CREDIT CARD FRAUD DETECTION

Student - Parth Bramhecha(33115)

Guide-Dr. Anant M. Bagade

Reviewer-Mrs. Prajakta S. Shinde

INTRODUCTION

AS MORE PEOPLE USE ONLINE SHOPPING AND DIGITAL BANKING, THE RISK OF FRAUD HAS INCREASED, MAKING STRONG DETECTION SYSTEMS VERY IMPORTANT. BANKS CAN LOSE A LOT OF MONEY, AND CUSTOMERS MAY LOSE TRUST IN USING DIGITAL PAYMENTS. CREDIT CARD FRAUD DETECTION IS CRUCIAL AS FRAUD TECHNIQUES BECOME MORE SOPHISTICATED WITH RISING CARD USAGE.

WHAT ARE TYPES OF CREDIT CARD FRAUDS ?

- **CARD-NOT-PRESENT (CNP) FRAUD:**
  UNAUTHORIZED TRANSACTIONS MADE WITHOUT THE PHYSICAL CARD, OFTEN ONLINE.
- **CARD-PRESENT FRAUD:**
  FRAUDULENT TRANSACTIONS USING A STOLEN OR CLONED PHYSICAL CARD AT A POS TERMINAL.
- **ACCOUNT TAKEOVER:**
  UNAUTHORIZED ACCESS TO A CARDHOLDER'S ACCOUNT TO MAKE CHANGES OR REQUEST A NEW CARD.
- **CARD SKIMMING:**
  CAPTURING CARD DETAILS VIA A SKIMMING DEVICE TO CREATE COUNTERFEIT CARDS.
- **PHISHING:**
  DECEPTIVE COMMUNICATION (EMAIL, PHONE) TO TRICK USERS INTO GIVING UP THEIR CARD INFORMATION.

