BUDT758T-Data Mining and Predictive Analytics

Group 20:

Ad Tracking Fraud Detection

Bharath Kumar Routhu • Manikanta Koneru
• Sai Vaishnav Vinjamuri • Jayasree Karthik Nandula
•Meghana Chenreddy

University of Maryland Robert H. Smith School of Business

Contents:

source: www.kaggle.com/c/fraud-detection-clicks

Introduction:

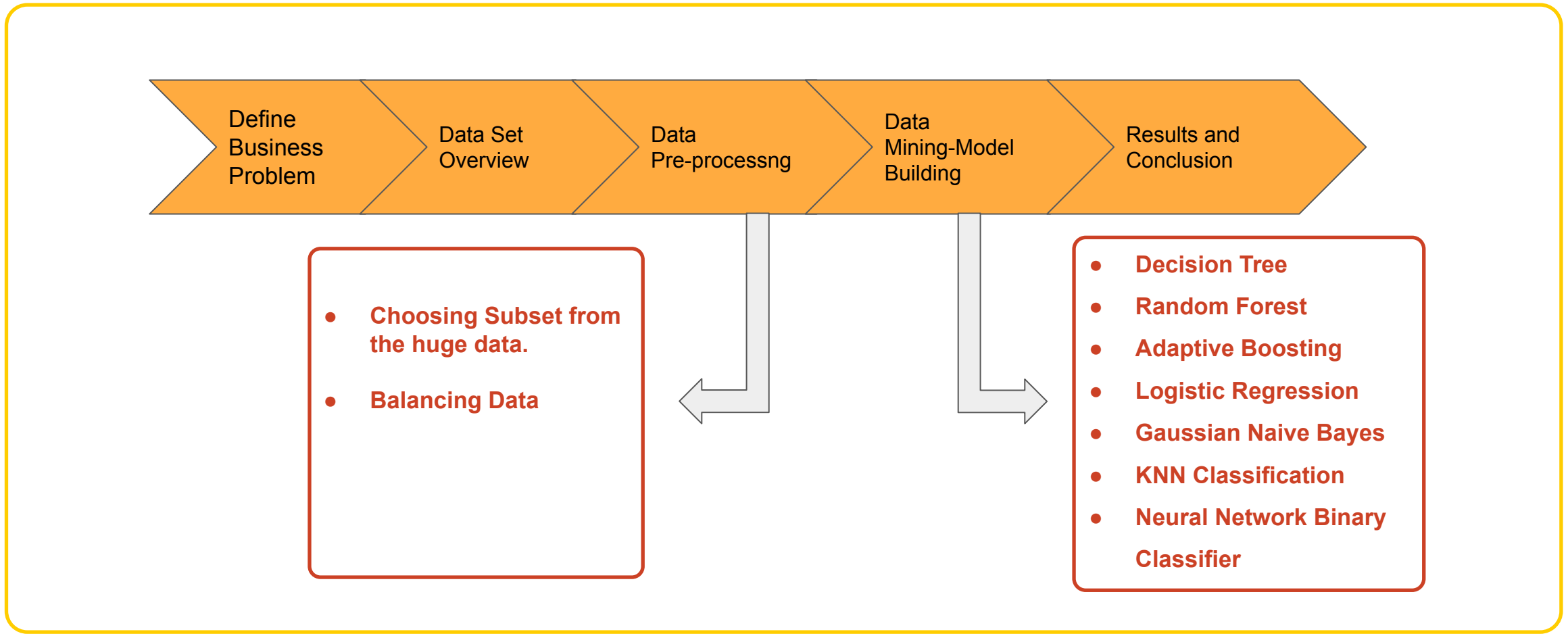
- Fraud risk is everywhere, but for companies that advertise online, click fraud can happen at an overwhelming volume, resulting in misleading click data and wasted money.

- Ad channels can drive up costs by simply clicking on the ad at a large scale.

- With over 1 billion smart mobile devices in active use every month, companies which invest large share into digital marketing therefore suffers from huge volumes of fraudulent traffic.

