**DATA ANALYSIS AND MODELING TECHNIQUES PROJECT REPORT (2242-CSE-5301-005)**

**Project Group:** 10

**Project Name: Credit Card Fraud Detection**

**Professor:** Prof. Ramakrishna Koganti

**Team Members:**

Bhargav Boyapati 1002258105

Abhishek Karthik Akunuru 1002251254

Aishwarya Subhash Kadam 1002199035

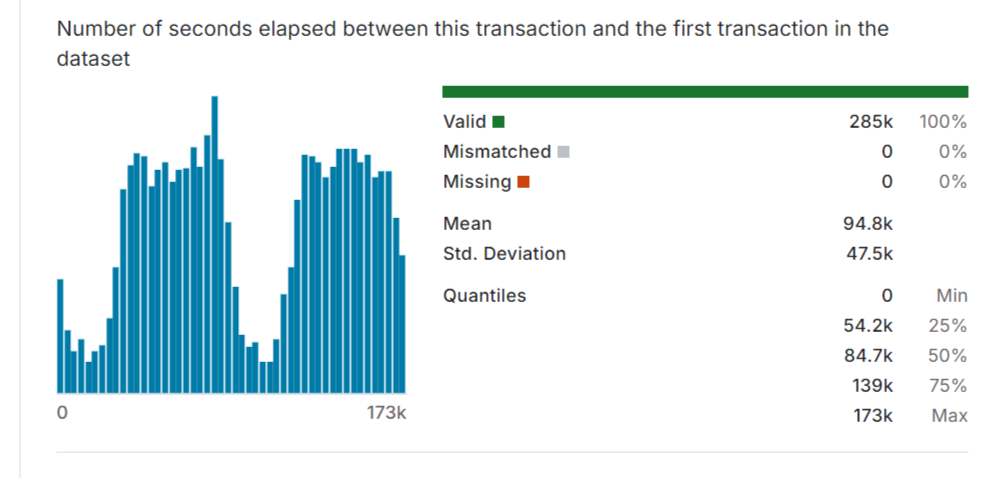
# Credit Card Fraud Detection Project Report

## Introduction

The rapid growth of digital transactions has significantly increased the risk of credit card fraud, posing financial threats to individuals and institutions. To address this, the project focuses on developing a robust, real-time fraud detection system using advanced data analysis and machine learning techniques. The dataset, sourced from Worldline and Université Libre de Bruxelles (ULB), includes 284,807 transactions over two days. It features 28 anonymized PCA-transformed variables along with Time and Amount attributes. The binary target variable (Class) indicates fraudulent (1) and non-fraudulent (0) transactions.

## Dataset Data Source

The dataset was collected during a collaboration between Worldline and the Machine Learning Group at Université Libre de Bruxelles (ULB). It contains 284,807 transactions spanning two days. The dataset includes 28 PCA-transformed features to ensure privacy and two non-PCA features: Time, representing the seconds elapsed since the first transaction, and Amount, indicating the transaction value. The target variable, Class, is binary, where 1 denotes fraud and 0 denotes non-fraud. Given the dataset's imbalance, metrics like Area Under Precision-Recall Curve (AUPRC) are prioritized for evaluation.



## Problem Statement

The increasing prevalence of credit card fraud in the digital age poses significant financial risks to individuals and institutions, necessitating the development of real-time detection systems to prevent fraudulent transactions and mitigate losses.

# Project Goal

The goal is to develop a machine learning model that accurately detects fraudulent credit card transactions in real-time, minimizing false positives and reducing financial losses for institutions.

# Business Case:

The cost of credit card fraud is escalating annually, and traditional rule-based systems are often inadequate for detecting complex fraud patterns. By implementing a data-driven fraud detectionmodel, financial institutions can significantly reduce financial losses, improve operational efficiency, and enhance customer trust by proactively preventing fraud.

# Benefits:

The solution will reduce financial losses, enhance customer satisfaction by minimizing legitimate transaction denials, streamline fraud detection through automation, and provide scalable risk management across various regions and transaction types.