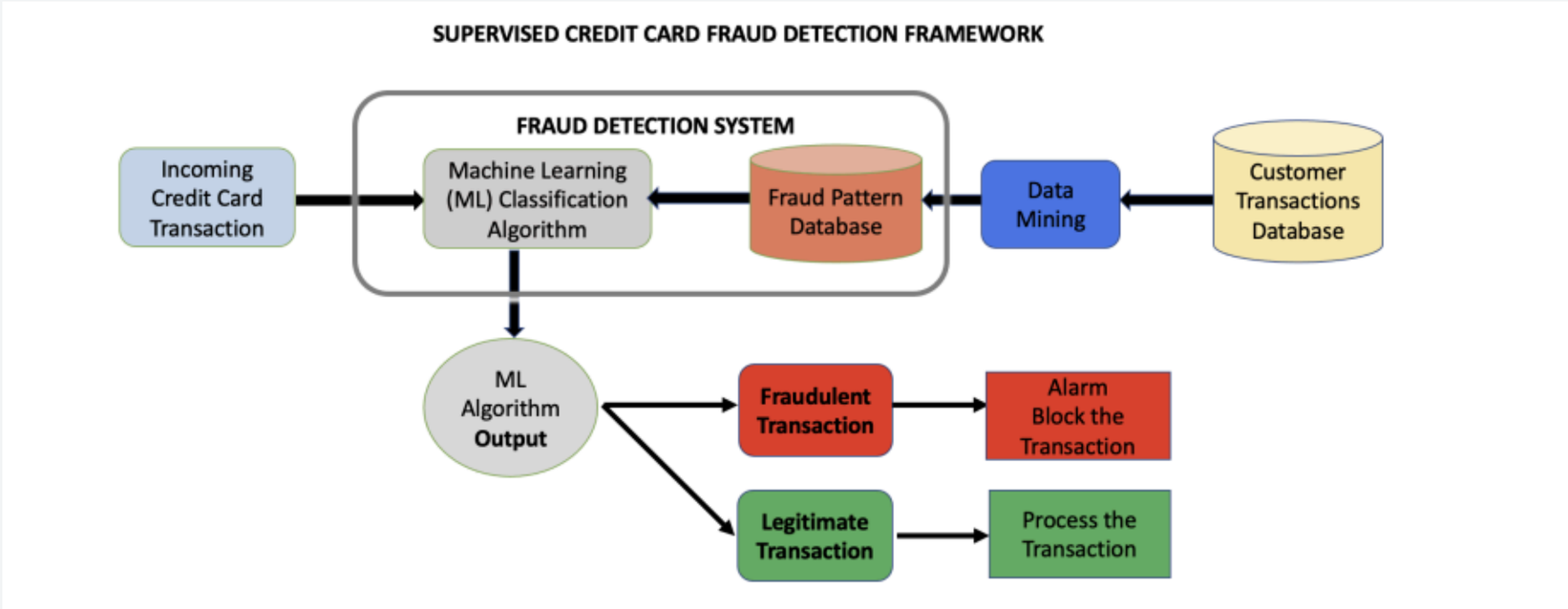
LITERATURE SURVEY

Title	Author	Publication Date	Aim/Objective
Credit Card Fraud Detection Using Machine Learning	Deep Prajapati, Ankit Tripathi, Jeel Mehta, Kirtan Jhaveri, Vishakha Kelkar	04 December 2021	The aim is to evaluate and compare machine learning algorithms like Random Forest, XGBoost, and ANN for their effectiveness in accurately detecting and preventing fraudulent credit card transactions.
Credit Card Fraud Detection using Machine Learning	Naga Ashwini Nayak V J , C. Suchika , N Sandhya , M Lakshmi , Roja J	7, April 2023	The aim is to identify credit card fraud by employing machine learning algorithms like Decision Tree, Random Forest, and Extreme Gradient Boosting, and to evaluate their effectiveness using both public and real-world financial data. The goal is to assess the robustness and accuracy of these models in detecting fraudulent activities.
Credit Card Fraud Detection System based on Operational & Transaction features using SVM and Random Forest Classifiers	C. Sudha; D. Akila	25 February 2021	The aim is to develop a Credit Card Fraud Detection system utilizing Operational and Transaction features, applying Random Forest and Support Vector Machine classifiers. The system aims to classify user features into benign or suspected categories and evaluate its performance using precision, accuracy, recall, and F1-score metrics, demonstrating high detection rates and accuracy for both classifiers.

