Comparative study on credit card fraud detection

Parth Bramhecha

Parth Bramhecha; PICT, (IT), Pune, Maharastra India, parth.bramhecha007@gmail.com

Abstract

This seminar introduces an advanced credit card fraud detection system that combines super vised and unsupervised machine learning techniques to accurately identify anomalous trans action patterns, mitigating significant financial risks to users and the global financial system. The solution meticulously preprocesses and analyzes credit card transaction data, incorporating temporal patterns and distribution analysis, to distinguish fraudulent from legitimate transactions. A comparative analysis of XGBoost, Logistic Regression, Decision Tree, and Random Forest algorithms is conducted, targeting common fraud types such as card-not-present fraud (unauthorized transactions conducted over the phone), card-present fraud (transactions using cloned or stolen physical cards), and account takeover fraud (where fraudsters gain unauthorized access to an account to make transactions). These algorithms are evaluated based on accuracy, sensitivity, specificity, F1-score, and receiver operating characteristic curve-area under curve, aiming to enhance fraud detection and provide a more secure banking environment at scale

Keywords: Machine learning, XGBoost, Logistic Regression, Decision Tree, Random Forest, Financial transactions security, Banking sector

A well-structured article should follow a standard pattern as given below. This comprises of:

- **Introduction & Literature Survey**
- **Proposed Methods**
- **Results & Discussion**
- **Conclusions**
- **Acknowledgement (funding Organization etc.)**

1. Introduction

In today's digital landscape, credit card fraud has emerged as a formidable challenge, driven by the exponential growth of online transactions and the widespread adoption of electronic payment systems. While this rapid shift toward digital finance has ushered in greater convenience for consumers, it has also introduced significant vulnerabilities, making the prompt detection of fraudulent activities imperative. Credit card fraud manifests in various forms, including unauthorized transactions, account takeover, and identity theft, leading to substantial financial losses for both consumers and financial institutions. As such, developing robust fraud detection techniques is crucial to safeguarding the integrity of digital transactions and maintaining trust in the financial system.

The importance of timely fraud detection cannot be overstated, as delays can exacerbate losses and compromise consumer trust. Traditional methods of fraud detection, often reliant on rule-based systems and historical data analysis, struggle to keep pace with sophisticated and evolving fraud tactics. In this context, machine learning (ML) algorithms offer a promising solution. This seminar will conduct a comparative study of four powerful ML techniques: XGBoost, Logistic Regression, Decision Trees, and Random Forest. Each of these algorithms will be evaluated based on their effectiveness in identifying fraudulent patterns, their adaptability to new fraud strategies, and their computational efficiency.

We will delve into the methodologies behind each algorithm, assessing their strengths and weaknesses in the realm of credit card fraud detection. XGBoost is known for its high performance and speed, while Logistic Regression

offers interpretability and simplicity. Decision Trees provide a clear decision-making process, and Random Forest enhances accuracy through ensemble learning. By highlighting case studies and current trends, this exploration seeks to illuminate the transformative potential of these machine learning techniques in safeguarding financial transactions and improving overall fraud detection efficacy

2. Proposed Methods

In our proposed algorithm for detecting fraudulent transactions in credit card datasets, we con duct a comparative study of four prominent machine learning algorithms: XGBoost, Logistic Regression, Decision Tree, and Random Forest. Each algorithm offers unique strengths and mechanisms tailored to address the challenges inherent in fraud detection.

2.1.1XGBoost

XGBoost (Extreme Gradient Boosting) employs a gradient boosting framework that optimizes model performance through iterative improvements. XGBoost handles high-dimensional data well and includes regularization to reduce overfitting. It also ranks features based on their contribution to the model.

2.1.2 Logistic Regression

Logistic Regression is a simple, interpretable binary classification algorithm. Despite assuming linear relationships, it performs well in cases of simple relationships. Regularization techniques such as L1 and L2 help improve its performance on imbalanced datasets.

2.1.3 Decision Tree

Decision Tree models split the data based on feature values and are highly interpretable. How ever, they tend to overfit on noisy data, which can be mitigated by pruning or using them within an ensemble.

2.1.4 Random Forest

Random Forest is an ensemble method that builds multiple Decision Trees and aggregates their predictions. This reduces overfitting while improving generalization, making it ideal for fraud detection with high-dimensional data.

2.2 Methodology

2.2.1 Preprocessing

Data preprocessing is a critical aspect of any machine learning project, transforming raw data into a suitable format for analysis and modeling. This stage typically involves the removal or modification of unnecessary data features, handling missing values, managing outliers, and converting textual data into numerical formats. For this dataset, we are not performing outlier treatment, as all columns are already PCA-transformed, implying that outlier values have been addressed during the transformation.