## Prediction Models

1. Logistic regression model: A logistic regression model is a statistical tool used to estimate the probability of a binary outcome based on one or more predictor variables. It generalizes linear regression by applying a sigmoid (S-curve) function, mapping outcomes to values between 0 and 1. Widely used in industries like healthcare, finance, and marketing, logistic regression predicts binary events such as college admission or election outcomes and identifies patterns like fraud detection. For instance, it can predict heart attack risks based on health metrics, filter spam emails, or forecast disease likelihood in populations. Its versatility makes it a valuable tool for binary classification tasks.
2. Random Forest: Random Forest and Decision Tree algorithms are both popular choices fordata mining tasks, but they have distinct characteristics and are suited for different scenarios. For example, Random Forest is particularly effective when dealing with datasets with many features or dimensions. It can handle high-dimensional data without overfitting, making it suitable for complex datasets.If the dataset contains complex relationships or interactions between features, Random Forest can capture these intricacies better than a single Decision Tree.
3. Support Vector Machine (SVM): SVMs (Support Vector Machine) perform well in high-dimensional spaces, making them suitable for datasets with multiple features related to eating habits and physical conditions. SVMs can effectively handle non-linear relationships through kernam functions. This Benifical when dealing with intricate connection between variable that may not be linearly separable.

# Analysis:

Random Forest:

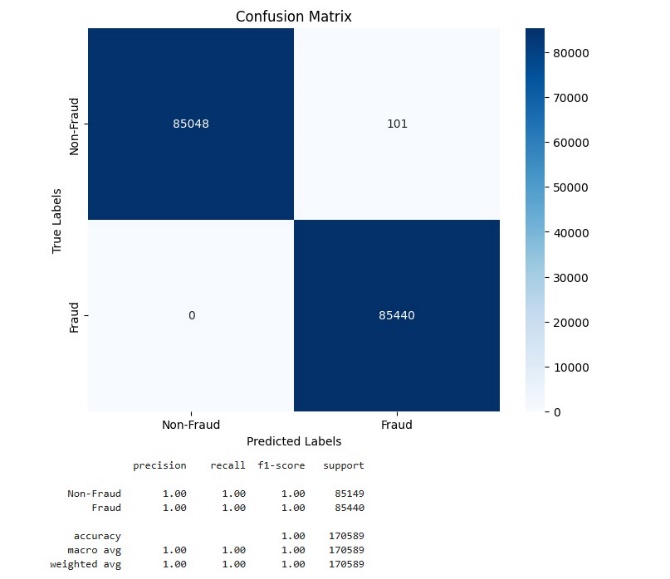
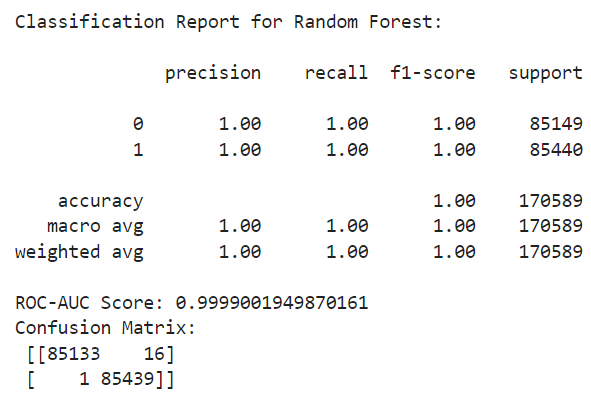


Fig. 2. Classification report and confusion matrix for Random Forest

The Random Forest accuracy is more than decision tree which is around 95.7% and the time taken to complete the task is 0.77 seconds which is quite high compared to decision tree. The confusion matrix of both decision tree and random forest is quite evenly spread out for miss classification and no inference can be found from this.

