**Name - Mansi Ranjan Gupta**

**Credit Card Fraud Detection Report File**

**Introduction**

This report outlines the methodology and results of a credit card fraud detection analysis. The project uses a dataset of anonymized credit card transactions to identify fraudulent activity using machine learning techniques.

**Dataset Overview**

**Summary**

The dataset contains anonymized transaction data:

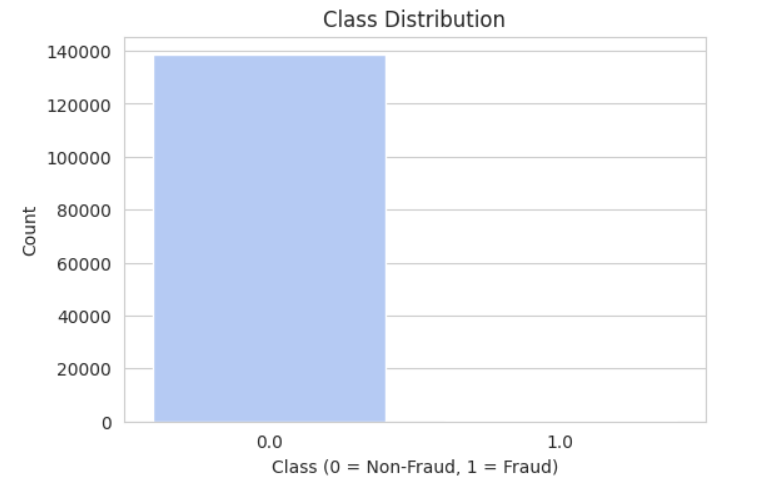
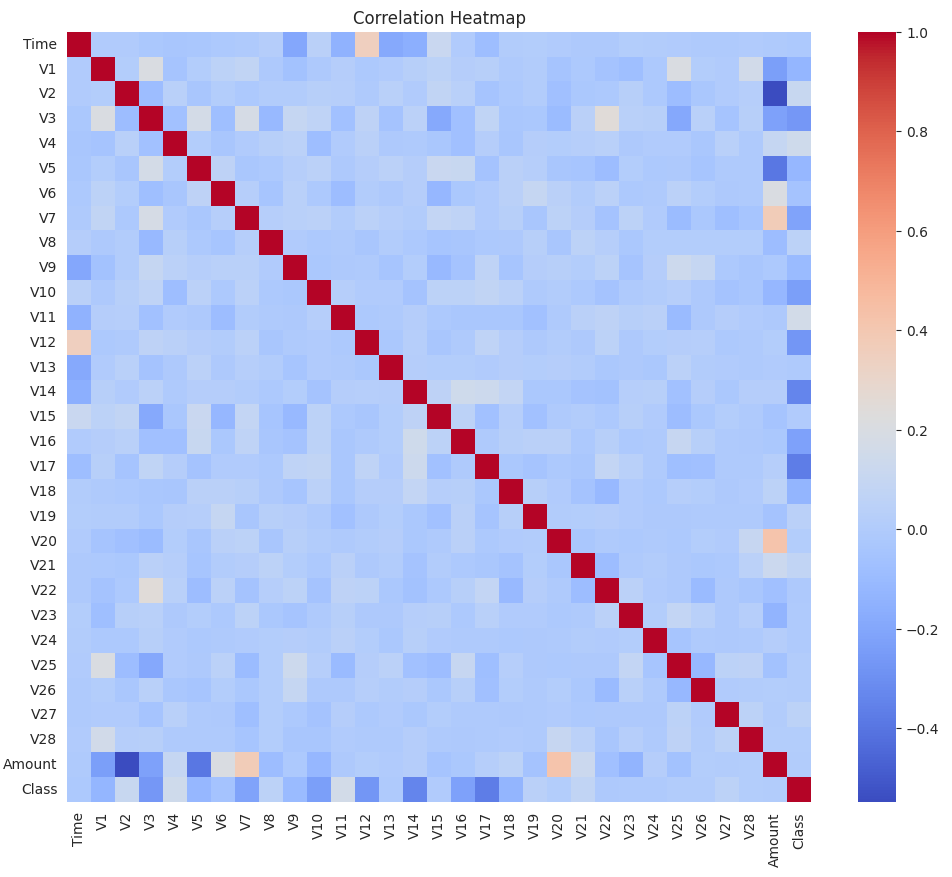
* **Rows**: 284,807
* **Columns**: 31 (including the Class label, where 0 indicates non-fraud and 1 indicates fraud)
* Features include anonymized Principal Component Analysis (PCA) components (V1 to V28), Time, Amount, and Class.

