Fraudulent e-commerce dataset

Yuran Belane

April 30, 2024

Introduction

In the rapidly evolving world of e-commerce, the security of transactions has become a paramount concern. With the increasing sophistication of fraudulent activities, it is crucial to develop robust and reliable fraud detection systems. This report presents a comprehensive study conducted on the synthetic dataset, "Fraudulent E-Commerce Transactions."

The dataset, which is available at Kaggle platform in two versions, is meticulously designed to simulate transaction data from an e-commerce platform, with a particular emphasis on fraud detection. It encompasses a wide array of features typically found in transactional data, providing a realistic representation of e-commerce activities. Furthermore, the dataset includes additional attributes specifically engineered to bolster the development and testing of fraud detection algorithms.

This study aims to delve into this dataset, analyze its characteristics, and explore its potential in enhancing fraud detection techniques. By doing so, I hope to contribute to the ongoing efforts in fortifying the security of e-commerce transactions and mitigating the risks associated with fraudulent activities.

1. Dataset Overview

The “Fraudulent E-Commerce Transactions” dataset is a synthetic dataset meticulously designed to simulate transaction data from an e-commerce platform. The dataset comprises 23,634 transactions and 16 distinct features, providing a comprehensive and realistic representation of e-commerce activities. Approximately 5% of the transactions in the dataset are labeled as fraudulent, offering a substantial sample for studying fraudulent activities.

1.1 Feature Details

The features in the dataset include:

Transaction ID: A unique identifier for each transaction.

Customer ID: A unique identifier for each customer.

Transaction Amount: The total amount of money exchanged in the transaction.

Transaction Date: The date and time when the transaction took place.

Payment Method: The method used to complete the transaction (e.g., credit card, PayPal, etc.).

Product Category: The category of the product involved in the transaction.

Quantity: The number of products involved in the transaction.

Customer Age: The age of the customer making the transaction.

Customer Location: The geographical location of the customer.

Device Used: The type of device used to make the transaction (e.g., mobile, desktop).

IP Address: The IP address of the device used for the transaction.

Shipping Address: The address where the product was shipped.

Billing Address: The address associated with the payment method.

Is Fraudulent: A binary indicator of whether the transaction is fraudulent (1 for fraudulent, 0 for legitimate).

Account Age Days: The age of the customer’s account in days at the time of the transaction.

Transaction Hour: The hour of the day when the transaction occurred.

1.2 Generation Method

The dataset is intended for use in developing and testing machine learning models for fraud detection in e-commerce transactions. It can also be used for exploratory data analysis, feature engineering, and benchmarking fraud detection algorithms.

The data is entirely synthetic, generated using Python’s Faker library and custom logic to simulate realistic transaction patterns and fraudulent scenarios. The dataset is not based on real individuals or transactions and is created for educational and research purposes.

2. Insights from the Fraudulent E-Commerce Transactions Dataset

Our analysis of the “Fraudulent E-Commerce Transactions” dataset revealed a total of 1,222 suspicious transactions that were potentially fraudulent. These transactions were observed across four primary payment methods: Bank Transfer, Credit Card, Debit Card, and PayPal.

Bank Transfer was associated with 326 fraudulent cases, resulting in sales of approximately 1.34 million. This was closely followed by Credit Card transactions, which accounted for 301 fraudulent cases and sales of around 1.35 million. Debit Card transactions were linked to 285 fraudulent cases, with sales totaling 1.36 million. Lastly, PayPal was used in 310 fraudulent cases, leading to sales of about 1.37 million.

