Click Through Rate Prediction
ML Analysis

- Name Vatsal Raicha
- Date 13th November 2022
- Course Master of Science in Data Science
  upGrad / University of Arizona
- Email vatsalraicha@arizona.edu
  vatsalraicha@outlook.com

Agenda

- Objective
- Approach
- Observation
- Recommendations

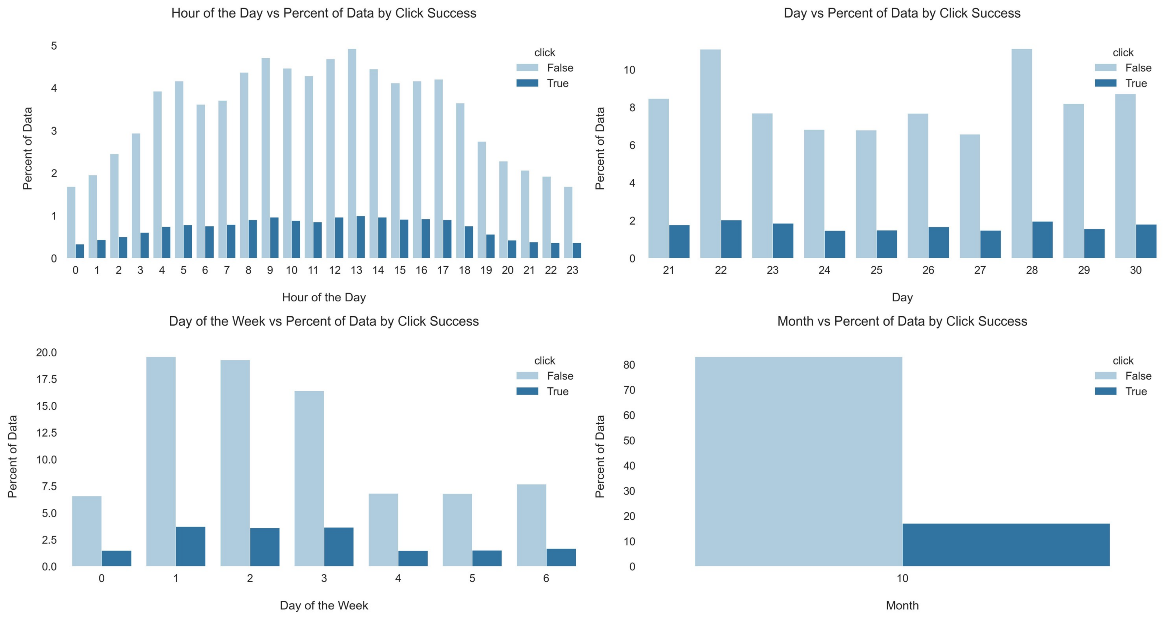
Objective

To predict whether a user will click on an ad or not.

