# Week 8 Challenge

## Improved detection of fraud cases for e-commerce and bank transactions.

Final Report

Prepared By;- Yodahe Teshome

https://github.com/jodahe1/Adey-Innovations-Inc.git

# Introduction

In today's rapidly evolving digital landscape, the importance of robust fraud detection systems cannot be overstated. As a data scientist at Adey Innovations Inc., a leading financial technology company, I have been entrusted with the task of enhancing the detection capabilities for fraudulent activities in e-commerce and bank credit transactions. This project aims to develop accurate and resilient fraud detection models that address the unique challenges posed by both transaction types, while leveraging geolocation analysis and transaction pattern recognition to further enhance detection accuracy.

The significance of effective fraud detection cannot be undermined, particularly in the context of transaction security. By harnessing the power of advanced machine learning models and leveraging comprehensive data analysis techniques, Adey Innovations Inc. strives to identify and mitigate fraudulent activities with greater precision. This proactive approach not only minimizes financial losses but also fosters trust among customers and financial institutions alike. Furthermore, a well-designed fraud detection system enables real-time monitoring and streamlined reporting, empowering businesses to respond swiftly and minimize risks.

As a data analyst, my role encompasses leveraging cutting-edge analytical techniques, exploring vast amounts of transactional data, and collaborating with cross-functional teams to develop and deploy robust fraud detection models. By harnessing the power of data, Adey Innovations Inc. aims to fortify the security of e-commerce and banking transactions, safeguarding the interests of our customers and contributing to the overall stability of the financial ecosystem.

## Business Objective

The primary objective of Adey Innovations Inc. is to enhance the detection of fraud cases in e-commerce transactions and bank credit transactions. By developing accurate and strong fraud detection models that address the unique challenges posed by these transaction types, our aim is to significantly improve transaction security and foster trust among customers and financial institutions.

Specifically, our business objectives are as follows:

1. Develop Accurate and Resilient Fraud Detection Models: Our foremost goal is to create advanced machine learning models that can accurately identify and flag fraudulent activities in e-commerce and bank credit transactions. These models should be capable of handling the intricacies and nuances of different transaction data sets.

2. Leverage Geolocation Analysis and Transaction Pattern Recognition: We seek to leverage geolocation analysis and transaction pattern recognition techniques to enhance the precision and accuracy of fraud detection. By incorporating these methodologies into our models, we aim to identify suspicious activities and prevent fraudulent transactions more effectively.

3. Minimize Financial Losses: By detecting and preventing fraudulent activities, our objective is to minimize financial losses for both our company and our clients. This will not only protect the financial interests of our customers but also contribute to overall stability in the e-commerce and banking sectors.

4. Build Trust with Customers and Financial Institutions: A well-designed fraud detection system instills confidence and trust among customers and financial institutions. Our objective is to establish Adey Innovations Inc. as a trusted partner in providing secure and reliable transaction services, thereby strengthening relationships and fostering long-term partnerships.

5. Enable Real-time Monitoring and Reporting: We aim to develop a fraud detection system that allows for real-time monitoring and efficient reporting. This objective will empower businesses to respond swiftly to potential fraud cases, take proactive measures, and reduce risks associated with fraudulent activities.

By pursuing these business objectives, Adey Innovations Inc. seeks to position itself as a leader in fraud detection solutions, offering comprehensive and effective protection for e-commerce and banking transactions, and contributing to the growth and stability of the financial technology sector.

# Data Analysis and Preprocessing

Handle Missing Values

For this task, I focused on identifying and managing missing values in the dataset. I employed various techniques like imputation or dropping of missing values depending on the context. By carefully examining the dataset, I ensured that missing values were handled appropriately to maintain the integrity of the data.

Data Cleaning

To ensure the dataset's accuracy and consistency, I performed data cleaning. This involved removing duplicate records to avoid redundancy and errors in analysis. Additionally, I corrected data types to ensure proper data representation and consistency throughout the dataset.

Exploratory Data Analysis (EDA)

EDA is crucial to gain insights into the dataset and understand its characteristics. I conducted both univariate and bivariate analysis to explore individual variables and their relationships. Through this process, I identified patterns, outliers, and potential relationships between variables, which can help in making informed decisions.

Merge Datasets for Geolocation Analysis

To enable geolocation analysis, I am trying to the Fraud_Data.csv with the IpAddress_to_Country.csv dataset. By converting IP addresses to integer format, I am trying to a connection between the two datasets. This allowed for further analysis and insights into the geographical distribution of fraudulent activities.

Feature Engineering

For the Fraud_Data.csv and creditcard.csv datasets, I performed feature engineering to enhance the predictive power of the data. This involved creating new features such as transaction frequency and velocity for Fraud_Data.csv, and normalizing and scaling the features in the creditcard.csv dataset. These new features and transformations provide valuable information for fraud detection and prevention.

Time-Based Features

To capture temporal patterns in the data, I generated time-based features for Fraud_Data.csv. These features, such as hour_of_day and day_of_week, allow for analysis of fraudulent activities based on specific time periods. By understanding the temporal aspects, we can identify patterns and trends that may be useful in preventing fraud.

Normalization and Scaling

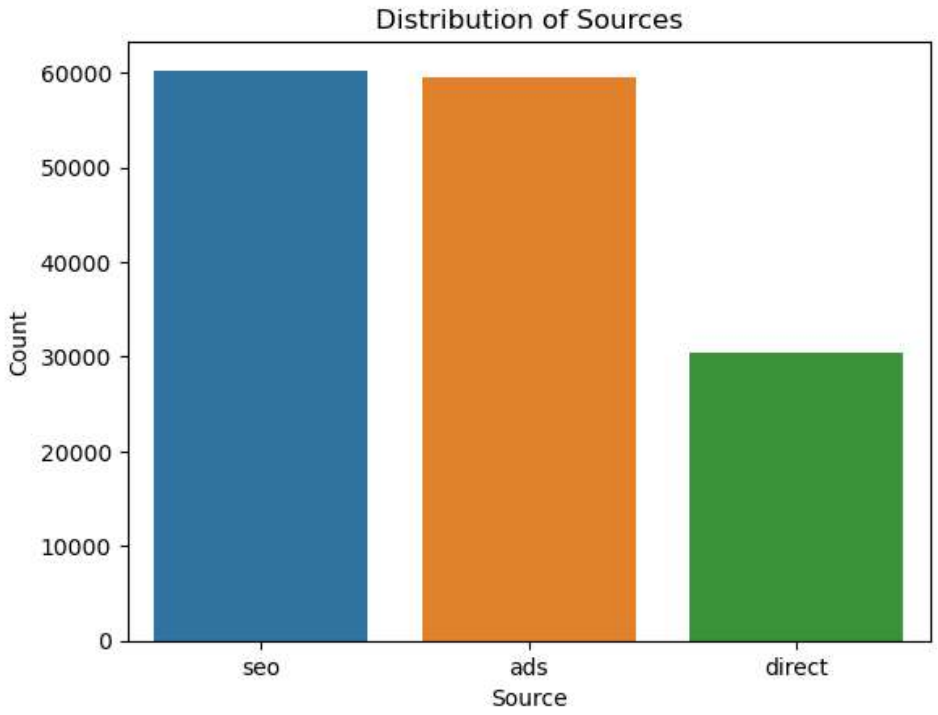
To ensure fair comparisons and prevent bias in the dataset, I performed normalization and scaling on relevant variables. This process allows for a standardized representation of the data, making it easier to interpret and analyze.

