COMPSCI 4AL3 - Project Proposal

Fall 2025

Group 34

Team Members

- Viransh Shah
- Ellen Xiong

Task Title: Fraud Detection in E-Commerce Transactions

Financial fraud is one of the most costly and persistent challenges faced by e-commerce platforms and financial institutions. Our project focuses on developing a machine learning model to detect fraudulent transactions. We will use a synthetic e-commerce dataset specifically designed for fraud detection, which provides rich features such as customer demographics, device information, transaction time, and location. The significance of this project lies in the real-world applicability, fraudulent transactions cause billions of dollars in losses annually. Detecting fraud is challenging due to; class imbalance (fraudulent transactions are rare), high-dimensional features (transaction amount, location, account age, device, etc.), and adaptive adversaries (fraudsters constantly change behavior).

Task Definition

- Type of data: Tabular data of e-commerce financial transactions.
- Learning type: Supervised classification.
- Number of classes: 2 (Fraudulent = 1, Legitimate = 0).
- Label type: Single-label binary classification.

Problem Description & Real-World Impact

The problem we aim to solve is determining whether a given e-commerce financial transaction is legitimate or fraudulent. Some examples of fraudulent transactions are credit card fraud, identity theft, refund fraud, card testing, and chargeback fraud. Financial fraud detection can be an extremely helpful application of machine learning. Even a small improvement in fraud detection rates can save institutions millions of dollars, while also protecting customers from identity theft and financial loss. To solve this problem, there are a few challenges that we have identified. A big challenge is the sparse data on fraudulent transactions (~5% of the synthetic dataset, <1% in real-world credit card data), so the amount of data for both classes are imbalanced. It will also be challenging to capture complex feature interactions, since suspicious behavior often emerges from unusual combinations rather than a single variable. Additionally, training an accurate model will be difficult, as models trained on one dataset may not generalize well to the real-world due to shifting fraud patterns across platforms and regions.

Data Sources and Collection Plan

We will use the Fraudulent E-Commerce Transactions Dataset available on Kaggle: https://www.kaggle.com/datasets/shriyashjagtap/fraudulent-e-commerce-transactions

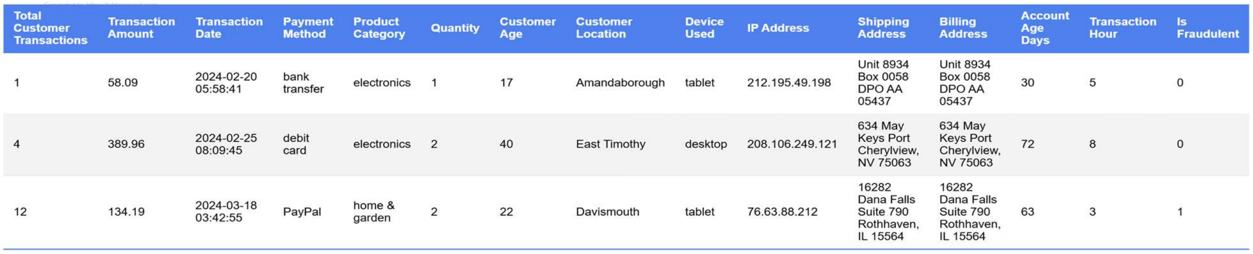
- Original Source: Created by Shriyash Jagtap using Python's Faker library and custom fraud-simulation logic.
- Dataset Versions: 1,472,952 rows (v1), 23,634 rows (v2).
- Features: 16 fields per transaction.
- Labels: Provided directly in the dataset via the Is Fraudulent column (~5% fraud rate).
- Collection Method: Fully synthetic data. No real individuals are represented; data mimics realistic e-commerce and fraudulent patterns using Faker + custom rules.

Our Plan:

- Download the dataset directly from Kaggle, following its licensing terms.
- Start with v2 for rapid prototyping, then scale to v1 for final training and evaluation.
- Use most features but may drop high-cardinality fields (e.g., raw addresses, IPs) if they do not improve model performance.
- No scraping, API access, or manual labeling is needed, as the dataset is already labeled.

Expected Dataset Size and Example

- Number of data points: 1,472,952 rows (v1), 23,634 rows (v2).
- Total dataset size: 392.82 MB.



Total Customer Transactions	Transaction Amount	Transaction Date	Payment Method	Product Category	Quantity	Customer Age	Customer Location	Device Used	IP Address	Shipping Address	Billing Address	Account Age Days	Transaction Hour	Is Fraudulent
1	58.09	2024-02-20 05:58:41	bank transfer	electronics	1	17	Amandaborough	tablet	212.195.49.198	Unit 8934 Box 0058 DPO AA 05437	Unit 8934 Box 0058 DPO AA 05437	30	5	0
4	389.96	2024-02-25 08:09:45	debit card	electronics	2	40	East Timothy	desktop	208.106.249.121	634 May Keys Port Cherylview, NV 75063	634 May Keys Port Cherylview, NV 75063	72	8	0
12	134.19	2024-03-18 03:42:55	PayPal	home & garden	2	22	Davismouth	tablet	76.63.88.212	16282 Dana Falls Suite 790 Rothhaven, IL 15564	16282 Dana Falls Suite 790 Rothhaven, IL 15564	63	3	1

Proposed Solution

We plan to solve this problem by treating the model as a binary classification problem, where there are two classes: yes/true and no/false, that categorize whether a transaction is fraudulent or not. The first key step we plan to take is to set up the dataset for our problem/model by splitting the data appropriately, hooking up the data to our code, and creating an additional feature variable. We plan to create a feature variable called “Total Customer Transactions” that is based on the existing columns “Transaction ID” and “Customer ID” in the dataset, where it indicates the total number of transactions (Transaction IDs) a customer (Customer ID) has made. The next key steps we plan to take are to preprocess the data by normalizing/standardizing the features. Afterwards, we will find the best model for our problem by evaluating our data with loss functions and experimenting with hypertuning. Once we have trained our model, we will evaluate it by using the test data we prepared and generalization. The models we plan to try are logistic regression, a linear model that handles binary classification, and random forest, an algorithm that combines the predictions of multiple decision trees for more accurate results. This decision was based on existing solutions to the problem of detecting transaction fraud, where we found examples of models that used logistic regression and random forest. Our target label would therefore be “yes” and “no”, indicating whether an e-commerce transaction is considered fraudulent (“Is Fraudulent”). There are 14 feature labels, and they are “Total Customer Transactions”, “Transaction Amount", "Transaction Date”, “Payment Method”, “Product Category”, “Quantity”, “Customer Age”, "Customer Location”, “Device Used”, “IP Address”, “Shipping Address”, “Billing Address”, “Account Age Days”, “Transaction Hour”. We will know if the model is good if it is able to accurately identify whether a given transaction (given 15 feature variables) is fraudulent. We will evaluate our model by letting it predict on a set of test data and comparing the model’s predicted target value with the true target value using a loss function.

The libraries we intend to use for this project are; pandas, numpy.

Proposed Solution Sources:
https://www.sciencedirect.com/science/article/pii/S2665917424001144
https://link.springer.com/article/10.1186/s40537-022-00573-8
https://www.sciopen.com/article/10.26599/BDMA.2023.9020023