**Dataset source**

The data was extracted from the Kaggle website - e-commerce fraud dataset. This is a synthetic dataset that is designed to simulate transaction data from an e-commerce platform with a focus on fraud detection. Version 1 of the data set, which has a labeled target: **Is fraudulent** was used in the data analysis.

**Dataset source:**

https://www.kaggle.com/datasets/shriyashjagtap/fraudulent-e-commerce-transactions?resource=download

**Dataset Overview**

* **Number of Transactions**: 1,472,952
* **Features**: 16
* **Fraudulent Transactions**: Approximately 5%

**Hypothesis Testing**

The objective of the paper is to test two hypotheses:

A high transaction amount is more likely to be fraudulent.

Less account age is more likely to be fraudulent.

**Preprocessing steps**

**1. Data reduction**

The file widget revealed that there were no missing values in the data. However, with over one million transaction data, it was not possible to use all the data for fraud detection due to limitations in the computational power of the computer. Therefore, the data size was reduced to ten percent of the original size(7365 instances) using the data set widget.

**2. Role assignment**

Using the File widget, the role and type of each variable were assigned as follows:

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Role** |
| Is Fradulent | categorical | target |
| Transaction Amount | numeric | feature |
| Transaction Date | datetime | feature |
| Payment Method | categorical | feature |
| Product Category | categorical | feature |
| Quantity | numeric | feature |
| Customer Age | numeric | feature |
| Device Used | categorical | feature |
| Account Age Days | numeric | feature |
| Transaction Hour | numeric | feature |
| Transaction ID | text | meta |
| Customer ID | text | meta |
| Customer Location | categorical | feature |
| IP Address | categorical | feature |
| Shipping Address | categorical | feature |
| Billing Address | categorical | feature |

Here, **Is Fraudulent** is set as the target variable. The Transaction ID, Customer ID has been assigned as meta because they are unique identifiers, making it irrelevant to making predictions.

**3. Rank evaluation**

All variables in the dataset might not be useful in building the prediction model. It is important to select variables that are important because using redundant variables for prediction reduces the generalizability of the model. (Aman, 2024) Given this, the importance of each independent variable was examined using the Rank widget. Information gain and gini decrease are two scoring metrics that were used to evaluate the importance of the independent variables. Information gain quantifies the reduction in entropy as a result of splitting the decision tree nodes. Similarly, Gini decrease measures the reduced probability of misclassifying the target in a decision tree. Another two methods used to calculate the importance of independent variables were Chi-square statistics and ReliefF. Chi-square statistics quantifies the importance of the relationship between variables by comparing the observed frequencies with how the model compares to actual observed data. On the other hand, the ReliefF algorithm selects the nearest neighbor samples from each of the samples in different categories, which is repeated several times to get the weighted importance score.

Interestingly, Billing Address, Shipping Address, IP Address, Account Age Day, and Customer Location were ranked as the five most important independent variables for prediction. When different scoring metrics were used(information gain, gini decrease, Chi-square, ReliefF), the order of the variables, which are sorted based on the importance score, changed. Despite the change in order, they consistently remained as the five most important variables. However, when predictions are made based on the variables, the prediction performance is poor. Even when different machine learning algorithms (Tree, Logistic Regression, Random Forest, kNN) are employed, the confusion matrix revealed that the true positive is 0% for each case. These findings indicate that the variables are not helpful for prediction purposes. One explanation could be that the Rank widget measures the importance of each variable in isolation. However, in reality, whether a transaction is fraud or not depends on the interaction of many different variables. Similarly, categorical variables > 100 unique values may decrease prediction performance because of overfitting. Given the limitations of one-hot encoding for high cardinality variables, we assigned Type: text to Billing Address, Shipping Address, IP Address, and Customer Location.

The Sieve Diagram widget was used to evaluate the significance of each variable based on a 5% significance level, which revealed only the following are significant:

|  |  |
| --- | --- |
| **Name** | **p-value** |
| Account Age Days | 0.000 |
| Transaction Hour | 0.000 |
| Transaction Amount | 0.000 |