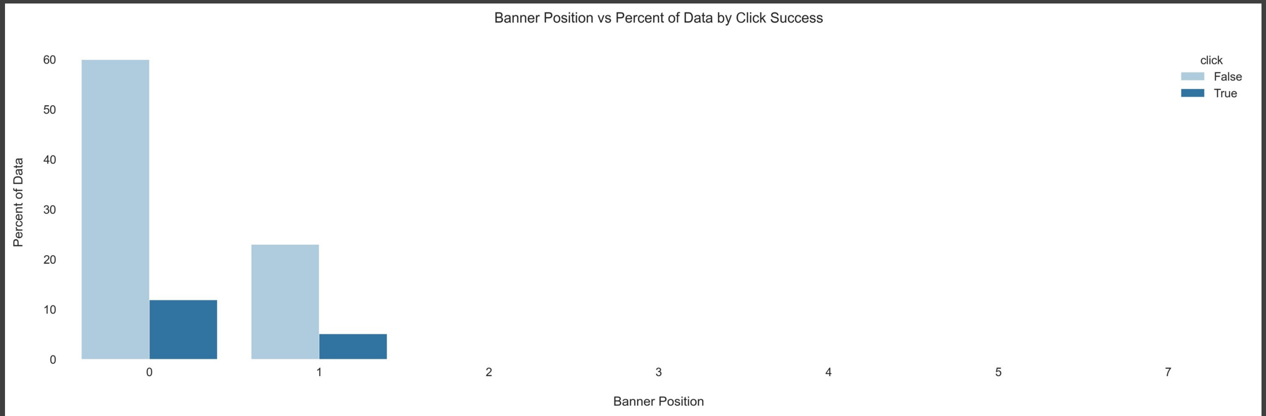
Approach

- Perform Exploratory Data Analysis
  - Most appearing categories for each column etc.
- Perform Data Cleaning
- Perform Feature removal (Remove Columns that may not be of use)
- Perform Scaling
- Apply Decision Tree Classifier
- Apply Random Forest Classifier
- Apply XGBoost Classifier
- Others - Gradient Boost Classifier, Adaboost and Bagging exercises
- Document Observations.
- Make recommendations.

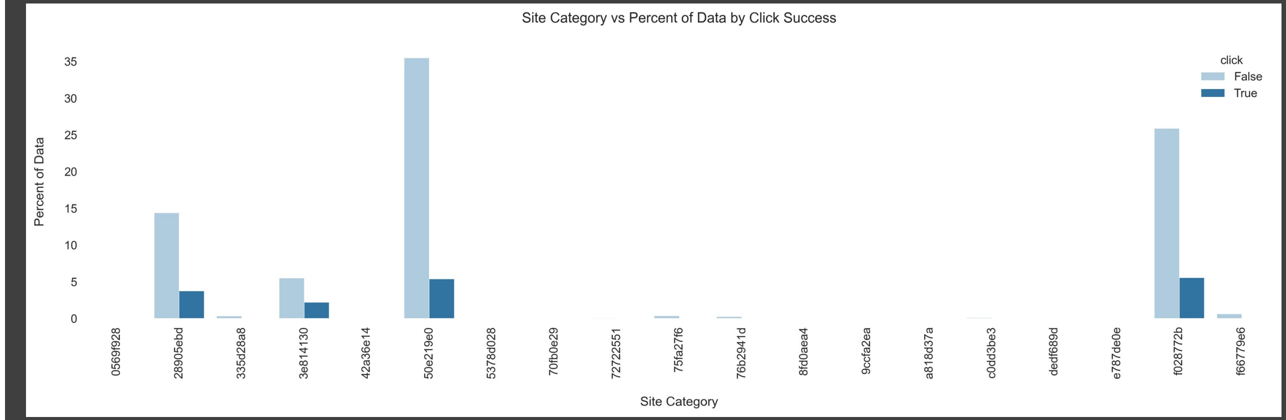
EDA



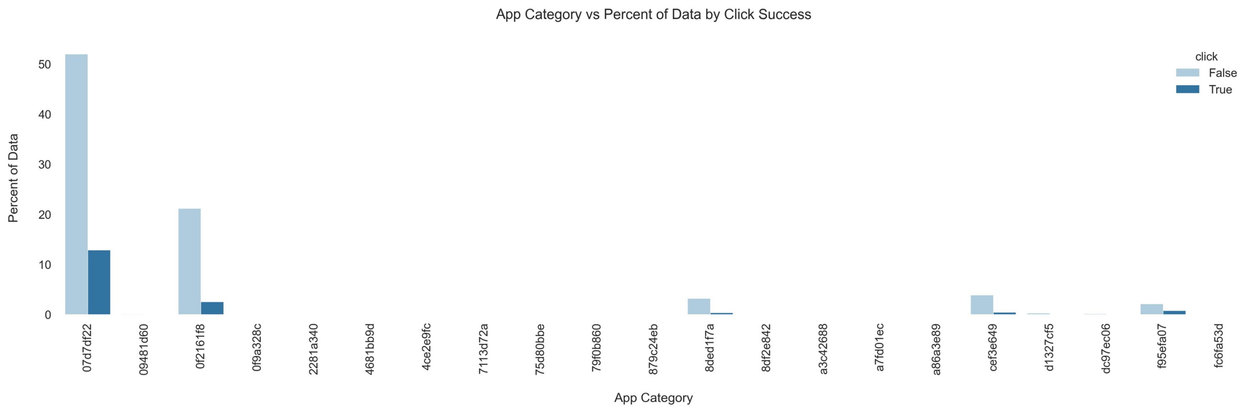
Hour of the Day vs Percent of Data by Click Success
click
False
True
Percent of Data
Hour of the Day
Day vs Percent of Data by Click Success
Day
Day of the Week vs Percent of Data by Click Success
Day of the Week
Month vs Percent of Data by Click Success
Month