- Preventive measures have the risk of losing potential customers

Agenda:

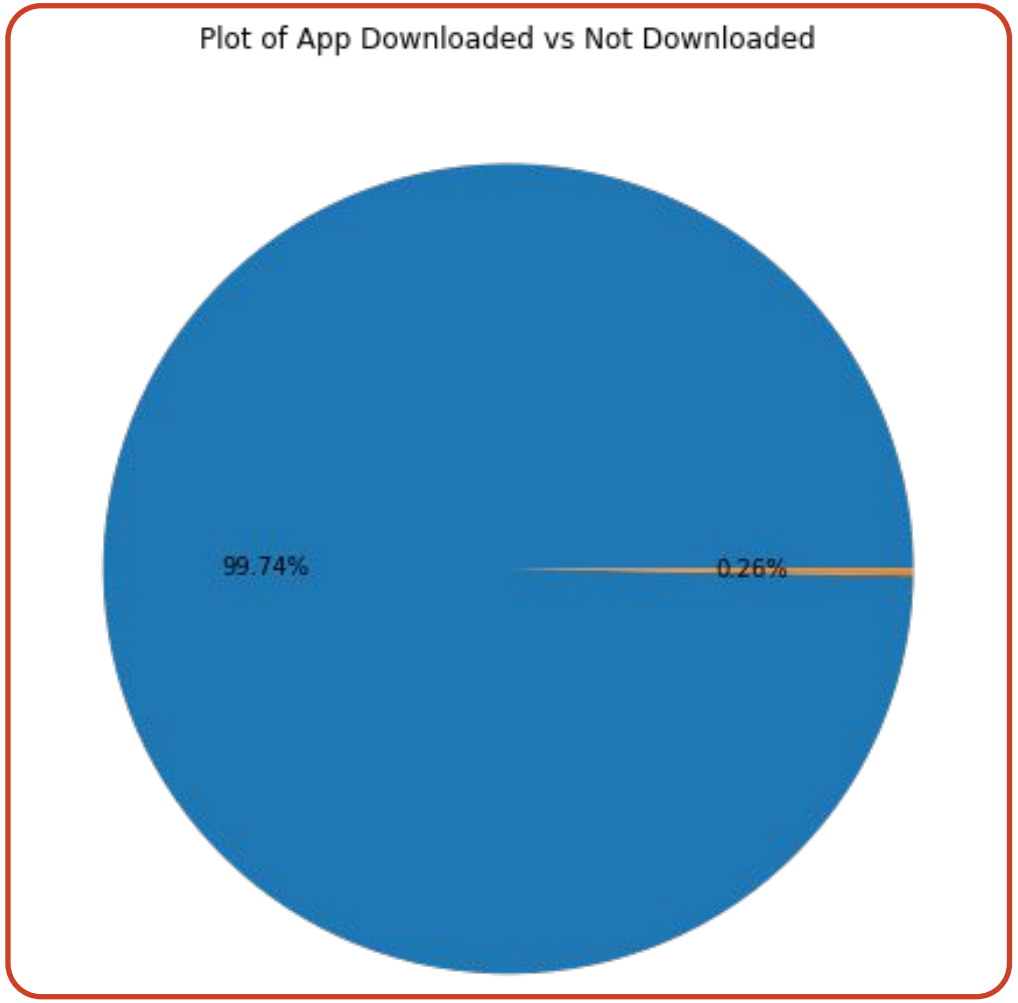

Define Business Problem
Data Set Overview
Data Pre-processng
Data Mining-Model Building
Results and Conclusion
Choosing Subset from the huge data.
Balancing Data
Decision Tree
Random Forest
Adaptive Boosting
Logistic Regression
Gaussian Naive Bayes
KNN Classification
Neural Network Binary Classifier

Objective:

- The main objective is to classify a particular transaction as fraudulent or genuine.
- The main challenge lies in misclassification of a transaction, specially marking a fraudulent as genuine (false positive)
- Thus, our challenge is to generate a model that provides high accuracy with least number of fraudulent misclassifications.

Challenges during EDA:

- Enormous data to analyze.
- Highly Imbalanced data.
- Cold Start Issue.
- Continuous Upgrade to the Model to incorporate new invitations in App, Device, OS, Channel

