Asian Institute of Technology (AIT), Thailand

Computer Programming for DSAI

Assignment

Problem Statements and Datasets

Team Duo:

st125457: Ulugbek Shernazarov

st124997: Suryansh Srivastava

08/22/2024

Assignment

Problem statement and datasets

Problem

Fraud in financial transactions is a significant challenge for banks, payment processors, and e-commerce platforms. Fraudsters use sophisticated techniques to exploit vulnerabilities in financial systems, leading to substantial financial losses and undermining customer trust. The problem lies in identifying fraudulent transactions in real time, given the large volume of legitimate transactions and the evolving nature of fraudulent tactics.

Possible Users:

1. Banks and Financial Institutions: To monitor and prevent fraudulent activities in customer accounts.
2. E-commerce Platforms: To detect and block fraudulent transactions, protecting both the platform and customers.
3. Payment Processors: To ensure secure processing of transactions and reduce chargebacks due to fraud.
4. Insurance Companies: To identify and investigate potentially fraudulent claims.

Effective fraud detection can significantly reduce financial losses, protect users from unauthorized transactions, and maintain the integrity of financial systems. For customers, it builds trust and ensures the safety of their assets. For businesses, it minimizes reputational damage and operational costs associated with fraud, enabling them to operate securely in a digital economy. Additionally, early detection of fraud can lead to faster responses, mitigating potential damages and deterring future fraudulent activities.

Datasets description

1. Synthetic Fraud Detection Datasets for Fraud Detection - Kaggle link

The PaySim dataset is a synthetic dataset generated by the PaySim simulator, which models mobile financial transactions based on real-world financial behavior. The data simulates the interactions of users in a mobile payment system, with a small fraction of transactions labeled as fraudulent.

Number of records: 6,362,620 transactions

Number of features: 11 features

The dataset includes **time step**, **transaction type**, **amount**, **nameOrig** (origin account), **oldbalanceOrg** (initial balance before transaction), **newbalanceOrig** (new balance after transaction), **nameDest** (destination account), **oldbalanceDest** (initial destination balance), **newbalanceDest** (new destination balance), **isFraud** (indicates if transaction is fraudulent), and **isFlaggedFraud** (indicates if transaction was flagged as potential fraud).

Target is **isFraud**, where 1 indicates a fraudulent transaction and 0 indicates a non-fraudulent transaction.

Notes: The dataset simulates various transaction types and scenarios, making it an ideal testbed for fraud detection algorithms, especially in mobile money environments.

2. Credit Card Fraud Detection - Kaggle link

This dataset contains transactions made by European cardholders over two days in September 2013. The goal is to detect fraudulent credit card transactions. The dataset is highly imbalanced, with only 492 fraud cases out of 284,807 transactions (approximately 0.17% fraud cases).

Number of records: 284,807 transactions

Number of features: 31

The dataset includes 28 anonymized features labeled **V1** to **V28**, which are the result of a Principal Component Analysis (PCA) transformation. The two remaining features are **time** and **amount**.

The **class** feature is a target (0 or 1).

Notes: The time feature contains the seconds that elapsed between this transaction and the first transaction in the dataset. The amount feature shows the transaction amount.

3. IEEE-CIS Fraud Detection - Kaggle link, github link

This dataset was created for a Kaggle competition by IEEE and the Consumer Identity Solutions (CIS) team. It contains a large set of anonymized transactions from an e-commerce platform, to detect fraud in these transactions. The data includes transactions with both fraudulent and non-fraudulent labels, and it is highly imbalanced.

Number of records: approximately 1.1 million transactions

Number of features: 433 features (two tables - transaction and identity)

The dataset contains a mix of numerical and categorical features, including information related to the payment, such as **transaction times, amounts, payment methods, card information**, and **user/device** information. Due to privacy reasons, most features are anonymized and labeled generically.

Target is **isFraud**, where 1 indicates a fraudulent transaction and 0 indicates a non-fraudulent transaction.

Notes: The dataset is split into two files: train_transaction.csv (transaction data) and train_identity.csv (identity data). These can be merged on the TransactionID.

Justification

1. Synthetic Financial Datasets for Fraud Detection (PaySim)

Interesting transactions: This dataset simulates various types of financial transactions, reflecting a wide range of scenarios where fraud might occur.

Timestamp: The step feature captures the time dimension, which is critical for identifying suspicious activity over time (e.g., multiple high-value transactions in a short period).

Account Behavior: By tracking both the origin (nameOrig) and destination (nameDest) accounts, along with their balances before and after transactions, the dataset allows models to detect anomalies in account behavior that are indicative of fraud.

The PaySim dataset provides a realistic simulation of a mobile money environment, with enough diversity in transaction types and account behaviors to effectively train and validate fraud detection models.

2. Credit Card Fraud Detection (Kaggle)

Feature Amount: The dataset contains 31 features, with 28 anonymized features derived from a principal component analysis (PCA). These features capture complex patterns in the data, making them suitable for identifying subtle correlations that could indicate fraudulent behavior.

Imbalance Awareness: Although the dataset is highly imbalanced (only 0.17% of the transactions are fraudulent), this mirrors real-world conditions where fraud is rare. Techniques like anomaly detection methods can be applied to ensure effective model training.

Transaction Context: The dataset includes the transaction amount and time, providing additional context that is essential for detecting abnormal transaction patterns.

The combination of anonymized principal components, transaction amount, and time provides a comprehensive view of the transaction, enabling models to identify fraudulent behavior with high precision.

3. IEEE-CIS Fraud Detection (Kaggle)

Feature's Amount: With 433 features, this dataset offers a vast amount of information, including details about payment methods, device types, and user identity. This diversity allows for the capture of intricate patterns associated with fraud.

Real-world Data: The dataset is derived from an actual e-commerce platform, simulating a real-world environment where fraud detection models will be deployed. This makes it ideal for training models that need to generalize well to unseen data.

The large number of features, combined with the realism of the data, provides a robust foundation for training models that can effectively differentiate between fraudulent and legitimate transactions.