MGT 6203 Project Proposal Team 36 - Fraud Detection

LESLIE FRANCIS, ALEJANDRO DE JESUS, TANIA DE PENA, VAISAG RADHAKRISHNAN

Abstract

E-commerce fraud is defined as any type of fraudulent transaction that occur on e-commerce platforms. Many data mining techniques and machine learning algorithms can be used to detect fraudulent transactions based on various features about the transaction. This study is intended to test several machine learning classification techniques for determining if a transaction is fraudulent or not. Fraud Detection dataset from Kaggle is used in the research. The algorithms that we are planning to use are classification techniques like logistic regression, KNN, Random Forest, SVM etc. The aim is to find the best model that predicts the test data accurately. The results will be measured by confusion matrix, F1 measure and/or mean classification error.

Introduction

According to the new "State of Fraud 2023" study from fraud prevention solution provider Signifyd, the total cost of e-commerce fraud in 2023 will reach $206.8 billion. These frauds affect both companies as well as customers. Although most banks will reimburse fraudulent transaction cost, it gets passed to the customer. To identify and stop potential financial loss for both the businesses and their clients, high accuracy fraud detection is essential. Effective fraud detection can protect customer data and maintain trust between companies and their clients. In this project, we are using e-commerce transaction data to detect fraudulent transactions. The aim is to find out whether a transaction is fraud or not. Identifying this will help banks and payment providers alert their customers in case of suspicious activities, so they can block these transactions.

Literature Review

The simplicity and ability to explain the model are typically ignored in favor of focusing on predictive performance. Most papers focus on predictive performance as opposed to the interpretation of the model - while predictive performance is important, interpreting a model is equally important to make critical business decisions. As fraud detection occurs in the financial services industry, many datasets will include masked columns for privacy concerns. Both Varmedja et al and Awoyemi et al used methods like Principal Component Analysis to keep the data anonymous. Due to this, they were able to make accurate predictions, but the model's interpretability suffered.

Fraud detection being a key factor in gaining and maintaining customer trust, businesses come up with various machine learning techniques to predict fraudulent transactions. Numerous researchers are trying their best to better the existing methods. Varmedja et al used machine learning techniques like logistic regression, Naïve Bayes, Random Forest and Multilayer perception to obtain high accuracy in credit card fraud detection. Awoyemi et al used Naïve Bayes, logistic regression, and K-nearest neighbor techniques to obtain high predictive performance.

Most fraud detection datasets are imbalanced. An imbalanced dataset is a dataset where the number of records belonging to one class is much higher than the other. This usually occurs in the fraud detection use case as the number of fraudulent transactions are much lesser than that of the number of

non-fraudulent ones. It is extremely important to balance the dataset to get good predictive performance. Varmedja et al used SMOTE technique to balance the dataset by a technique called oversampling where the records in the smaller class is increased in a balanced way to make sure that the samples of both classes are equal. While other research papers like Awoyemi et al explored a combination of oversampling and under sampling to achieve balance in the dataset.

Problem Statement and Data Sources

The datasets for this project are from Kaggle. There are two datasets used here. First, a transaction dataset that contains the fraudulent as well as non-fraudulent transactions. Next, ip address to country cross reference table to find out the country associated with transactions in the transaction dataset. The objective of the assignment is to predict whether an e-commerce transaction is fraudulent or not based on various available features. Accurately predicting this will help increase the customers' trust on the companies and help reduce losses for both the customers as well as the companies.

Planned Approach

Regarding data preparation, we plan to do some Exploratory Data Analysis to understand the raw data better. As a result, we may identify outliers, erroneous data or some cleaning tasks that need to be done prior to model training. After that, we will create features that may be found useful in model training. Indicator variables and encoding text variables as numbers or dummy variables may be done at this point.

Fraud detection datasets tend to be imbalanced as there are considerably more legitimate transactions than fraudulent ones. For this reason, we are thinking about balancing by under sampling legitimate transactions or by oversampling fraudulent ones. This serves the purpose of making more relevant the class that we are interested in identifying (frauds). We will experiment with different combinations of these techniques, especially SMOTE to synthesize new positive cases.