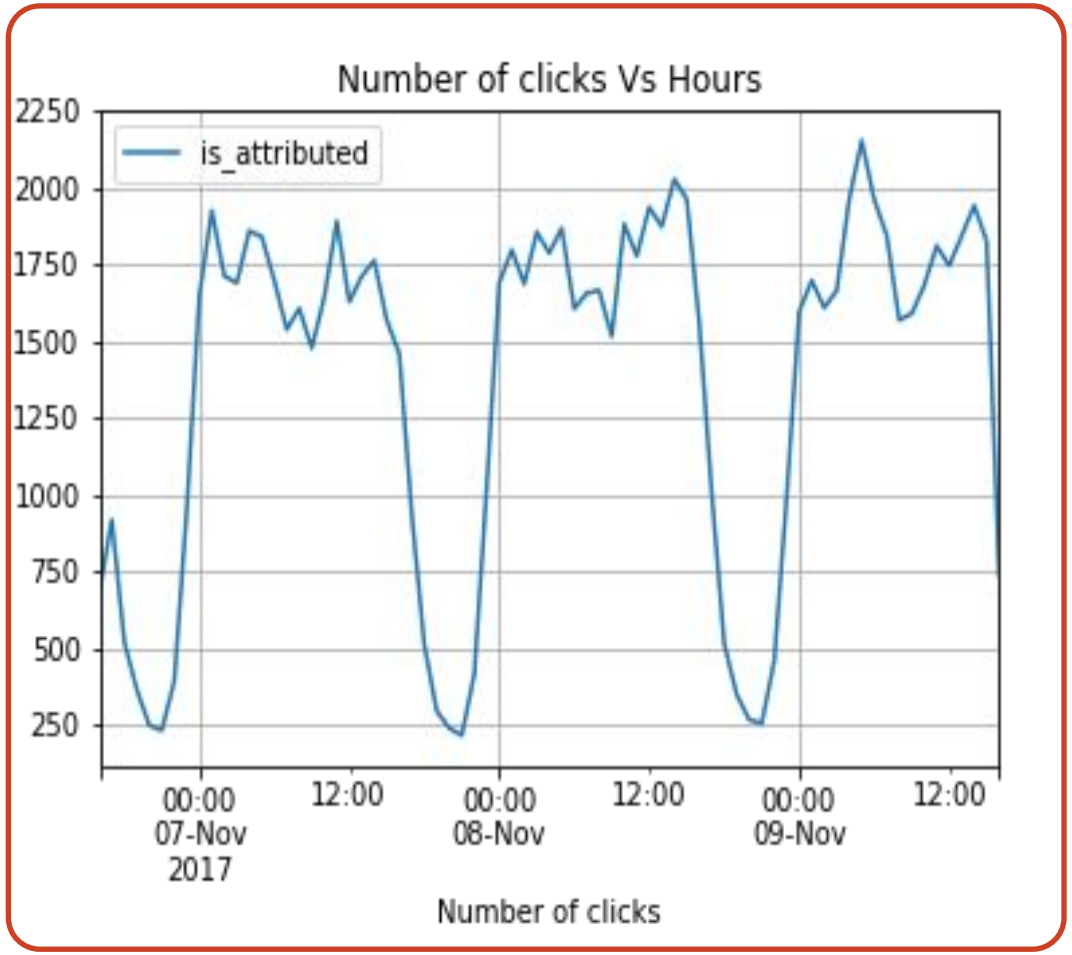


Approach :

- The Main Challenge was the enormous data to be processed.

- From the visualization we can observe a trend across days.

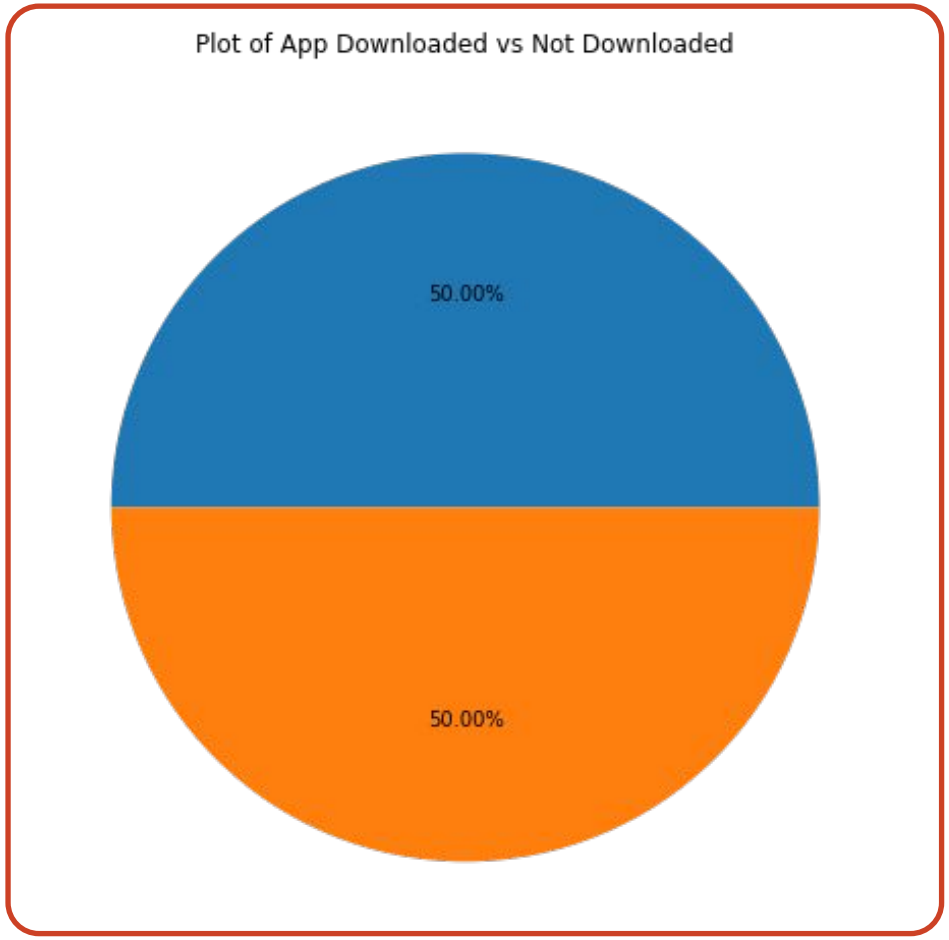
- Hence we have considered data of a Single day (7th November) for Analysis.



