💳🔍 Credit Card Fraud Detection Report

🎯 Objective

The goal of this project is to detect fraudulent credit card transactions using machine learning techniques. Early detection of fraud can save significant financial losses for both consumers and financial institutions.

📚 Dataset Used

The dataset consists of anonymized credit card transactions made by European cardholders over two days in September 2013. Important features include:

- **Time**: Seconds elapsed between this transaction and the first transaction in the dataset.
- **V1–V28**: Result of a PCA (Principal Component Analysis) transformation to protect confidentiality.
- **Amount**: Transaction amount.
- **Class**: Target variable where 0 = legitimate and 1 = fraud.

Target Variable:

- Class (0 for non-fraud, 1 for fraud)

🤖 Models Chosen

For this project, we used:

- **Logistic Regression** — as a baseline for binary classification.
- **RandomForestClassifier** — to enhance accuracy and manage class imbalance.

- **XGBoostClassifier** — to leverage advanced boosting techniques for better performance.

🔎 Methodology

1. **Exploratory Data Analysis (EDA)**
2. **Data Preprocessing** (handling missing values, resampling techniques like SMOTE)
3. **Feature Scaling** (StandardScaler for 'Amount' and 'Time')
4. **Model Building** (Training on different classifiers)
5. **Model Evaluation** (using multiple metrics)

📊 Performance Metrics

- **Accuracy**: Correctly predicted transactions over all transactions.
- **Precision**: Ratio of true frauds among all predicted frauds.
- **Recall**: Ratio of correctly detected frauds over all actual frauds.
- **F1-Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Visual evaluation of prediction errors.

📊 Model Evaluation Results:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**0.976**	**0.76**	**0.60**	**0.67**
Random Forest Classifier	**0.999**	**0.91**	**0.76**	**0.83**
XGBoost Classifier	**0.999**	**0.94**	**0.78**	**0.85**

Challenges Faced

- Extreme class imbalance (fraud transactions were less than 0.2% of all transactions).
- Avoiding overfitting on the minority class.
- Ensuring the model maintains high precision and recall, not just accuracy.

Learnings

- Learned techniques to handle imbalanced datasets (SMOTE, undersampling, class weighting).
- Gained deeper understanding of model evaluation beyond accuracy (precision, recall, F1-score).
- Understood the impact of feature scaling on algorithms like Logistic Regression.
- Built end-to-end fraud detection pipelines.
- Learned to deploy models using joblib for real-world applications.

Conclusion

This project successfully built an effective fraud detection system using Logistic Regression, Random Forest, and XGBoost classifiers. The models achieved strong evaluation metrics, with Random Forest and XGBoost delivering exceptional performance on unseen data. This solution can help in identifying fraudulent transactions early, ensuring better financial security and trust among customers and institutions.