Credit Card Fraud Detection using Logistic Regression

Author: Durga *Vani Moka*

Date: November 2025
Tools: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
Environment: Jupyter Notebook

Objective

The goal of this project is to build a machine learning model that accurately identifies fraudulent credit card transactions.
Since fraudulent transactions are extremely rare, the focus is on achieving a high recall (detect most frauds) while maintaining reasonable precision (minimize false alarms).

Problem Statement

Credit card companies process millions of transactions every day, and detecting fraud in real time is critical to prevent financial loss.
Fraudulent transactions represent less than 1% of total activity, making this a highly imbalanced classification problem.
Traditional accuracy metrics can be misleading, so the model must balance precision, recall, and overall interpretability.

Dataset Overview

- **Source:** Kaggle – "Credit Card Fraud Detection Dataset"
- **Shape:** 284,807 transactions × 31 columns
- **Target Column:** Class (0 = Legit, 1 = Fraud)
- **Features:**
    - V1–V28: anonymized numeric features (PCA transformed)
    - Time: seconds elapsed since the first transaction
    - Amount: transaction amount in euros

Class distribution:

Legitimate (0): 284,315 → 99.83%

Fraudulent (1): 492 → 0.17%

Strong imbalance — handled using Undersampling to balance the dataset.

Exploratory Data Analysis (EDA)

Dataset information:

- In Credit card fraud detection dataset there is null values
- Class =0 is legit transactions
- Class=1 is fraud transactions

Fraud vs Legit Comparison

Statistic	**Legit**	**Fraud**
Mean Amount	88.29	122.21
Median Amount	22.0	9.25
Std. Dev.	250.1	256.6

Fraud transactions show higher variability, indicating inconsistent patterns typical of suspicious activity.

Correlation:

The correlation analysis revealed that features V17, V14, V12, and V10 have strong negative correlation with fraud,

indicating they are the most influential predictors.

A few features like V11, V4, and V2 show weak positive correlation, contributing moderately.

Columns such as Amount, Time, and several PCA components (V19–V28) have negligible correlation,

implying they do not directly impact fraud prediction but may still offer minor support when combined with other features.

Overall, the dataset shows no multicollinearity, and the presence of both strong and weak predictors makes it well-suited for logistic regression modeling.

Features are PCA-transformed and uncorrelated, so I did not remove any features due to multicollinearity.

Outlier Detection:

Outlier detection was skipped in this project because fraudulent transactions themselves are outliers — rare and extreme by nature.

Removing them could lead to the loss of important fraud patterns.

Additionally, since the dataset's features (V1–V28) are PCA-transformed, the effect of extreme values is already minimized.

Instead of removing outliers, feature scaling (using StandardScaler or RobustScaler) was applied to handle data variations and maintain model stability for Logistic Regression.

Data Preprocessing

- Split data into 80% training and 20% testing (stratified to preserve fraud ratio).
- Standardized features using RobustScaler.

RobustScaler was used to scale numerical features, as it is less affected by outliers compared to StandardScaler or MinMaxScaler.
This helps stabilize model coefficients and prevents large transaction values from dominating training.

Model Building

Algorithm: Logistic Regression
Reason:

- Interpretable and computationally efficient
- Provides probabilistic outputs
- Works well as a baseline for fraud detection

Model Evaluation

Metric	**Legit (0)**	**Fraud (1)**
Precision	0.91	0.97
Recall	0.97	0.9
F1-score	0.94	0.93

Accuracy **0.93**

Interpretation:

- The model correctly identified **~**90**%** of frauds.
- Only **~**10 fraud cases were missed.
- Few legitimate transactions were incorrectly flagged (false positives).

Confusion Matrix

	Predicted Legit	**Predicted Fraud**
Actual Legit	97%	2%
Actual Fraud	12% missed	86% detected

ROC–AUC and PR–AUC

- **ROC–AUC:** 0.975 → excellent separation ability

- **PR–AUC:** 0.981 → strong performance even under imbalance

These curves confirmed the model's reliability and robustness.

Insights and Interpretability

- Logistic Regression coefficients revealed that certain anonymized variables (e.g., V14, V17) had higher positive weights, indicating stronger correlation with fraud likelihood.

Conclusion

The Logistic Regression model achieved **93% accuracy** with **balanced precision (89%) and recall (90%)** for fraud detection.
The model successfully detected most fraudulent transactions while minimizing false alerts.
With its interpretability and stability, it serves as a strong baseline model for credit card fraud detection.

This project demonstrates a **complete end-to-end data science workflow**:

Data Understanding → Feature Engineering → Model Building → Evaluation → Interpretation.

The chosen model balances simplicity, interpretability, and effectiveness, making it an ideal starting point for real-world fraud detection systems.
