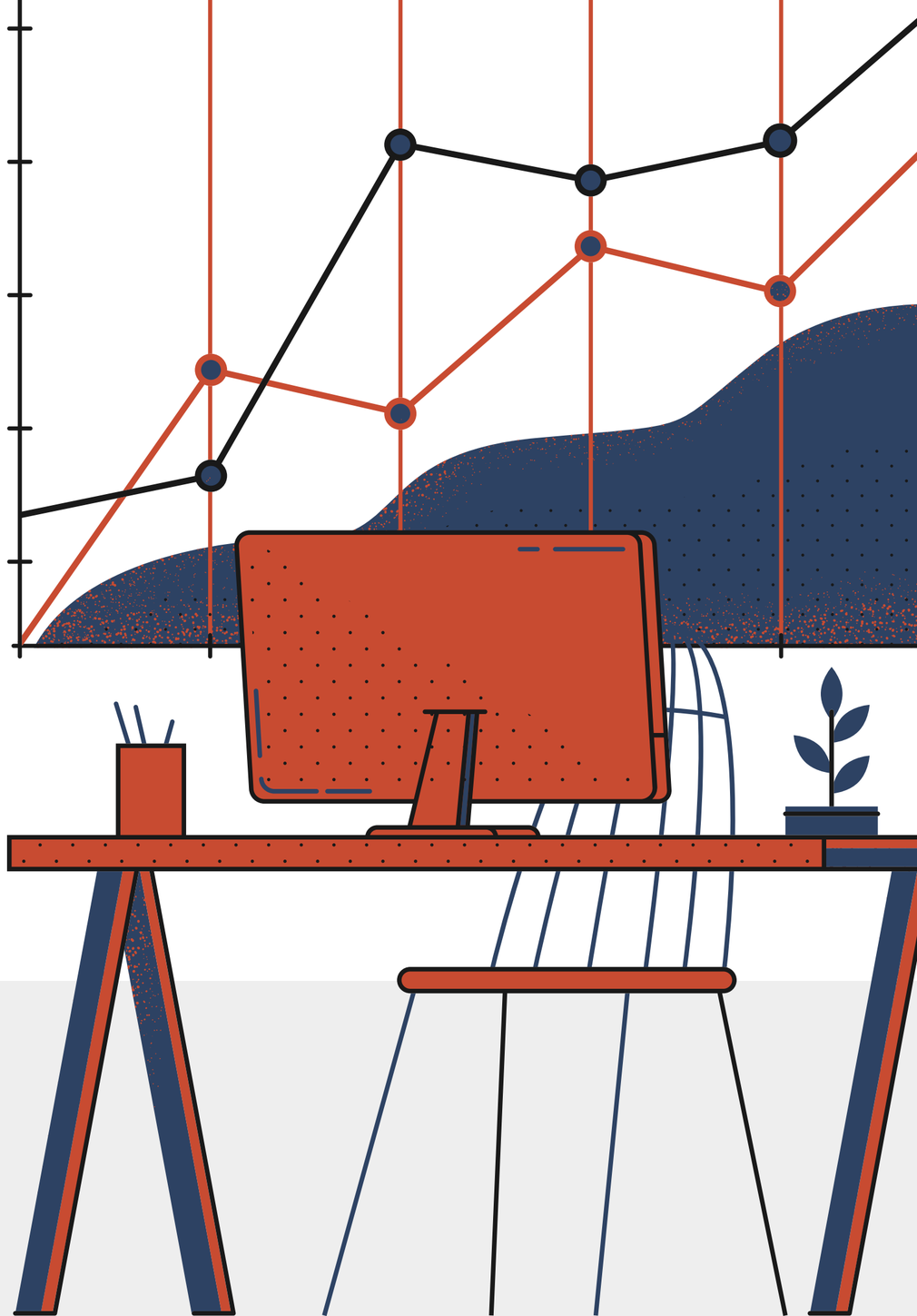
American Express Makeathon 2023

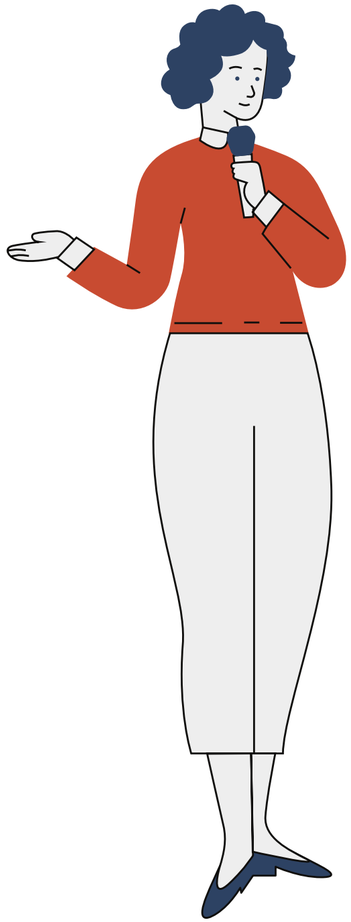
ML-based fraud detection



SUBMITTED BY

Purnima Kumar

Table of Contents



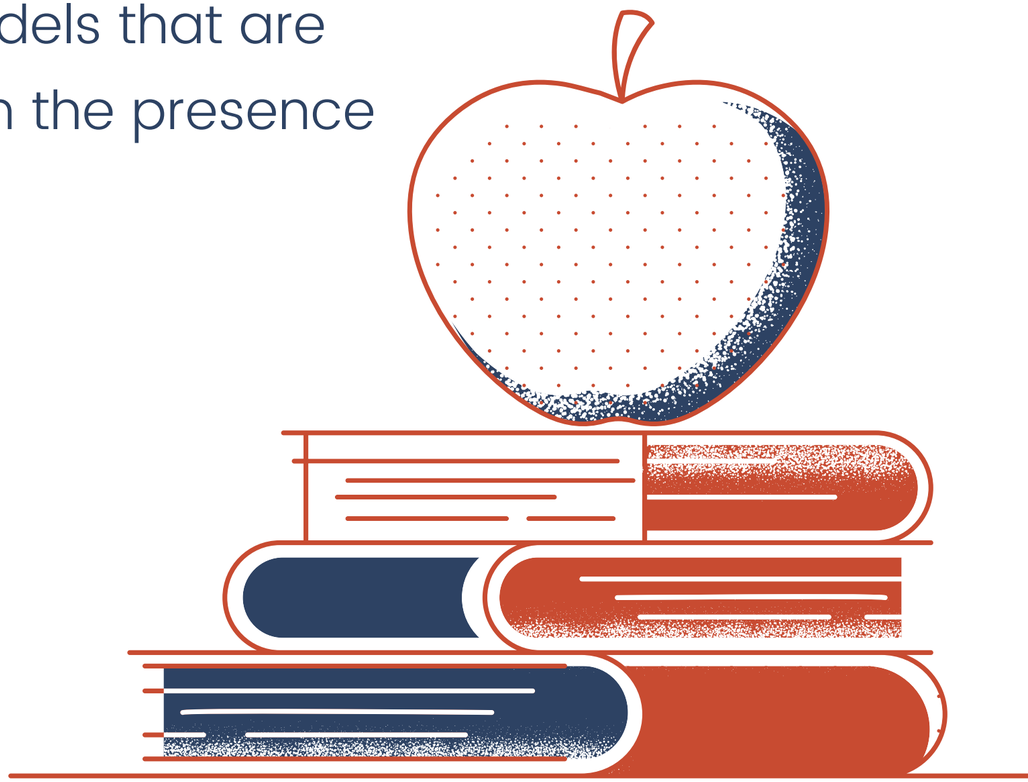
Purnima Kumar

Data Science| Machine Learning

- I am pursuing bachelor's degree in Computer Science and Engineering with specialization in Data Science from the Noida Institute of Engineering and Technology, Greater Noida.
- My main area of focus lies in Machine Learning and Data Science..
- My core skills include statistical analysis, machine learning, data visualization.
- I have extensive experience with Python Programming Language. I am also proficient with SQL.