Approach:

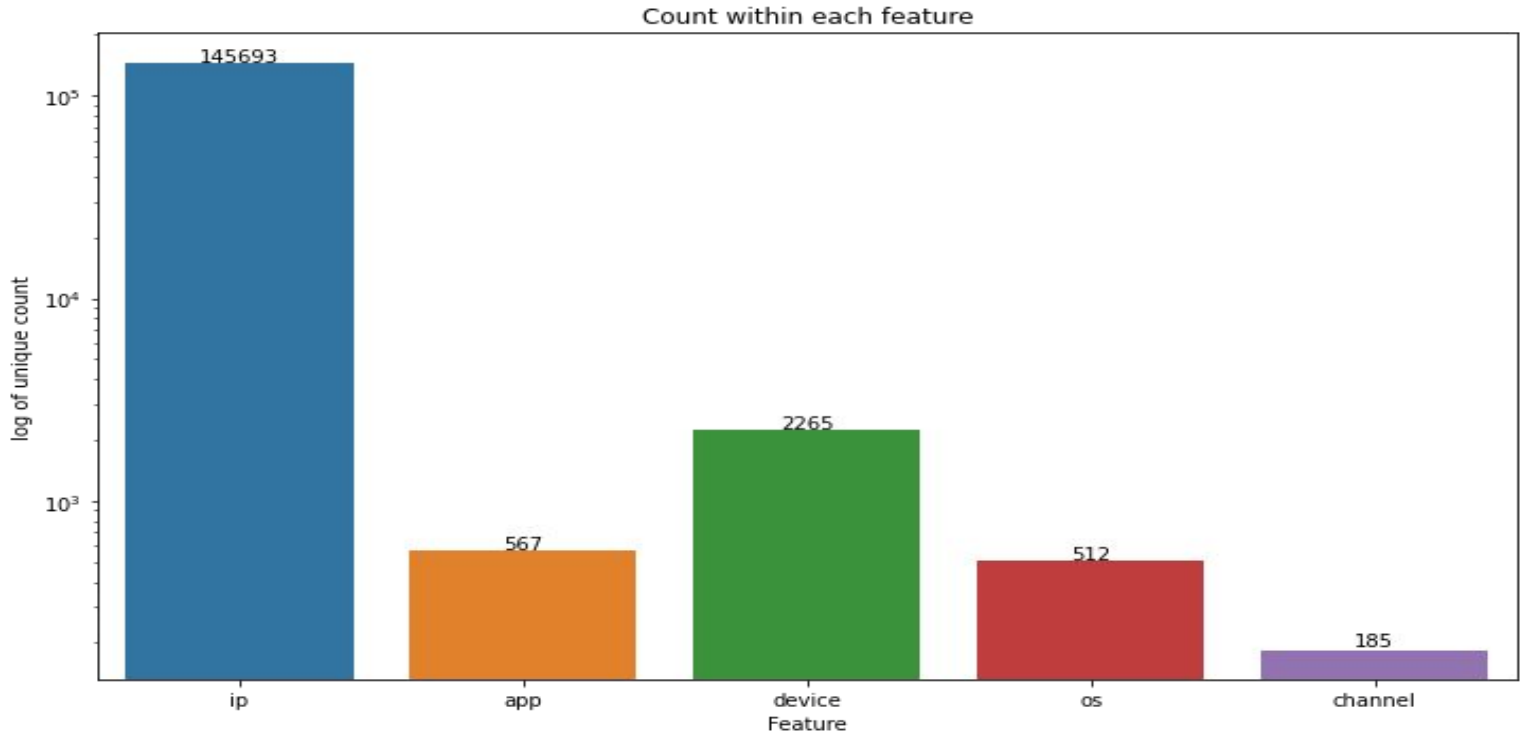
- One other challenge that hindered the reliability of the model was the highly imbalanced data set.