Banner Position vs Percent of Data by Click Success
click
False
True
Percent of Data
Banner Position

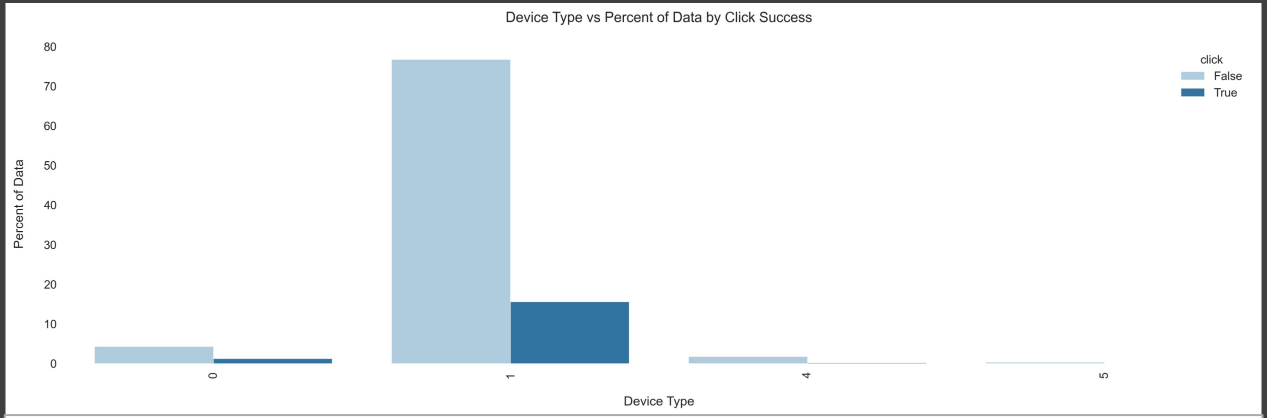


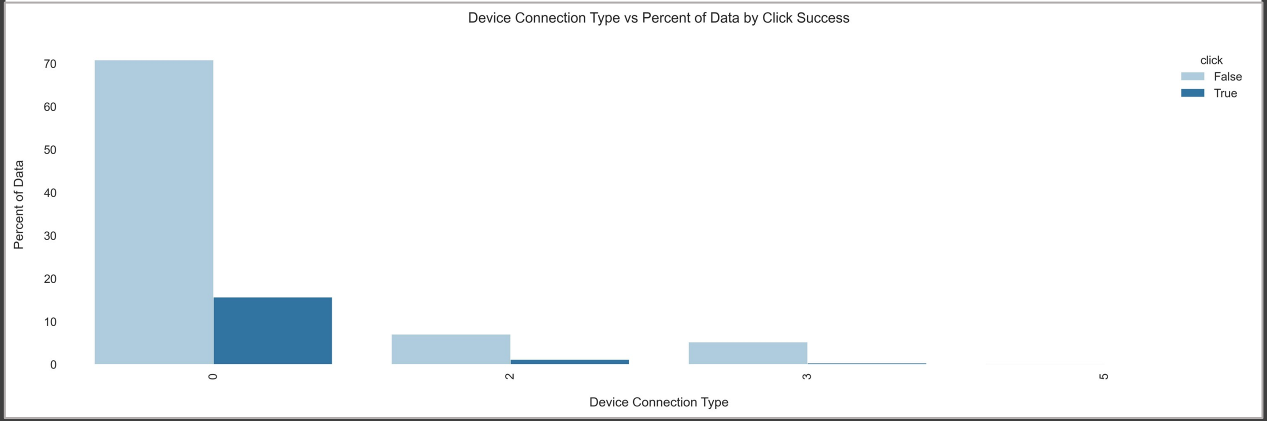
Site Category vs Percent of Data by Click Success
click
False
True
Percent of Data
0569f928
28905ebd
335d28a8
3e814130
42a36e14
50e219e0
5378d028
70fb0e29
72722551
75fa27f6
76b2941d
8fd0aea4
9ccfa2ea
a818d37a
c0dd3be3
dedf689d
e787de0e
f028772b
f66779e6
Site Category

