Team 3

FRAUD DETECTION IN E-COMMERCE

➤ *What is E-commerce?*

- ➤ Electronic commerce - e-commerce is a business model that facilitates the transactions of goods, services, funds or data over an electronic network, specifically the internet.
- ➤ E- commerce business transactions can be sub-divided into four categories – Business to Business, Business to Consumer, Consumer to Business, and Consumer to Consumer.
- ➤ E- commerce has allowed firms to establish a market presence, by providing efficient distributed chain for their products and services.
- ➤ The upsurge of e-commerce is also accompanied by a drastic increase in Fraud.

Why is it important to detect fraud?

- Most of the eCommerce businesses focus on acquiring more customers, in-order to generate more revenue and achieve their targets. Hence, it is important to keep a tab on all kind of fraudulent activities that their business is prone to.
- The Fraud Detection System can help e-commerce companies in various ways -
  - Increase customer retention
  - Increase company revenue
  - Help increasing company's brand value

Role of Analytics in E-Commerce/How do we detect the fraud

- An important early step in fraud detection is to identify factors that can lead to fraud.
- What specific phenomenon typically occurs before, during, or after a fraudulent incident?
- What characteristics are generally seen in fraud?
- When these aspects, factors and characteristics are pinpointed, predicting and detecting fraud becomes a much more manageable task.
- It is important to both correctly identify fraudulent behavior when it arises and to not flag normal user as fraudulent one, thereby alienating customer base and achieving high recall.
- All the patterns can captured by various Machine learning algorithms.

Objective

- The approach to evaluate and ascertain what “normal” user behavior is in terms of time spent by the user browsing the website, how fast they move through the billing, no.of users per device, the IP address, time of the year, etc.
- These characteristics are all represented by numbers and tend to be uniform for a “normal” user.
- However for a fraudulent user, some of these numbers many deviate from the norm, which can be captured by the model.

Data Explanation

➤This particular project uses two data sets which are as follows:

1) Fraud data

2) Ip_address_Country

➤ In the first data set we have five of the features were numerical; the remaining six were categorical features.

➤The features included:

userid

signup-time

purchase-time

purchase-value

device-id, source

browser

sex,

age,

ip- address

class.

➤In the second data set we have two numerical feature and one categorical feature.

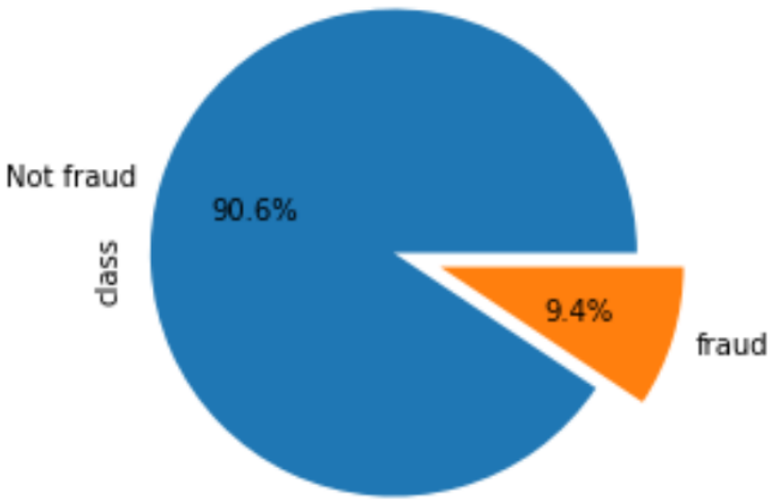
➤The features included:

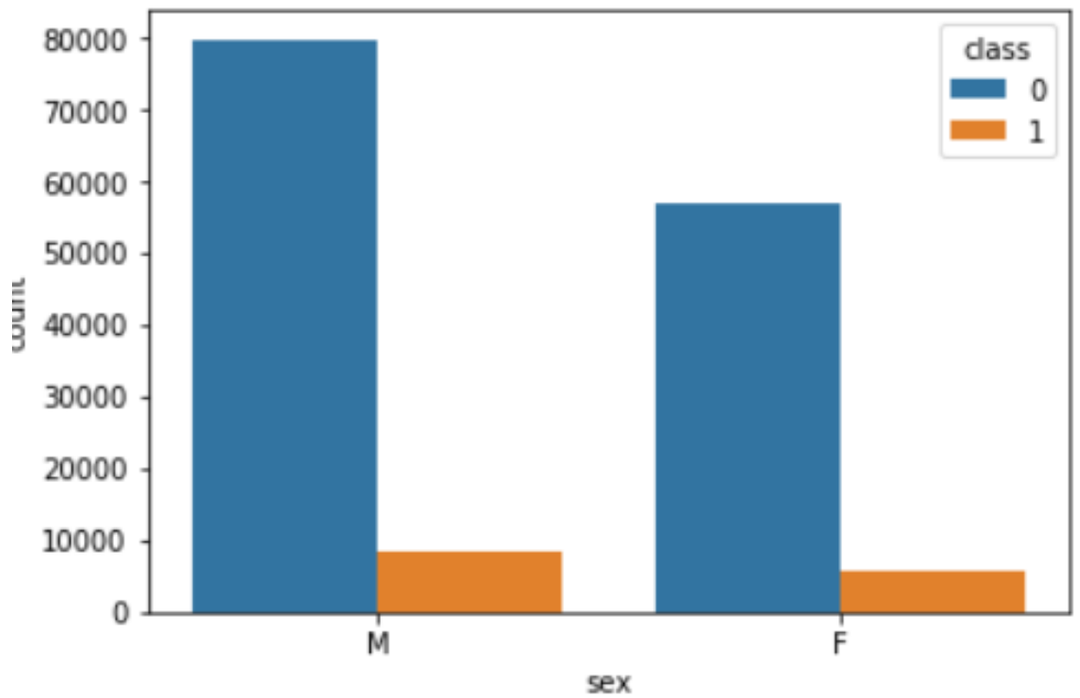
Lower-bound IP

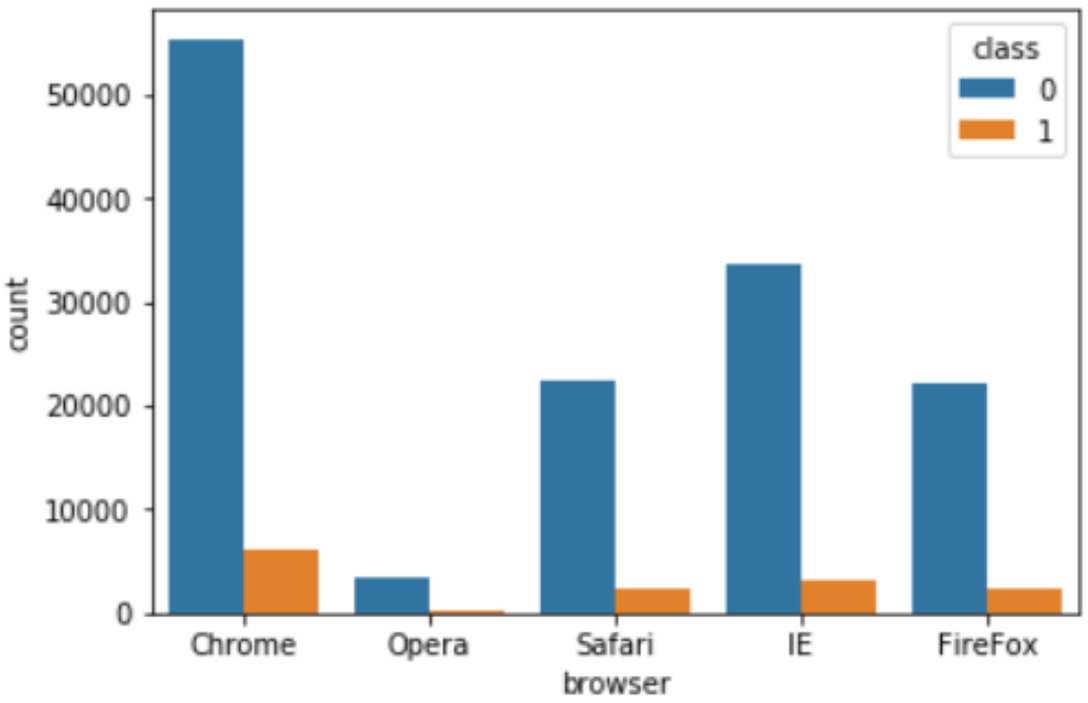
Upper-Bound IP