Principal Component Analysis (PCA) is a statistical technique used to reduce dataset dimensionality while preserving the most important patterns or relationships between variables, without prior knowledge of the target variables. It serves as a feature extraction technique that aims to retain as much original information as possible. PCA is widely utilized in exploratory data analysis and predictive modeling. The primary goal of PCA is to map data from a higher-dimensional space to a lower-dimensional space while maximizing variance in the lower-dimensional space.

It is commonly applied in various AI applications, including computer vision and image compression, and is utilized across fields such as finance, data mining, and psychology.

2.2.2 Handling Missing Values

Handling missing values is crucial in data preprocessing. Missing data can lead to errors in data exploration and yield incorrect results. Several techniques for managing missing values include using the mean or median of the data, completely removing the affected rows or columns, and employing imputation methods to estimate missing values based on other observations.

2.2.3 Distribution of Classes with Time

In the context of credit card fraud detection, analyzing data distribution helps identify suspicious patterns and clusters of data points. This analysis allows for a better understanding of the data, guiding the selection of features or observations useful for predicting fraud. Various visualization methods, such as histograms, box plots, and scatter plots, can be employed, but we have chosen curves for clearer insights.

The analysis revealed no specific patterns distinguishing fraudulent and non-fraudulent transactions over time, leading us to drop the Time column from consideration.

2.2.4 Data Distribution Analysis

Understanding the patterns associated with credit card fraud, particularly their temporal nature, is essential. We performed an analysis of class distributions over time by plotting the occurrences of fraudulent versus non-fraudulent transactions. Comparing these distributions provides insights into the timing of fraudulent activities.

2.2.5 Distribution of Classes with Amount

In parallel with time distribution analysis, we examined the distribution of fraud and non-fraud transactions concerning the transaction amount. This analysis enhances our understanding of how transaction amounts influence fraud risk.

Our findings indicate that fraudulent transactions are primarily concentrated in the lower range of amounts, whereas non-fraudulent transactions are more evenly distributed across a broader range.

2.2.6 Train-Test Split

Splitting the dataset into training and testing sets is a vital step in model selection and evaluation. The training set develops the model, while the test set assesses its performance. This split should be executed randomly to ensure equal representation of all data points in each set.

2.2.7 Feature Scaling

Feature scaling is a preprocessing technique that adjusts the range of independent variables or features to ensure similar scales across all features, preventing any single feature from dominating the modeling process. It is particularly important for algorithms sensitive to feature scales, such as gradient-based optimization algorithms.

Importance of Feature Scaling:

- **Gradient Descent:** Optimization algorithms, including gradient descent, converge faster when features are on a similar scale, reducing the time to reach the minimum.

• **Distance-Based Algorithms:** Algorithms like k-nearest neighbors and support vector machines are sensitive to feature scales. Scaling ensures each feature contributes proportionally to distance calculations.

• **Regularization:** In models like linear regression and support vector machines, regularization terms penalize large coefficients. Feature scaling helps apply regularization uniformly across all features.

For our preprocessing, we utilized Standard-Scaler from scikit-learn, which standardizes features by removing the mean and scaling to unit variance, transforming features to have a mean of 0 and a standard deviation of 1. The standardization formula is given by:

$$X_{standardized} = \frac{(X - \mu)}{\sigma}$$

where X is the original value, μ is the mean, and σ is the standard deviation.

2.2.8 Different Models

With the data preprocessed and split, we are now ready to build the model. We will employ both supervised and unsupervised machine learning algorithms, testing different models such as XGBoost, Logistic Regression, Decision Tree, and Random Forest. For each algorithm, we will utilize a range of performance metrics, including the confusion matrix, classification report, accuracy, sensitivity, specificity, F1 score, and ROC-AUC score, to evaluate performance.

3. Results & Discussion

3.1Dataset study

Fig. 1 Figure showing the number of fraudulent data is very small compared to nonfraudulent data

