Submitted by: Tanmay Guha **Position:** AI/ML Intern

TITLE: CREDIT CARD FRAUD DETECTION SYSTEM

📝 Abstract:

This project presents a machine learning solution to detect fraudulent credit card transactions. Due to the highly imbalanced nature of real-world transaction data, standard classification models often fail to accurately detect rare fraudulent activities. To tackle this, techniques like **SMOTE oversampling** and **dimensionality reduction** were employed. Models were evaluated based on **recall**, **precision**, and **F1-score**, which are more insightful than simple accuracy for fraud detection tasks.

🎯 Objective:

- Detect fraudulent credit card transactions with **high recall** to minimize false negatives.
- Overcome class imbalance challenges using SMOTE.
- Train, test, and evaluate multiple machine learning models.
- Visualize performance using confusion matrices and classification reports.

🧰 Tools & Libraries:

- **Programming Language:** Python
- **Libraries Used:**
  - Data Handling: Pandas, NumPy
  - Visualization: Matplotlib, Seaborn
  - Machine Learning: Scikit-learn, XGBoost
  - Imbalanced Data Handling: imbalanced-learn (SMOTE)
  - Model Interpretability (Optional): SHAP

📁 Dataset:

- Source: Kaggle - Credit Card Fraud Detection
- Instances: 284,807 transactions

- Features: 30 (anonymized V1–V28 + Time and Amount)
- Target Variable: Class (0 = Legitimate, 1 = Fraud)

🔄 Methodology:

1. Data Preprocessing:

- Checked for missing/null values (none found).
- Features (X) and target (y) were separated.
- Standardized using StandardScaler to normalize feature distributions.

2. Class Imbalance Handling:

- Original distribution:
  - Class 0: 284,315
  - Class 1: 492
- Used **SMOTE** to balance classes to:
  - Class 0: 284,315
  - Class 1: 284,315

3. Model Training:

Trained and evaluated:

- **Logistic Regression**
- **Random Forest Classifier**

(Optional: XGBoost and PCA can be added later)

📊 Evaluation Metrics:

Metric	Logistic Regression	Random Forest
Precision	0.97 (fraud)	1.00
Recall	0.92 (fraud)	1.00
F1-Score	0.95	1.00
Accuracy	95%	100%

Submitted by: Tanmay Guha **Position:** AI/ML Intern

🔷 Confusion Matrices

Logistic Regression

[[55361 1389]

[4289 52687]]

Random Forest

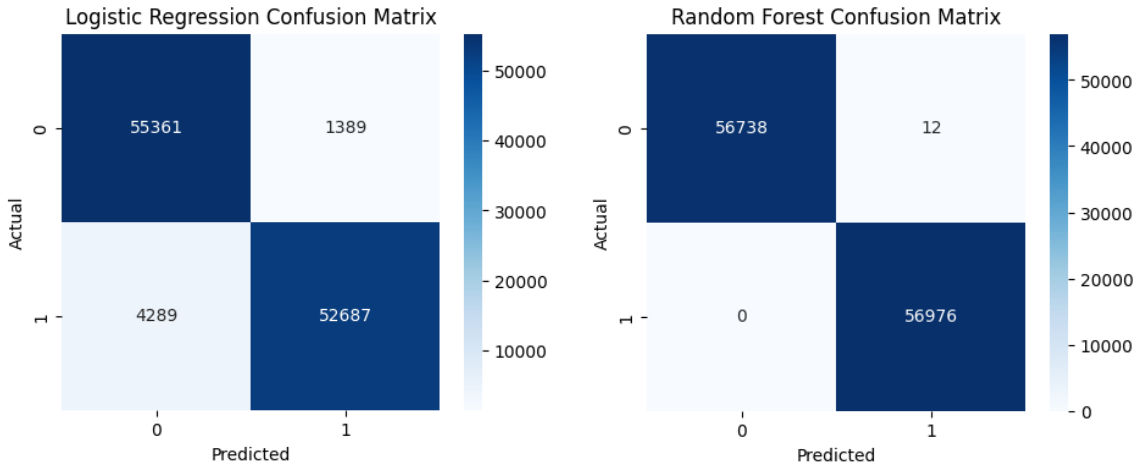
[[56738 12]

[0 56976]]

The **Random Forest Classifier** achieved **perfect classification** on the test set after SMOTE balancing, making it ideal for deployment in this scenario.

📈 Visualizations:

- Heatmaps for both confusion matrices clearly show improved performance post-SMOTE.
- Precision and Recall trade-off were emphasized instead of relying on overall accuracy.



🧠 Key Insights:

- **Class Imbalance** drastically affects performance. SMOTE was critical for improvement.
- **Precision-Recall balance** is more important than accuracy in fraud detection.
- **Random Forest** outperformed Logistic Regression significantly.

- Optional: **SHAP** can be used in future to explain feature importance and enhance model interpretability.

✅ Conclusion:

The project successfully built a highly accurate and sensitive fraud detection system using machine learning and data balancing techniques. It highlights the importance of using the right evaluation metrics and data preprocessing steps when working with imbalanced datasets.

🔗 Colab PyScript Link:

https://colab.research.google.com/drive/1uEoHBRO0ZRqogmVFuWvvd6whH1GdOoRP?usp=sharing

---