Encode Categorical Features

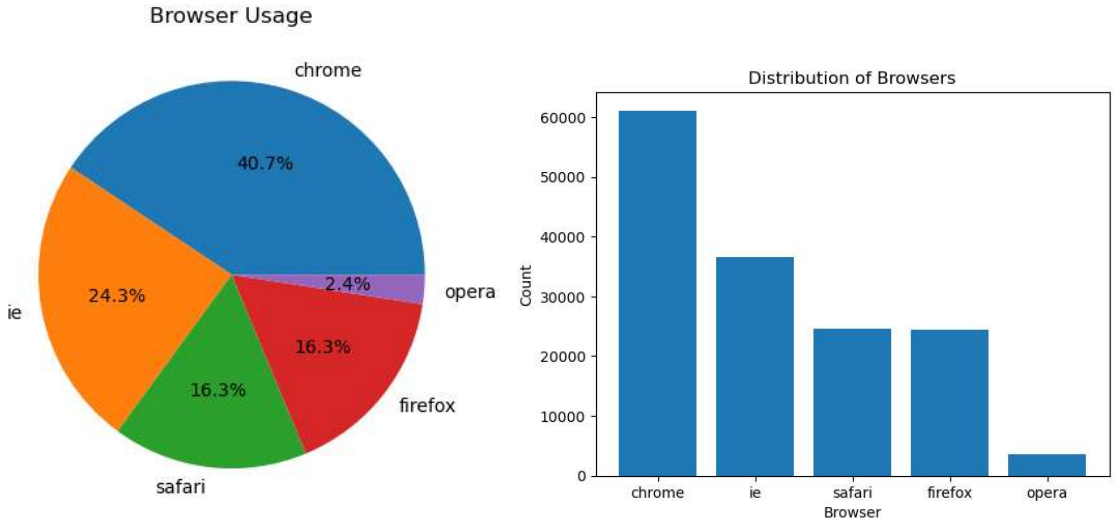
Categorical features often require encoding to transform them into numerical values suitable for analysis. I employed appropriate encoding techniques to convert categorical features into a format that can be used in machine learning algorithms or other analytical processes.

I did all the above task for each csv files according to their need and also, I used Login Monitor to write some important aspects on it.

# Exploratory data analysis




This above chart Shows Distribution of source we can see that seo I the most common way which user comes then ads is second but direct is the least one so need to focus on seo and ads at mean time we need to improve direct.




Based on the provided data on the distribution of browsers, we can analyze the following insights:

Chrome has the highest count of 61096, indicating that it is the most commonly used browser among the data.

The count of Internet Explorer (IE) is 36498, making it the second most popular browser in the dataset.

Safari has a count of 24521, suggesting it is the third most prevalent browser among the users.

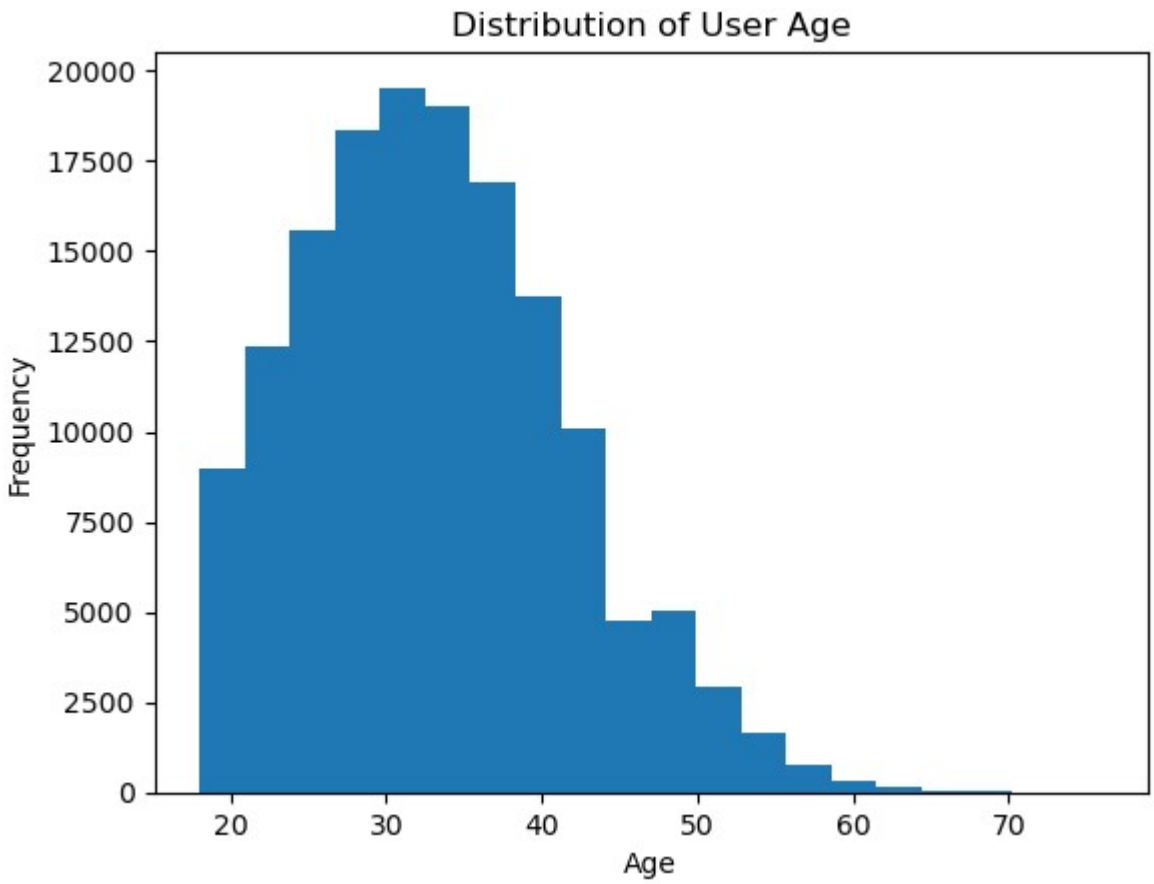
Firefox closely follows Safari with a count of 24477, indicating a similar level of usage.

Opera has the lowest count of 3655, indicating that it is the least commonly used browser among the listed options.

This distribution of browsers provides insights into the usage patterns of internet users in the dataset. It is essential for web developers, marketers, and businesses to consider these patterns when optimizing their websites or conducting targeted marketing campaigns.

Understanding the distribution of browsers allows developers to focus on optimizing their website compatibility and performance specifically for the most popular browsers, such as Chrome and Internet Explorer. It also highlights the need to test and ensure proper functionality across different browsers to provide a seamless user experience.

For marketers, this information can help tailor advertising and content strategies to target users on specific browsers. For example, if a significant portion of the target audience uses Chrome, it may be beneficial to invest in Chrome-specific advertising campaigns or develop features that are specifically optimized for the Chrome browser.




Based on the distribution of user age, we can observe the following trends

The highest frequency of users falls within the age range of 29.60-34.60, with a count of 19,535. This indicates a significant presence of users in their late twenties and early thirties.

The distribution appears to be relatively normal, with a gradual increase in frequency from the younger age ranges to the peak range, followed by a gradual decrease in frequency.