- In order to tackle the problem we have downsampled the number of records of the failure scenario to match the number of success scenarios



Exploratory Data Analysis (EDA):

- Level Count for each Categorical Variables.



Feature Engineering:

- The Date time Variable is partitioned into independent variables of *Day*, *Hour*, *Minutes* and *Seconds* .
- The "*clicks*" Column captures the number of clicks generated by a particular IP in a day.
- The "*ip_hour_clicks*" captures clicks generated by an IP in the respective Hour

app	device	os	channel	day	hour	min	sec	ip_hour_clicks	clicks
23	1	37	153	1	0	0	0	189	2943
39	1	13	101	1	0	0	0	2133	43939
35	1	13	274	1	0	0	0	34	259
35	1	13	274	1	0	0	0	55	720
3	1	14	442	1	0	0	2	21	403
...	...	...	...	...	...	...	...	...	...

Modeling Techniques:

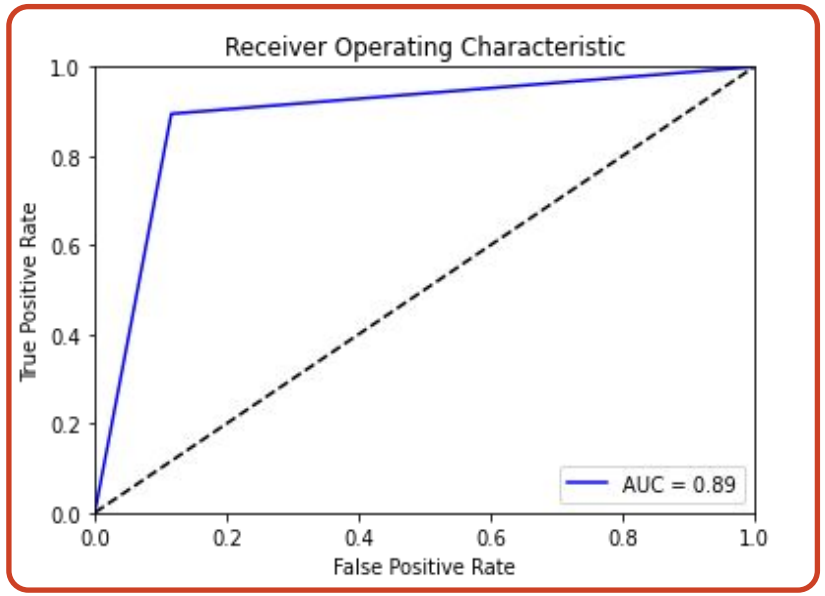
We have Modeled and Implemented several classification techniques:

- *Decision Tree*
- *Logistic Regression*
- *Gaussian Naive Bayes*
- *KNN Classification*
- *Random Forest*
- *Adaptive Boosting*
- *Neural Network Binary Classifier*

Decision tree:

Decision tree is explained by the entities - decision nodes & leaves. A decision tree has:

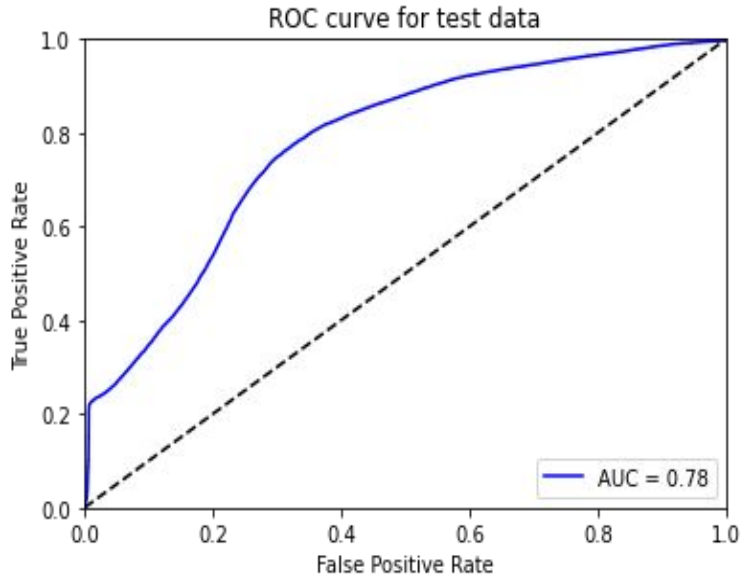
- Internal node - represents a feature test.
- Leaf node - represents a class label.
- Pathways - represent the categorization criteria .
- Branches - represent feature combinations that lead to those class labels.