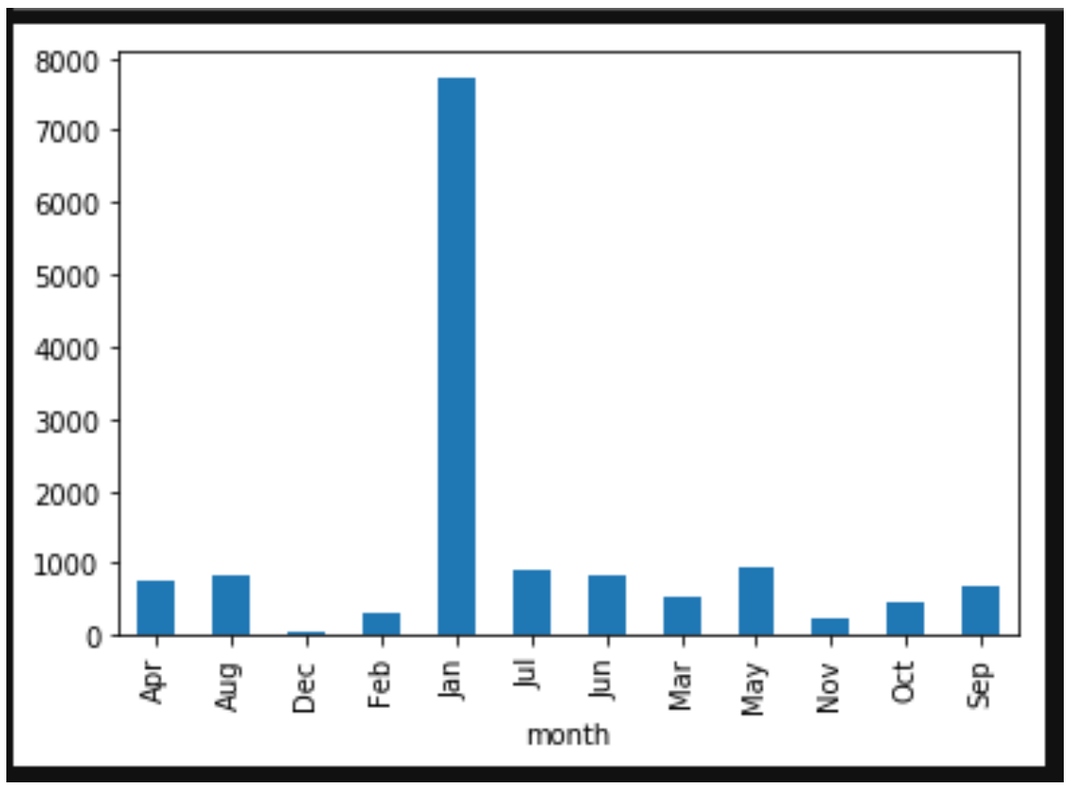
Country

Insights from our data. Fraud Vs sex, browser and month