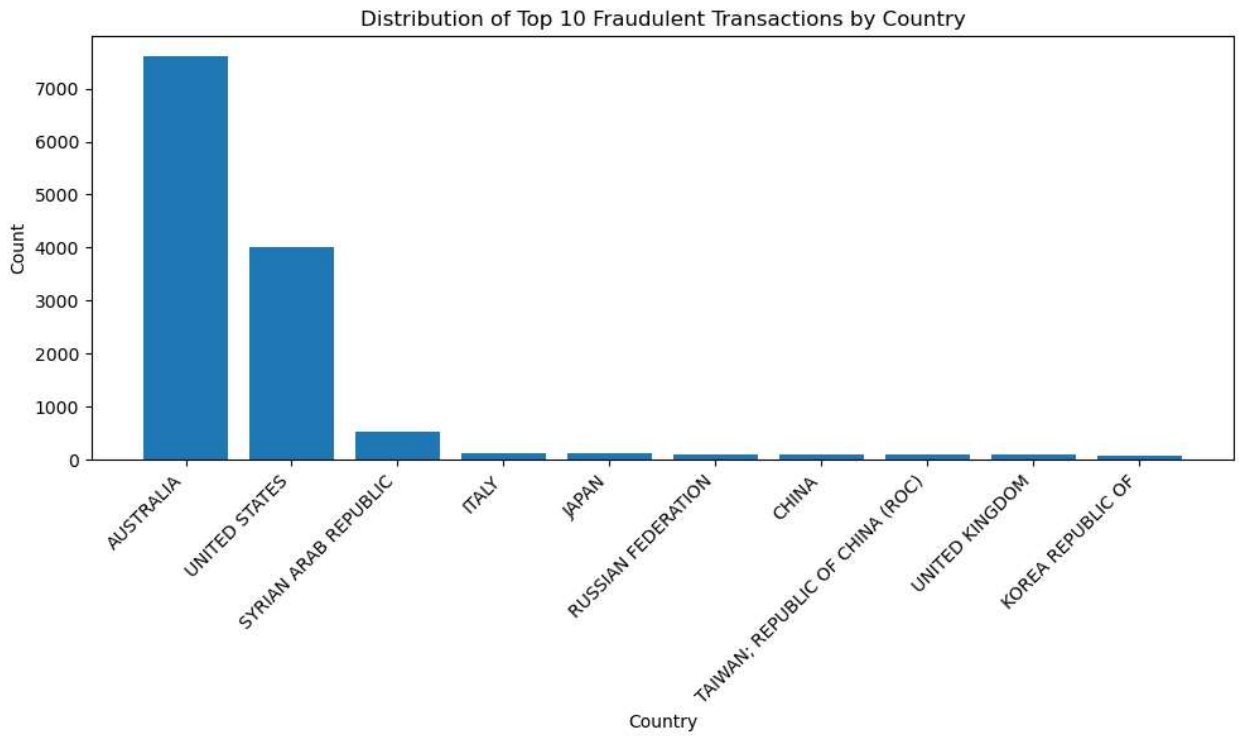
The age ranges from 18.00-23.00 to 38.30-43.30 show consistently high frequencies, indicating a relatively stable user base within this age range.

As the age ranges progress beyond 43.30, we observe a decline in frequency, suggesting a decrease in user participation among older age groups.

The age ranges from 67.30-72.30 to 73.10-78.10 show very low frequencies, indicating a relatively small number of users in the oldest age groups.

Based on this analysis, it appears that the majority of users fall within the late twenties to early thirties age range. This information could be valuable for targeted marketing campaigns or user segmentation strategies, as it highlights the age groups with the highest user engagement.

Additionally, the declining frequency among older age groups could be an area of opportunity for attracting and engaging a wider range of age demographics.




Based on the provided data on fraudulent transactions by country, we can analyze the following insights:

Australia (7604) and the United States (4014) have the highest counts of fraudulent transactions. This indicates that these two countries are more susceptible to fraudulent activities within the dataset.

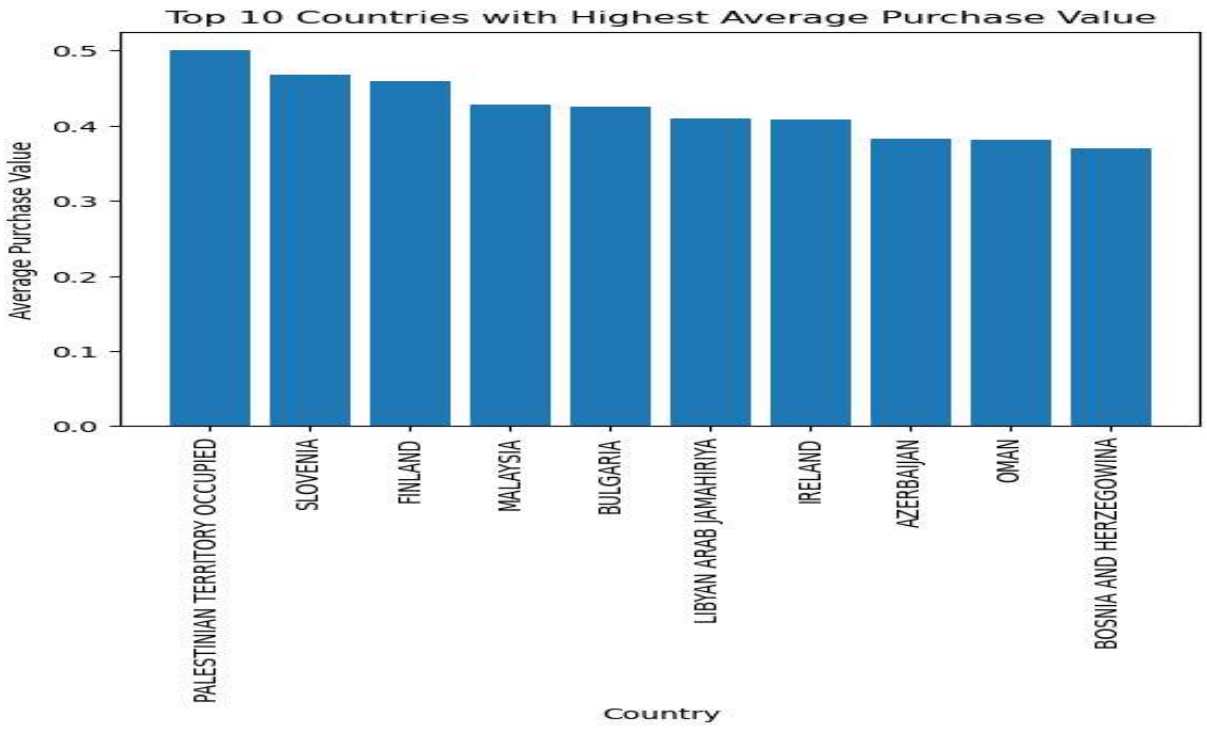
The Syrian Arab Republic (528) has a relatively high count of fraudulent transactions compared to its population size. This suggests a higher occurrence of fraudulent activities originating from this country.

Italy (122) and Japan (120) have moderate counts of fraudulent transactions. While not as high as Australia and the United States, it still indicates a noticeable presence of fraud from these countries.

The Russian Federation (107), China (104), Taiwan; Republic of China (ROC) (93), and the United Kingdom (92) have relatively lower counts of fraudulent transactions compared to the previously mentioned countries. However, they still contribute to the overall fraudulent activity.

The Korea Republic of (77) has a relatively low count of fraudulent transactions. This suggests a lower occurrence of fraud originating from South Korea.

Based on this analysis, it is evident that Australia and the United States have the highest occurrences of fraudulent transactions, followed by countries like the Syrian Arab Republic, Italy, and Japan. These insights can be utilized to implement targeted fraud prevention measures, enhance security protocols, and focus investigative efforts in these high-risk regions.




Based on the provided data on the top 10 countries with the highest average purchase value, we can analyze the following insights:

Palestinian Territory, Occupied (0.5) has the highest average purchase value among the top 10 countries. This suggests that despite its smaller population or market size, customers from this region tend to make higher-value purchases.