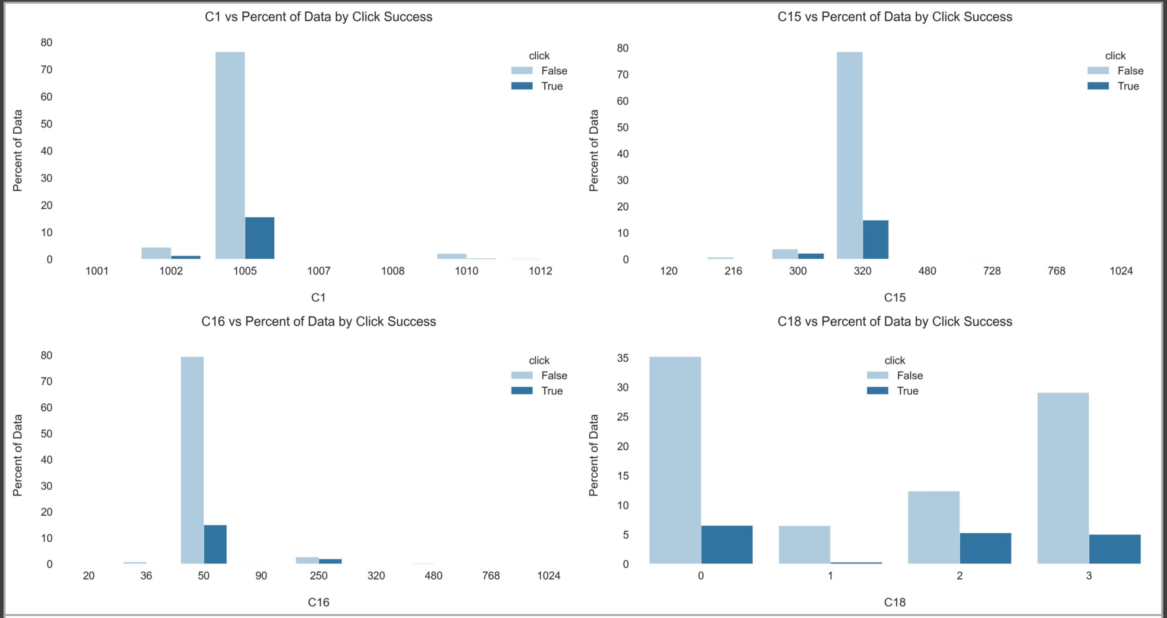


App Category vs Percent of Data by Click Success
click
False
True
Percent of Data
07d7df22
09481d60
0f2161f8
0f9a328c
2281a340
4681bb9d
4ce2e9fc
7113d72a
75d80bbe
79f0b860
879c24eb
8ded1f7a
8df2e842
a3c42688
a7fd01ec
a86a3e89
cef3e649
d1327cf5
dc97ec06
f95efa07
fc6fa53d
App Category

EDA







	Column	Type	NumOfUniqueValues	PercentUnqOfTotal	MostAppearing	MostAppearing%
10	device_ip	object	78013	78.01	6b9769f2	0.521005
9	device_id	object	16801	16.8	a99f214a	82.579826
11	device_model	object	3145	3.15	8a4875bd	6.04106
14	C14	int64	1722	1.72	4687	2.306023
3	site_id	object	1485	1.49	85f751fd	36.02236
6	app_id	object	1354	1.35	ecad2386	63.97764
4	site_domain	object	1331	1.33	c4e18dd6	37.362374
17	C17	int64	399	0.4	1722	10.99111
20	C20	int64	154	0.15	-1	46.767468

EDA

- Columns - day, dayoftheweek, hour, month will be retained
- Columns – app_id, device_ip, device_model, site_id and site_id have too many unique values and their most appearing value doesn’t account for a lot of data, hence I would drop these.