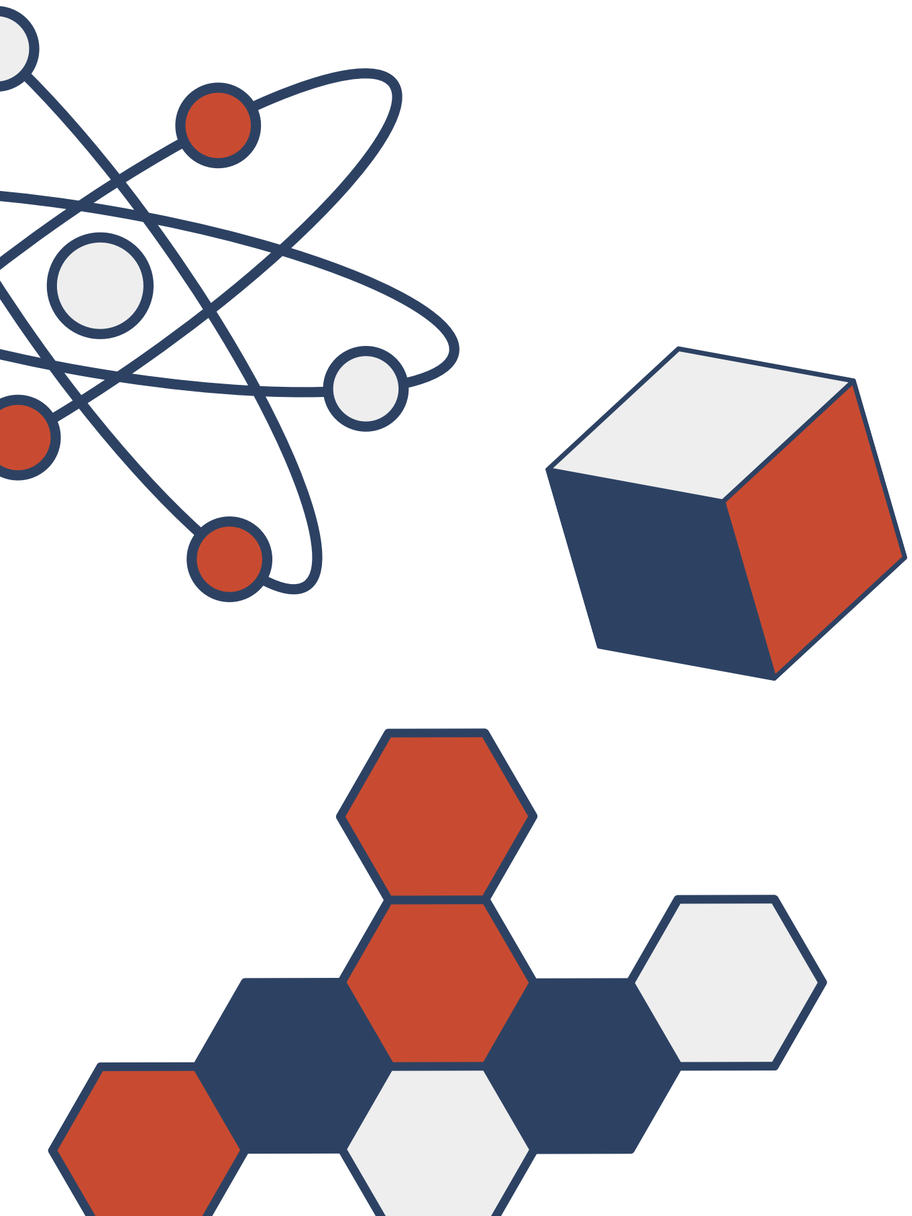
ML-based fraud detection

Develop and maintain ML-based fraud detection models that are effective at identifying evolving fraud patterns even in the presence of imbalanced data.