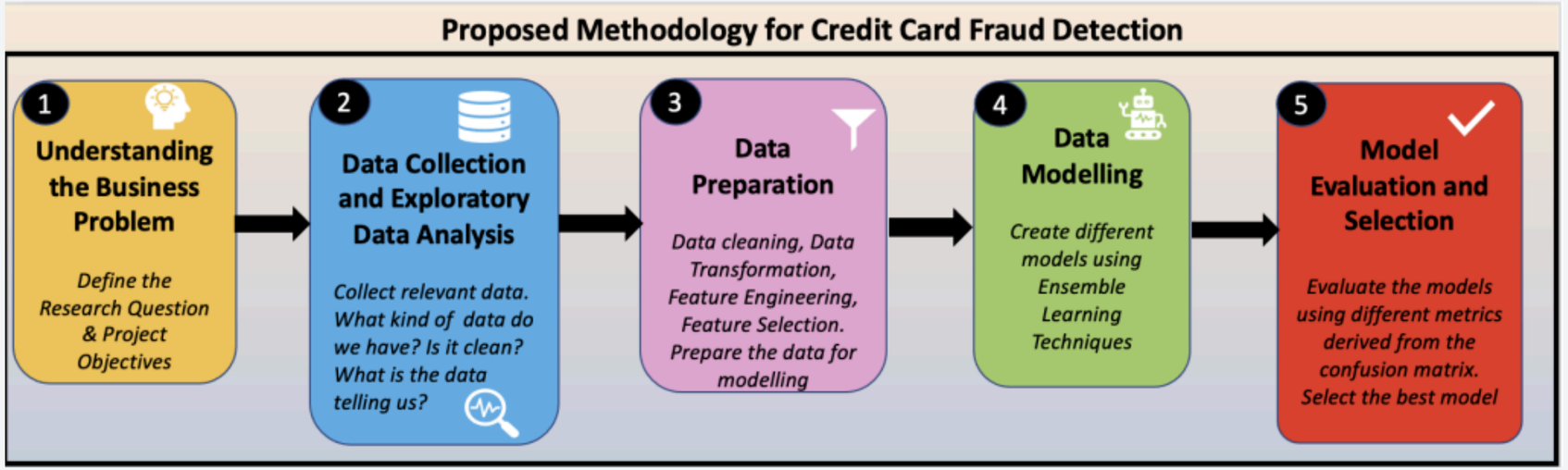
LITERATURE SURVEY

Title	Author	Publication Date	Aim/Objective
Credit Card Fraud Detection Using Machine Learning	Sanjay Bharadwaj	21 March 2024	The aim is to compare the effectiveness of two low-cost machine learning techniques, Random Forest and K-Nearest Neighbor (K-NN), in predicting credit card fraud using a highly unbalanced dataset. The research evaluates the models based on key machine learning metrics, ease of implementation, and cost-effectiveness to draw conclusions applicable to real-world scenarios.
Credit Card Fraud Detection Using Machine Learning	Neethu Tressa, V Asha, Govindaraj M, Sangamesh Padanoor, Rahila Tabassum, Desai Vatsal Dharmesh, Binju Saju	10 October 2023	The aim is to develop a system that uses machine learning techniques to detect fraudulent credit card transactions, ensuring customers are not charged for unauthorized purchases. The goal is to model historical transaction data, including fraudulent cases, and use this pattern to identify and prevent future fraudulent transactions.

PROPOSED ARCHITECTURE



PROPOSED METHODOLOGY



DATASET STUDY


```
Class
0    284315
1       492
Name: count, dtype: int64
```

- Highly unbalanced dataset
- Shape of dataset:(284807,31)
- No missing values
- Have implemented PCA no need to implement outliers treatment
- There is no clear pattern forbased on the "Time" feature.
- Fraudulent transactions are concentrated in the lower amount range, while non-fraudulent transactions are spread across both low and high amounts.
- Only the Amount column was scaled, as the other features are PCA transformed
- Oversampling the minority class (e.g., SMOTE - Synthetic Minority Over-sampling Technique) needs to be applied

METHODLOGY

Data Preprocessing

- Cleanse data: Handle missing values, outliers, inconsistencies.
- Create features: Extract meaningful patterns (e.g., time, amount).
- Normalize data: Scale features for consistent range.

Class Imbalance

- Unbalanced data: Fewer fraudulent cases.
- Balancing techniques:
  - Over-sampling: Increase rare cases.
  - Under-sampling: Decrease common cases.
  - Synthetic data: Generate new minority class examples.

Machine Learning Algorithms

- Logistic Regression: Simple, interpretable, good baseline.
- Decision Tree: Understandable rules, prone to overfitting.
- Random Forest: Multiple trees for better accuracy, reduced overfitting.
- XGBoost: Powerful, often top performer, requires tuning.

Model Evaluation Metrics

- Beyond accuracy: Consider imbalanced data.
- Precision vs. recall: Focus on catching fraud (recall).
- Confusion matrix: Detailed performance breakdown.
- ROC-AUC curve: Overall model performance.

Cross-Validation

- Evaluate model performance: Split data into folds, train on most, test on one.
- Robust evaluation: Average performance across folds for reliable results.

Feature Importance

- Feature importance: Identify key factors driving predictions.
- SHAP values: Explain individual predictions based on features.

IMPLEMENTATION

Data Preprocessing

- Cleanse data: Handle missing values, outliers, inconsistencies.
- Create features: Extract meaningful patterns (e.g., time, amount).
- Normalize data: Scale features for consistent range.

Class Imbalance

- Unbalanced data: Fewer fraudulent cases.
- Balancing techniques:
  - Over-sampling: Increase rare cases.
  - Under-sampling: Decrease common cases.
  - Synthetic data: Generate new minority class examples.

Machine Learning Algorithms

- Logistic Regression: Simple, interpretable, good baseline.
- Decision Tree: Understandable rules, prone to overfitting.
- Random Forest: Multiple trees for better accuracy, reduced overfitting.
- XGBoost: Powerful, often top performer, requires tuning.

IMPLEMENTATION

Data Preprocessing

- Cleanse data: Handle missing values, outliers, inconsistencies.
- Create features: Extract meaningful patterns (e.g., time, amount).
- Normalize data: Scale features for consistent range.

