Capstone Project

Final Project

Contents:

Problem Statement.

Business Objective:

An aviation company that provides domestic as well as international trips to the customers now wants to apply a targeted approach instead of reaching out to each of the customers.

This time they want to do it digitally instead of tele calling. Hence, they have collaborated with a social networking platform, so they can learn the digital and social behaviour of the customers and provide the digital advertisement on the user page of the targeted customers who have a high propensity to take up the product.

[Propensity of buying tickets is different for different login devices. Hence, you have to create 2 models separately for Laptop and Mobile. [Anything which is not a laptop can be considered as mobile phone usage.]

The advertisements on the digital platform are a bit expensive; hence, you need to be very accurate while creating the models.

Introduction of the business problem:

Introduction of the business problem

Defining problem statement:

In today's time a large amount of population living in metro cities spends most of their free time in different social media platforms like FB, Insta, Tweeter, google, ect. Therefore, it has become a great way to promote the products to the direct consumers based on their likes, dislikes. It is equally important for a travel company to participate in such campaigns and help growing their business.

In this particular data we will evaluate the some of the behaviour of users and identify the potential clients to target with their offerings.

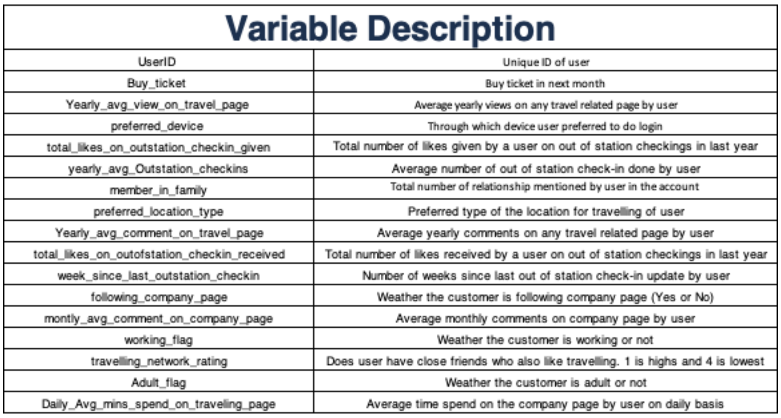
Need of the study/project:

It is important to study the project file and pen down the important variables to determine the right trends and audience.

Understanding business/social opportunity:

Social media is one of the most popular medium today to do analysis on multiple users having similar kind of behavioural pattern and help company to targets a larger sets of people instead of individual approach.

We will evaluate the data by performing multiple activities like, information, shape, duplicate values, null values, outliers and fix the same without losing any important variable and data. EDA will give us more insight by performing univariate, bivariate and multivariate analysis.



Variable Description

UserID	Unique ID of user
Buy_ticket	Buy ticket in next month
Yearly_avg_view_on_travel_page	Average yearly views on any travel related page by user
preferred_device	Through which device user preferred to do login
total_likes_on_outstation_checkin_given	Total number of likes given by a user on out of station checkings in last year
yearly_avg_Outstation_checkins	Average number of out of station check-in done by user
member_in_family	Total number of relationship mentioned by user in the account
preferred_location_type	Preferred type of the location for travelling of user
Yearly_avg_comment_on_travel_page	Average yearly comments on any travel related page by user
total_likes_on_outofstation_checkin_received	Total number of likes received by a user on out of station checkings in last year
week_since_last_outstation_checkin	Number of weeks since last out of station check-in update by user
following_company_page	Weather the customer is following company page (Yes or No)
montly_avg_comment_on_company_page	Average monthly comments on company page by user
working_flag	Weather the customer is working or not
travelling_network_rating	Does user have close friends who also like travelling. 1 is highs and 4 is lowest
Adult_flag	Weather the customer is adult or not
Daily_Avg_mins_spend_on_traveling_page	Average time spend on the company page by user on daily basis

Data Report:

Understanding how data was collected in terms of time, frequency and methodology

Data includes:

- Used_ID- List of customers doing various activities on company's social media page.
- Taken_product: customers taken product and not.
- Using different device to surf the site, ratings, members of the family and how their travel trends are ect..
- Based on that we need to establish the potential customers which can help company in increase on selling.

Visual inspection of data (rows, columns, descriptive details)

Understanding of attributes (variable info, renaming if required)

- looks like there are missing values present.
- Descriptive analysis suggests the data is skewed and that also indicates the data has outliers.
- Column "member_in_famly will requited further cleaning.
- Preferred device column needs to be fixed as Laptop and Mobile attributes.
- Data frame has some 17 columns and 11760 rows in it.

Exploratory data analysis:

1. Columns userI_d and yearly_avg_Outstation_checkins looks of no use so have drop them.

2. On doing descriptive analysis, noticed few columns such as ‘total_likes_on_outstation_checkin_given’, ‘total_likes_on_outofstation_checkin_received’, ‘montly_avg_comment_on_company_page’, ‘Daily_Avg_mins_spend_on_traveling_page’ looks right skewed which suggest the presence of outliers.

