Fraud Detection in Online Transactions - Project Proposal

As we move towards a growing digital world, many of our shopping habits shift from in person to online. As a result, we rely more and more on online transactions, opening up new opportunities for ecommerce fraud. Especially over the last few years due to the Covid-19 pandemic, people have become more reluctant to shop in stores for fear of spreading and catching the virus. It is especially important for ecommerce businesses to have a strong fraud detection system to protect their customers and improve sales.

There are many types of ecommerce fraud that can occur on an online shopping platform. These range from using a stolen credit card, using a false identity, claiming chargeback to receive free items, etc. Web services like Google Analytics help to identify and prevent these types of fraud by tracking website traffic and user activity on merchant sites. They collect data including where users originated from, interactions with different URLs on the site, browser information, transactional data, and more. All of this data is used to determine whether a transaction is legitimate or not, and can determine how likely a user with similar traffic would be a fraudster.

The dataset I am using for this project is from Kaggle [link here] and contains user information for each transaction made, and whether that transaction is legitimate or fraudulent. I would like to take this data and build a model to predict the likelihood that a transaction is fraudulent. I would split the data into training and testing sets, making sure there are equal proportions of fraudulent and legitimate transactions in each set. I would then train the training set on various models, tune parameters, and use the model to predict fraud in the test set.

There are certain limitations with this dataset, including the class imbalance of the legitimate vs fraudulent sets, which can be remedied by using a tree model. We would still need to be careful about the amount of data available for fraudulent cases to make sure it is sufficient to produce an accurate model. The dataset only ranges over the span of one year, so there might be seasonal spikes in fraudulent transactions that cannot be accounted for. Also, the types of fraud may have changed since the data was gathered, so even if the model is accurate with the current dataset, it may not be for future data.