Class Imbalance

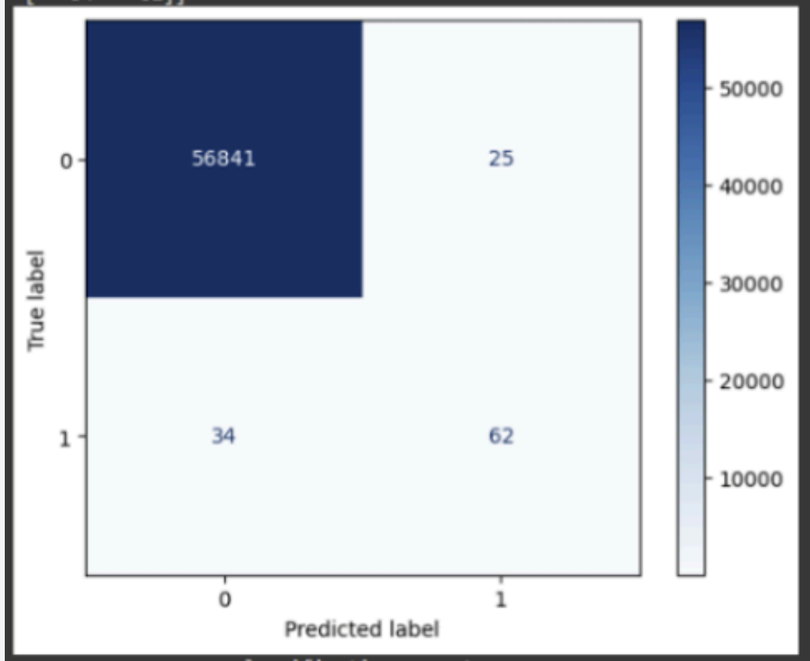
- Unbalanced data: Fewer fraudulent cases.
- Balancing techniques:
  - Over-sampling: Increase rare cases.
  - Under-sampling: Decrease common cases.
  - Synthetic data: Generate new minority class examples.

Machine Learning Algorithms

- Logistic Regression: Simple, interpretable, good baseline.
- Decision Tree: Understandable rules, prone to overfitting.
- Random Forest: Multiple trees for better accuracy, reduced overfitting.
- XGBoost: Powerful, often top performer, requires tuning.

RESULTS

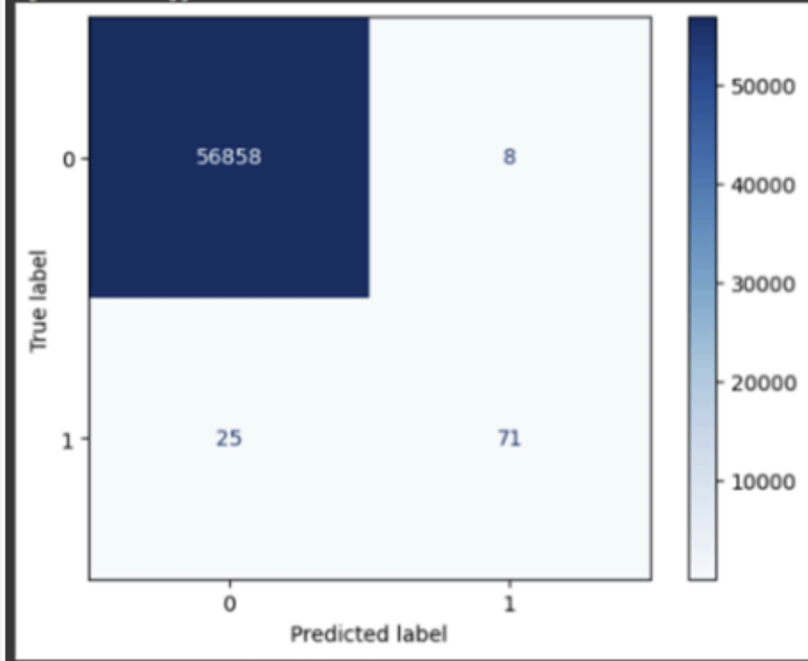
```
------------------ Confusion Matrix --------------------
[[56841    25]
 [   34    62]]
```




```
------------------ classification_report --------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56866
           1       0.71      0.65      0.68        96

    accuracy                           1.00     56962
   macro avg       0.86      0.82      0.84     56962
weighted avg       1.00      1.00      1.00     56962

------------------ More Specific classification_report --------------------
Accuracy:- 0.9989642217618764
Sensitivity:- 0.6458333333333334
Specificity:- 0.9995603699926142
F1-Score:- 0.6775956284153006
```