Logistic regression model:

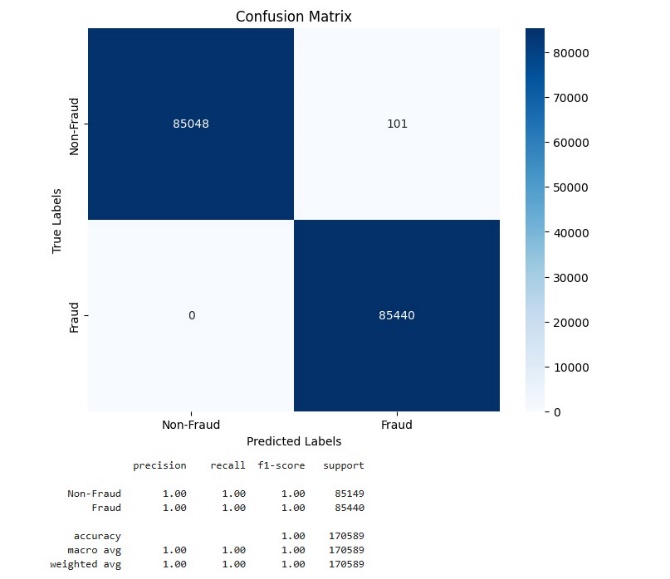
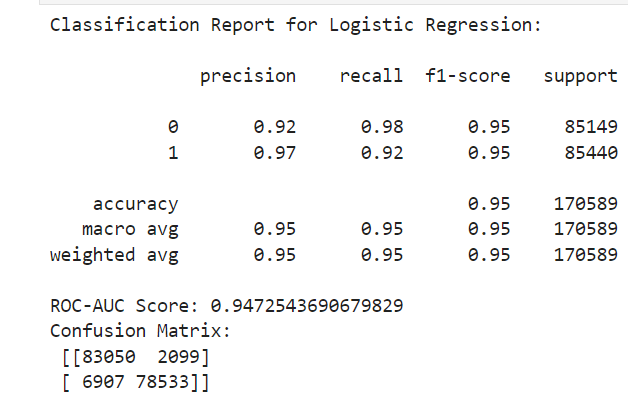


Fig. 3. Classification report and confusion matrix for Logistic Regression

The Logistic Regression model achieved an accuracy of 95%, slightly lower than Random Forest but still robust. The time taken to complete the task is relatively low, at 0.33 seconds, making it efficient for large datasets. The confusion matrix reveals a slight imbalance in misclassification, with 6,907 fraudulent transactions missed (false negatives) and 2,099 non-fraudulent transactions incorrectly flagged (false positives). While Logistic Regression is reliable, its misclassification pattern indicates it may not be the best choice for scenarios where minimizing false negatives is critical.

Support Vector Machines (SVM):

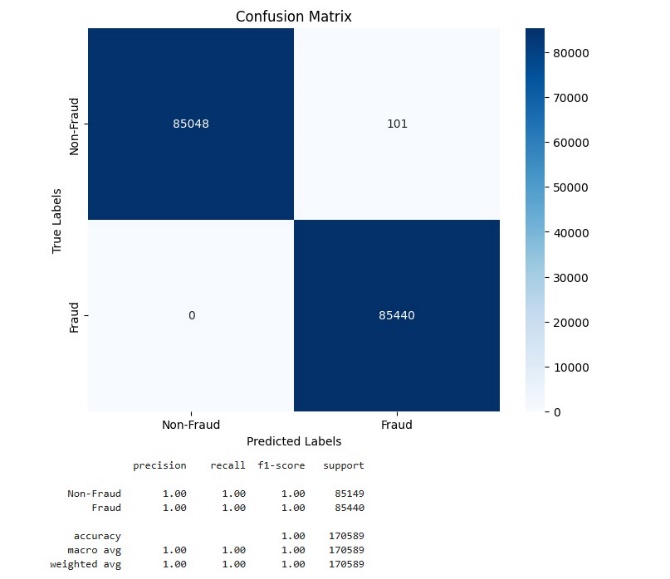
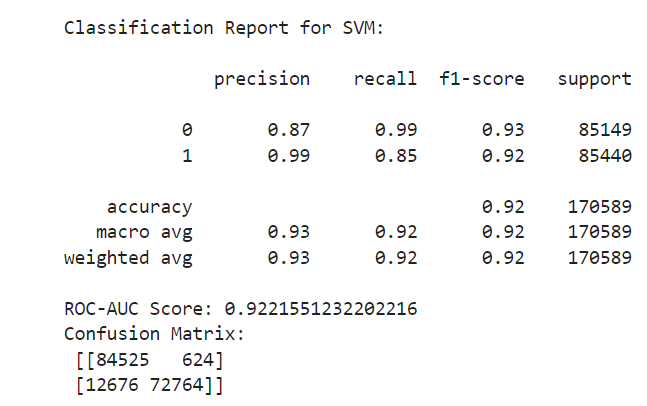