It can be observed the dataset is highly unbalanced

3.2 Random Forest

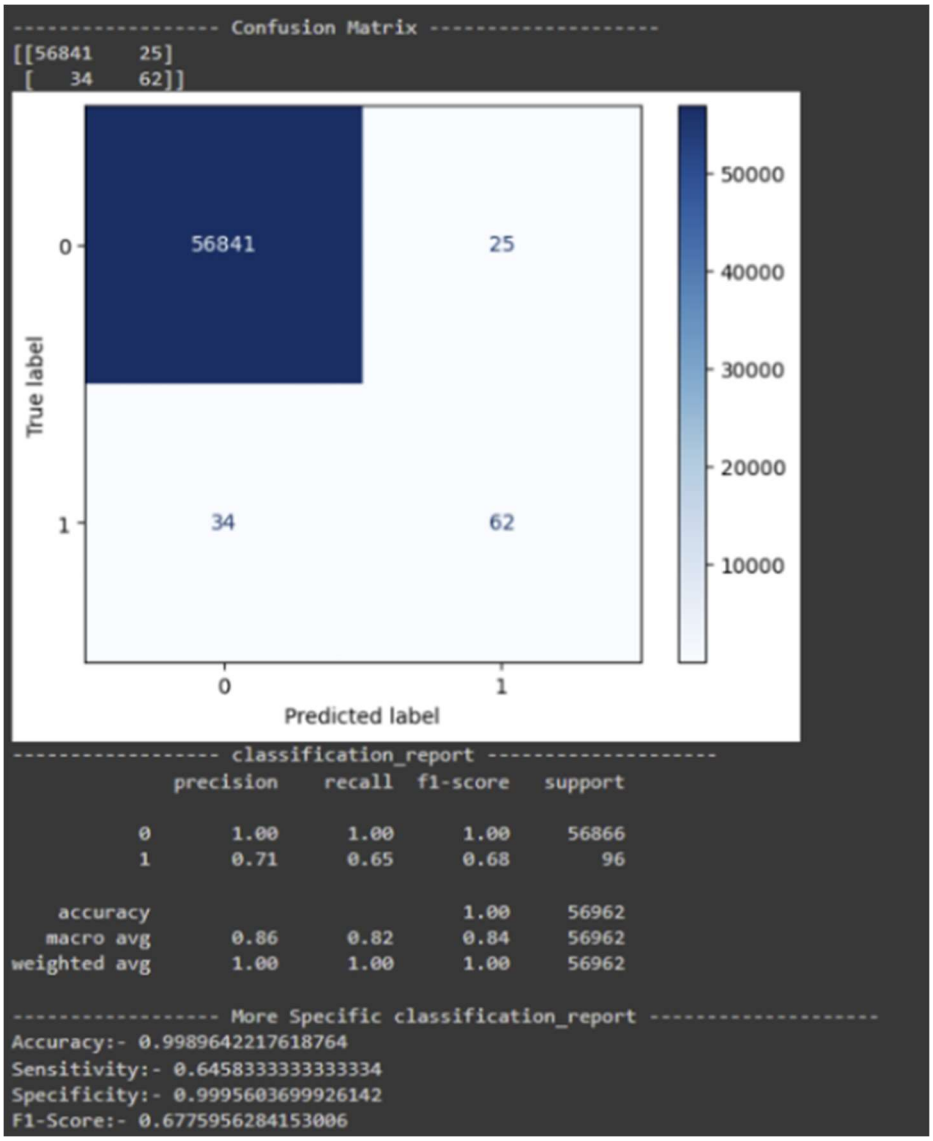

Fig. 2 Figure showing the output of training the Random Forest classifier

3.3 XG Boost

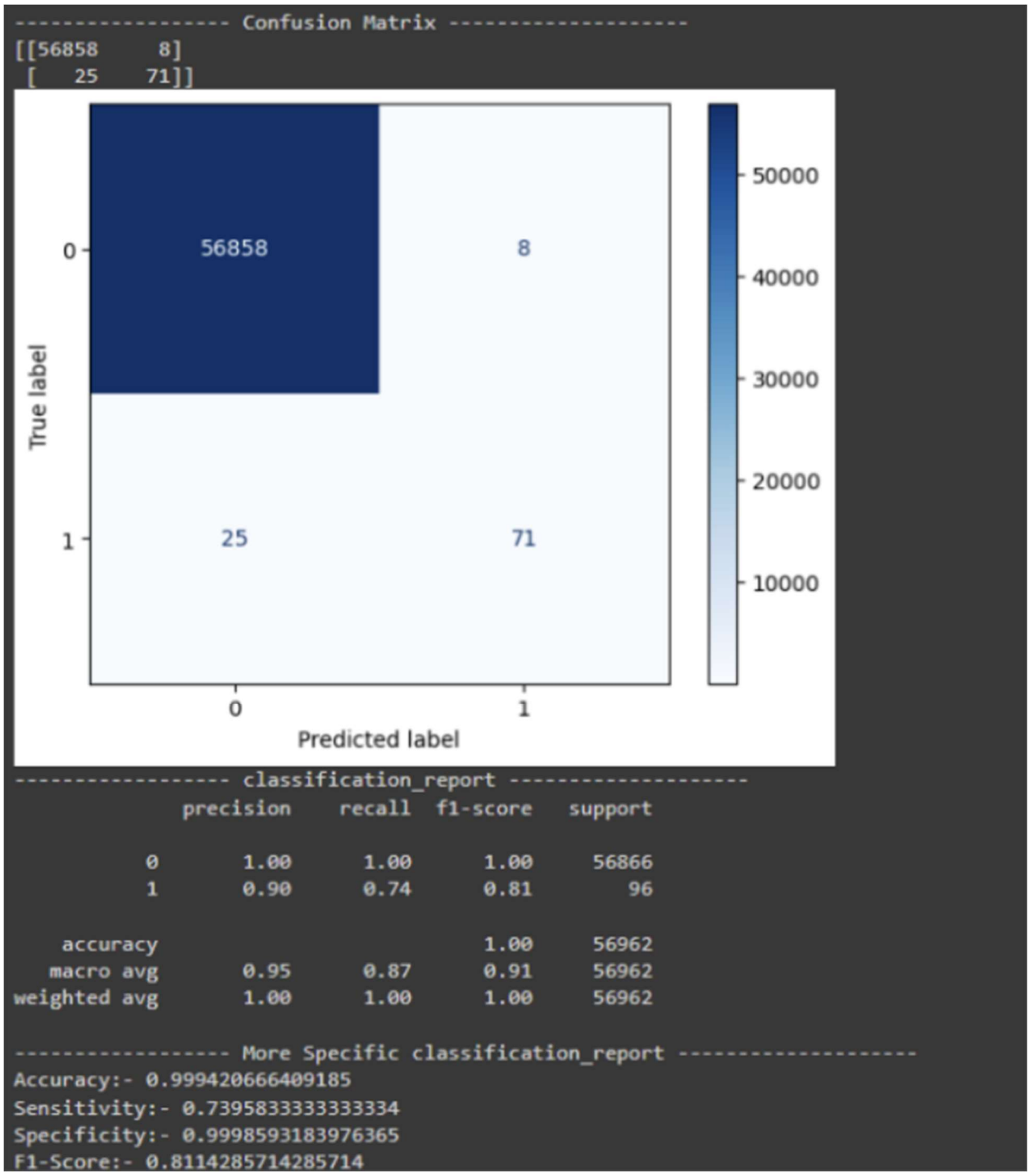