RANDOM FOREST

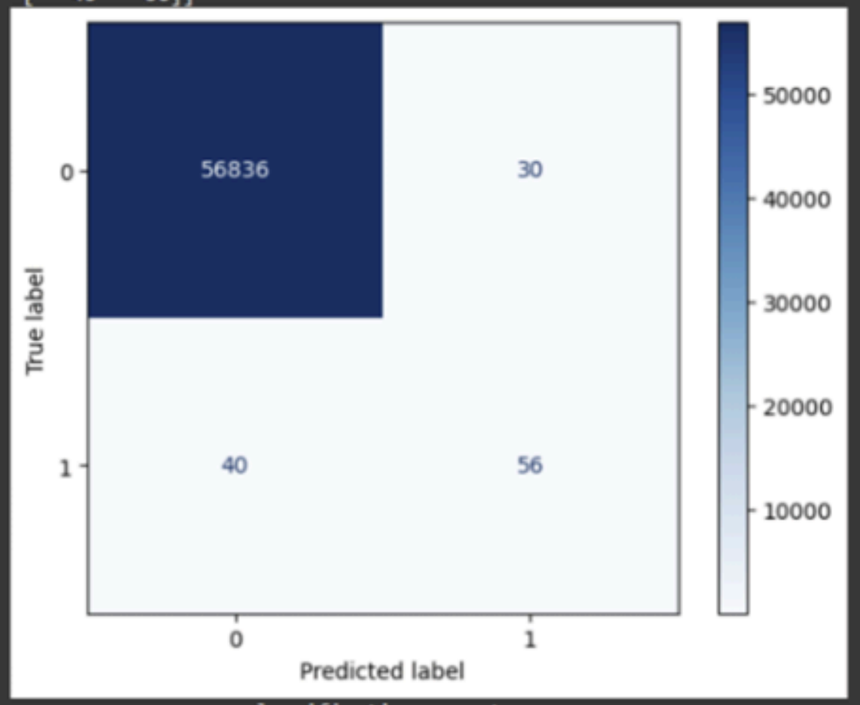
```
------------------ Confusion Matrix --------------------
[[56858     8]
 [   25    71]]
```




```
------------------ classification_report --------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56866
           1       0.90      0.74      0.81        96

    accuracy                           1.00     56962
   macro avg       0.95      0.87      0.91     56962
weighted avg       1.00      1.00      1.00     56962

------------------ More Specific classification_report --------------------
Accuracy:- 0.999420666409185
Sensitivity:- 0.7395833333333334
Specificity:- 0.9998593183976365
F1-Score:- 0.8114285714285714
```

XG BOOST

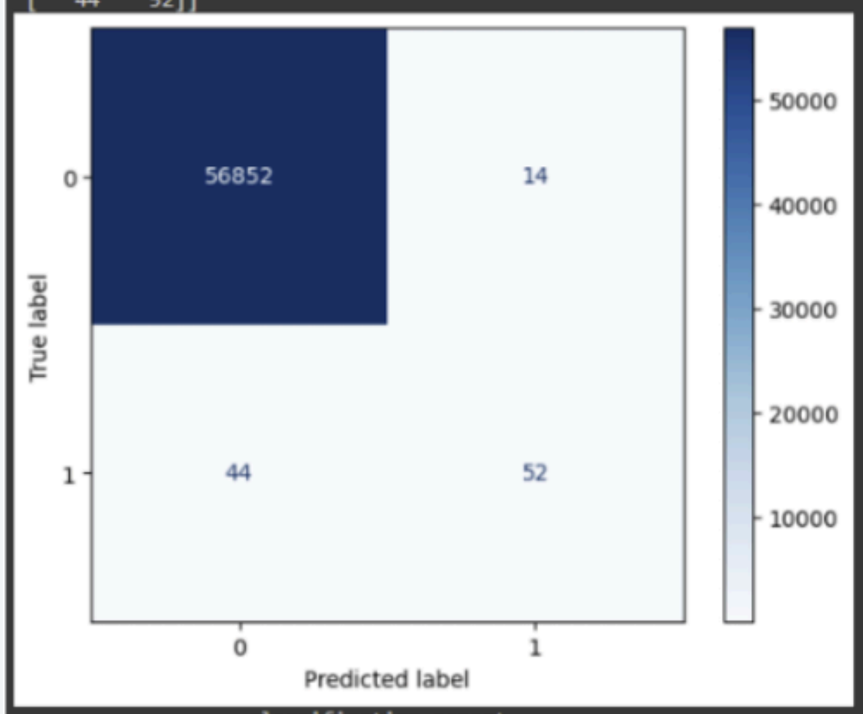
```
------------------ Confusion Matrix --------------------
[[56836    30]
 [   40    56]]
```




```
------------------ classification_report --------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56866
           1       0.65      0.58      0.62        96

    accuracy                           1.00     56962
   macro avg       0.83      0.79      0.81     56962
weighted avg       1.00      1.00      1.00     56962

------------------ More Specific classification_report --------------------
Accuracy:- 0.9987711105649381
Sensitivity:- 0.5833333333333334
Specificity:- 0.9994724439911371
F1-Score:- 0.6153846153846155
```