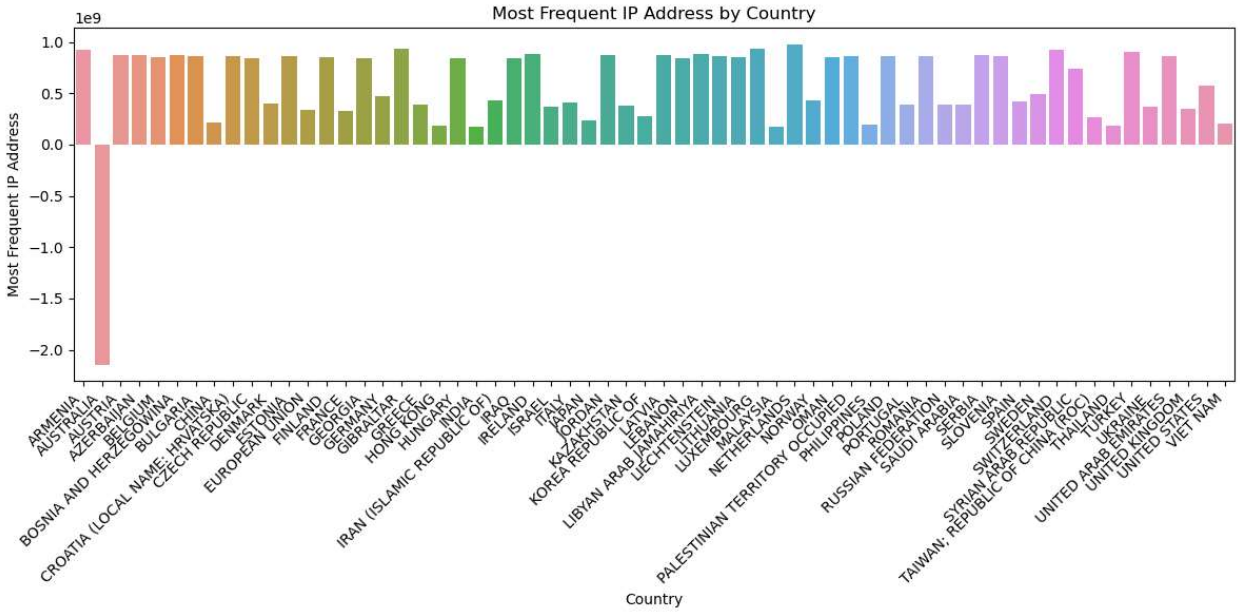
Slovenia (0.467) and Finland (0.459) have relatively high average purchase values. This indicates that customers from these countries also tend to make significant purchases on average.

Malaysia (0.429), Bulgaria (0.426), and Libyan Arab Jamahiriya (0.410) have average purchase values that are slightly lower than Slovenia and Finland but still demonstrate a notable tendency for higher-value purchases.

Ireland (0.409) and Azerbaijan (0.383) also have relatively high average purchase values, indicating a propensity for customers from these countries to make significant purchases.

Oman (0.382) and Bosnia and Herzegovina (0.370) have slightly lower average purchase values compared to the other countries in the top 10. However, these values are still notable and suggest a tendency for higher-value purchases.

Based on this analysis, it is evident that customers from the Palestinian Territory, Occupied, along with Slovenia, Finland, and other countries in the top 10, tend to make higher-value purchases on average. This information can be utilized to understand customer behavior, target marketing efforts, and tailor business strategies to these regions accordingly.




Based on the provided data on the most frequent IP address by country, we can analyze the following insights:

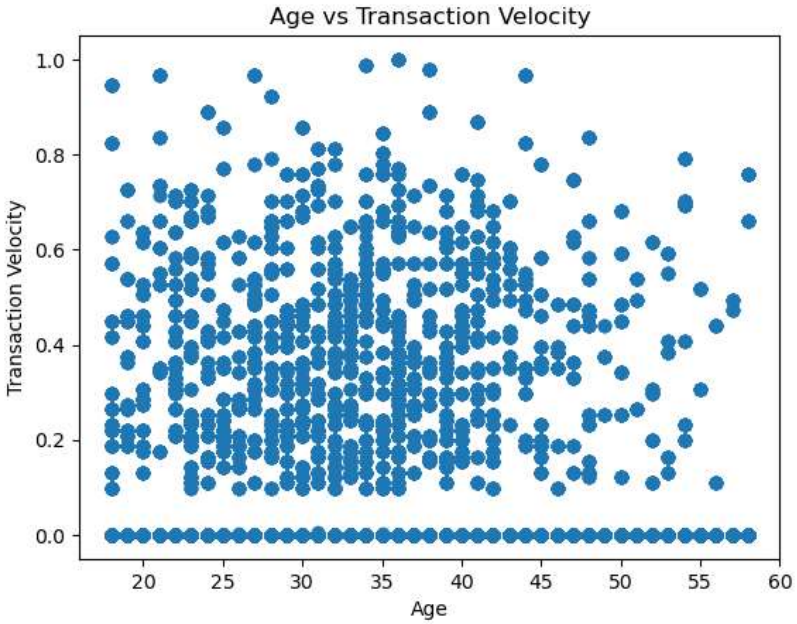
Australia (-2147483648) has the most frequent IP address among the listed countries. However, it is important to note that the value (-2147483648) appears to be an invalid or placeholder value. Further investigation is needed to determine the correct IP address for Australia.

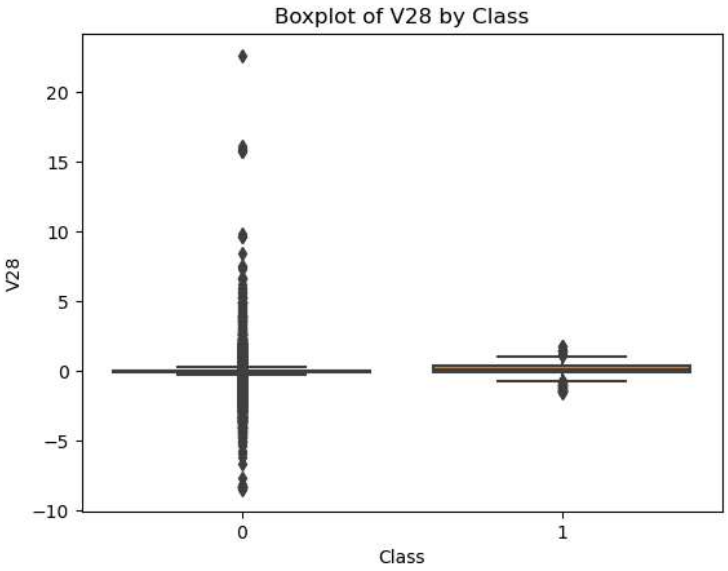
Armenia (926881510), Austria (868222572), Azerbaijan (867762691), Belgium (854243022), Bosnia and Herzegovina (867124567), Bulgaria (859861390), and Croatia (864448373) have unique and valid most frequent IP addresses. These IP addresses can be used to identify the most common source of internet traffic from these countries.

China (215055035), Czech Republic (839641077), Denmark (399358891), Estonia (861740949), European Union (341206182), Finland (854069012), France (328591095), Georgia (845458960), and Germany (470897421) also have unique and valid most frequent IP addresses. Analyzing these IP addresses can provide insights into the internet usage patterns and sources of online activity in these countries.