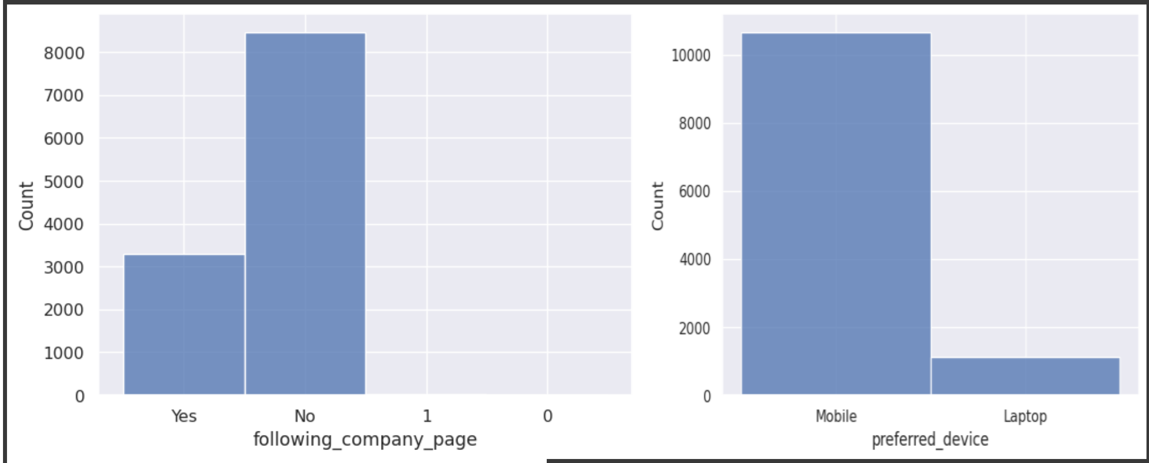
3. Missing values are treated with median and mode for both numerical and categorical columns respectively.

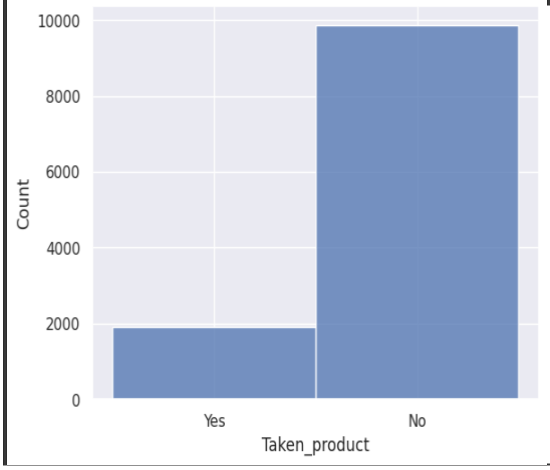
4. Outliers will be treated as per IQR.

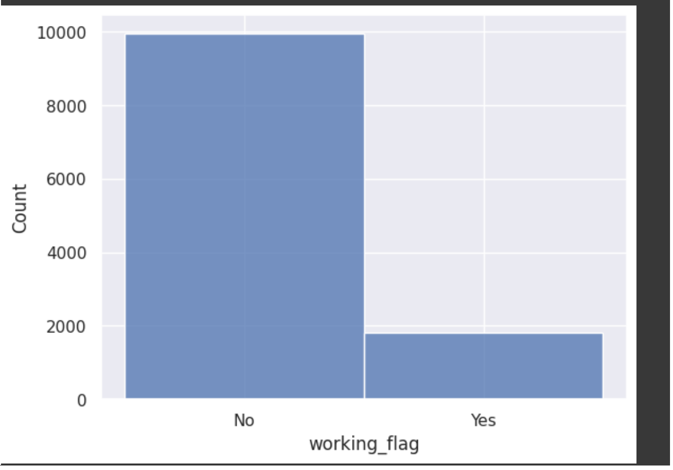
5. Target Variable "Taken_Product has been transformed as "Laptop and Mobile" attributes as advised.

Univariate analysis.

Some of the plots for reference:

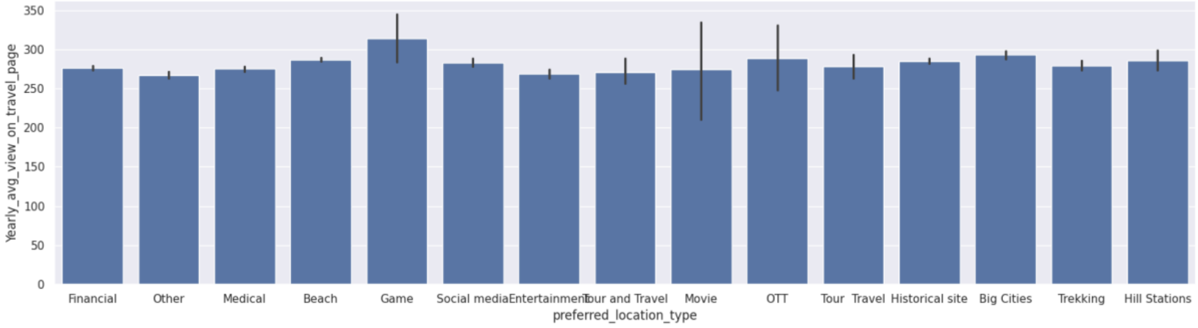


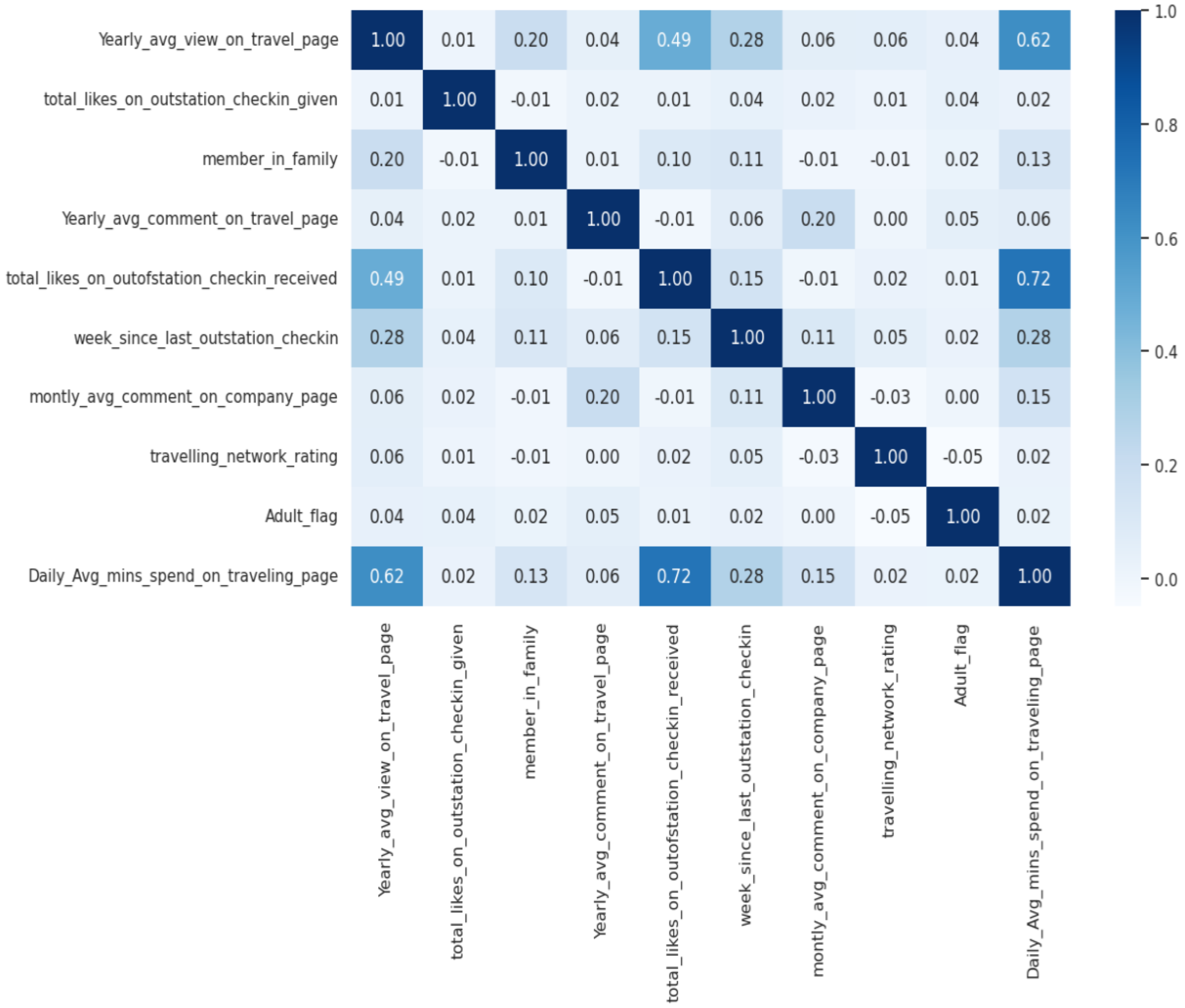




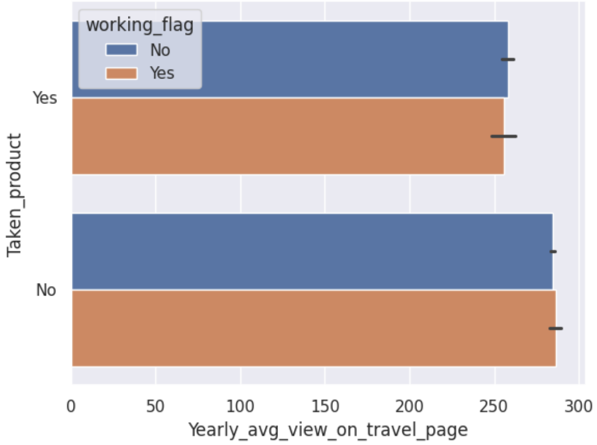
- With the above analysis it shows that more people uses mobile and less using laptops
- Product has been taken by less people.
- Most of the people visiting sites seems to be not working.
- Also, less people following the company.