	Column	MostAppearingValue	Percentage	Col_Mean	Col_Max	Col_Min	ColRepresentationOfData	Col_UniqueValues
17	device_ip	6b9769f2	0.521005	0.001282	0.521005	0.001	1.547015	78013
18	device_model	8a4875bd	6.041060	0.031797	6.041060	0.001	16.482165	3145
24	site_id	85f751fd	36.022360	0.067340	36.022360	0.001	63.162632	1485
11	app_id	ecad2386	63.977640	0.073855	63.977640	0.001	74.446744	1354
23	site_domain	c4e18dd6	37.362374	0.075131	37.362374	0.001	67.180672	1331

Data Cleaning/Preparation/Formatting

- For Columns – app_domain, app_category, site_category, device_id, lets look at their top 5 unique values

Column Values	app_domain MostAppearingValue	Percentage	app_category MostAppearingValue	Percentage	site_category MostAppearingValue	Percentage	device_id MostAppearingValue	Percentage
0	7801e8d9	67.464675	07d7df22	64.769648	50e219e0	40.839408	a99f214a	82.579826
1	2347f47a	12.893129	0f2161f8	23.644236	f028772b	31.408314	c357dbff	0.062001
2	ae637522	4.701047	cef3e649	4.300043	28905ebd	18.107181	0f7c61dc	0.051001
3	5c5a694b	2.850029	8ded1f7a	3.519035	3e814130	7.668077	afeffc18	0.034000
4	82e27996	1.889019	f95efa07	2.868029	f66779e6	0.634006	936e92fb	0.027000

- We will replace the remaining values in these columns while maintaining the proportion of the spread of these 5 unique values. Result is -

Column Values	app_domain MostAppearingValue	Percentage	app_category MostAppearingValue	Percentage	site_category MostAppearingValue	Percentage	device_id MostAppearingValue	Percentage
0	7801e8d9	77.666777	07d7df22	65.668657	50e219e0	41.391414	a99f214a	99.825998
1	2347f47a	12.893129	0f2161f8	23.644236	f028772b	31.833318	c357dbff	0.062001
2	ae637522	4.701047	cef3e649	4.300043	28905ebd	18.354184	0f7c61dc	0.051001
3	5c5a694b	2.850029	8ded1f7a	3.519035	3e814130	7.775078	afeffc18	0.034000
4	82e27996	1.889019	f95efa07	2.868029	f66779e6	0.646006	936e92fb	0.027000

- These unique values will now simply be replaced with 0,1,2,3,4, so that they are not strings anymore.

via this I was able to avoid unnecessary hashing and represent the data in almost the same way

Scaling and prepping data for Model Building

- Rest of the columns will be kept, scaled
- Train Test Datasets will be created

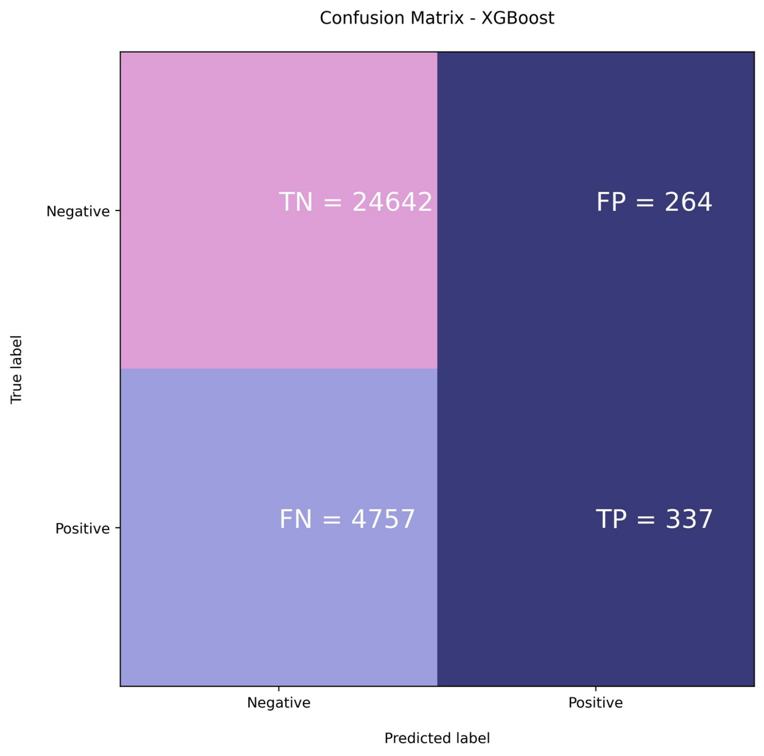
Scaling and prepping data for Model Building

