Credit Card Fraud Detection Using Machine Learning

By
Krisi Doshi
Samriddhi Singh
Cas Wang
Jake Moscovitz

PROBLEM STATEMENT

Problem We Are Tackling:

- **Financial Losses**: Credit card fraud causes billions in losses annually, making it a critical concern.
- **Detection Challenges**: Fraudulent transactions are rare (<1%), making detection difficult, and traditional rule-based systems are no longer sufficient.
- **Operational Issues**: High false positives impact customer experience, while false negatives lead to financial risks, and real-time detection is hindered by massive transaction volumes.

Objectives

- **Develop Accurate Detection Models**: Build machine learning models that effectively identify fraudulent transactions.
- **Address Data Imbalance**: Tackle the class imbalance to improve detection performance.
- **Enhance Real-Time Detection**: Create scalable solutions that minimize false positives and handle high transaction volumes efficiently.

Dataset - creditcard.csv

Overview: Transactions from European card holders in 2013. Features broken down into Principal Components.

Columns: Time, Amount, Class, v1 - v28

Entries: 284,807 transactions (492 fraudulent)

Source: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download

Relevant or Prior Work

Feature Selection with Genetic Algorithms (GA)

- Optimizes feature subsets for improved fraud detection.
- Research shows GA enhances model accuracy (*Journal of Big Data*).

Addressing Class Imbalance with SMOTE

- Generates synthetic minority class samples to balance data.
- Proven to improve fraud detection rates in imbalanced datasets (*ArXiv Study*).

Ensemble Learning for Robust Detection

- Combines multiple classifiers (e.g., AdaBoost, XGBoost, Gradient Boosting).
- Enhances fraud identification while reducing false positives.

Integration of These Methods in Our Approach

- Leveraging feature selection techniques for feature optimization.
- Using SMOTE to handle imbalanced fraud data.
- Implementing ensemble models to improve detection accuracy and adaptability.

Approach

1. Step-1: Data Preprocessing - Load Dataset, Handle missing Values (if any), Extract useful feature from Time
2. Step-2: Implement ML Algorithms from Scratch, Logistic Regression, Decision trees, KNN, naive Bayes
3. Step-3: Train and Test Models, Split data into training (80%) and testing (20%) sets
4. Step-4: Implement an ensemble model, Combine multiple classifiers and evaluate performance
5. Step-5: Evaluate with Precision and Recall, Since data is imbalanced, accuracy is misleading

1. Logistic Regression:
    - Baseline classifier for binary classification.
    - Assumes a linear relationship between features and probability of fraud.
2. Decision Trees:
    - Helps analyze feature importance and detect complex patterns.
    - Recursively splits the data based on the most informative feature.
3. K-Nearest Neighbors (KNN):
    - Detects fraud by comparing transaction similarity.
4. Isolation Forest:
    - Unsupervised learning technique to detect rare fraud cases.
    - Assumes fraud transactions are outliers in high-dimensional space.
    - Isolation Forest identifies fraud by randomly partitioning the data.

Challenges

- Highly Imbalanced Dataset: Fraud cases are extremely rare, affecting training.
- Feature Engineering: Raw data is anonymized, requiring effective feature extraction.
- Balancing False Positives & Negatives: Preventing excessive blocking of legitimate transactions.