**Data Analysis and Insights Generation for Fraud Detection**

Data analysis and insights generation constituted the fifth critical component of the fraud detection project. During this phase, patterns were extracted from transactional data that have business implications and have consequences on identifying and mitigating fraudulent activity. We were trying to convert raw transactional data into intelligence that can help us make better decisions (in real time) and quickly detect likely fraud.

The first stage of data analysis was the exploratory data analysis (EDA) stage. When we think of data manipulation, we should immediately search for the closest data scientist to help with the task, but the fact is that EDA is an integral part of understanding the data’s structure, finding anomalies and properly preparing it for further statistical analysis. In this phase, we applied data cleanings techniques such as removing missing values, duplicates, etc. Before we could proceed further, it was crucial that the data remain intact as double transactions or missing data would skew the results.

I cleaned the data and identified key metrics used to detect fraudulent transactions. Customer transaction behavior of transaction frequency, total spending and refund rates were analyzed. I wanted to find patterns that could suggest unusual activity – the precursor to fraud. A suspect transaction might be flagged if it had unusually high purchase amount, frequent refunds or if multiple high value items were bought. Further, these flagged transactions were subjected to additional analysis to see to what extent they matched with fraudulent practice patterns.

It also featured a particularly insightful analysis of refund behavior. A common fraud indicator, but one that occurs right after a purchase, is the refund. They then did a detailed analysis of refund rates by product and segment of customer. On the basis of this analysis, I was able to show that some products had a higher-than-expected refund rate, and some customers seemed to have a pattern of buying something then returning the product very quickly. Further, these behaviors were investigated as potential fraud indicators.

We also analyzed transaction velocity in order to detect "carding" fraud wherein, fraudsters make a bunch of small transactions to validate a list of stolen credit card details. The analysis followed the frequency of number of transactions customers made over a short period of time. High risk was marked as rapid purchases or multiple orders from the same account in a short period. This made the system automatically flag suspicious accounts for manual review and thus eliminates further fraudulent transactions.

Analysing product categories and the combinations of items also provided another critical insight. Fraudsters are known to go after specific products with a high resale value or are easy to sell. Based on that, the team analyzed purchasing patterns to pick out products that had a greater chance of being part of fraudulent transactions. For example, fraud cases tended to disproportionately contain high value electronics or luxury items. This informed the team to quadruple security efforts around particular item types and to observe more exacting verification standard at high value exchanges.

Forecasting future fraudulent activities, based on historical data, a key element in the process involved predictive modeling. Decision trees and random forests were used to train machine learning algorithms which adapt to predictions regarding the likelihood of a given transaction being fraudulent. These models were trained on transaction data such as amount, refund rate, frequency and customer demographics. The models generated probability for each transaction allowing proactive identification of fraudulent transactions before they did a lot of damage.

Charts and graphs that made the results easier to interpret for stakeholders were created by the team to communicate the analysis results. An important visualization was a heat map that showed where the regions had high concentration of fraudulent transactions. One other important visualization was a scatter plot plotting transaction amount to refund rate, aimed at showing outliers that may point to fraudulent behavior.

Fraud detection was also driven by customer segmentation. The team grouped customers by characteristics like purchasing behavior, transaction volume and refund history to discover segments more likely to be involved in fraudulent activity. For example, they found cases where customers who had a history of making high value purchases, or who frequently returned items, were more likely to be involved in fraud. These customers were marked as high risk and were kept closer under watch.

The analysis also studied seasonal patterns of fraudulent activity. Because there are known periods when fraudulent transactions spike (think holiday season, Black Friday, Cyber Monday, etc) without a penalty uplift for your DSA contract, a competitor could pay less and have the potential to enjoy an easier path to profitability. Using historical data, the team determined periods of time when this risk was high, and took preventative measures — additional verification and monitoring was employed during those high-risk time periods.

From the data analysis phase, we generated a lot of insights which are helpful in modulating our fraud detection approach. The team used what they found to fine tune their detection systems and concentrate on the highest risk areas. The papers showed that this approach enabled fast detection of fraudulent activity with real time monitoring reducing friction time – the lag between the fraudulent activity and detection.