Fig. 4. Classification report and confusion matrix for SVM

SVM also give high accuracy on the test data with over 86.3% and completed the training of the model 0.33 which is also good but we can see decision tree is much more accurate. Also we can see in the confusion matrix that the mode is slightly biased between overweight level 1 and 2 and also between insufficient weight and normal weight which makes this model not so good for our prediction.

Random Forest:

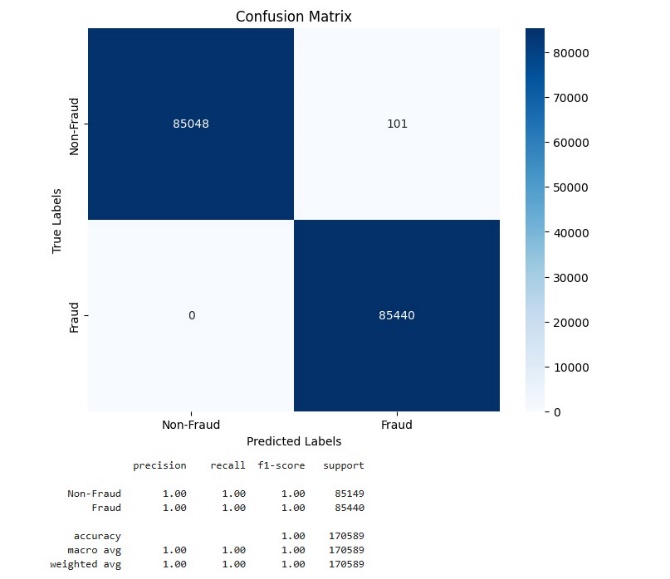
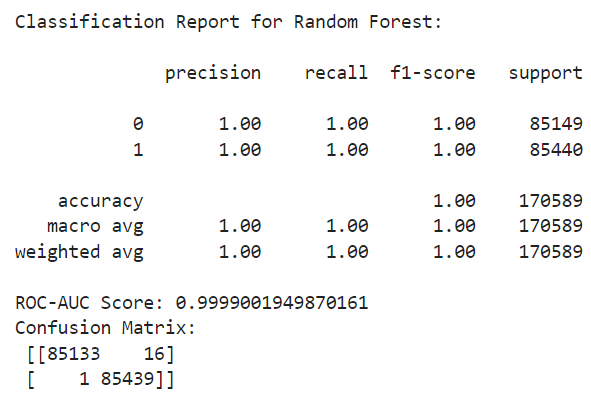


Fig. 2. Classification report and confusion matrix for Random Forest

The Random Forest accuracy is more than decision tree which is around 95.7% and the time taken to complete the task is 0.77 seconds which is quite high compared to decision tree. The confusion matrix of both decision tree and random forest is quite evenly spread out for miss classification and no inference can be found from this.