DECISION TREE

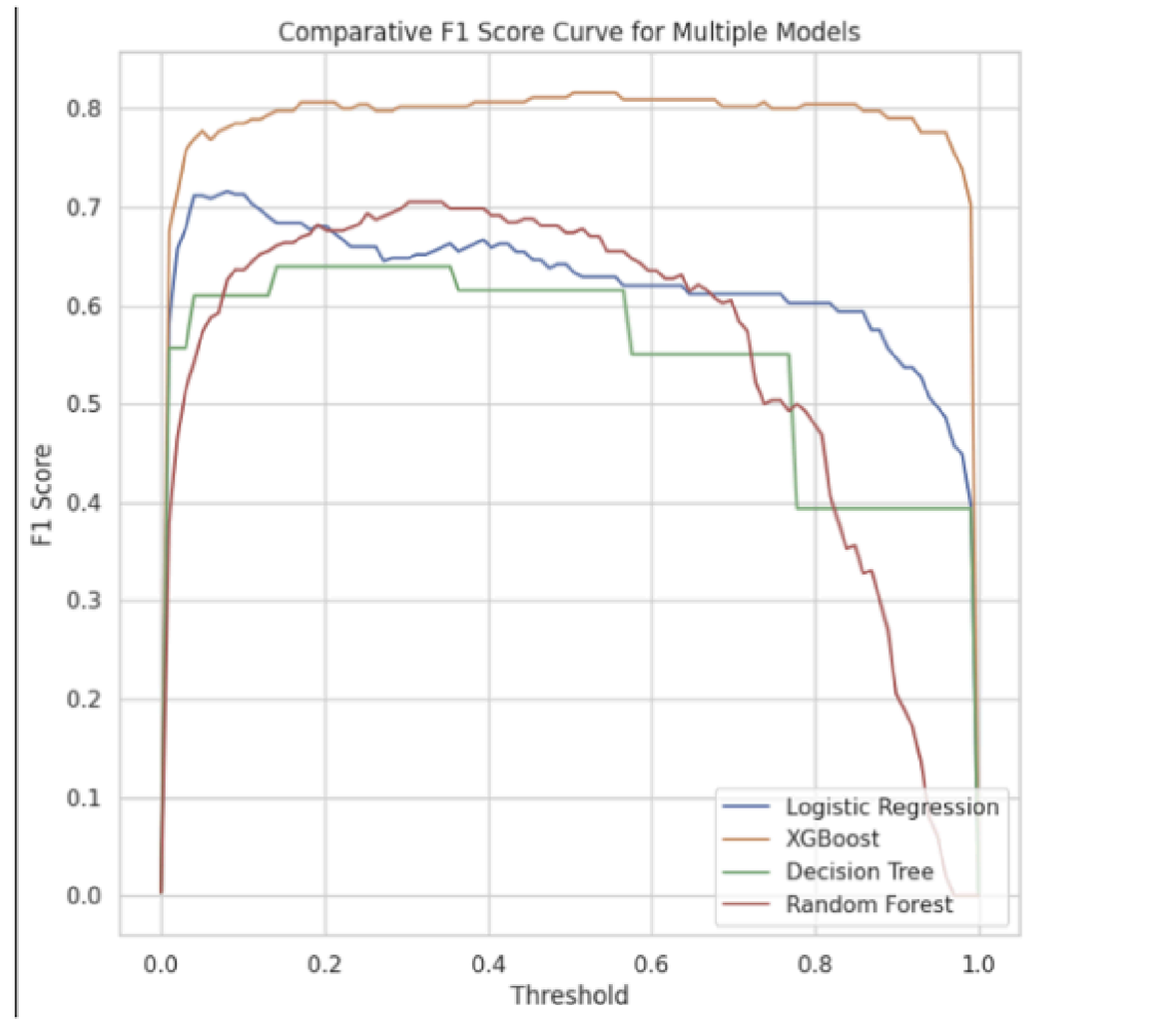
```
------------------ Confusion Matrix --------------------
[[56852    14]
 [   44    52]]
```