Task

- Data collection and preparation
- Model training
- Use imbalance data to train the model
- Model evaluation



Credit Card Fraud Detection

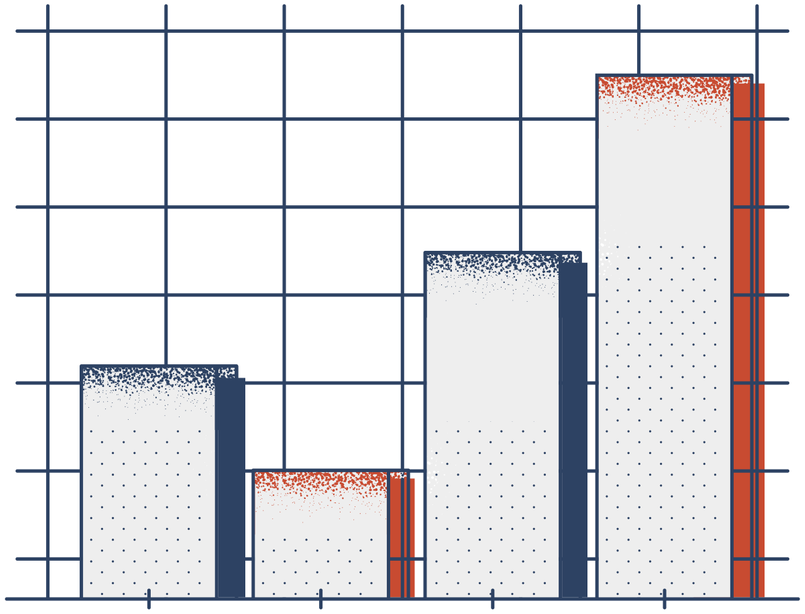
- The dataset contains transactions made by credit cards in random two days of September 2013 by European cardholders.
- We have 492 frauds out of 284,807 transactions.
- The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numerical input variables which are the result of a PCA transformation, the only features which have not been transformed with PCA are 'Time'(sec) and 'Amount'.
- Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Python Libraries Used

- Numpy - to perform a wide variety of mathematical operations on arrays and matrices.
- Pandas - functions for analyzing, cleaning, exploring, and manipulating data
- Sci-kit Learn - an open-source data analysis, and Machine Learning (ML) LIbrary
- XGBoost - scalable, distributed gradient-boosted decision tree (GBDT) machine learning library
- Seaborn - a library for making statistical graphics
- Imblearn - Toolbox for imbalanced dataset