To test and compare various models, we will use K-fold Cross Validation (KCV) and F1 measure as the comparison metric. To use KCV we will need to split the data into K chunks instead of only splitting it into train and test sets as done when only one model will be tested. The decision to use F1 measure as metric comes from the article "A Gentle Introduction to Threshold-Moving for Imbalanced Classification" by Jason Brownlee. The precision-recall curve focuses on correctly identifying positive cases (which is what is desired in fraud detection models) and F1 measure is the harmonic mean between these 2 metrics.

A higher F1 indicates a better balance between precision and recall. This balance is the goal because we want to correctly identify a high percentage of frauds to mitigate losses, but we do not want to saturate clients by blocking their accounts/credit cards when they are making legitimate transactions.

This problem calls for classification models. We will start by testing logistic regression, random forest, KNN and SVM models. Many variations of these models may be tested along the way.

Progress, Analysis and Results

We are working with two datasets. First, we have the Fraud transaction dataset that contains the fraudulent and non-fraudulent transactions. Also, we have information of the transaction characteristics such as:

- Customer ID
- Amount of the purchase
- Date and time of the purchase
- Used Browser for the transaction.
- Source (Direct, Ad, etc.)
- Date and time of the sign up
- Demographic info of the customer like age and gender
- Ip address

The second dataset includes the Ip address boundaries for each country.

We are working with 151,000 unique transactions made from January 2015 to December 2015 across different locations. The dataset contains 14K fraudulent transactions and 138K non-fraudulent transactions, meaning we are working with an imbalanced dataset, since only the 9% of the transactions are marked as class 1 (fraudulent trans.) *Figure 1* shows the distribution of the fraudulent(class = 1) and non-fraudulent(class = 0) transactions.

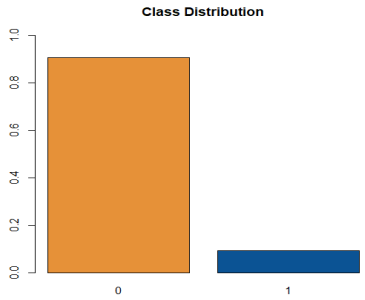

Figure 1: Class Distribution of the response

After an initial exploration on the dataset and completing the joining of the two tables, we moved to the data cleaning process. The initial search was for null values; however, the dataset contains data in every column. The next step was to search for outliers and create some graphs to show how the data is distributed.

- The most used Browser is Chrome and the least used one is Opera. See *Figure 2*.
- The least used source of purchase is Direct purchases. This is expected, as companies spend money on advertisements and do search engine optimization to appear on top of the search results. See *Figure 3*.
- The variables age and purchase value have a skewed distribution to the right. There are no outliers in both of these numeric variables. *Figure 4* and *Figure 5* show these distributions.