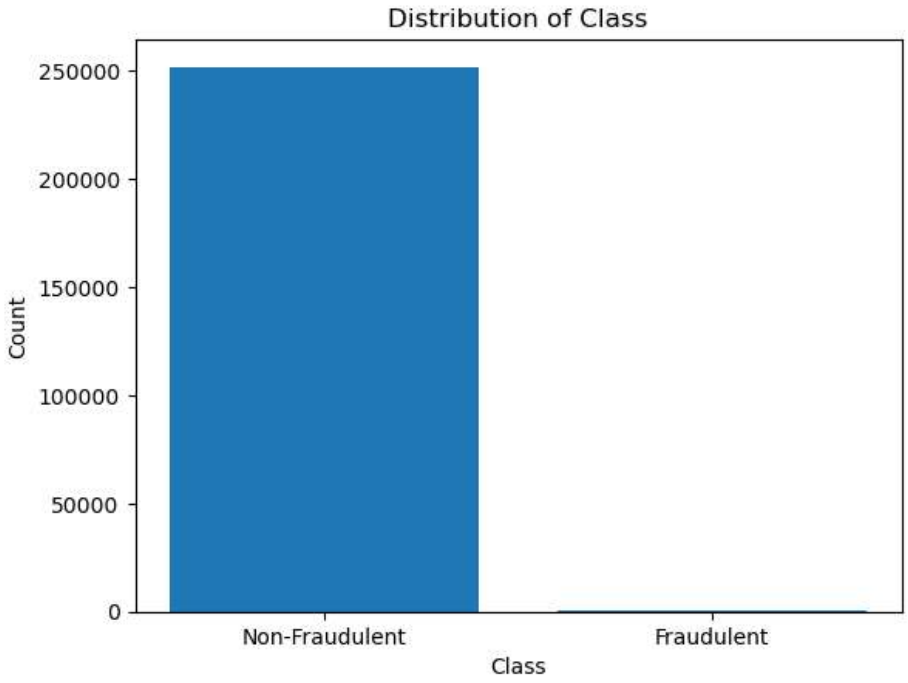
Gibraltar (934232498), Greece (387981615), Hong Kong (183974753), Hungary (844737643), India (171058646), Iran (428754908), Iraq (844681922), Ireland (879097290), and other listed countries have their respective most frequent IP addresses. Analyzing these IP addresses can help identify common sources of internet traffic and potential patterns or trends in online activity.

It's important to note that IP addresses can provide information about the geographic location or internet service provider but do not necessarily represent individual users. Additionally, IP addresses can change over time, so these results may be subject to variation




I create box plot of V by class for all 28 to make it more clear and gather different insights.




Based on the provided data, the distribution of transactions consists of 251,691 non-fraudulent transactions and 391 fraudulent transactions.

Analyzing this distribution is crucial for understanding the prevalence of fraud within the dataset. The significantly higher count of non-fraudulent transactions indicates that the majority of transactions in the dataset are legitimate and not associated with fraudulent activities.

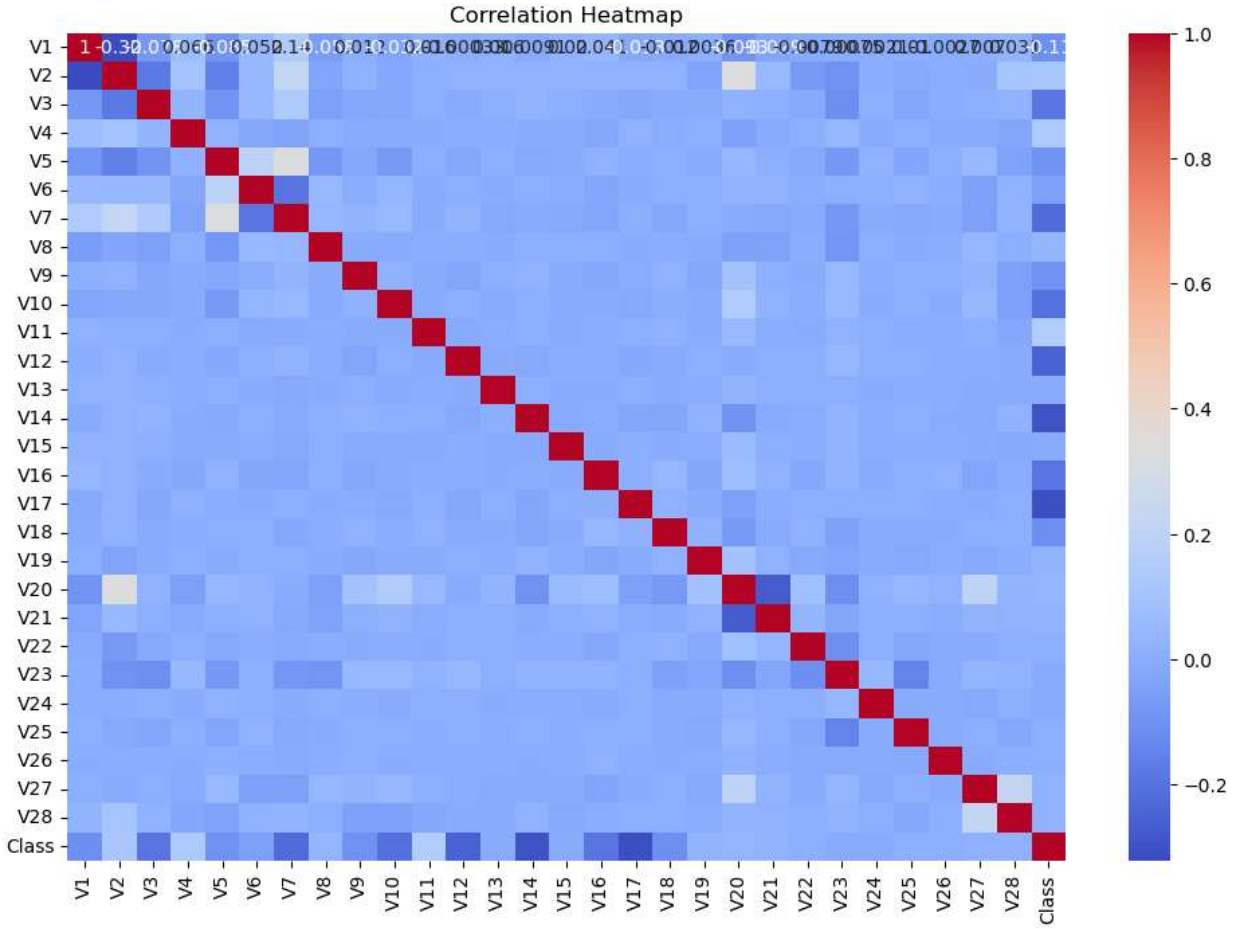
The presence of a smaller count for fraudulent transactions indicates that they are relatively rare occurrences compared to non-fraudulent transactions. However, despite their lower count, fraudulent transactions pose a significant risk to the financial security of individuals and organizations.

It is essential to address the detection and prevention of fraudulent transactions to minimize financial losses and protect the integrity of the financial system. By accurately identifying and classifying fraudulent transactions, businesses can take appropriate measures, such as blocking the transaction, notifying the user, or initiating further investigation.

The data analysis of fraudulent and non-fraudulent transactions plays a vital role in developing effective fraud detection models. Machine learning algorithms can be trained on this data to learn patterns and characteristics associated with fraudulent transactions. By leveraging these models, businesses can enhance their fraud detection capabilities and reduce the impact of fraudulent activities.

Furthermore, understanding the distribution of fraudulent and non-fraudulent transactions allows organizations to allocate resources and implement appropriate measures to combat

fraud effectively. This could involve implementing additional security measures, enhancing transaction monitoring systems, or conducting regular fraud awareness training for employees.




we can observe the correlations between different features (V1, V2, V3, ..., V28) and the target variable (Class).