Fig. 3 Figure showing the output of training the XG Boost classifier

3.4 Decision Tree

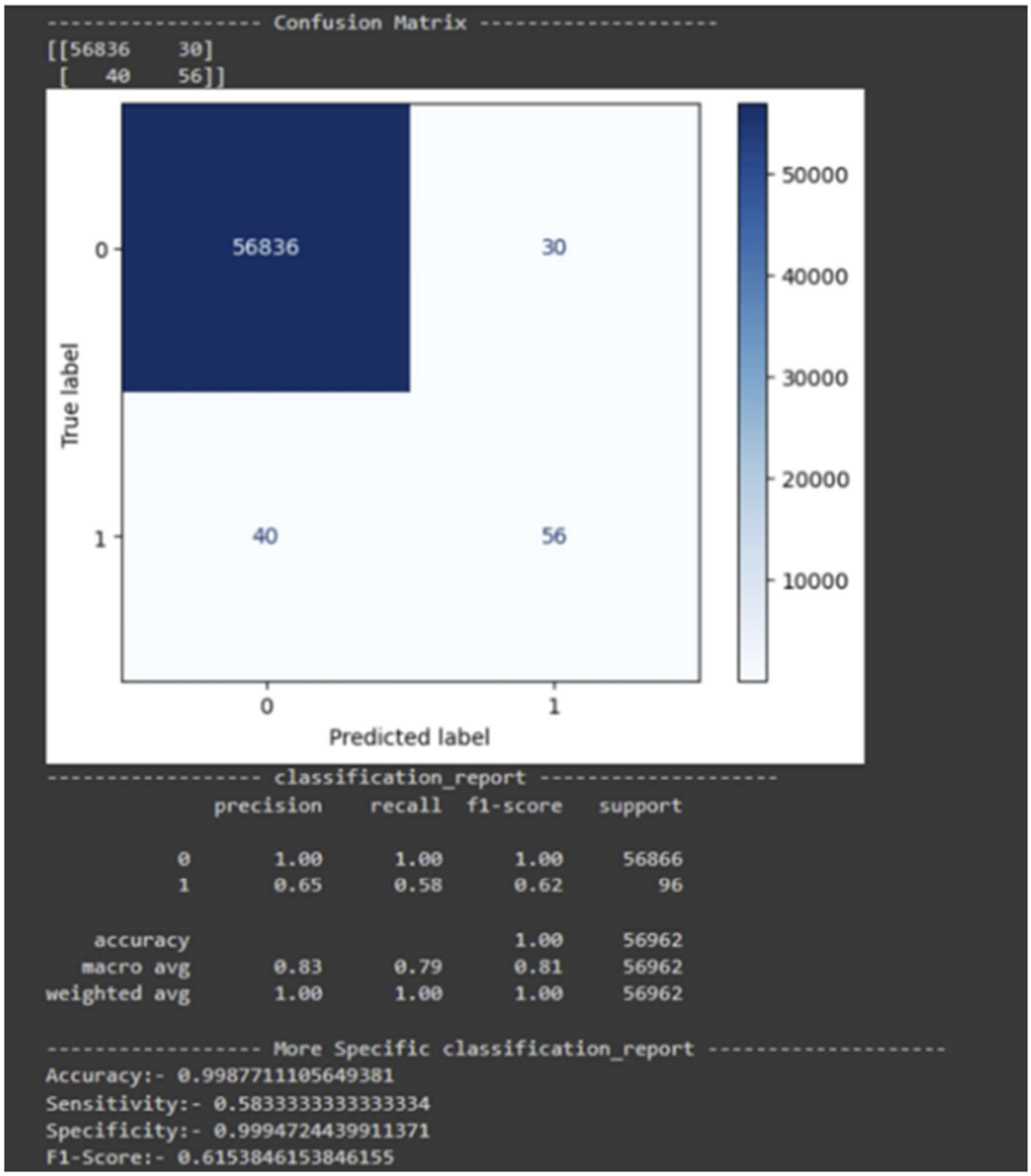

Fig. 4 Figure showing the output of training the Decision tree classifier

3.5 Logistic Regression