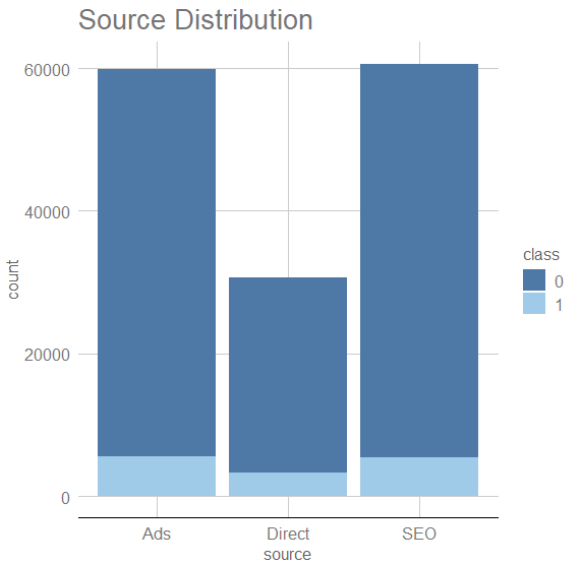

Figure 2: Distribution of the Source of the transaction

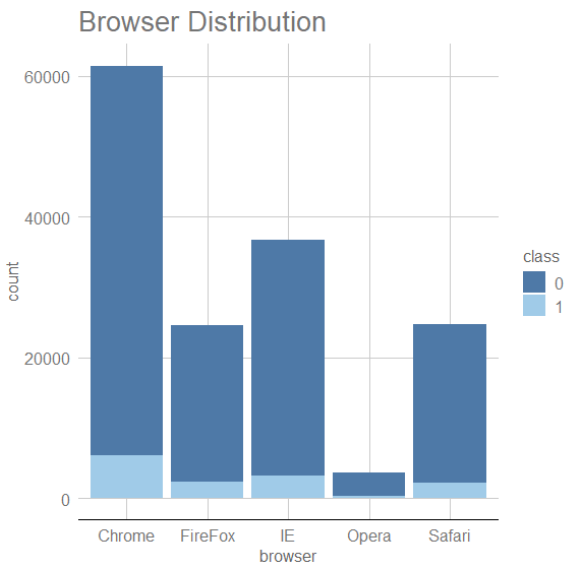

Figure 3: Distribution of the Browser used

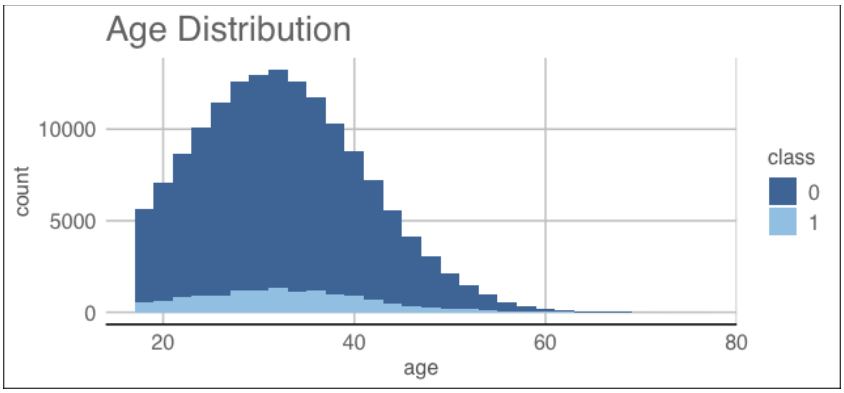

Figure 4: Age Distribution

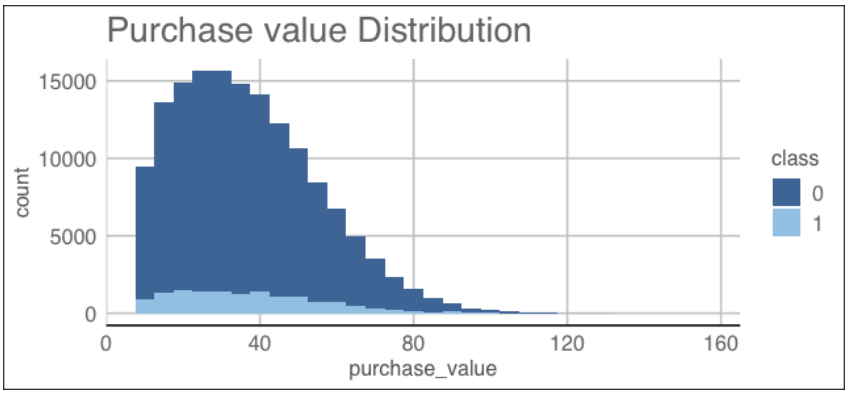

Figure 5:Purchase Value Distribution

Based on the original variables from the dataset, some extra variables were constructed:

- Day of purchase: Day of the week on which the purchase occurred
- Weekend or weekday: whether the purchase was made on weekends or not
- Conversion time: The time taken between signup and the purchase. See Figure 6 for conversion time vs transaction outcome. It can be observed that, on average, fraudulent transactions occur on cases where the signup and purchase are on the same day. This is expected as these fraudsters tend to create a new account just for committing a fraud.
- Time of transaction: if the transaction was made on the morning, afternoon, or night.
- Month of the transaction. See Figure 7 for the distribution of the purchases and the outcomes. Looks like January had the most fraudulent transactions compared to other months.
- Age range: If the person is young, middle aged or senior
- Dummy variables for each of the categorical variable for ease of modeling