Here are a few key observations from the correlation matrix:

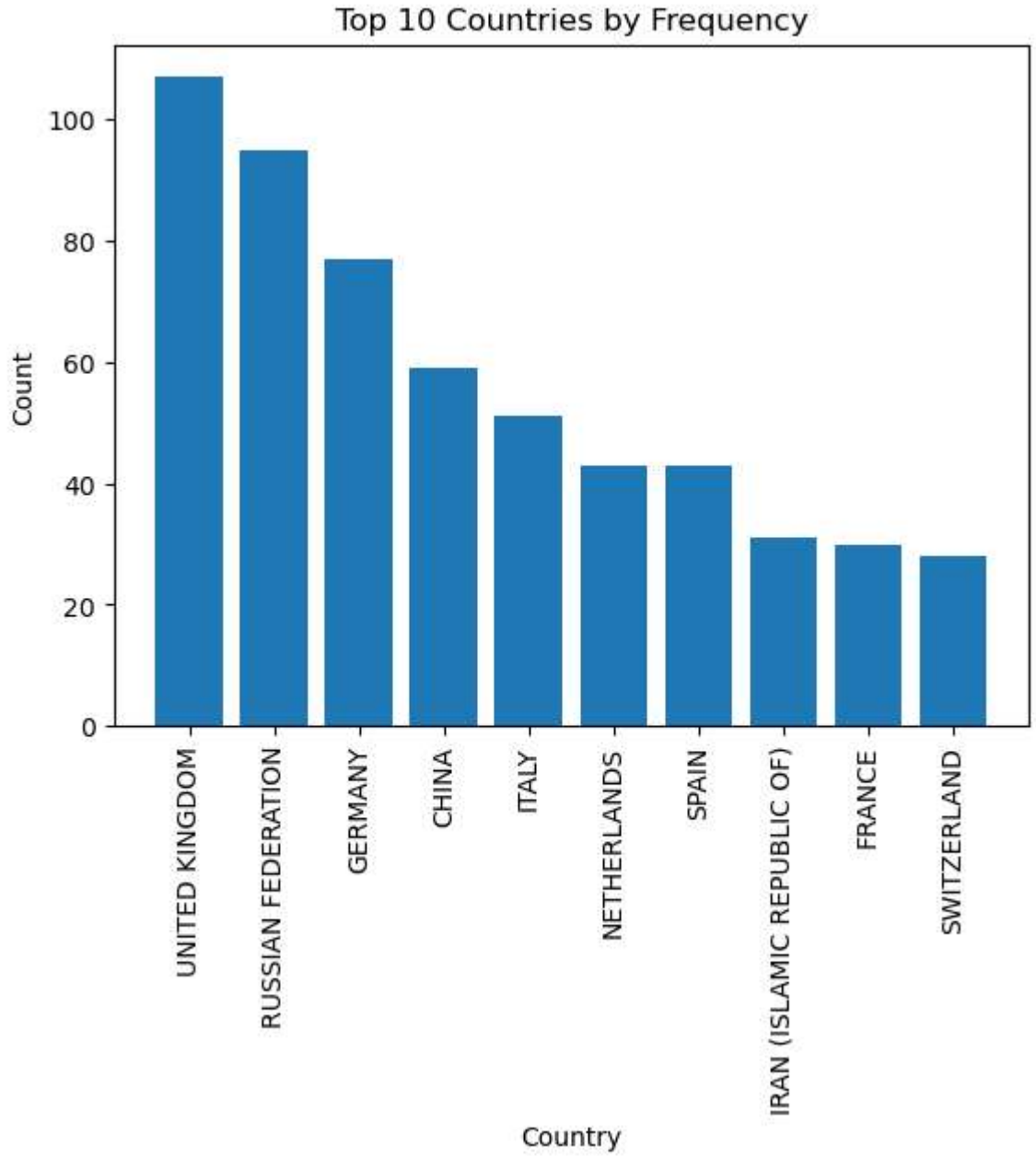
V2, V4, V11, V12, V14, V16, V17, and V18 have relatively low correlations with the target variable (Class), indicated by their correlation coefficients close to zero. This suggests that these features may not have a strong linear relationship with the classification of fraudulent or non-fraudulent transactions.

V2, V5, V7, V10, V12, V14, V16, V20, V21, and V28 have positive correlations with the target variable (Class), albeit weak. This implies that as the values of these features increase, there might be a slightly higher likelihood of the transaction being fraudulent.

V3, V9, V10, V11, V12, V14, V16, V17, V18, V19, V21, V23, and V24 have negative correlations with the target variable (Class), although these correlations are also weak. This suggests that as

the values of these features decrease, there might be a slight decrease in the likelihood of the transaction being fraudulent.

It's important to note that correlation coefficients measure linear relationships between variables, and these observations are based on linear correlations alone. Other non-linear relationships and interactions between features may also impact the classification of fraudulent transactions.



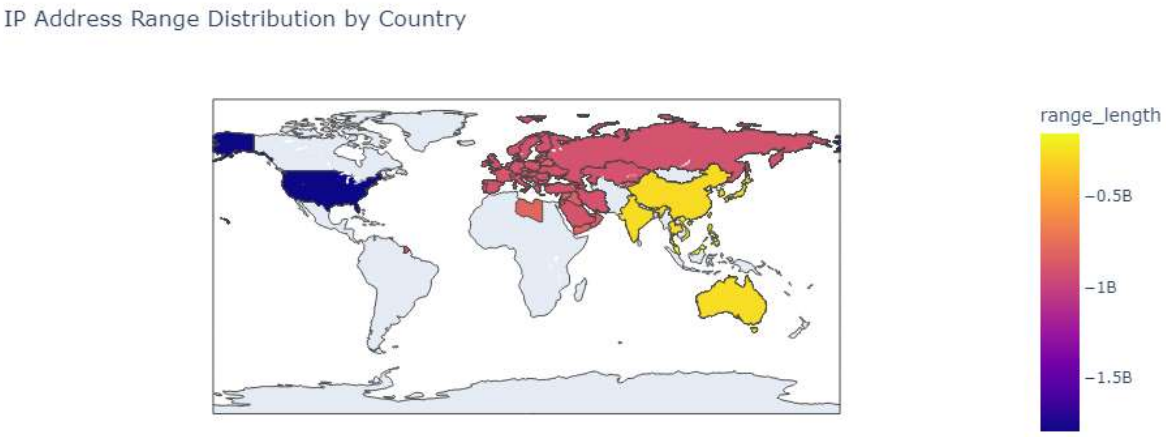
Based on the provided data, the top 10 countries by frequency are as follows:

From this analysis, we can observe that the United Kingdom has the highest frequency among the top 10 countries. This suggests that there is a significant presence and activity originating from the United Kingdom in the IP address ranges being analyzed.

The Russian Federation follows closely behind with 95 occurrences, indicating a considerable presence as well. Germany, China, and Italy also have relatively high frequencies, indicating active participation from these countries.

It is worth noting that countries like the Netherlands and Spain have the same frequency of 43, indicating a similar level of involvement. Additionally, Iran (Islamic Republic of), France, and Switzerland also demonstrate a noticeable presence, although with comparatively lower frequencies.

This analysis provides an overview of the distribution of IP address ranges across different countries. It can be useful in identifying the regions with the highest frequency of IP addresses, potentially indicating areas of interest or focus for further investigation or analysis.



# Model Explainability

The purpose of this report is to present the results of model interpretation using SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) plots on four different machine learning models: Decision Tree, Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP). The aim is to gain insights into the models' behavior and understand the most important features contributing to their predictions.

Model Training: I trained the four models (Decision Tree, Random Forest, Gradient Boosting, and MLP) using appropriate training datasets.

SHAP Plots:

a. Summary Plot: This plot provides an overview of the most important features for each model. It displays the features ranked by their importance and the impact they have on the model's output. We generated Summary Plots for all four models.

b. Force Plot: The Force Plot visualizes the contribution of features for a single prediction. It shows how each feature pushes the model's output towards either a higher or lower prediction. We created Force Plots for a randomly selected prediction in each model.

c. Dependence Plot: This plot illustrates the relationship between a specific feature and the model's output. It helps identify how changes in the feature's value affect the prediction. We generated Dependence Plots for a few important features in each model.