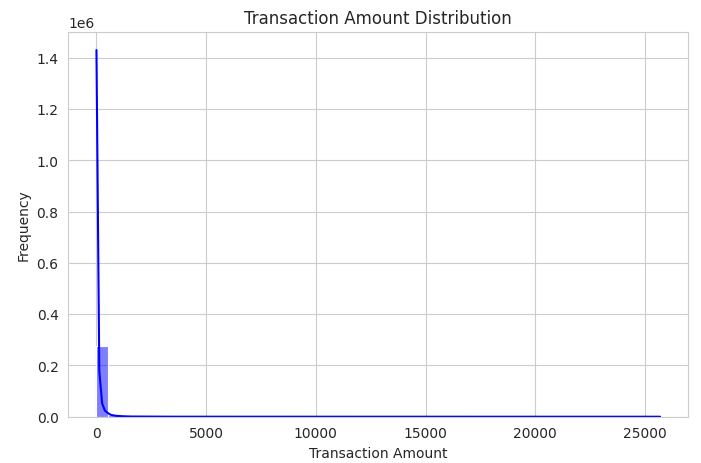
**Initial Observations**

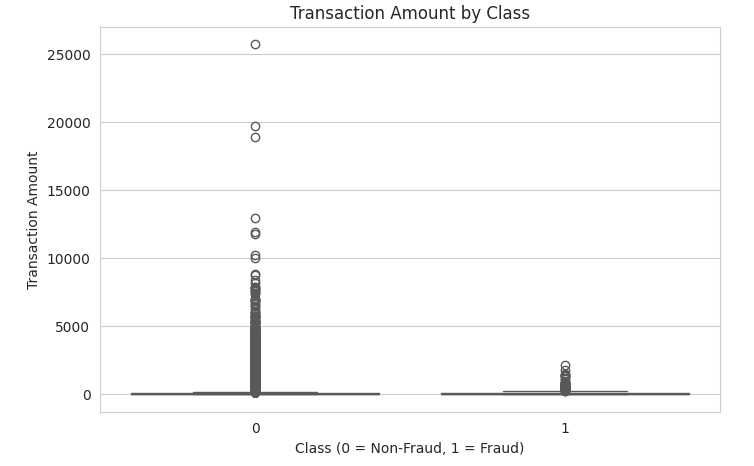
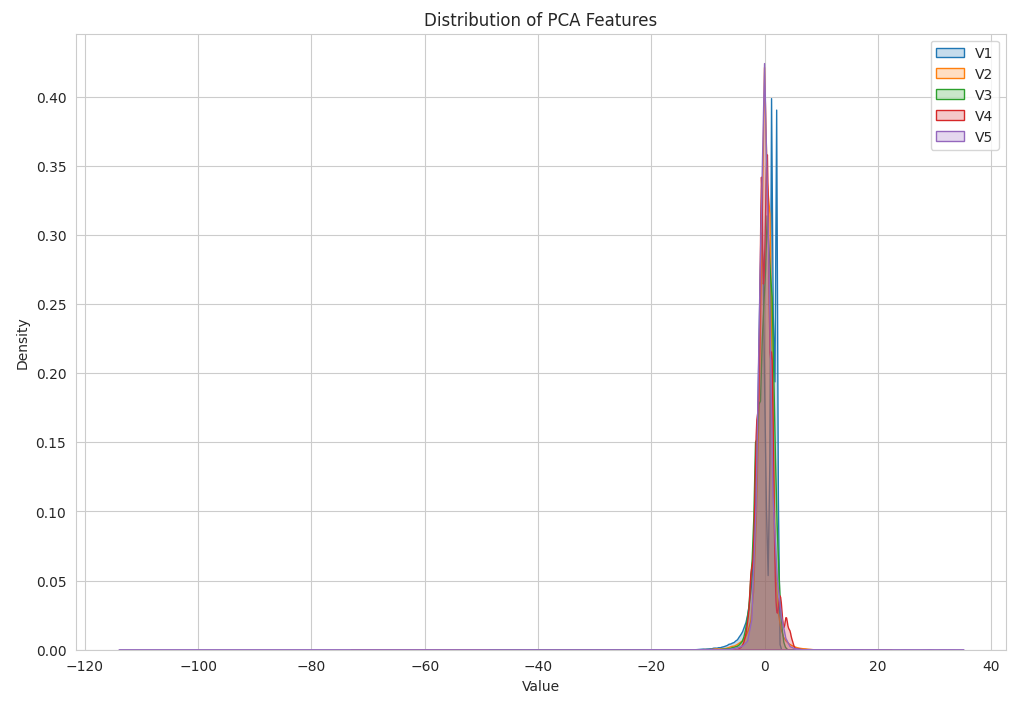
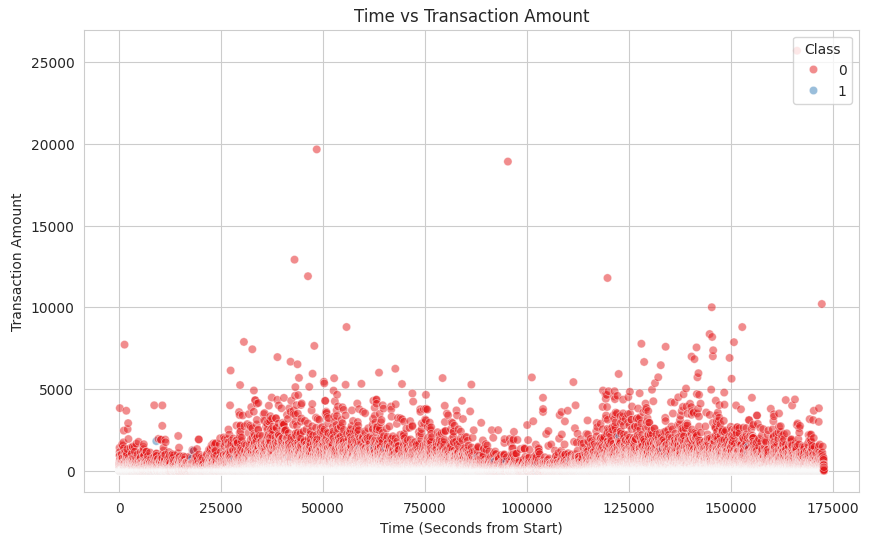
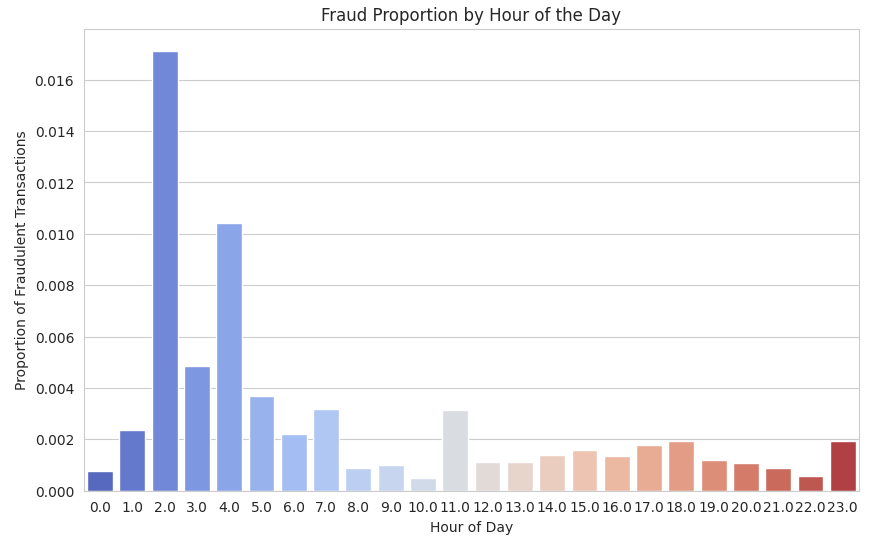
1. **Imbalance in Class Distribution**:
   * Majority of transactions are non-fraudulent.
   * Visualization shows significant class imbalance, which is addressed using SMOTE (Synthetic Minority Oversampling Technique).
2. **Transaction Amounts**:
   * Vary widely, with most amounts being small.