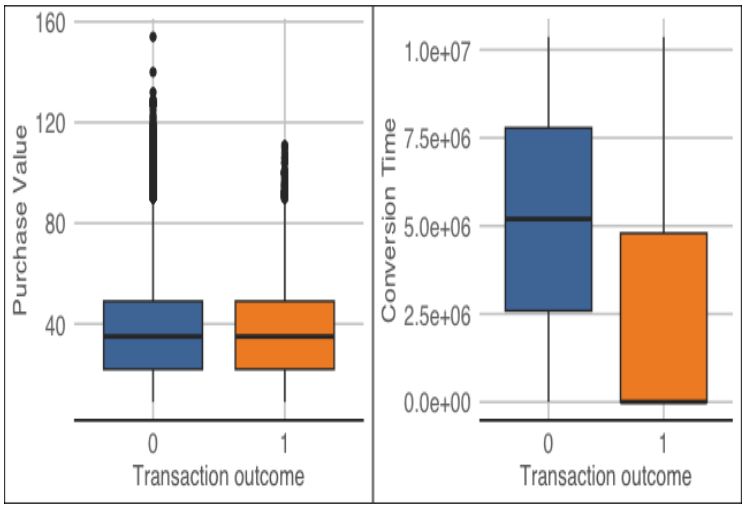

Figure 6: Boxplot of Purchase Value and Conversion Time vs Transaction Outcome

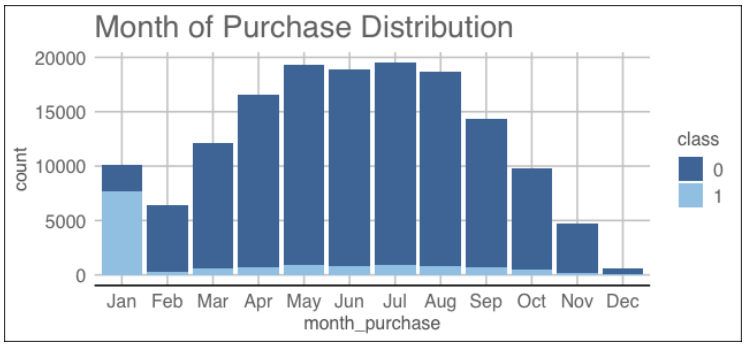

Figure 7: Distribution based on Month of Purchase

Next, we analyzed the difference between the fraudulent and non-fraudulent transaction, by looking at the propensity of both groups and comparing it with the average of the dataset. See Figure 8 for detailed analysis on various fields. Below are some of the insights on this analysis.

- The browser with most fraudulent transaction's propensity is chrome (43%) vs the average (40.7%). See Figure 8.1
- On weekends we have a higher propensity for fraudulent transactions than the rest of the week (46.4% weekend propensity for fraud transactions vs 43.1% the average). See Figure 8.2
- Tuesdays and Wednesdays are the days of the week with lowest propensity for fraudulent transactions. See figure 8.3
- As discussed earlier, purchases made the same day of sign up have higher propensity for fraudulent transactions. See figure 8.4