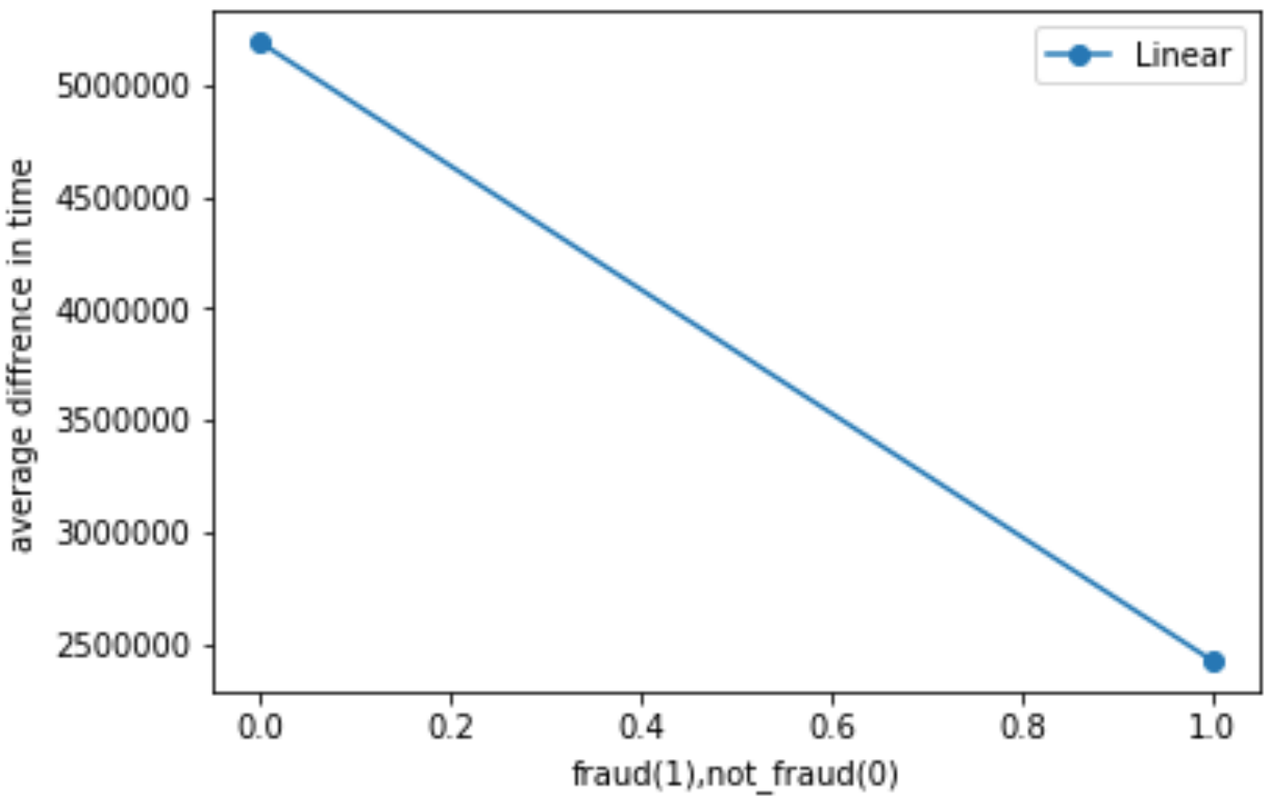




Feature engineering

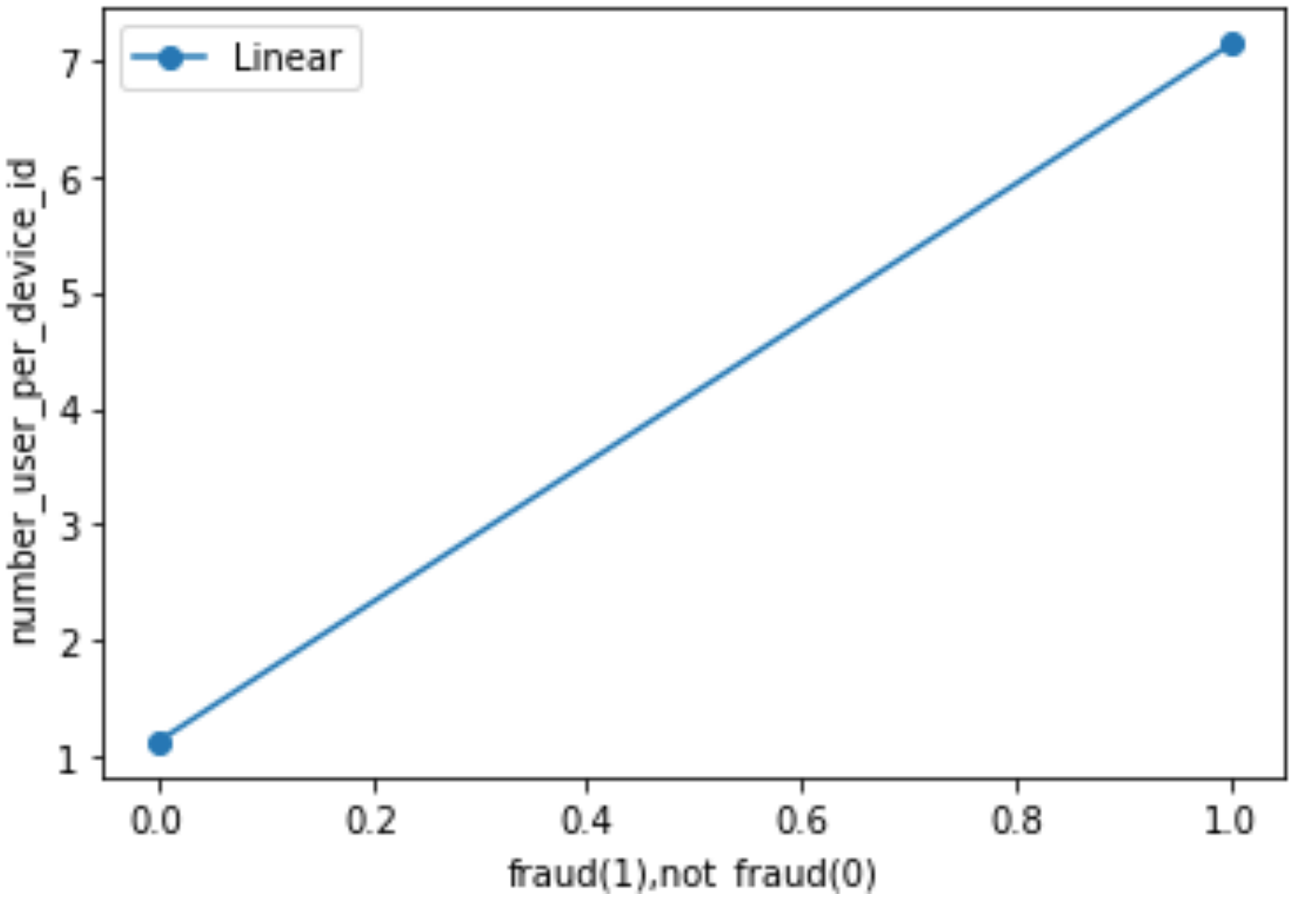
- New columns created in-order to identify the pattern in case of fraud from the data sets are as follows:

    1. “time_difference” which is derived after converting the difference from sign-up time to purchase time in seconds.
    2. “Number_of_users_per_device” derived from count of users per device.
    3. “Country” derived and added to the data by comparing the IP range.



Data Visualization

- Sign-up to purchase time in seconds versus Fraud as shown in Figure 1
- Number of users per device versus Fraud as shown in Figure 2

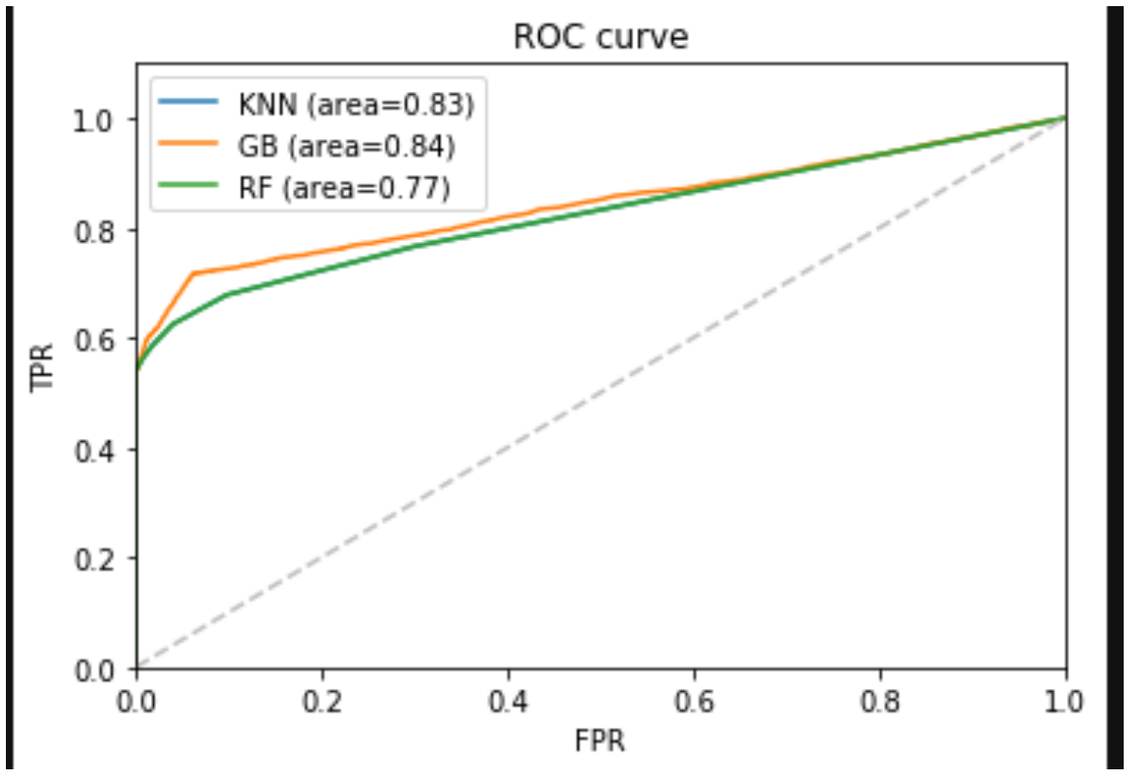


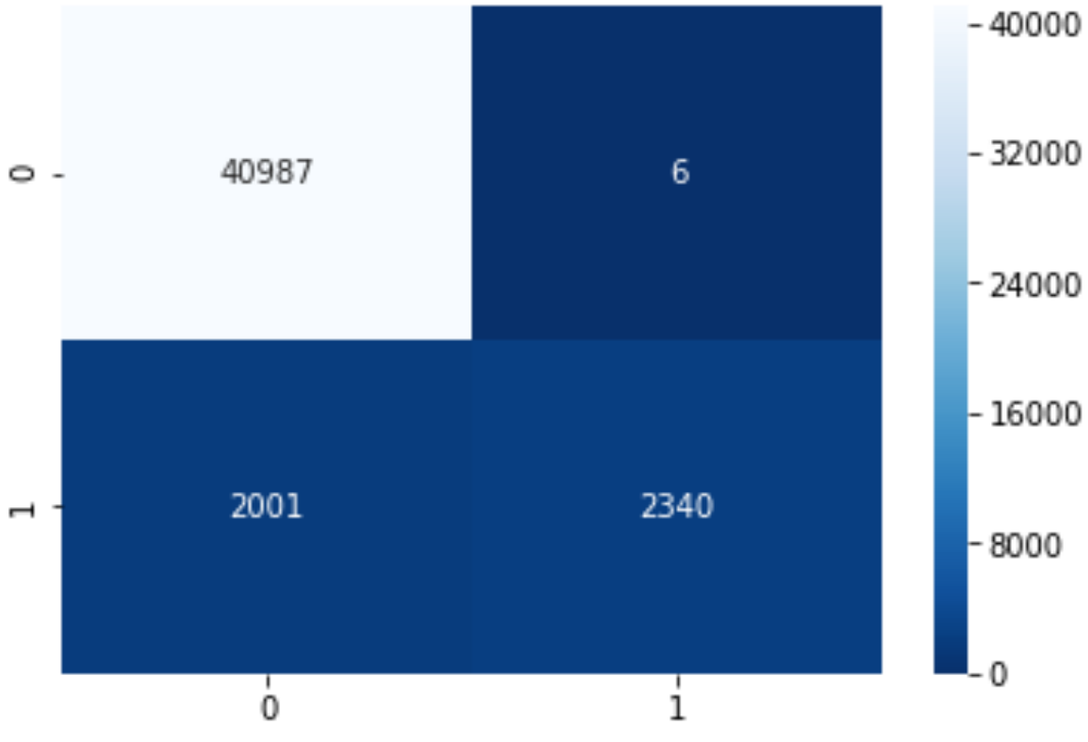
- These images provides vital insight in order to detect fraud in the given dataset

Models Used

- The goal of this project is to build an anomaly detection model that predicts the probability that the transaction of a user is fraudulent.
- We have cross validated the data on various model which can be seen from the screenshot below.
- Random forest algorithm has given us the highest accuracy with less number of false positives
- The model was the Random Forest: supervised learning algorithm based on building several decision trees and combining them to form the ensemble tree. The splitting is based on the purity of the node and the Gini index in order to identify aberrant data in the dataset has given us the highest accuracy of all the models

