and balanced precision-recall scores. Logistic Regression is simpler but may not handle complex patterns as well, while Logistic & multiple Regression could outperform other models with additional tuning but may require higher computational costs.

Recommendations



Ensemble modeling, class imbalance mitigation, hyperparameter optimization, and the incorporation of supplementary features can enhance fraud detection capabilities. Furthermore, the implementation of real-time detection frameworks and the continuous evaluation of models are crucial for ensuring accuracy and adapting to evolving fraud patterns.

- Real-Time Monitoring and Deployment
- Stream Processing: Implement real-time fraud detection by integrating machine learning models with stream-processing frameworks like Apache Kafka or Spark.
- Automated Alerts: Develop an alerting system to automatically notify users or financial institutions when suspicious transactions are detected.
- Data Enhancement Techniques

Thank You

team member -
Abhishek Shrivas -20035356
Fatih Husnu Basbagi - 20046748
Anjali Raj - 20042704
Kingsley Ogidi - 20038983
Rohith Puthanveedu - 20041033