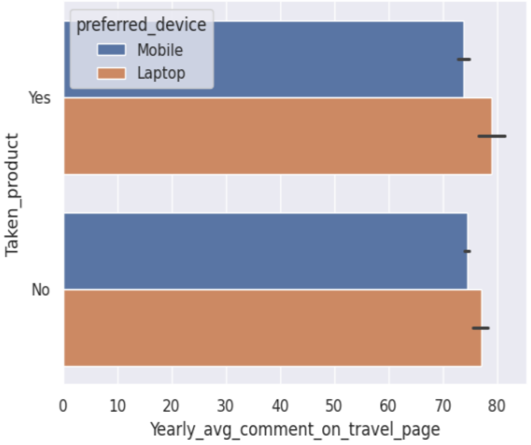
Bivariate Analysis:





	Yearly_avg_view_on_travel_page	total_likes_on_outstation_checkin_given	member_in_family	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin	montly_avg_comment_on_company_page	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
Yearly_avg_view_on_travel_page	1.00	0.01	0.20	0.04	0.49	0.28	0.06	0.06	0.04	0.62
total_likes_on_outstation_checkin_given	0.01	1.00	-0.01	0.02	0.01	0.04	0.02	0.01	0.04	0.02
member_in_family	0.20	-0.01	1.00	0.01	0.10	0.11	-0.01	-0.01	0.02	0.13
Yearly_avg_comment_on_travel_page	0.04	0.02	0.01	1.00	-0.01	0.06	0.20	0.00	0.05	0.06
total_likes_on_outofstation_checkin_received	0.49	0.01	0.10	-0.01	1.00	0.15	-0.01	0.02	0.01	0.72
week_since_last_outstation_checkin	0.28	0.04	0.11	0.06	0.15	1.00	0.11	0.05	0.02	0.28
montly_avg_comment_on_company_page	0.06	0.02	-0.01	0.20	-0.01	0.11	1.00	-0.03	0.00	0.15
travelling_network_rating	0.06	0.01	-0.01	0.00	0.02	0.05	-0.03	1.00	-0.05	0.02
Adult_flag	0.04	0.04	0.02	0.05	0.01	0.02	0.00	-0.05	1.00	0.02
Daily_Avg_mins_spend_on_traveling_page	0.62	0.02	0.13	0.06	0.72	0.28	0.15	0.02	0.02	1.00





- People with interest in Movies, OTT, Games, Tour & Travel are amongst the highest viewers of the page.
- People with viewing average of 255 taking product however the not working count is almost same to ones working.
- People with laptops are more likely to but product.

Business insights from EDA:

- Is the data unbalanced? If so, what can be done? Please explain in the context of the business.
- The data is not unbalanced however more information would be helpful.
- Only 20% of customers are buying product which is way to less.
- A targets based approach would help.
- As cities with beaches, Financials, historical sites and medical are high selling however, people visiting page also have great interest in game, OTT, Movies and tour travel.
- There could be more information added so that it can attract more customers to buy product.

Model building and interpretation.

Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes) (Laptop Users)

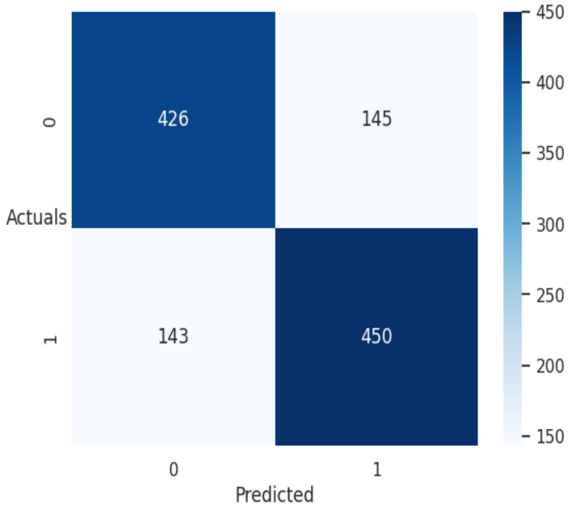
After scaling the data and dividing in to X & y will perform various model to make evolution on the data.

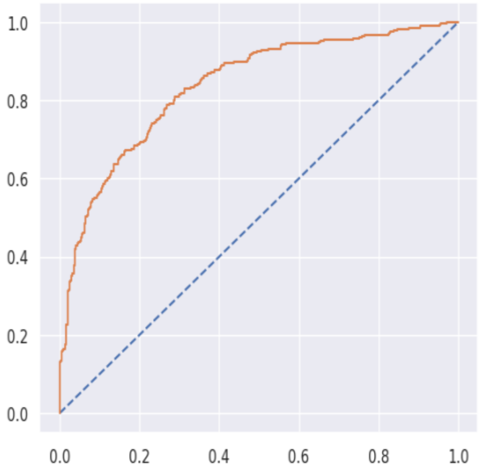
Logistic Regression:

Data has been split in 70:30 ratio-Prediction on Train Set:

```
The classification report for Logistic Regression training set is
               precision    recall  f1-score   support

         0.0       0.75      0.75      0.75       571
         1.0       0.76      0.76      0.76       593

    accuracy                           0.75      1164
   macro avg       0.75      0.75      0.75      1164
weighted avg       0.75      0.75      0.75      1164
```

AUC for Train set: The AUC score for Logistic Regression training set is: 0.837



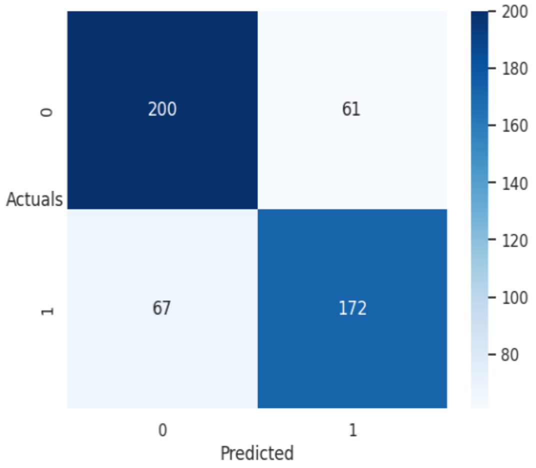


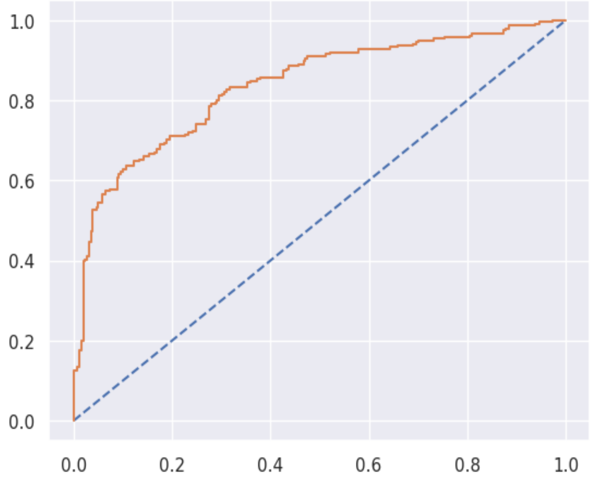
- Model has performed with 75% accuracy with similar recall and precision values.
- AUC score is 0,837.

Prediction on Train Set:

```
The classification report for Logistic Regression training set is
              precision    recall  f1-score   support

         0.0       0.75      0.77      0.76       261
         1.0       0.74      0.72      0.73       239

    accuracy                           0.74       500
   macro avg       0.74      0.74      0.74       500
weighted avg       0.74      0.74      0.74       500
```

The AUC score for Logistic Regression test set is: 0.837

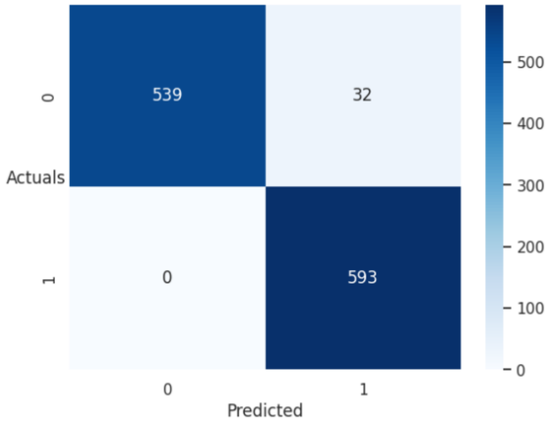


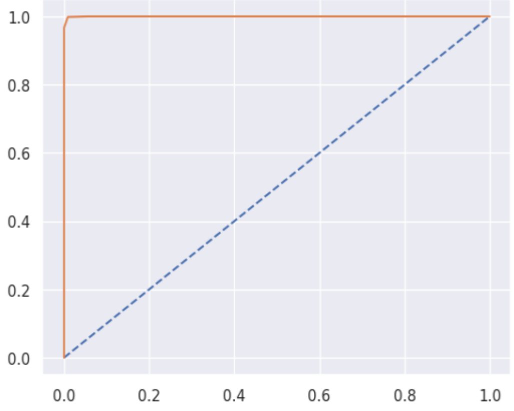


On test set accuracy has dropped a bit however recall and precision has gone up. AUC score remains same as 0.837.