- Rest of the columns will be kept, scaled
- Train Test Datasets will be created

Model Performance – XGBoost

- XGBoost Slightly outperformed Random Forest

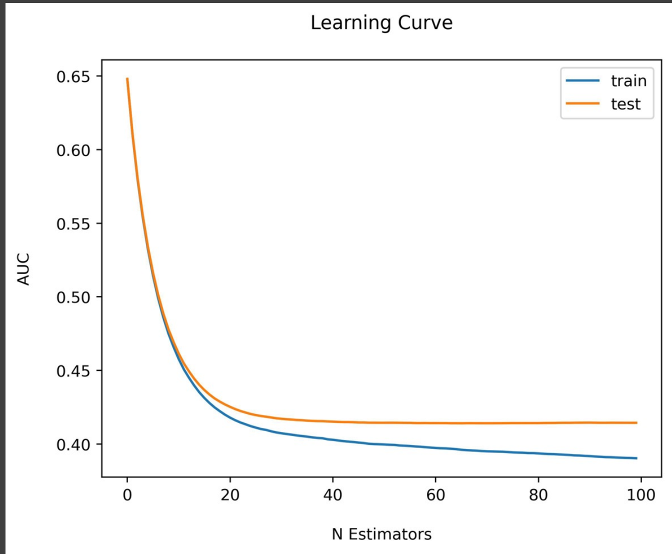
Confusion Matrix



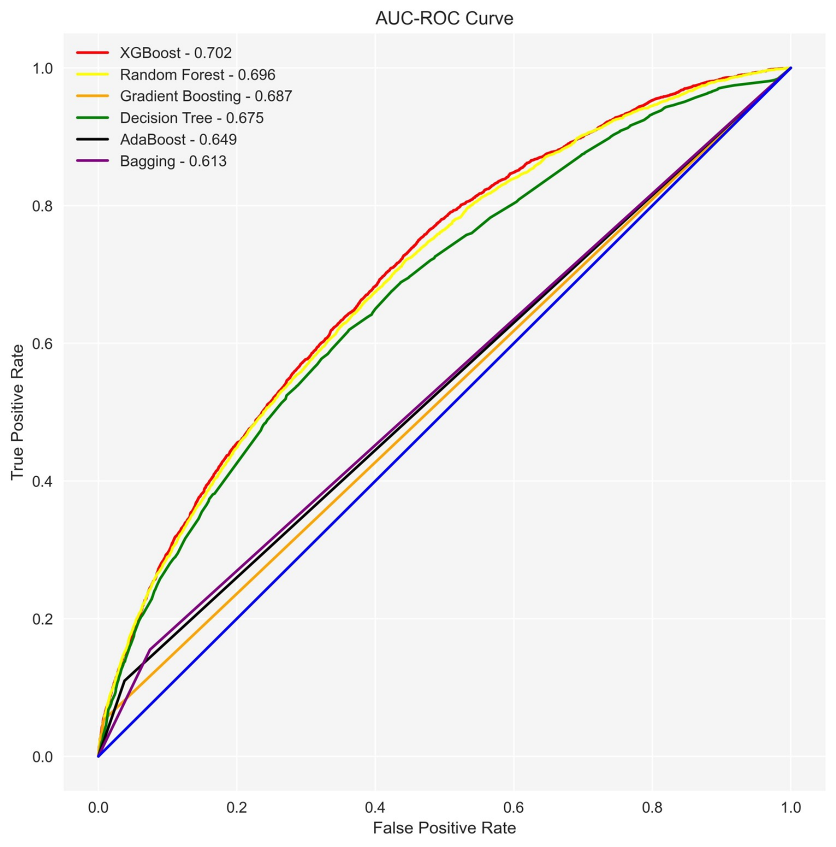
Classification Report

	precision	recall	f1-score	support
0	0.84	0.99	0.91	24906
1	0.56	0.07	0.12	5094
accuracy			0.83	30000
macro avg	0.70	0.53	0.51	30000
weighted avg	0.79	0.83	0.77	30000