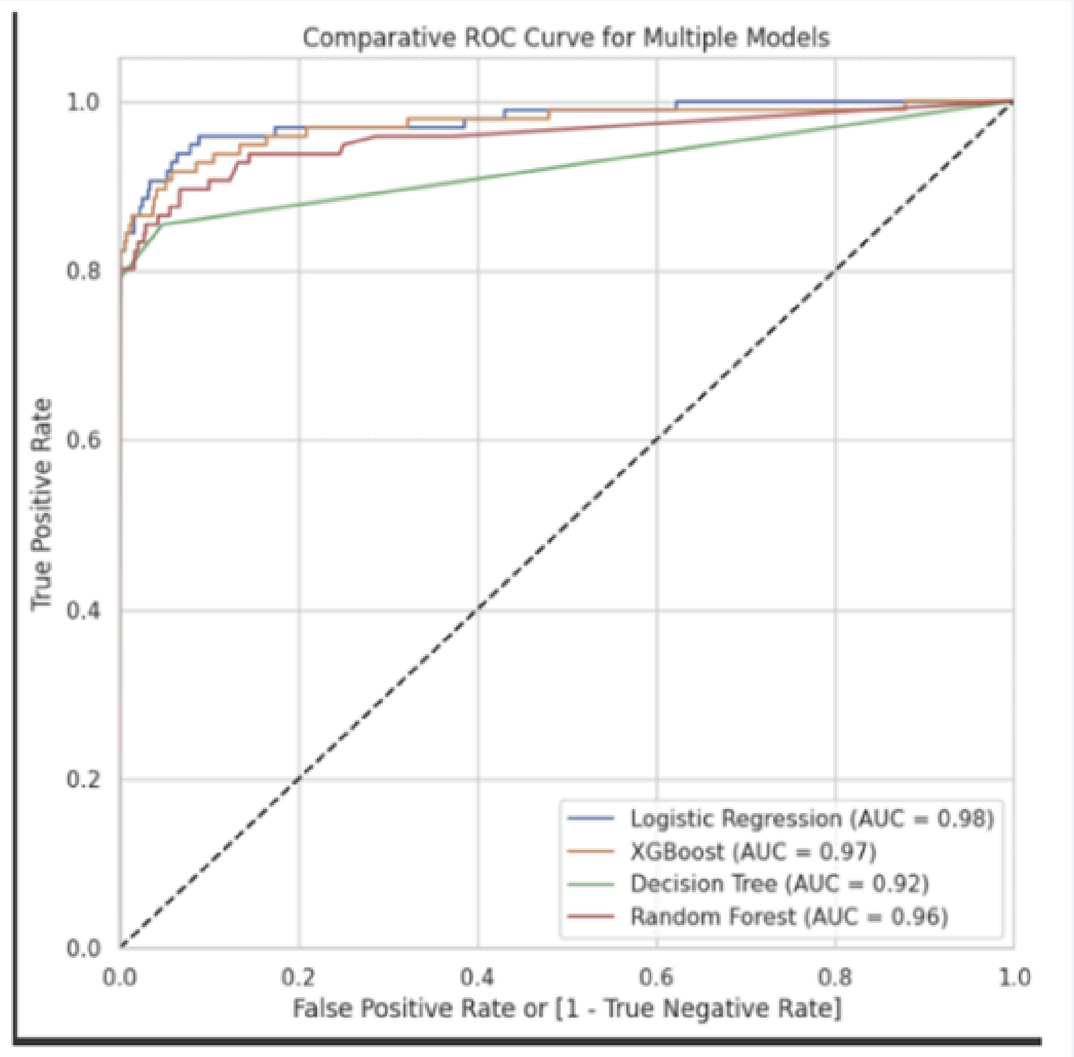

```
------------------ classification_report --------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56866
           1       0.79      0.54      0.64        96

    accuracy                           1.00     56962
   macro avg       0.89      0.77      0.82     56962
weighted avg       1.00      1.00      1.00     56962

------------------ More Specific classification_report --------------------
Accuracy:- 0.9989817773252344
Sensitivity:- 0.5416666666666666
Specificity:- 0.9997538071958639
F1-Score:- 0.6419753086419753
```