KNN Model: Will test both data set on KNN.

```
The classification report for KNN set is
              precision    recall  f1-score   support

         0.0       1.00      0.94      0.97       571
         1.0       0.95      1.00      0.97       593

    accuracy                           0.97      1164
   macro avg       0.97      0.97      0.97      1164
weighted avg       0.97      0.97      0.97      1164
```

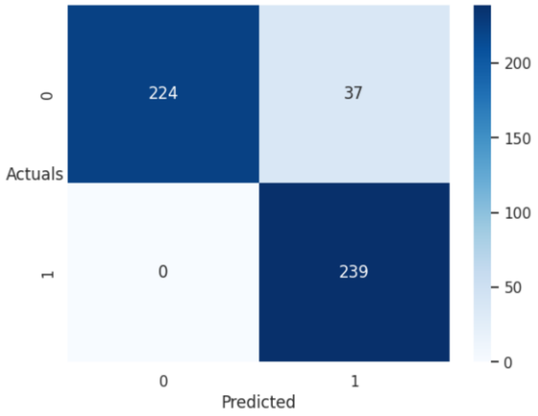


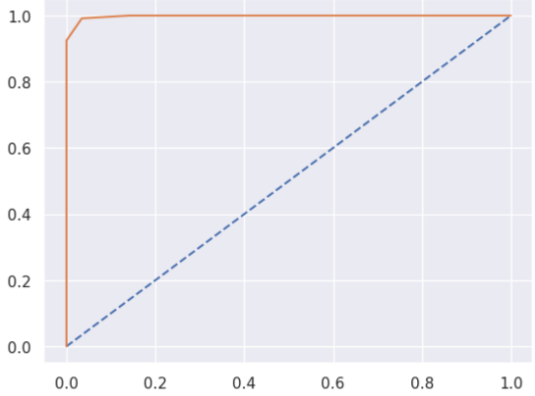


KNN model performing good with 97% of accuracy and making 100% prediction on recall and precision. AUC score also 100.

Let’s do the test data analysis:

```
The classification report for KNN set is
                  precision    recall  f1-score   support

         0.0         1.00      0.86      0.92       261
         1.0         0.87      1.00      0.93       239

    accuracy                             0.93       500
   macro avg         0.93      0.93      0.93       500
weighted avg         0.94      0.93      0.93       500
The AUC score for KNN test set is: 0.998
```

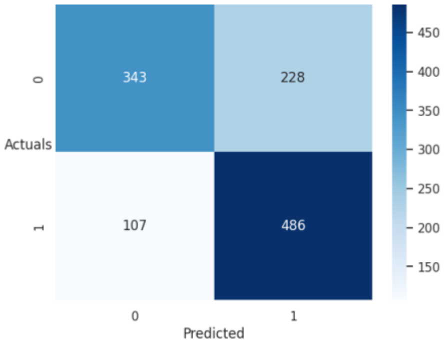


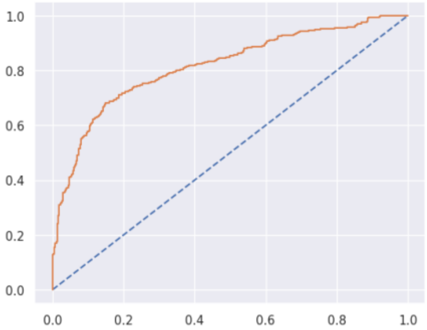


Model has performed with 93% of accuracy and both recall and precision predicted lower than train set.

Naive Bayes Model: Train Set-

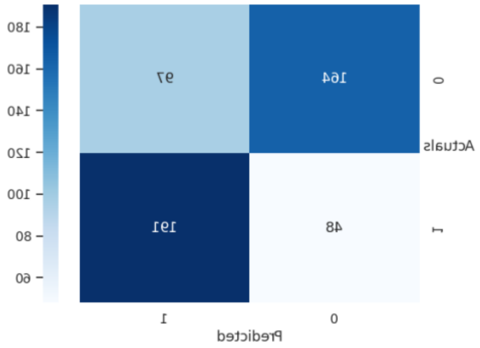
```
The classification report for Naive Bayes Model set is
                  precision    recall  f1-score   support

         0.0         0.76      0.60      0.67       571
         1.0         0.68      0.82      0.74       593

    accuracy                             0.71      1164
   macro avg         0.72      0.71      0.71      1164
weighted avg         0.72      0.71      0.71      1164
The AUC score for Naive Bayes training set is: 0.816
```

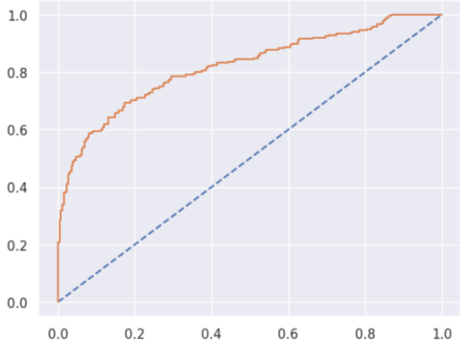




Naive Bayes Model: Test Set-

```
The classification report for Naive Bayes Model set is
              precision    recall  f1-score   support

         0.0       0.77      0.63      0.69       261
         1.0       0.66      0.80      0.72       239

    accuracy                           0.71       500
   macro avg       0.72      0.71      0.71       500
weighted avg       0.72      0.71      0.71       500
The AUC score for Naive Bayes test set is: 0.820
```