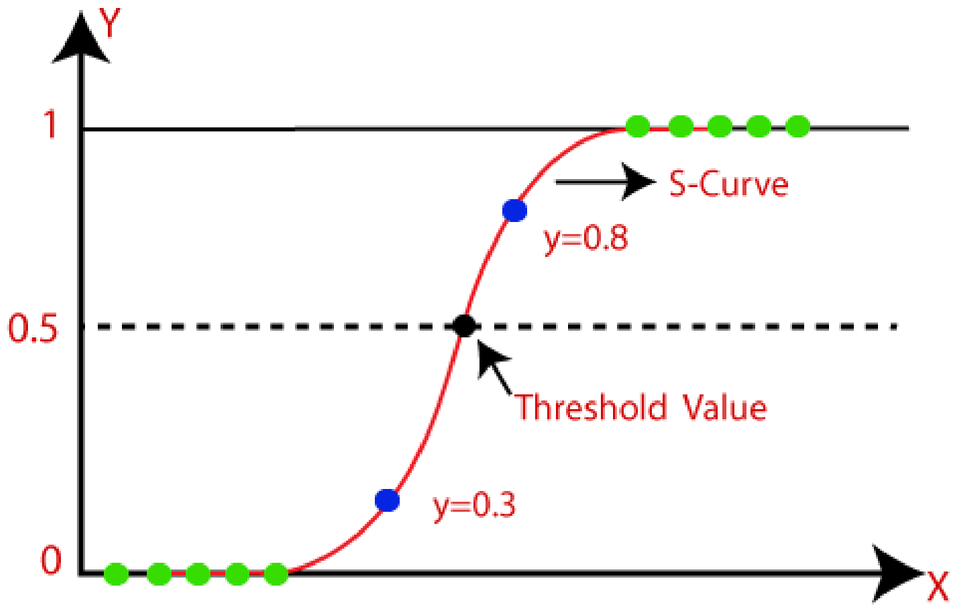
Under Sampling

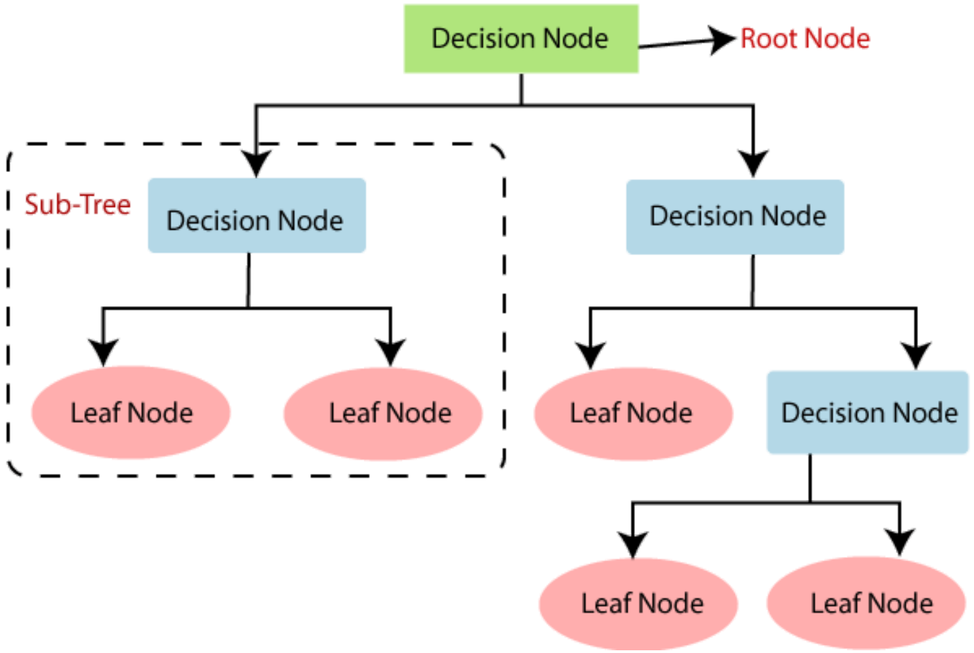
- The Dataset is highly unstable.
- After deleting duplicate values shape of data is 283726 x 31.
- Number of Fraud happened accounts of 473 times.
- With help of under sampling, we choose random 473 values out of 283253 legit data rows.
- Then, split the data with test data to be 20%, stratify as Y and random state as 42.



Logistic Regression

- Statistical method for predicting binary classes
- Estimation is done through maximum likelihood
- Provides a constant output
- Types of Logistic Regression: Binary Logistic Regression, Multinomial Logistic Regression and Ordinal Logistic Regression.
- In order to map predicted values to probabilities, we use the Sigmoid function.



