Credit Card Fraud Detection Report

1. Introduction

Credit card fraud is a significant problem in financial institutions. Detecting fraudulent transactions early helps prevent financial losses and protect customers. In this project, we use machine learning algorithms to analyze transaction data and classify whether a transaction is fraudulent or not. The dataset used for this study is taken from kaggle, which contains transactions made by European cardholders in September 2013.

2. Objective

The main objective of this project is to build a machine learning model that can accurately detect fraudulent credit card transactions using classification algorithms like Naïve Bayes, Decision Tree and Logistic Regression.

3. Steps Taken

1. **Data Collection:** The dataset was obtained from Kaggle's Credit Card Fraud Detection dataset. It contains 284,807 transactions, out of which 492 are fraudulent.
2. **Data Processing:** The data was cleaned and prepared for model training. Missing values were handled, and the data was standardized to ensure uniformity.
3. **Exploratory Data Analysis (EDA):** Various graphs were plotted to understand the data distribution, correlation, and imbalance between classes.
4. **Handling imbalanced Data:** The dataset was highly imbalanced, so we used the SMOTE (Synthetic Minority Oversampling Technique) method to balance the number of fraud and non-fraud samples.
5. **Features Selection:** The most relevant features were used to train the model.
6. **Model Building:** Three algorithms were implemented – Naïve Bayes, Decision Tree and Logistic Regression. Each model was trained on the training dataset and tested on the test dataset.
7. **Model Evaluation:** Models were evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC score. Confusion matrices were plotted to visualize the performance of each algorithm.

4. Results

- **Naïve Bayes:** Gave quick results but had lower precision due to oversimplified assumptions.
- **Decision Tree:** Showed good accuracy and interpretability but slightly overfitted on the training data.

- **Logistic Regression**: Performed best overall, achieving a balance between precision, recall, and F1-Score, Overall, Logistic Regression was found to be the most effective algorithm for fraud detection in this dataset.

5. Conclusion

This project successfully demonstrates how machine learning algorithms can be applied to detect fraudulent credit card transactions. Logistic Regression provided the most reliable results for the dataset. Future improvements can include using ensemble techniques like Random Forest or Gradient Boosting and deploying the model into a real – time fraud detection system.