Credit Card Fraud Detection

Made by: SAMAD KHAN

Problem Statement

Credit card fraud is a significant issue in the financial sector, leading to financial losses and eroding consumer trust. Detecting fraudulent transactions efficiently is crucial to minimizing risks and enhancing security. This project aims to develop a machine learning model capable of accurately identifying fraudulent credit card transactions while addressing the challenge of class imbalance.

Dataset Description

Source: creditcard.csv
The dataset contains credit card transactions made by European cardholders in September 2013. It includes **284,807** transactions, with only **492 fraudulent cases** (0.172% of the total). Due to confidentiality, features have been transformed using **Principal Component Analysis (PCA)** to anonymize sensitive data.

Key Features:

- **Time:** Seconds elapsed since the first transaction.
- **Amount:** Transaction amount.
- **V1-V28:** Anonymized numerical features derived from PCA.
- **Class:** Target variable (0 = Legitimate, 1 = Fraudulent).

Exploratory Data Analysis (EDA)

- **Checked for missing values and data consistency.**
- **Identified class imbalance** using visualizations.
- **Analyzed statistical distributions** of features.

Findings:

- No missing values in the dataset.
- The dataset is **highly imbalanced** (fraudulent transactions are extremely rare).
- Most features follow a **Gaussian distribution** due to PCA transformation.

Data Preprocessing

1. **Feature Scaling:**
   - Standardized the **Amount** feature using `StandardScaler`.
   - Dropped the **Time** feature as it does not contribute significantly to fraud detection.
2. **Handling Class Imbalance:**
   - Applied **Synthetic Minority Over-Sampling Technique (SMOTE)** to balance the fraudulent and non-fraudulent transaction classes.
3. **Train-Test Split:**
   - Split the dataset into **80% training and 20% testing** while preserving class distribution.

Feature Engineering

- Retained **PCA-transformed features** (V1-V28).
- Scaled the **Amount** feature for consistency.

Model Selection & Training

Several machine learning models were evaluated:

- **Logistic Regression**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**
- **XGBoost**

Best Performing Model:

- **Random Forest Classifier**
- Hyperparameter tuning was performed using **GridSearchCV** to optimize performance.

Hyperparameter Tuning

Used **GridSearchCV** to optimize key parameters:

- **n_estimators** = 100
- **max_depth** = None

Model Evaluation & Results

Metrics Used:

- **Accuracy:** Measures overall correctness.
- **Precision:** Measures how many detected frauds were actual frauds.
- **Recall (Sensitivity):** Measures how well frauds were detected.
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC Score:** Evaluates the classifier's ability to distinguish between fraudulent and non-fraudulent transactions.

Final Model Performance:

- **Accuracy:** 99.6%
- **Precision:** 92%
- **Recall:** 91%
- **F1-Score:** 91%
- **ROC-AUC Score:** 98%

Conclusion & Future Work

Key Takeaways:

- The **Random Forest Classifier with SMOTE balancing** was the most effective model.
- **Feature scaling and data preprocessing** significantly improved model performance.
- The **F1-score of 91%** indicates strong predictive capability.

Future Improvements:

- **Deploy the model** as a real-time fraud detection system using streaming data.
- Experiment with **deep learning models** such as **LSTMs and Autoencoders** for fraud detection.
- Develop a **cloud-based API** for real-time fraud detection in financial systems.

Project Deliverables

✅ **Jupyter Notebook** containing the source code.
✅ **Trained Model File** for further use.
✅ **Project Report (PDF)** detailing the methodology and results.

Acknowledgments

This project was developed as part of a **capstone initiative** to detect fraudulent transactions using machine learning.

🔗 **GitHub Repository:** [https://github.com/SamadK10/Capstone-Project.git]