Logistic regression model:

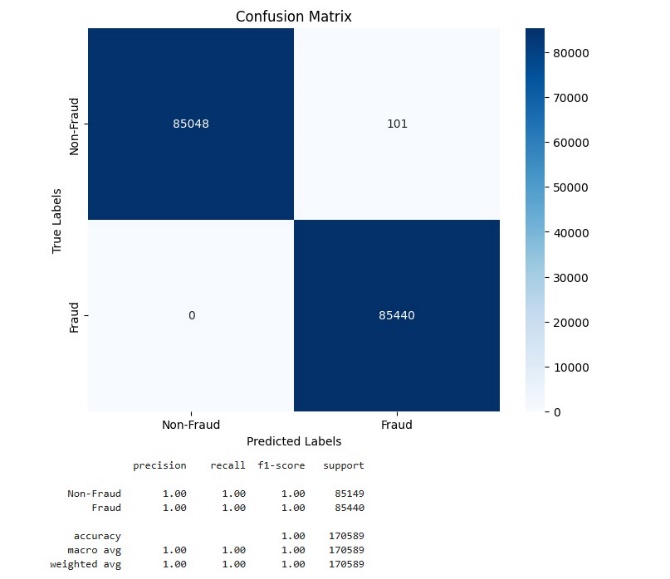
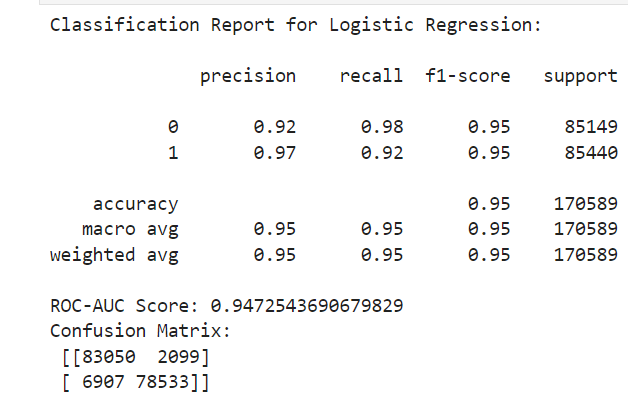
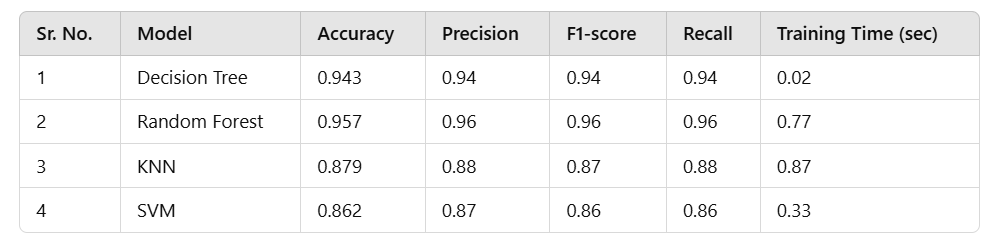


Fig. 3. Classification report and confusion matrix for Logistic Regression

The Logistic Regression model achieved an accuracy of 95%, slightly lower than Random Forest but still robust. The time taken to complete the task is relatively low, at 0.33 seconds, making it efficient for large datasets. The confusion matrix reveals a slight imbalance in misclassification, with 6,907 fraudulent transactions missed (false negatives) and 2,099 non-fraudulent transactions incorrectly flagged (false positives). While Logistic Regression is reliable, its misclassification pattern indicates it may not be the best choice for scenarios where minimizing false negatives is critical.

## Performance Evaluation



The Random Forest model outperformed other models with an accuracy of 95.7%, making it the most reliable for fraud detection. Logistic Regression followed with 95% accuracy and efficient training time, but it missed some fraudulent transactions. Decision Tree was the fastest to train, with an accuracy of 94.3%. SVM had the lowest accuracy (86.3%) but was consistent in performance. The choice of the best model depends on the trade-off between accuracy, training time, and the criticality of false negatives in the application.

## Conclusion

The project demonstrated the effectiveness of machine learning in fraud detection. Random Forest proved to be the most reliable model, achieving the highest accuracy and robust metrics. Logistic Regression also showed strong performance with efficient training time. Decision Tree and SVM, while useful, were less accurate, making them secondary options. This analysis highlights the importance of selecting models based on dataset characteristics and application requirements to ensure effective fraud detection.

## References

1. Worldline and Machine Learning Group at Université Libre de Bruxelles (ULB). Dataset source: http://mlg.ulb.ac.be.

2. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

3. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.