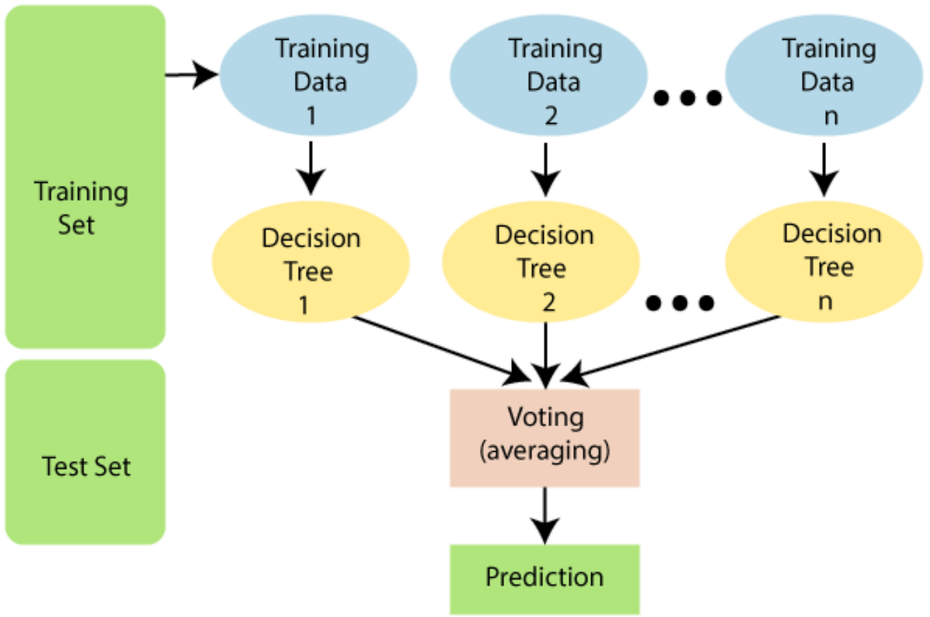


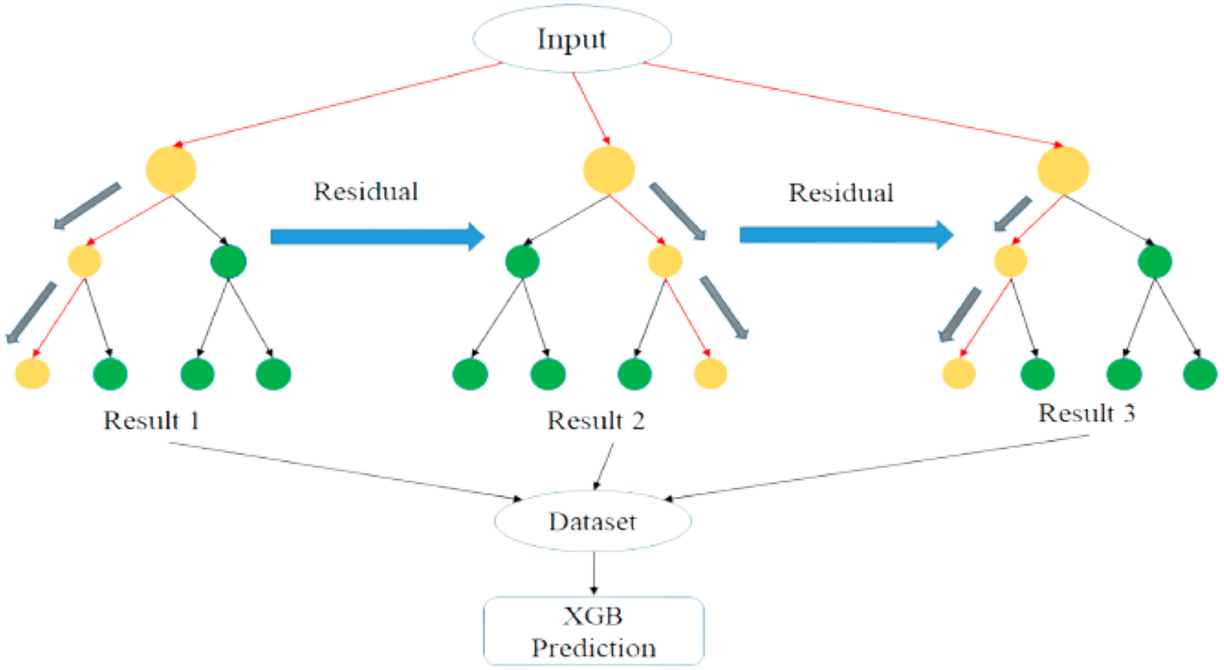
Decision Tree

- Flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.
- The topmost node is known as the root node.
- It learns to partition on the basis of the attribute value.
- It partitions the tree in recursively manner call recursive partitioning.
- Easy to understand and interpret.

Random Forest

- Contains number of decision trees on various subsets
- Based on the concept of ensemble learning
- Combining multiple classifiers to solve a complex problem
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- Capable of performing both Classification and Regression tasks.





XGBoost

- Extreme Gradient Boosting
- Fast than other gathering classifiers.
- Center calculation is parallelizable
- Reliably outflanks other technique calculations
- Wide assortment of tuning boundaries

Comparing Accuracy

Accuracy on Training data : 91.80%

Accuracy score of Test Data for Logistic Regression: 94.73%

Accuracy score of Test Data for Decision Tree: 91.57%

Accuracy score of Test Data for Random Forest: 94.73%

Accuracy score of Test Data for XGBoost: 95.26%

Conclusion

XGBoost have highest accuracy on testing data, i.e., 95.26%. Based on the performance of this model with the test data, we can conclude that it is effective in accurately predicting the outcome of credit card transactions and detecting fraudulent activity.

The End

Thank you