Hardi Nagaria

CREDIT CARD FRAUD DETECTION



Contents

[**1. Introduction** 2](#_Toc162908475)

[**1.1 Background** 2](#_Toc162908476)

[**1.2 Objectives** 2](#_Toc162908477)

[**2. Data Collection and Preprocessing** 2](#_Toc162908478)

[**2.1 Data Source** 2](#_Toc162908479)

[**2.2 Data Description** 2](#_Toc162908480)

[**2.3 Data Cleaning** 2](#_Toc162908481)

[**3. Exploratory Data Analysis (EDA)** 3](#_Toc162908482)

[**3.1 Summary Statistics** 3](#_Toc162908483)

[**3.2 Visualization** 3](#_Toc162908484)

[**3.3 Correlation Analysis** 3](#_Toc162908485)

[**4. Model Development** 3](#_Toc162908486)

[**4.1 Model Selection** 3](#_Toc162908487)

[**4.2 Model Training** 3](#_Toc162908488)

[**4.3 Model Evaluation** 4](#_Toc162908489)

[**Logistic Regression Model:** 4](#_Toc162908490)

[**RandomForestClassifier Model:** 4](#_Toc162908491)

[**5. Results and Discussion** 5](#_Toc162908492)

[**5.1 Model Performance** 5](#_Toc162908493)

[**5.2 Key Findings** 5](#_Toc162908494)

[**5.3 Limitations** 5](#_Toc162908495)

[**6. Conclusion and Recommendations** 5](#_Toc162908496)

[**6.1 Conclusion** 5](#_Toc162908497)

# **1. Introduction**

## **1.1 Background**

Financial institutions face significant challenges in detecting fraudulent activities, particularly in credit card transactions. Timely identification of fraudulent transactions is crucial to mitigate financial losses and protect customers’ assets.

## **1.2 Objectives**

The objective of this project is to develop a robust credit card fraud detection system using machine learning techniques. By leveraging advanced analytics and predictive modeling, the aim is to enhance the detection capabilities of financial institutions and minimize the impact of fraudulent activities.

# **2. Data Collection and Preprocessing**

## **2.1 Data Source**

The dataset used in this project comprises credit card transaction data obtained from a financial institution.

## **2.2 Data Description**

The dataset contains transaction records, including features such as transaction amount, time, and anonymized principal components derived from PCA transformation. It consists of 284,807 transactions, with a significant class imbalance between fraudulent (492) and non-fraudulent (284,315) transactions.

## **2.3 Data Cleaning**

The data cleaning process involved several crucial steps to enhance the quality and reliability of the dataset:

1. **Handling Missing Values**: Any missing data points were addressed by imputing them with appropriate values (e.g., mean, median, or mode) or removing the corresponding records.
2. **Outlier Detection and Treatment**: Outliers, which could distort the model’s performance, were identified and either adjusted or removed. Techniques like z-score analysis or interquartile range (IQR) were employed.

# **3. Exploratory Data Analysis (EDA)**

## **3.1 Summary Statistics**

Descriptive statistics were computed to understand the central tendency and dispersion of features within the dataset. Additionally, the presence of missing values was examined, ensuring data integrity.

## **3.2 Visualization**

Data visualization techniques, such as histograms and correlation matrices, were employed to visualize patterns and relationships within the data.

## **3.3 Correlation Analysis**

Correlation analysis was conducted to identify potential predictors and explore relationships between variables.

# **4. Model Development**

## **4.1 Model Selection**

Two models were selected for comparison: logistic regression and RandomForestClassifier.

## **4.2 Model Training**

Both models were trained on the preprocessed dataset, with parameter tuning and cross-validation to optimize performance.

## **4.3 Model Evaluation**

The performance of each model was evaluated using metrics such as accuracy, precision, recall, and F1-score, demonstrating their effectiveness in detecting fraudulent transactions.

**Model Evaluation Results:**

### **Logistic Regression Model:**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 97.92% |
| Precision | 7.00% |
| Recall | 89.00% |
| F1-Score | 12.00% |

### **RandomForestClassifier Model:**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 99.96% |
| Precision | 93.00% |
| Recall | 85.00% |
| F1-Score | 89.00% |

# **5. Results and Discussion**

## **5.1 Model Performance**

Both models exhibited promising results in detecting fraudulent transactions. The **RandomForestClassifier** outperformed the logistic regression model across various evaluation metrics:

* **Accuracy**: RandomForestClassifier achieved an impressive **99.96%** accuracy.
* **Precision**: It demonstrated a precision of **93.00%**.
* **Recall**: The recall rate was **85.00%**.
* **F1-Score**: The F1-score stood at **89.00%**.

## **5.2 Key Findings**

The analysis revealed a clear advantage for the **RandomForestClassifier** over logistic regression. This underscores the importance of exploring diverse algorithms when tackling fraud detection tasks.

## **5.3 Limitations**

During the project, we encountered a few limitation:

1. **Class Imbalance**: Addressing class imbalance issues is crucial for robust fraud detection.

# **6. Conclusion and Recommendations**

## **6.1 Conclusion**

The **RandomForestClassifier** model demonstrated exceptional accuracy and effectiveness in credit card fraud detection, surpassing the logistic regression model.