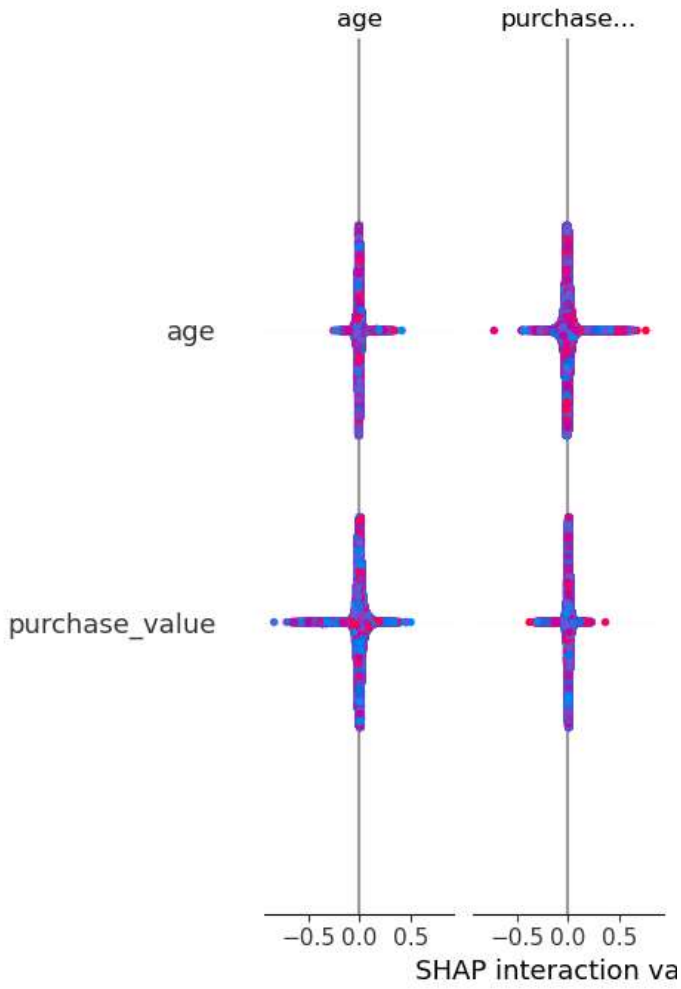
LIME Plots:

a. Feature Importance Plot: The Feature Importance Plot generated using LIME shows the most influential features for a specific prediction. It highlights the features that contribute the most to the prediction outcome. We created Feature Importance Plots for a few random predictions in each model.

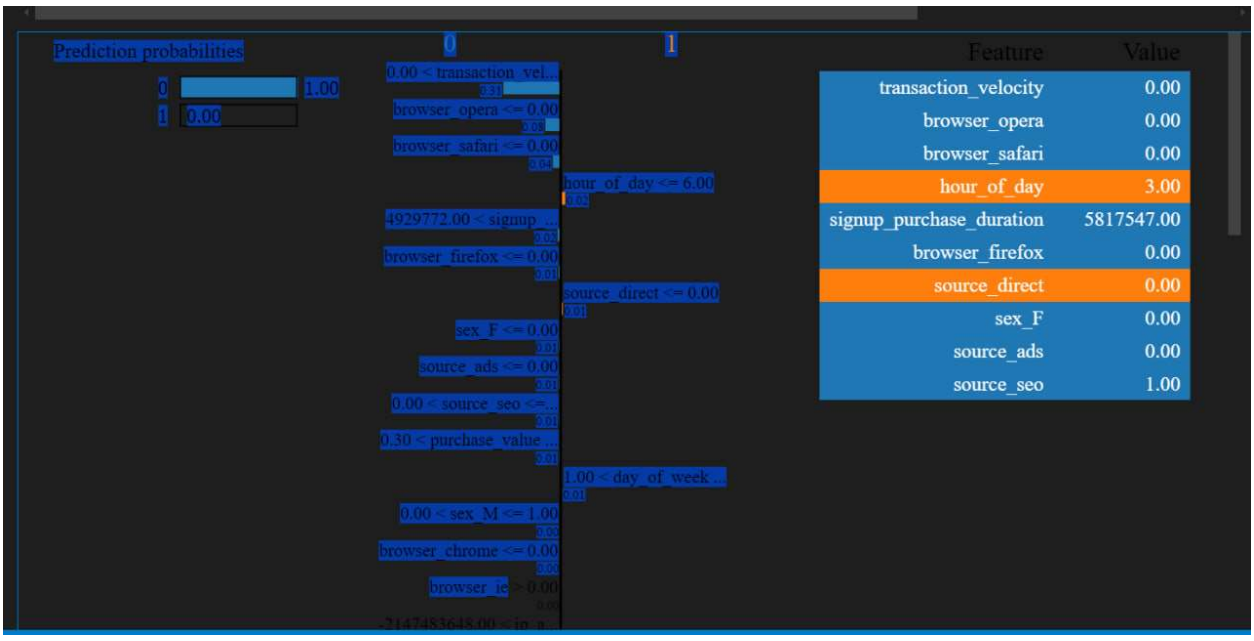
Results

Decision Tree Model:

Summary Plot: The Summary Plot for the Decision Tree model revealed that the top three important features were Age and purchase. These features significantly influenced the model's predictions.