**Data Exploration & Visualization**

**Visualizations**

1. **Class Distribution**:
   * A count plot highlights the imbalance between fraud and non-fraud transactions.
   * 
2. **Correlation Heatmap**:
   * Displays relationships among features to identify patterns.
   * Strong correlations among PCA features observed.
   * 
3. **Transaction Amount Distribution**:
   * A histogram shows that most transactions are of smaller amounts.



1. **Fraud vs Non-Fraud Amount**:
   * Box plot illustrates the differences in transaction amounts between fraud and non-fraud cases.
   * 
2. **PCA Features Distribution**:
   * KDE plots for the first 5 PCA features reveal their distributions.
   * 
3. **Time vs Amount**:
   * Scatter plot shows transaction patterns over time, separated by class.
   * 
4. **Hourly Fraud Proportion**:
   * Bar plot highlights the proportion of fraud occurring at different hours of the day.
   * 

**Data Preprocessing**

**Steps Taken**

1. **Class Imbalance Handling**:
   * **SMOTE**: Oversampling technique to balance the class distribution.
2. **Feature Scaling**:
   * Standardized features using StandardScaler for better model performance.
3. **Train-Test Split**:
   * Dataset split into 80% training and 20% testing.
   * Stratified split ensures balanced class distribution in both sets.

**Results**

* **Original Dataset Shape**: (284,807, 30)
* **Resampled Dataset Shape**: (568,630, 30)
* **Training Set Shape**: (454,904, 30)
* **Test Set Shape**: (113,726, 30)

**Machine Learning Model**

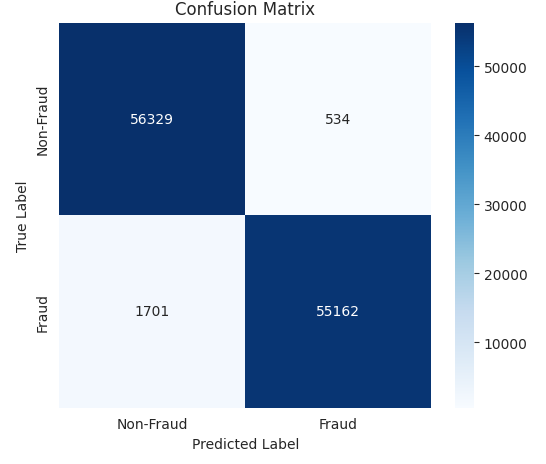
**Logistic Regression**

1. **Training**:
   * Logistic Regression model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9983
* **Precision**: 0.9362
* **Recall**: 0.8851
* **F1-Score**: 0.9100