- We get the Accuracy value for the ***Decision tree*** Model as 88.94%.

Logistic Regression:

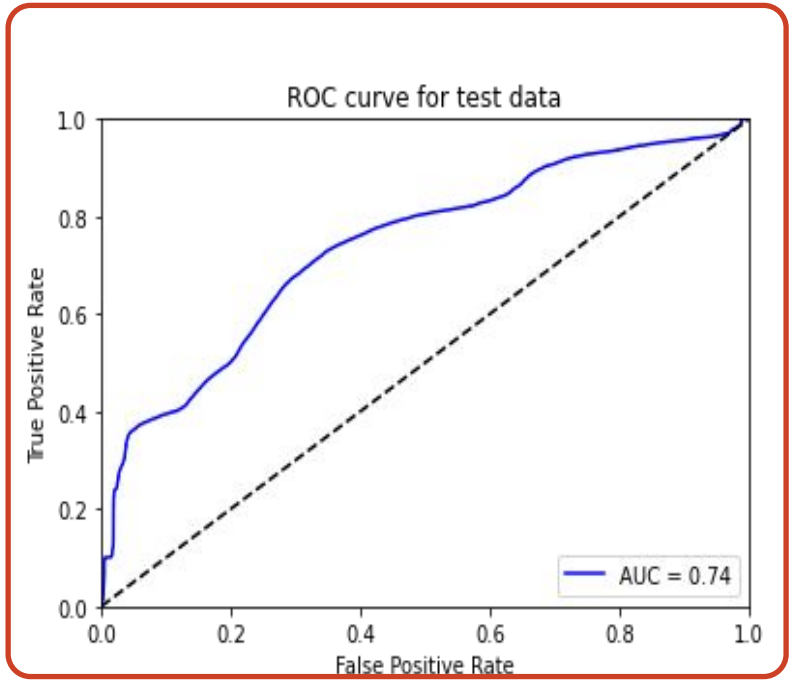
- This is used to predict the probability of a target variable.
- Helps in predicting the likelihood of an event happening or a choice being made.
- It has a binary outcome.



- We get the Accuracy value for the ***Logistic Regression*** Model as 72.31%.

Gaussian Naive Bayes:

- Gaussian Naive Bayes accepts continuous valued features and models them all as Gaussian (normal) distributions.



- To build a basic model, assume the data is characterized by a Gaussian distribution with no covariance (independent dimensions) between the parameters.

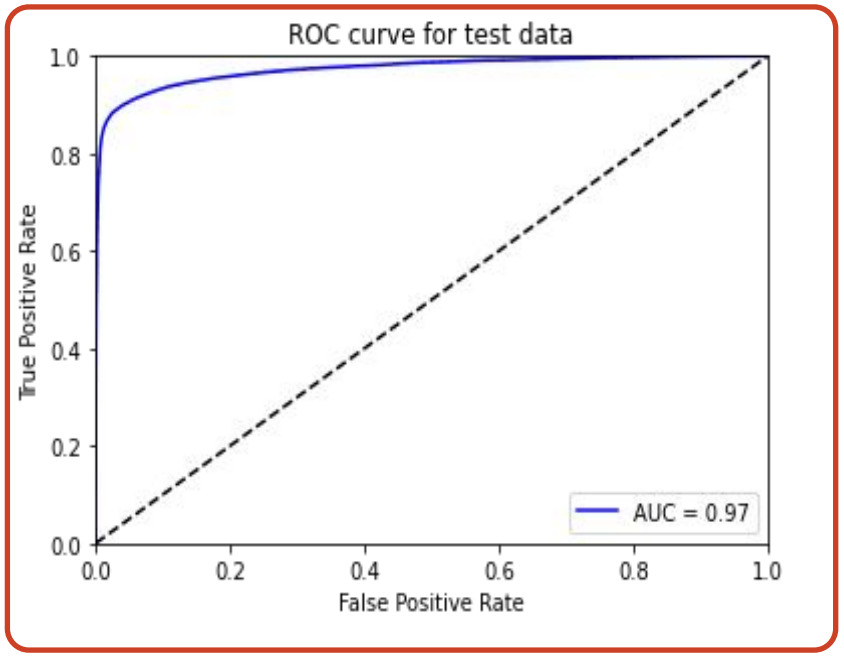
- We get the Accuracy value for the ***Gaussian Naive Bayes*** Model as 61.52%.