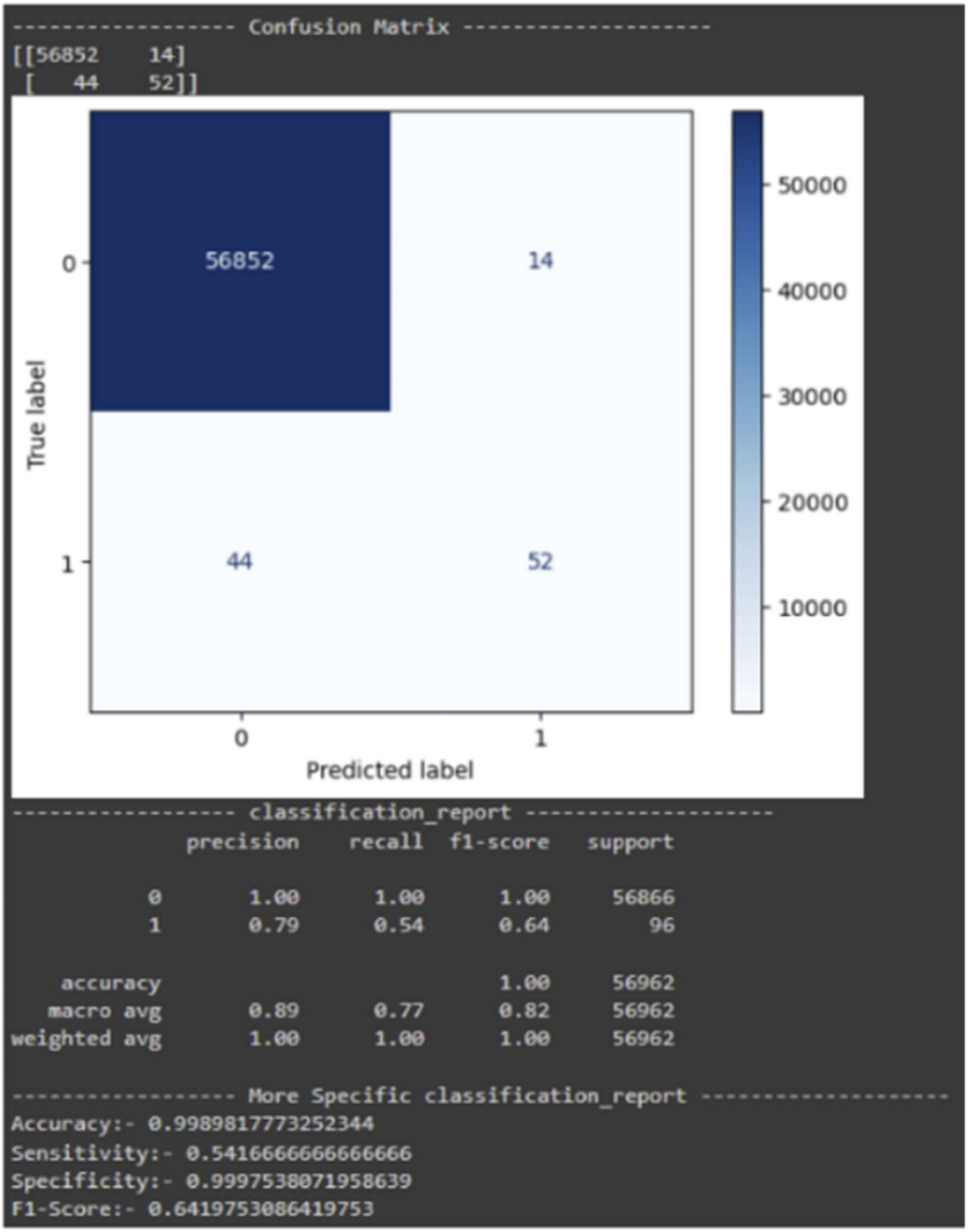

Fig. 2 Figure showing the output of training the Logistic regression classifier

3.6 Comparative F1 Score

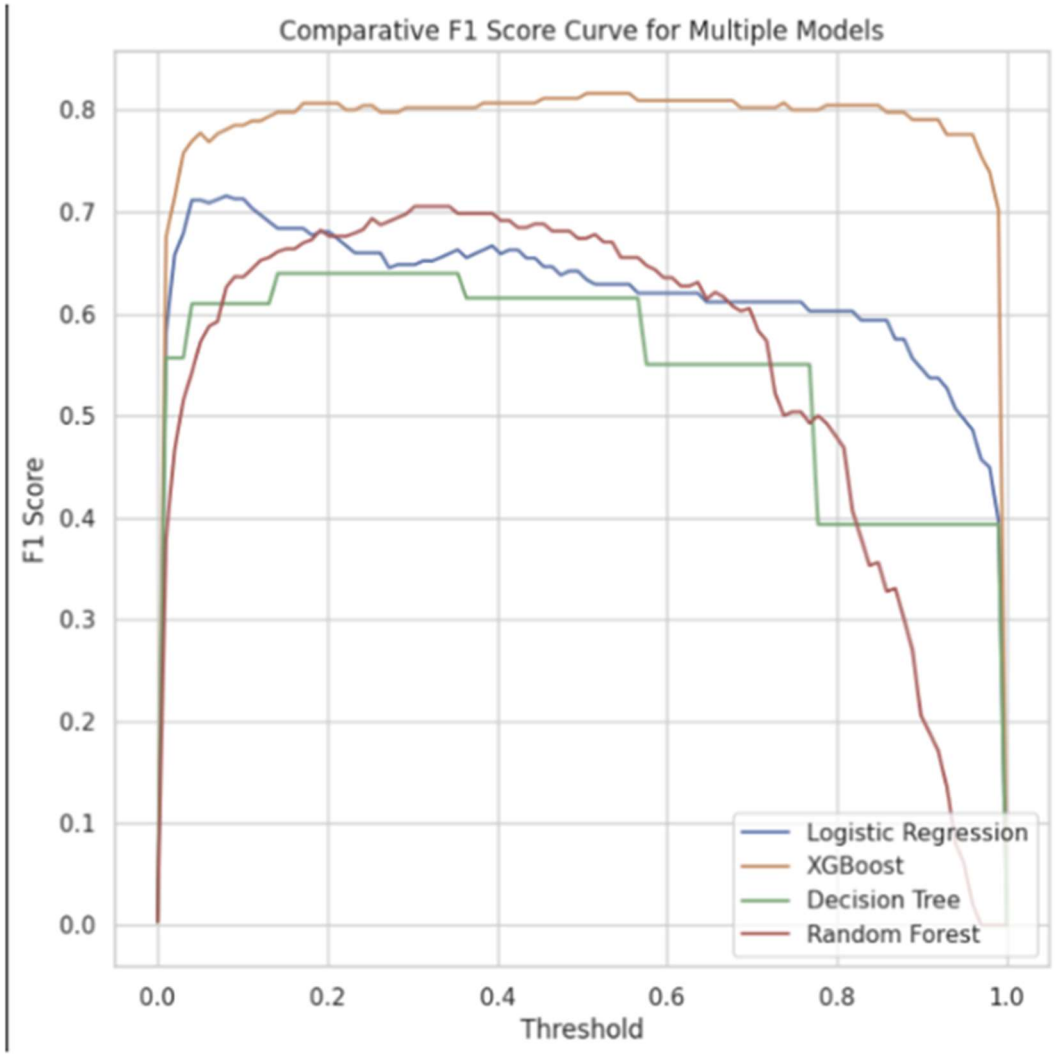

Fig. 6 The image shows the plot of F1 score of various algorithms

F1-score

The F1-score, which balances precision and recall, varies more significantly across models:

• **XGBoost:** Highest F1-score (0.811429), indicating a good balance between precision and recall.

