CREDIT CARD FRAUD DETECTION

Credit Card Fraud Detection is a machine learning project focused on identifying fraudulent credit card transactions from a real-world financial dataset. It leverages supervised classification techniques and data balancing strategies to distinguish between legitimate and fraudulent activities effectively.

PROBLEM STATEMENT AND REAL-WORLD RELEVANCE

Credit card fraud leads to billions in losses every year. The major challenge? **Class imbalance** —fraudulent cases are exceedingly rare. This project tackles that imbalance to accurately detect fraudulent activity, helping prevent financial losses and improve customer trust.

🎯 OBJECTIVE OF THE PROJECT

- Analyze transactional data
- Address class imbalance
- Train and evaluate classification models
- Derive business insights

DATASET DESCRIPTION

🔗 SOURCE

- [Kaggle – Credit Card Fraud Detection]

DATA SUMMARY

Feature Name	Description
V1–V28	PCA-transformed components
Amount	Transaction amount
Time	Time since first transaction
Class	Target variable: 0 = Legit, 1 = Fraud

⚠️ CLASS IMBALANCE

Class Label	Description	Count
0	**Legitimate**	**284,315**
1	**Fraudulent**	**492**

Only **0.17%** of transactions are fraudulent!

DATA PREPROCESSING

MISSING VALUES

- No missing or null values detected.

CLASS IMBALANCE HANDLING

- **Undersampling** applied:
  - Random sample from Class = 0 to match fraud count.
  - Created a **balanced** dataset for fair model training.

SCALING / ENCODING

- PCA components were pre-scaled.
- No categorical encoding required.

EXPLORATORY DATA ANALYSIS (EDA)

📊 INSIGHTS DISCOVERED

- Fraudulent transactions usually have **lower amounts**.
- Distribution was highly skewed toward legitimate transactions.
- Aggregated stats like mean and standard deviation compared for both classes.

FEATURE ENGINEERING

- No manual features added.
- All 28 PCA-transformed features retained.
- Amount included directly.

MODELING AND CLASSIFICATION

🧪 ALGORITHM USED

- ✅ **Logistic Regression**
  - Simple, interpretable, and effective for binary classification.

🛠️ TRAINING DETAILS

- **Train/Test Split**: 80/20
- **Stratified sampling**: Ensured class balance across splits
- **Iterations**: max_iter = 5000 to ensure convergence
- **Random State**: Used for reproducibility

MODEL EVALUATION

📊 PERFORMANCE METRICS

Metric	Training Set	Test Set
Accuracy	~95.5%	~93.9%
Precision	High	High
Recall	High	High

📌 Evaluation also included confusion matrix and predicted accuracy.

RESULTS AND BUSINESS INSIGHTS

- The model detected fraud **with high accuracy**.
- **Undersampling** proved effective but may oversimplify real-world imbalance.
- Logistic Regression offers **interpretability**, ideal for finance-related decisions.

LIMITATIONS AND HOW THEY WERE HANDLED

Limitation	Handling Strategy
Severe class imbalance	Undersampling of majority class
Lack of explainability	Chose Logistic Regression
No domain-based features	Preserved anonymized PCA features

TOOLS, LIBRARIES, AND FRAMEWORKS USED

Category	Tools Used
💻 Programming	Python
📓 Notebook	Jupyter Notebook
🛠️ ML Libraries	scikit-learn (LogisticRegression)
📊 Data Handling	pandas, numpy
📈 Evaluation	accuracy_score, confusion_matrix

CONCLUSION AND SUMMARY

This project establishes a solid baseline for detecting credit card fraud using Logistic Regression. It navigates class imbalance and achieves high accuracy through simple yet effective preprocessing. The framework is easily extensible for more complex use cases.

ABOUT ME

Saran S

Email: saranselvaraj2401sgmail.com

Github: github.com/saran2007s

linkedin:linkedin.com/in/saranselvaraj2401