KNN Classification:

- Using this algorithm, a new data point is classified based on similarity in the specific group of neighboring data points.
- The algorithm calculates the distances between a specific data point in the set and any other K numbers of data points in the dataset that are near to it.
- Then vote for the category with the highest frequency.
- Typically, Euclidean distance is used to calculate distance. As a result, the final model is just labeled data in a space.
- We got the Accuracy value for the ***KNN Classification*** Model as 87.32%.

Random Forest:

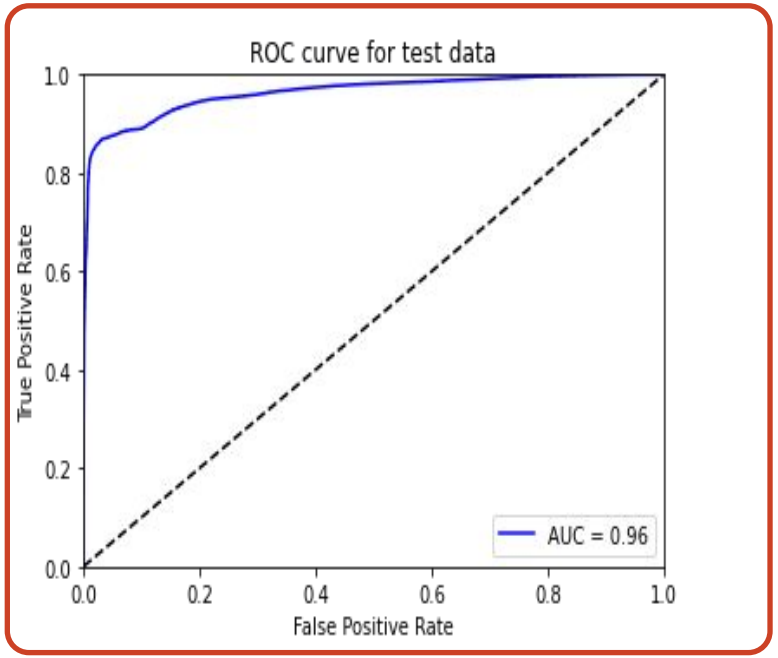
- A random forest has several separate decision trees that work together as an ensemble.
- When constructing each individual tree, it employs bagging and randomization.
- In a random forest, each tree may only choose from a random subset of characteristics.
- This creates even more variance among the trees in the model, resulting in decreased correlation and increased diversification.



- We get the Accuracy value for the ***Random Forest*** Model as 92.92%.