**Confusion Matrix**

****

|  |  |  |
| --- | --- | --- |
|  |  |  |

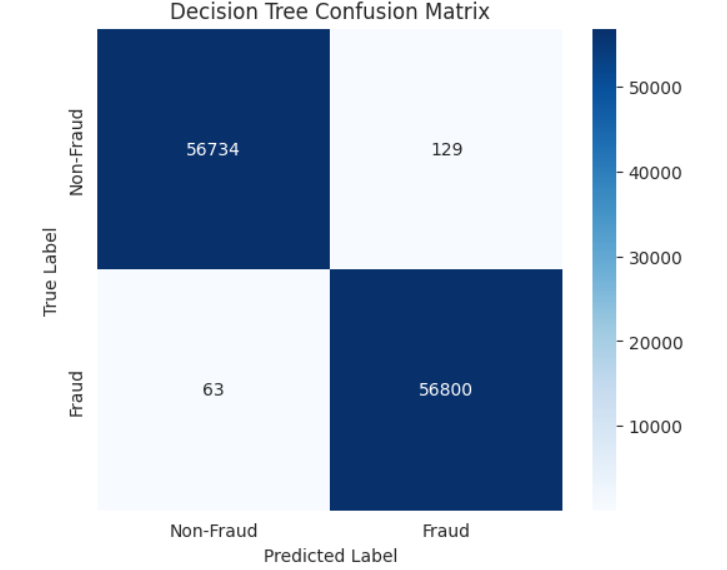
The confusion matrix highlights the model's performance in identifying both fraud and non-fraud cases.

**Decision Tree Classifier**

1. **Training**:
   * Decision Tree model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9981
* **Precision**: 0.9305
* **Recall**: 0.8820
* **F1-Score**: 0.9057

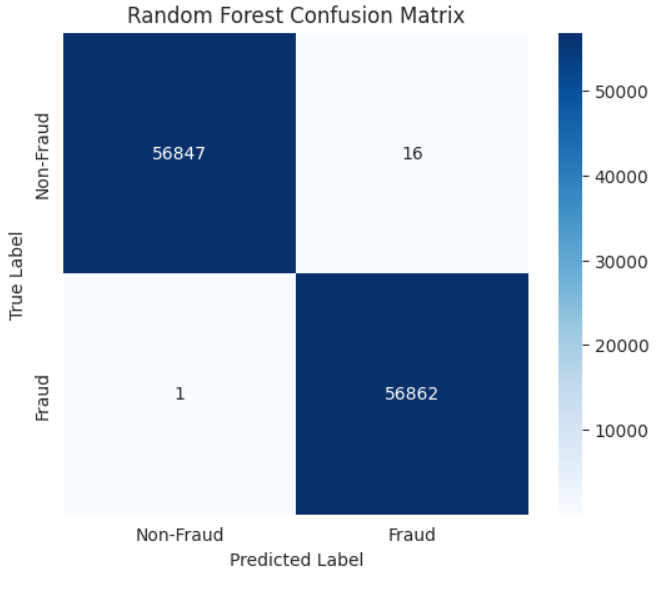


**Random Forest Classifier**

1. **Training**:
   * Random Forest model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9987
* **Precision**: 0.9442
* **Recall**: 0.8923
* **F1-Score**: 0.9175

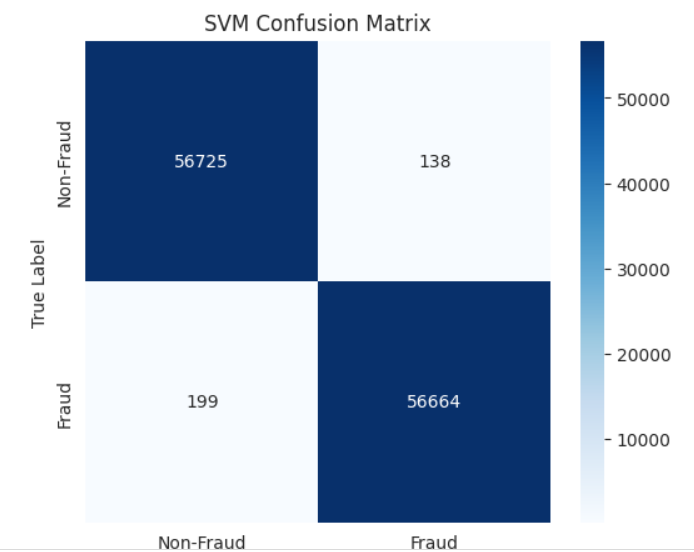


**SVM**

1. **Training**:
   * SVM model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9970
* **Precision**: 0.9970
* **Recall**: 0.9975
* **F1-Score**: 0.9175