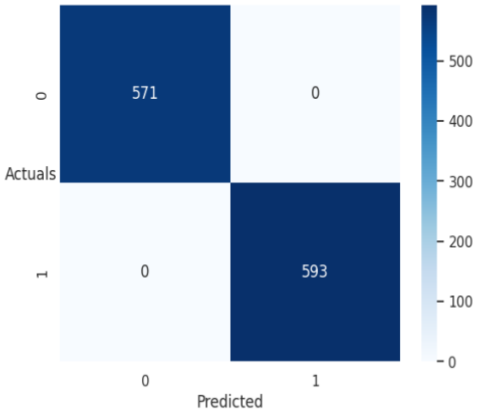


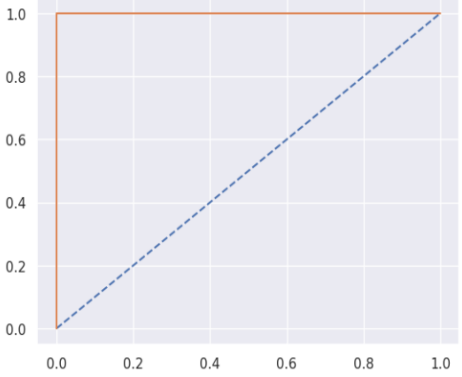


As noticed Naïve Bayes model has performed below par with the previous both the models.

Decision_Tree_Classifier model prediction on Train Set:

```
The classification report for Decision Tree training set is
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       571
         1.0       1.00      1.00      1.00       593

    accuracy                           1.00      1164
   macro avg       1.00      1.00      1.00      1164
weighted avg       1.00      1.00      1.00      1164
The AUC score for Decision Tree training set is: 1.000
```

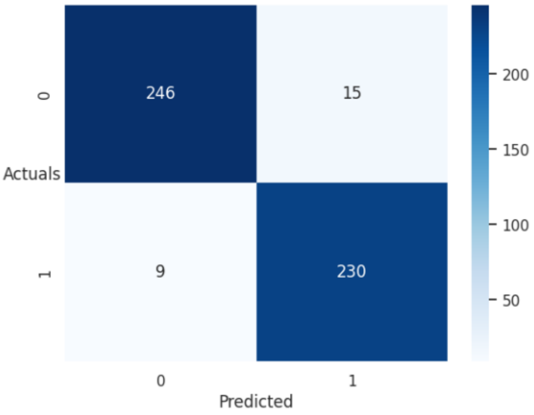


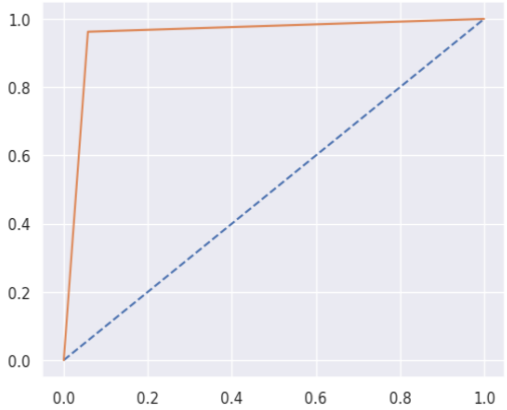


As observed model has performed 100% on train set and by far the best model performance. Let us check the same on test set.

```
The classification report for Decision Tree training set is
```

```
              precision    recall  f1-score   support

         0.0       0.96      0.94      0.95       261
         1.0       0.94      0.96      0.95       239

    accuracy                           0.95       500
   macro avg       0.95      0.95      0.95       500
weighted avg       0.95      0.95      0.95       500
```

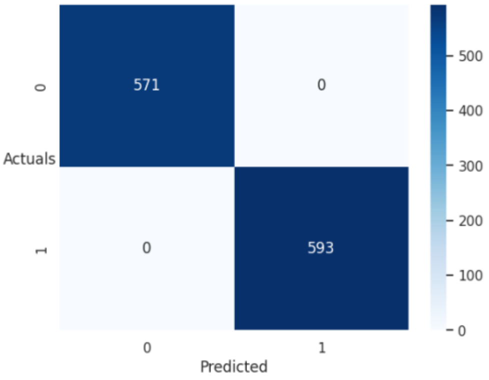


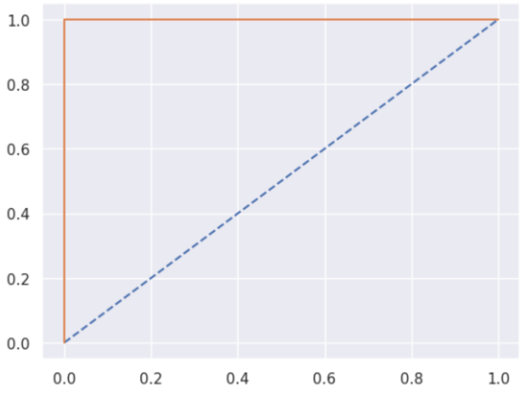


The performance has dropped on test set with the accuracy of 95%

Random Forest Classifier: Model predictions on Train Set

```
The classification report for RFC training set is
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       571
         1.0       1.00      1.00      1.00       593

    accuracy                           1.00      1164
   macro avg       1.00      1.00      1.00      1164
    weighted avg       1.00      1.00      1.00      1164
    The AUC score for RFC training set is: 1.000
```