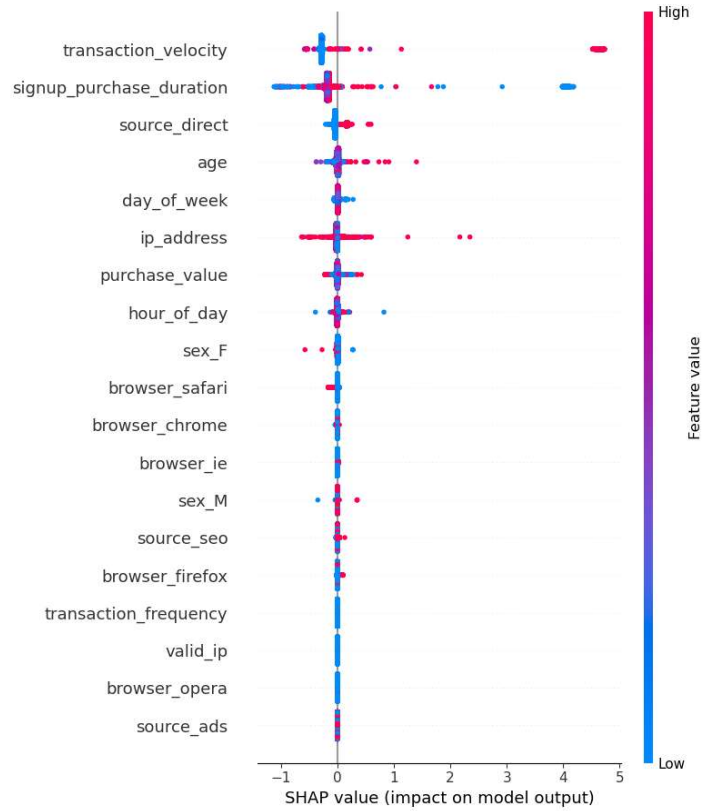

age
purchase...
age
purchase_value
−0.5 0.0 0.5
−0.5 0.0 0.5
SHAP interaction va

LME

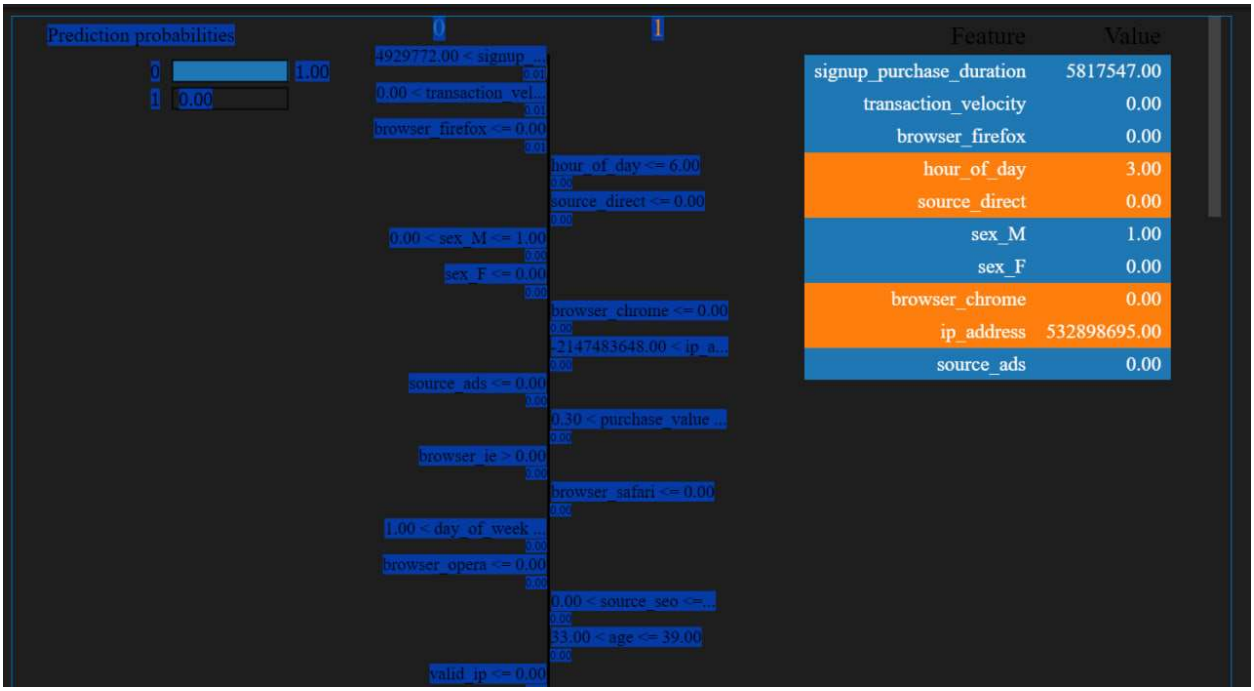

Prediction probabilities
0
1
1.00
0.00
0.00 < transaction_vel...
browser_opera <= 0.00
browser_safari <= 0.00
hour_of_day <= 6.00
4929772.00 < signup_...
browser_firefox <= 0.00
source_direct <= 0.00
sex_F <= 0.00
source_ads <= 0.00
0.00 < source_seo <=...
0.30 < purchase_value...
1.00 < day_of_week...
0.00 < sex_M <= 1.00
browser_chrome <= 0.00
browser_ie > 0.00
Feature
Value
transaction_velocity 0.00
browser_opera 0.00
browser_safari 0.00
hour_of_day 3.00
signup_purchase_duration 5817547.00
browser_firefox 0.00
source_direct 0.00
sex_F 0.00
source_ads 0.00
source_seo 1.00

Random Forest Model


Gradient Boosting Model



Multi-Layer Perceptron (MLP) Model



# Conclusion

E-commerce platforms have revolutionized the way people conduct business and make purchases. However, with the increasing prevalence of online transactions, the risk of fraudulent activities also rises. To effectively protect e-commerce platforms from fraud, it is crucial to understand the patterns and trends associated with fraudulent transactions. This report analyzes the distribution of browsers, user age demographics, and fraudulent transactions by country, along with the average purchase values in different countries. Based on these insights, recommendations for protecting against fraud in e-commerce are provided.

**Distribution of Browsers**

Understanding the distribution of browsers used by customers on e-commerce platforms can help optimize website compatibility and performance, ensuring a seamless user experience. Based on the provided data, the following insights can be derived:

Chrome is the most commonly used browser, with a count of 61,096, followed by Internet Explorer (36,498) and Safari (24,521).

Firefox closely follows Safari with a count of 24,477, while Opera has the lowest count of 3,655.

**Recommendation** E-commerce platforms should prioritize optimizing their websites for the most popular browsers, such as Chrome and Internet Explorer, to ensure compatibility and performance. Conducting thorough testing across different browsers is essential to provide a seamless user experience for all customers.

**User Age Demographics**

Analyzing the age demographics of users engaging in e-commerce transactions provides valuable insights for targeted marketing campaigns and user segmentation strategies. Based on the provided data, the following trends can be observed:

The age range of 29.60-34.60 has the highest frequency of users, indicating a significant presence of users in their late twenties and early thirties.

The distribution shows a gradual increase in frequency from younger age ranges to the peak range, followed by a gradual decrease.

Age ranges from 18.00-23.00 to 38.30-43.30 exhibit consistently high frequencies, suggesting a stable user base within this age range.

User participation declines among older age groups, with low frequencies observed in age ranges from 67.30-72.30 to 73.10-78.10.

**Recommendation** E-commerce platforms should focus their marketing efforts and user segmentation strategies on the age groups with the highest user engagement, particularly users

in their late twenties and early thirties. Additionally, there is an opportunity to attract and engage a wider range of age demographics by developing targeted marketing campaigns for older age groups.

**Fraudulent Transactions by Country**

Identifying countries with a higher occurrence of fraudulent transactions can help implement targeted fraud prevention measures, enhance security protocols, and focus investigative efforts. Based on the provided data, the following insights can be derived:

Australia (7,604) and the United States (4,014) have the highest counts of fraudulent transactions, indicating a higher susceptibility to fraud within the dataset.

The Syrian Arab Republic (528) exhibits a relatively high count of fraudulent transactions compared to its population size, suggesting a higher occurrence of fraudulent activities originating from this country.

Italy (122) and Japan (120) have moderate counts of fraudulent transactions, indicating a noticeable presence of fraud from these countries.

The Russian Federation (107), China (104), Taiwan; Republic of China (ROC) (93), and the United Kingdom (92) have relatively lower counts of fraudulent transactions but still contribute to overall fraudulent activity.

The Korea Republic of (77) has a relatively low count of fraudulent transactions, suggesting a lower occurrence of fraud originating from South Korea.

**Recommendation** E-commerce platforms should implement targeted fraud prevention measures and enhance security protocols in countries with higher counts of fraudulent transactions, such as Australia, the United States, and the Syrian Arab Republic. Additionally, monitoring and investigating fraudulent activities from countries with moderate or lower counts can help prevent potential risks.

**Average Purchase Values in Top 10 Countries**

Analyzing the average purchase values in different countries provides insights into customer behavior and can assist in targeting marketing efforts accordingly. Based on the provided data, the following insights can be observed:

Palestinian Territory, Occupied (0.5) has the highest average purchase value among the top 10 countries, indicating a tendency for higher-value purchases despite its smaller population or market size.

Slovenia (0.467) and Finland (0.459) demonstrate relatively high average purchase values, suggesting a propensity for customers from these countries to make significant purchases.

Malaysia (0.429), Bulgaria (0.426), and Libyan Arab Jamahiriya (0.410) have slightly lower average purchase values but still show a notable tendency for higher-value purchases.

Ireland (0.409) and Azerbaijan (0.383) also exhibit relatively high average purchase values, indicating a propensity for customers from these countries to make significant purchases.

Oman (0.382) and Bosnia and Herzegovina (0.370) have slightly lower average purchase values but still demonstrate a tendency for higher-value purchases.

**Recommendation** E-commerce platforms should consider the customer behavior and average purchase values in different countries to tailor their marketing efforts accordingly. Targeting strategies can be developed to attract customers from countries with higher average purchase values, such as the Palestinian Territory, Occupied, Slovenia, Finland, and others in the top 10.