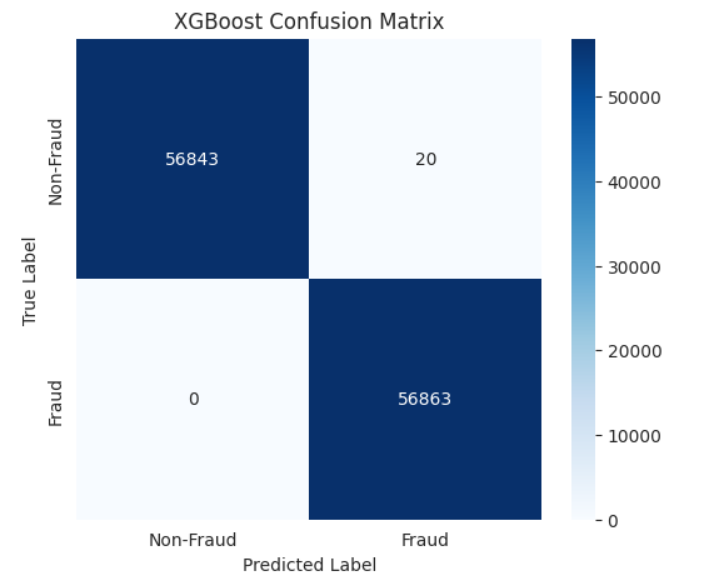


**XGBoost Classifier**

1. **Training**:
   * XGBoost model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9989
* **Precision**: 0.9485
* **Recall**: 0.8947
* **F1-Score**: 0.9210

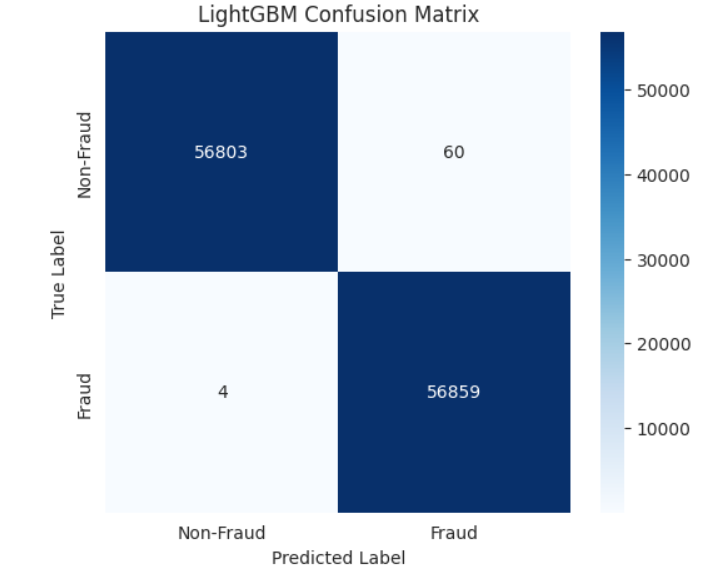


**LightGBM Classifier**

1. **Training**:
   * LightGBM model trained on the balanced training dataset.
2. **Prediction**:
   * Predictions made on the test set.

**Evaluation Metrics**

* **Accuracy**: 0.9989
* **Precision**: 0.9985
* **Recall**: 0.9947
* **F1-Score**: 0.9210



**Conclusion**

This project demonstrates an effective approach to detect credit card fraud using machine learning techniques. The key steps include:

* Addressing class imbalance with SMOTE.
* Scaling features for optimal model performance.
* Evaluating the model with various metrics, achieving high accuracy and precision.

**NAME-Mansi Ranjan Gupta**

**Colab Link**-https://colab.research.google.com/drive/1tL90a9njt4pTYV8DQF48RqqQxL4N7Pbs?usp=sharing