LOGISTIC REGRESSION

COMPARATIVE ANALYSIS



COMPARATIVE-F1 SCORE



COMPARATIVE-ROC

Algorithm	Accuracy (%)	F1 Score (%)	ROC (%)
Random Forest	99.8964	67.7596	95.7105
Decision Tree	99.8771	61.5385	92.1750
Logistic Regression	99.8982	64.1975	97.7696
XGBoost	99.9421	81.1429	97.2360

Algorithm	Accuracy (%)	F1 Score (%)	Strengths
Random Forest	99.8964	67.7596	Reduces overfitting, handles noisy data well, interpretable through feature importance.
Decision Tree	99.8771	61.5385	Easy to interpret, fast to train and deploy.
Logistic Regression	99.8982	64.1975	Fast to train, easy to interpret.
XGBoost	99.9421	81.1429	Highly efficient, often provides top performance on structured data, effective for large datasets.

COMPARISION OF ALGORITHMS

Probable Application

01

SECURITY FOR E-COMMERCE

Deploy fraud detection algorithms in online payment gateways to protect against card-not-present fraud during online purchases, reducing the risk of fraudulent transactions.

02

BEHAVIOR ANALYTICS

Analyze customer transaction data to understand typical spending patterns, allowing the system to detect deviations that may indicate fraudulent behavior.

03

RISK ASSESSMENT

Use machine learning to assess transaction risk by analyzing historical data, customer behavior, and patterns, enabling financial institutions to make informed decisions.

04

ACCOUNT TAKEOVER

Use machine learning to detect account takeovers by monitoring unusual activities like multiple logins, behavioral changes, and irregular transaction patterns.

Probable BE Problem statement

HEALTH INSURANCE FRAUD DETECTION

Identify fraudulent health insurance claims by analyzing patient treatment history and provider behavior to prevent financial losses and protect healthcare systems.

SUBSCRIPTION FRAUD DETECTION

Detect subscription fraud by identifying suspicious account creation patterns and usage behaviors to prevent losses from unpaid services and device theft.

REAL-TIME CREDIT CARD FRAUD DETECTION

Detect fraudulent credit card transactions in real-time to minimize unauthorized activities while reducing false positives to ensure a seamless customer experience.

E-COMMERCE FRAUD PREVENTION

Detect and prevent unauthorized account takeovers and fraudulent transactions by analyzing user behavior and transaction patterns to ensure secure online shopping.

CONCLUSION

Credit card fraud will continue to be a significant challenge for the financial industry. This seminar has showcased the potential of machine learning to mitigate these risks. By leveraging algorithms like XGBoost, Logistic Regression, Decision Trees, and Random Forest, future fraud detection systems will be capable of more accurately identifying fraudulent transactions.

Data preprocessing and feature engineering are crucial for optimizing model performance. As fraudsters evolve their tactics, ongoing research and development will be essential to stay ahead. By integrating emerging technologies and refining existing models, we anticipate a future where credit card transactions are increasingly secure.

REFRENCES

i. G. Rushin, C. Stancil, M. Sun, S. Adams and P. Beling, "Credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree,"2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2019, pp. 117-121.

ii. A.A.Taha, S.J.Malebary, An intelligent approach to credit card fraud detection using-An optimized light gradient boosting machine, IEEE Access 8(2020) 25579–25587, doi:10.1109/ACCESS.2020.2971354.

iii. Y. Lucas and J. Jurgovsky, ‘‘Credit card fraud detection using machine learning: A survey,’’ 2020 arXiv:2010.06479.

iv. A. Rb and S. K. Kr, ‘‘Credit card fraud detection using artificial neural network,’’ Global Transitions Proc., vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/j.gltp.2021.01.006.

v. Emmanuel Ileberi1*, Yanxia Sun1 and Zenghui Wang2 A machine learning based credit card fraud detection using the GA algorithm for feature selection Proc., vol. 2, no. 1, pp. 35–41, 2022

THANK YOU