Evaluation metrics

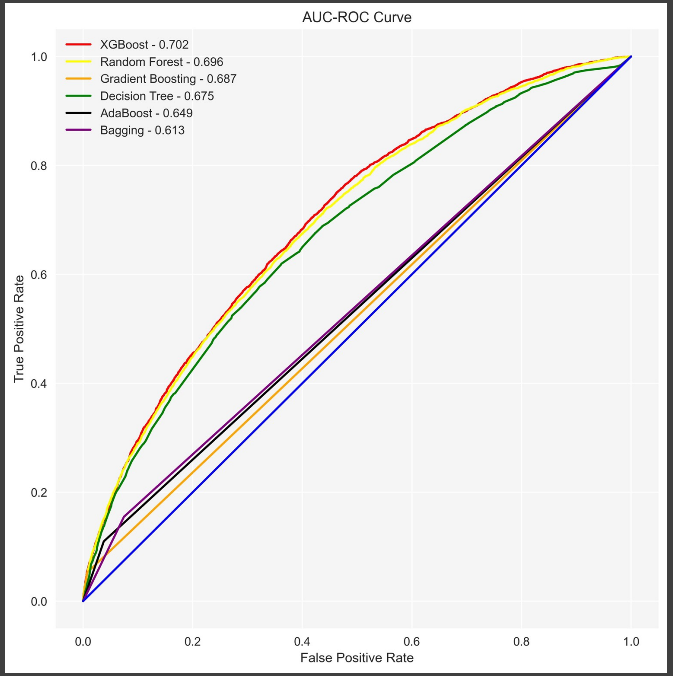


ROC AUC Score of 0.709



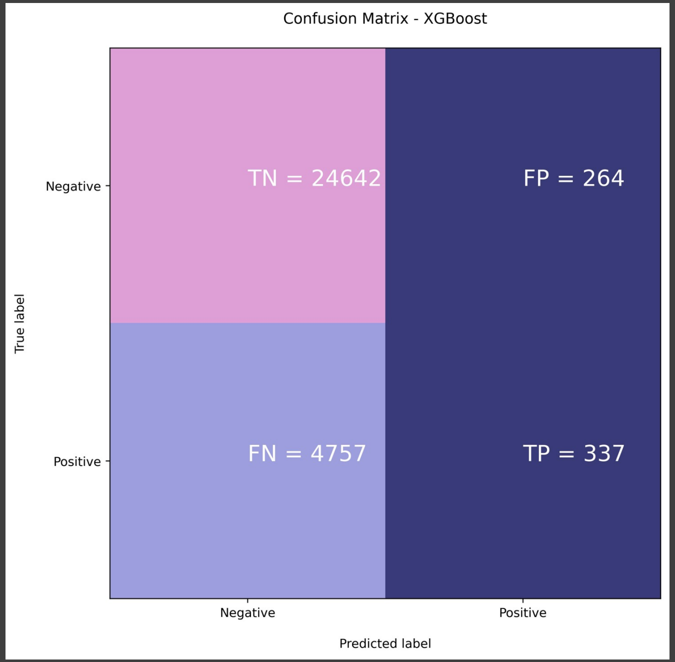
ROC Curve

- The figure shows the ROC curve of all the models tested. Higher the value of AUC, better the model is at predicting our classes.
- We can infer that ADA boost classifier has the highest AUC curve, which makes it the best among the others
- Single decision tree has the lowest AUC, as expected.



Confusion Matrix

- The confusion matrix clearly shows that the model is able to predict majority of class 0 and 1 correctly
- However further improvements can be made to reduce the false negative rate
- This can be achieved by feeding more class 1 examples for our model to learn from



What to look out for?

- A false negative predicted by our model indicates that the user has actually clicked the ad but the model predicted otherwise.
- This could potentially cause loss of revenue as we will not be able to target our audience with relevant ads and solutions.
- A false positive predicted by our model indicates that the user has not clicked ad yet our model predicted otherwise.
- This could lead to wrong ads pushed to our target audience, which indirectly would lead to loss of business.
- Further data and analysis must be invested to reduce FPR and FNR.

Thank You!!!

Appendix

- Python Notebook built using DataSpell attached.