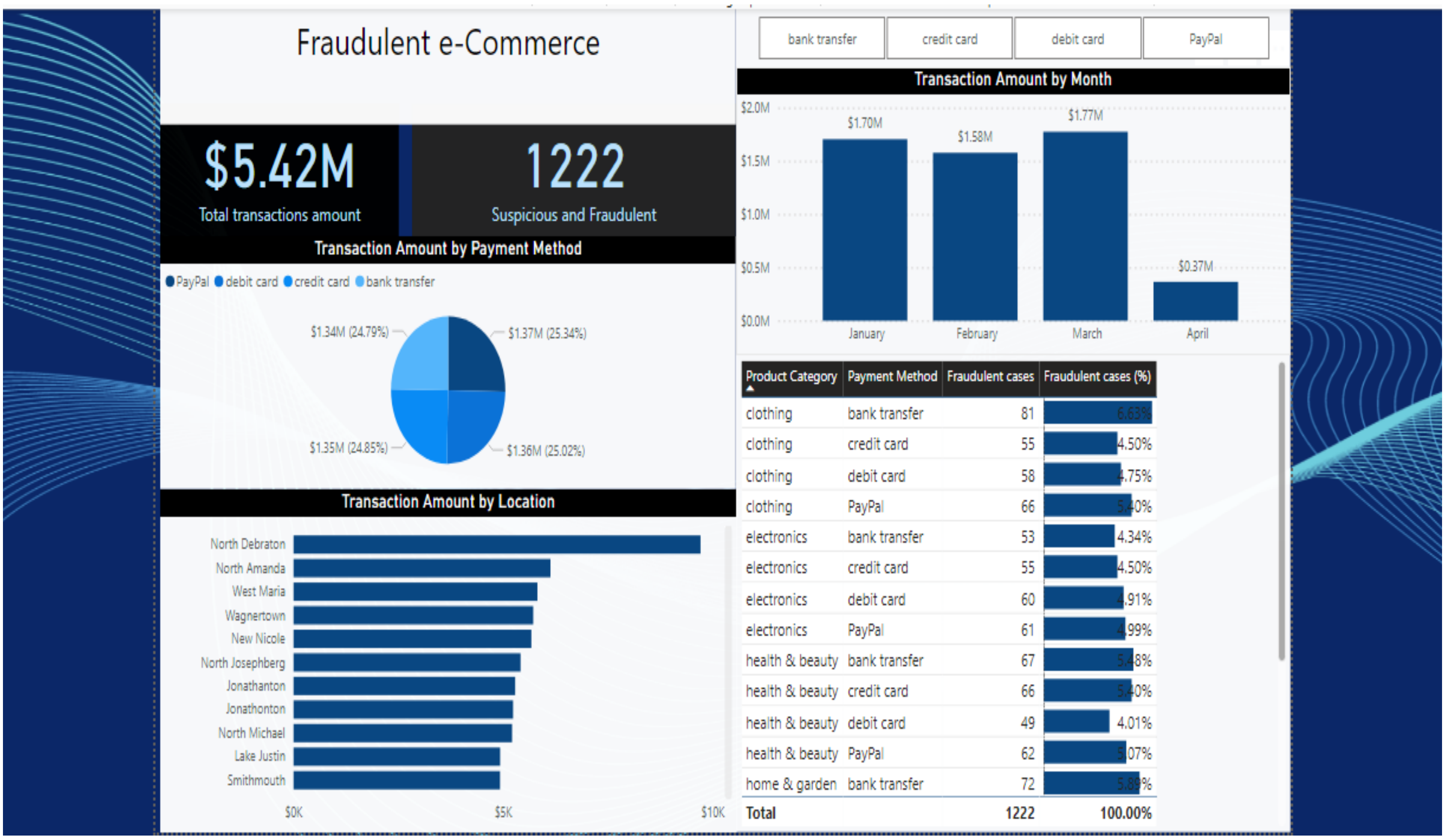
These findings highlight the pervasive nature of fraudulent activities across all major payment methods in e-commerce transactions. It underscores the need for robust and comprehensive fraud detection systems that can effectively monitor and flag suspicious activities across diverse payment platforms.

2.1 Why Most Fraudulent Cases Happen in Payment Methods

Fraudulent activities are often concentrated in payment methods due to several reasons. Firstly, payment methods are the final and crucial step in the completion of a transaction, making them a prime target for fraudsters. Secondly, each payment method has its unique vulnerabilities that can be exploited by fraudsters. For instance, credit and debit cards can be cloned or stolen, bank account details can be hacked, and PayPal accounts can be breached.

Moreover, the anonymity provided by certain payment methods, such as PayPal, can be attractive to fraudsters. Also, the speed and convenience of online transactions, which are typically seen as advantages, can also facilitate fraudulent activities by making it easier for fraudsters to carry out their operations before they are detected.

Dashboard



Product Category	Payment Method	Fraudulent cases	Fraudulent cases (%)
clothing	bank transfer	81	6.63%
clothing	credit card	55	4.50%
clothing	debit card	58	4.75%
clothing	PayPal	66	5.40%
electronics	bank transfer	53	4.34%
electronics	credit card	55	4.50%
electronics	debit card	60	4.91%
electronics	PayPal	61	4.99%
health & beauty	bank transfer	67	5.48%
health & beauty	credit card	66	5.40%
health & beauty	debit card	49	4.01%
health & beauty	PayPal	62	5.07%
home & garden	bank transfer	72	5.89%
Total		**1222**	**100.00%**

Credits

Dataset from Kaggle by Shriyash Jagtap: Fraudulent E-Commerce Transactions (kaggle.com)