**Project Summary: Credit Card Fraud Detection**

**Objective:**

To detect fraudulent transactions from credit card data using machine learning techniques, improving the ability to differentiate between genuine and fraudulent behavior.

**Dataset Overview:**

* **Records:** ~284,807 transactions
* **Features:** 30 total (28 anonymized PCA features, Time, Amount)
* **Target Column:** Class (0 = Non-fraud, 1 = Fraud)
* **Highly imbalanced dataset**:
  + ~99.8% Non-fraud
  + ~0.2% Fraud

**Key Insights from Exploratory Data Analysis (EDA):**

1. **Class Imbalance:**  
   Fraudulent transactions are extremely rare (~0.17%). This requires special handling in modeling (e.g., SMOTE).
2. **Feature Scaling:**  
   Time and Amount needed scaling due to different distributions compared to the anonymized features.
3. **Correlation:**  
   Some features (like V14, V10, V17) showed strong correlation with the fraud class — useful signals for the model.
4. **Amount & Time Trends:**  
   No strong visual patterns in time or amount that distinguish fraud, confirming the need for multivariate analysis.

**Data Preprocessing:**

* Scaled Amount and Time using StandardScaler.
* Dropped original Time and Amount columns.
* Used **SMOTE** (Synthetic Minority Over-sampling Technique) to address class imbalance during training.

**Modeling Results:**

**i. Logistic Regression:**

* Simple and interpretable.
* Good baseline model.
* Missed some frauds, lower recall compared to RF.

**ii. Random Forest Classifier:**

* **Best performance among tested models.**
* High recall and precision.
* Detected frauds more accurately with fewer false positives.

**iii. Evaluation Metrics for Random Forest:**

* **Precision (Fraud):** High → few false positives
* **Recall (Fraud):** High → correctly identified most frauds
* **F1 Score:** Balanced trade-off between precision and recall
* **ROC AUC Score:** Close to **1.0**, indicating strong discriminative ability.

**ROC Curve Analysis:**

* The ROC curve for Random Forest was well above the diagonal (baseline), showing excellent performance.
* **AUC ≈ 0.99**: Model is very good at distinguishing between fraud and non-fraud.

**Conclusions:**

* **Random Forest** is highly effective for this dataset, especially when combined with SMOTE to handle class imbalance.
* With proper preprocessing and resampling, even traditional ML models perform well on fraud detection.
* Fraud detection systems should optimize for **high recall** (i.e., catching as many frauds as possible) while balancing false alarms.