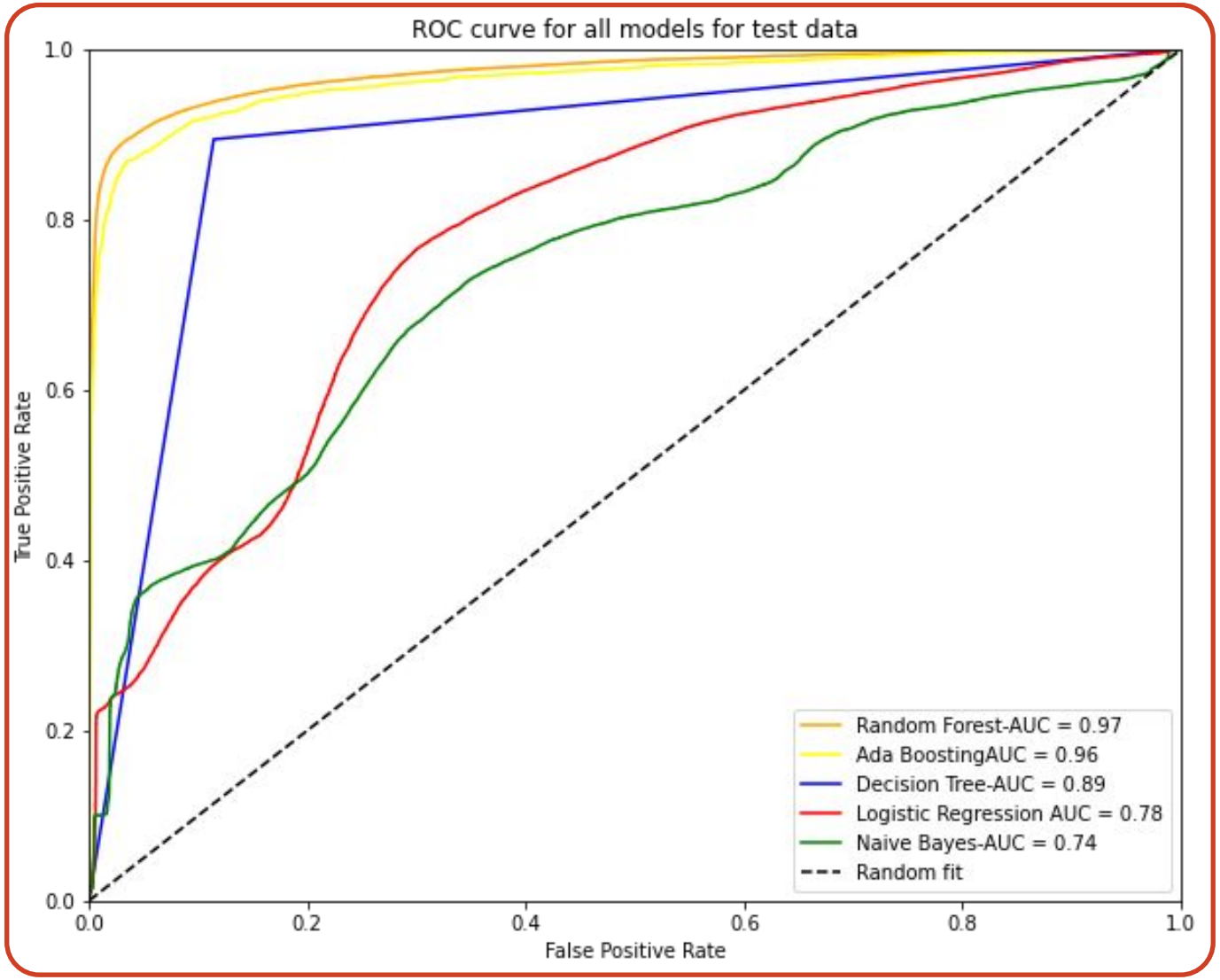
Adaptive Boosting:

- Adaboost generates a large number of Stumps, or poor learners (decision trees with one level of nodes).
- The stumps are usually associated with one data feature and its leaf nodes.
- The stumps are built in a sequential order, with the preceding stump's performance influencing or influencing the development of the following stump.
- All the stumps' developed aid in developing a sturdy or more accurate learner or model.
- We get the Accuracy value for the ***Ada Boost*** Model as 91.66%.



Best Model-“Random Forest”

- 92.92% Accuracy.
- Not Prone to *Overfitting*



Neural Network Binary Classifier:

- Neural nets are a means of doing machine learning, in which a computer learns to perform some task by analyzing training examples.
- Neural nets are a means of doing machine learning, in which a computer learns to perform some task by analyzing training examples.
- These layers have multiple nodes based on the independent and dependent variables and are interconnected with weighted neurons.
- We get the Accuracy value for the ***Neural network Binary Classifier*** Model as 92.16%.

Business Perspective:

Model evaluation is based on the below business problem:

- An ad agency wishes to implement a model that evaluates the clicks and download trends generated through their digital marketing platform.
- They wish to use the model to assess their profit margins they earn on each click that leads to a successful download.
- In these models, Misclassifications errors can lead to wrong estimates.
- Hence in order to come up with a Baseline amount we should make sure the number of False Positives are minimized.

Best Model:

- *Random Forests, Adaptive Boosting* and *Neural Network Classifier* would be a great choice.

Results and Inferences:

Model Type	Accuracy	Percentage of False negatives in Misclassification	Percentage of False positives in Misclassification	Sensitivity	Specificity
Decision Tree	88.94%	47.7%	52.3%	0.894	0.883
Random Forests	92.92%	75.9%	24.1%	0.893	0.965
Adaptive Boosting	91.66%	79.1%	20.9%	0.868	0.965
Logistic Regression	72.31%	47.3%	52.7%	0.738	0.707
Gaussian Naive Bayes	61.52%	21.3%	78.7%	0.835	0.392
KNN Classification	87.32%	77.8%	22.2%	0.798	0.942
Neural network Binary classifier	92.16%	70.3%	29.7%	0.888	0.952

Conclusions:

- We have implemented a model which can predict the outcome of a ad click based on IP, app, device, OS and channel.
- In order to equip the model with the capability to detect fraudulent clicks from a certain IP, we have created calculated fields like clicks and ip_hour_clicks.
- This as a result induces a sense on dependency on the complete data.
- However, this might not pinpoint a particular IP, the features provide the general sense of flagging such concurrent request as fraudulent.

Questions ?

Thank You!