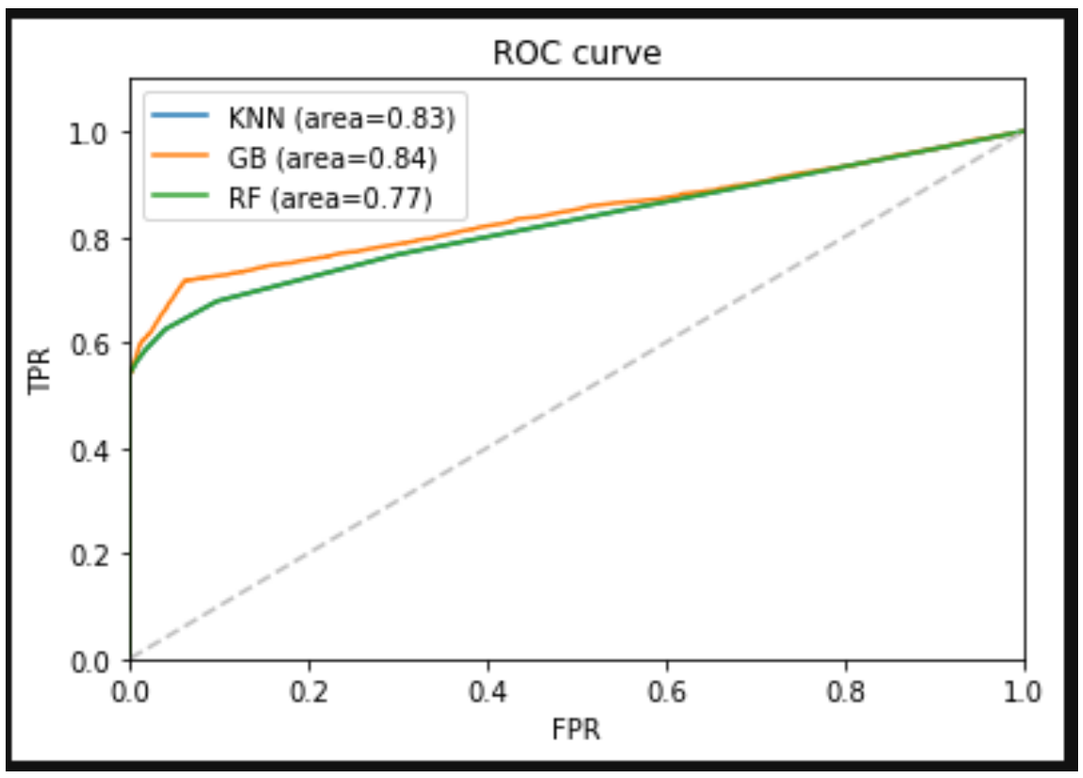
```
Logistic Regression: 0.907259 (0.002493)
Decision Trees: 0.915654 (0.002439)
Random Forest: 0.955756 (0.001860)
KNN: 0.943221 (0.002331)
Naive Bayes: 0.907259 (0.002493)
```





Random Forest Algorithm

- The criterion to train the model for Random Forest Classifier was ‘gini’, along with 10 estimators.
- Hence, the splitting decision implemented in the project was based on the Gini Index.



- The classifier was able to detect the anomaly with an accuracy of 95.55%.
- This prediction model produced ~ 91% true positives, ~4.4% false positives, 0.01% false negatives and ~6% true negatives.

Market Size

➢Propelled by rising smartphone penetration, the launch of 4G networks and increasing consumer wealth, the Indian e-commerce market is expected to grow to US$ 200 billion by 2026 from US$ 38.5 billion in 2017 .

➢Online retail sales in India are expected to grow by 31 per cent to touch US$ 32.70 billion in 2019, led by Flipkart, Amazon India and other apparel websites.

➢During 2018, electronics is currently the biggest contributor to online retail sales in India with a share of 48 per cent, followed closely by apparel at 29 per cent.

Observation/Conclusion

- Brief summary of the results generated by applying the above methods to the E-commerce data set is mentioned in this section.
- The dataset was split into two portions, where 2/3 of the dataset was used for training and the rest 1/3 of the dataset was used for testing.
- The most important features in spotting fraudulent transactions from the given e-commerce dataset were found to be:

- The speed through which the anomaly traversed from sign-up to purchase.
- Number of user ids associated which a device.

- References:

Predicting Fraud in Electronic Commerce: Fraud Detection Techniques in E-Commerce by Amitha Raghava-Raju

https://pdfs.semanticscholar.org/3535/6e66343eb909697bc2eef28673dd7d66f061.pdf