• **Logistic Regression, Random Forest, and Decision Tree:** Lower F1-scores (0.615385 to 0.677596), suggesting limitations in identifying positive instances or avoiding false positives

3.7 Comparative ROC-AUC Score

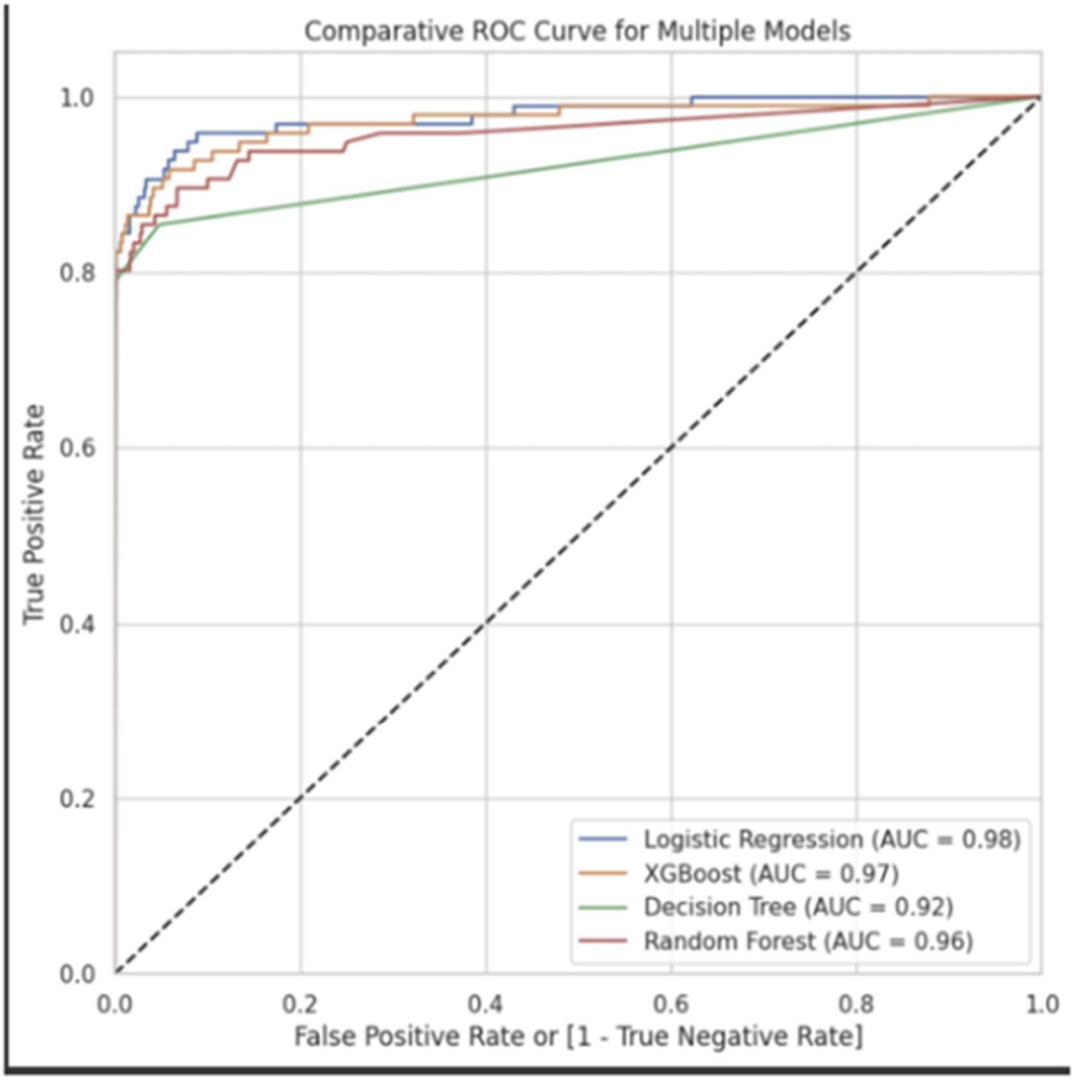

Fig. 7 The image shows the plot of ROC-AUC score of various algorithms

ROC-AUC:

Logistic Regression is the best at distinguishing between the positive and negative classes, as indicated by its high ROC-AUC.