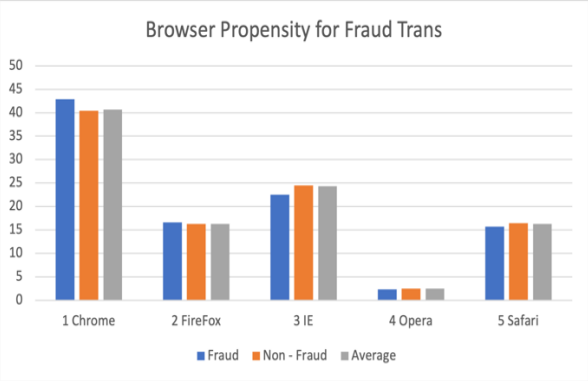

Figure 8.1 Browser Propensity

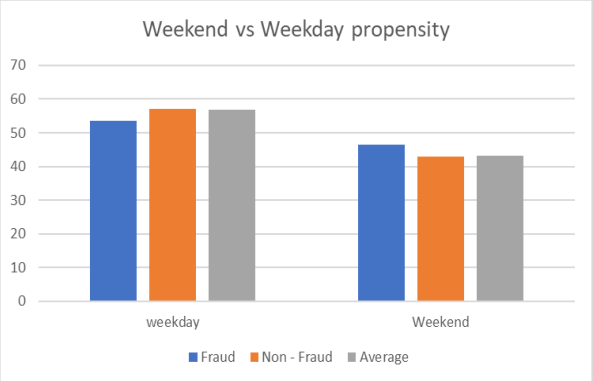

Figure 8.2 Weekend vs Weekday Propensity

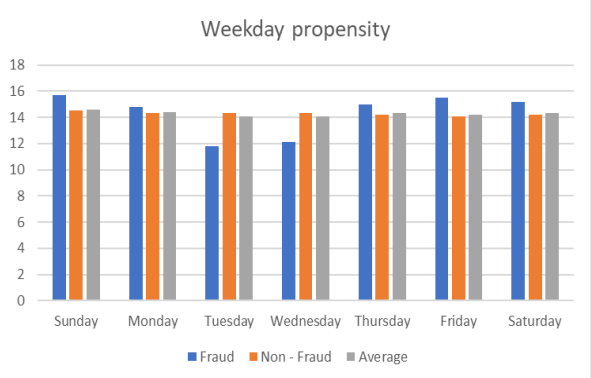

Figure 8.3 Days of the week propensity

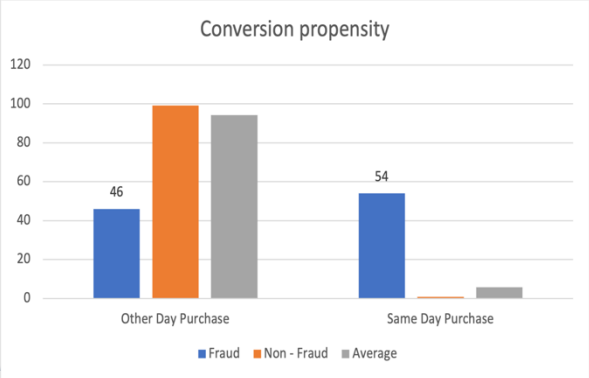

Figure 8.4 Conversion Propensity

Figure 8: Various Propensity Analysis

Unexpected Problems and challenges:

1. Joining the two datasets: The key variable between the two dataset is the ip address. However, the IP address dataset only have the boundaries for each country (lower bound and upper bound). We had to figure out how to join the two tables because the description of the variables wasn't very clear. Also, the processing time of the join was too long and end up crashing the R

studio program. We solved this by splitting the Transaction dataset in smaller dataset and joining them separately.

2. We have very few transactions marked as class 1, meaning we are working with an imbalanced dataset. To train our models and perform a correct predictive analysis we will need to use different techniques to balance the dataset. We have already tried balancing the data set with an oversampling technique which tries to create new positive cases on the dataset. This technique is taking too long to process, and we haven't been able to balance the dataset so far. We will continue to try different approaches and to find more R functions that can do the same.

Next Steps

After finishing our exploratory analysis, we will start to build our predictive models. We plan to use Logistic Regression, Random Forest and KNN to see which one performs better. Also, we will need to perform multiple iterations using variable selection techniques to find the best model using K-fold Cross Validation (KCV) or Monte Carlo Cross Validation(MCCV) and with F1 measure, accuracy, precision, or mean classification error as the comparison metrics.

References

1. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.
2. J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.
3. "New FTC Data Show Consumers Reported Losing Nearly $8.8 Billion to Scams in 2022." Federal Trade Commission, 23 Feb. 2023, www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022.
4. "New Data Shows FTC Received 2.8 Million Fraud Reports From Consumers in 2021." Federal Trade Commission, 22 Feb. 2022, www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0
5. Berthiaume, Dan. "Study: Total E-commerce Fraud in 2023 Will Exceed $200 Billion." Chain Store Age, 25 Jan. 2023, https://chainstoreage.com/study-total-e-commerce-fraud-2023-will-exceed-200-billion#:~:text=According%20to%20the%20new%20%E2%80%9CState,tangible%20losses%20for%20online%20retailers
6. Brownlee, Jason. "Cost-Sensitive Learning for Imbalanced Classification", 7 Feb. 2020, https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/
7. Brownlee, Jason. "A Gentle Introduction to Threshold-Moving for Imbalanced Classification", 10 Feb. 2020, https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/
8. Brownlee, Jason. "SMOTE for Imbalanced Classification with Python", 17 Jan. 2020, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/