XGBoost offers a good balance of ROC-AUC and F1-score, making it a competitive choice for balanced predictions.

Random Forest performs moderately well in ROC-AUC but not as well as the other two models.

Decision Tree is the weakest performer in terms of ROC-AUC, suggesting it may struggle more

with class separation compared to the other models.

Table 1 Summary of all the findings

Algorithm	Accuracy (%)	F1 Score (%)	ROC (%)	Strengths
Random Forest	99.8964	67.7596	95.7105	Reduces overfitting, handles noisy data well, interpretable through feature importance.
Decision Tree	99.8771	61.5385	92.1750	Easy to interpret, fast to train and deploy.
Logistic Regression	99.8982	64.1975	97.7696	Fast to train, easy to interpret.
XGBoost	99.9421	81.1429	97.2360	Highly efficient, often provides top performance on structured data, effective for large datasets.

4. Conclusion

In the realm of credit card fraud detection, this study addressed the critical issue of identifying fraudulent transactions amidst legitimate ones. By leveraging advanced machine learning algorithms and real-time monitoring techniques, we proposed a comprehensive solution aimed at enhancing fraud detection efficacy.

The advanced credit card fraud detection system presented in this seminar demonstrates the potential of leveraging both supervised and unsupervised machine learning techniques to identify fraudulent transactions with higher precision. By carefully analyzing temporal patterns, transaction distributions, and comparing multiple algorithms, this solution addresses the dynamic and complex nature of financial fraud. The performance evaluation based on key metrics like accuracy, sensitivity, specificity, F1-score, and ROC-AUC highlights that a multi-faceted approach, including XGBoost, Logistic Regression, Decision Tree, and Random Forest, can significantly improve fraud detection. Implementing such a robust system at scale would not only reduce financial losses but also ensure greater trust and security within the banking ecosystem. Ultimately, this study contributes to enhancing transaction security, mitigating risks, and providing a more secure environment for consumers and financial institutions alike.

Throughout the paper, we outlined how machine learning can proactively detect fraud, analyze user behavior, and implement verification processes. We have successfully accomplished a detailed exploration of these methodologies, offering insights into their practical applications in banking.

Acknowledgements

I would like to express my deepest gratitude to everyone who supported and guided me throughout the preparation and completion of this paper.

I am immensely grateful to my seminar guide, Dr Anant M Bagade, for her invaluable guidance, encouragement, and insightful feedback at every stage of this seminar. Her expertise and unwavering support were instrumental in

shaping the direction and outcome of this work. I also wish to thank Mrs.Prajakta S., the reviewer, for his constructive evaluation and suggestions, which greatly contributed to enhancing the quality of this paper.

I would like to acknowledge my peers and friends for their valuable input and support, which motivated me throughout the seminar preparation. Lastly, I extend my heartfelt appreciation to my family for their continuous encouragement, patience, and unwavering support. Thank you all for your contributions and support

References

[1] Sanjay Bharadwaj, Credit Card Fraud Detection Using Machine Learning, March 2024. Compares the effectiveness of two low-cost machine learning techniques, Random Forest and K-Nearest Neighbor (K-NN), in predicting credit card fraud. Evaluates models based on key machine learning metrics, ease of implementation, and cost-effectiveness.

[2] Neethu Tressa, Credit Card Fraud Detection Using Machine Learning, October 2023. Pro poses a system for detecting credit card fraud using Decision Tree and Random Forest algorithms. The goal is to ensure fraudulent transactions are detected, preventing unauthorized charges.

[3] Naga Ashwini Nayak, Credit Card Fraud Detection Using Machine Learning, April 2023. Utilizes Decision Tree, Random Forest, and Extreme Gradient Boosting to evaluate their effectiveness in detecting fraudulent activities in both public and real-world financial datasets.

[4] Deep Prajapati, Credit Card Fraud Detection Using Machine Learning, December 2021. Evaluates and compares Random Forest, XGBoost, and ANN for their effectiveness in detecting and preventing fraudulent credit card transactions.

[5] C. Sudha and D. Akila, Credit Card Fraud Detection System based on Operational & Transaction features using SVM and Random Forest Classifiers, February 2021. Develops a credit card fraud detection system using SVM and Random Forest classifiers. Classifies transactions into benign or suspected categories based on operational features. Evaluates performance with metrics such as precision, accuracy, recall, and F1-score.