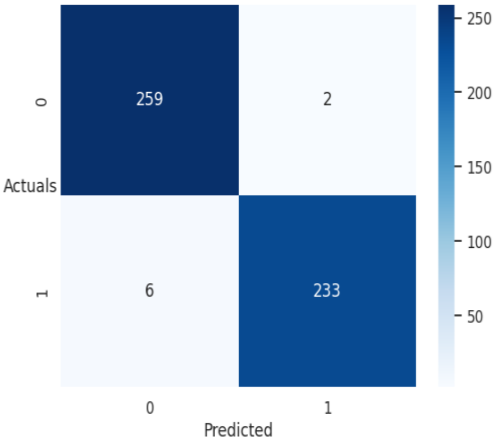


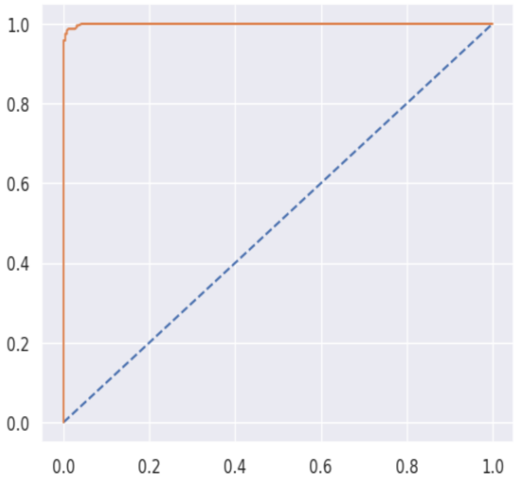


Again model has performed a00% on train set, lets check the test data.

Random Forest Classifier on Test Set:

```
The classification report for RFC training set is
              precision    recall  f1-score   support

         0.0       0.98      0.99      0.98       261
         1.0       0.99      0.97      0.98       239

    accuracy                           0.98       500
   macro avg       0.98      0.98      0.98       500
     weighted avg       0.98      0.98      0.98       500
     The AUC score for RFC test set is: 0.999
```



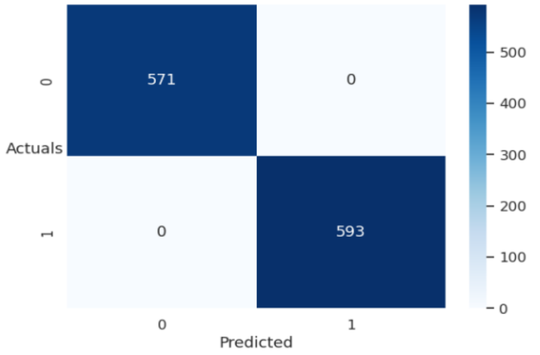


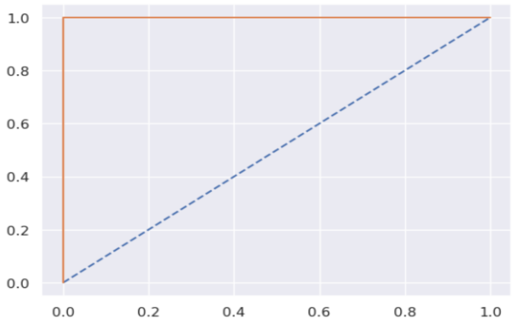
As observed that model has slight less on test but it is still the best performed model by far.

Model Tuning and business implication:

Bagging technique: On Train:

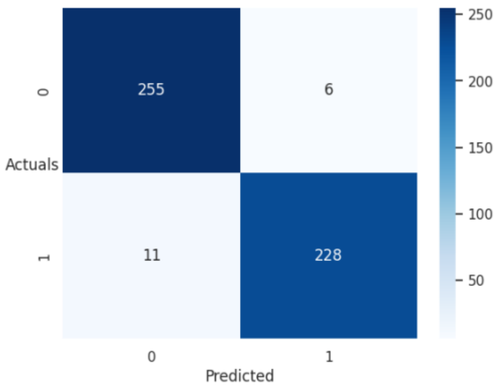
```
The classification report for Bagging training set is
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       571
         1.0       1.00      1.00      1.00       593

    accuracy                           1.00      1164
   macro avg       1.00      1.00      1.00      1164
weighted avg       1.00      1.00      1.00      1164
The AUC score for Bagging training set is: 1.000
```

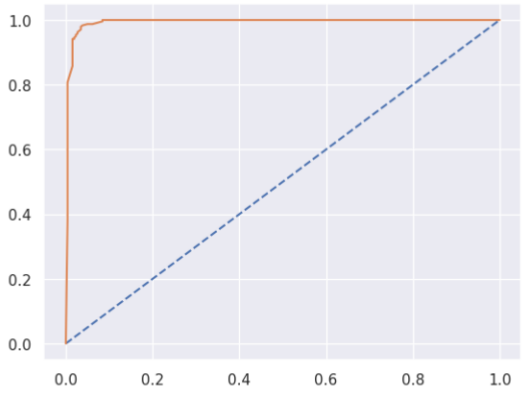




Bagging on Test set:

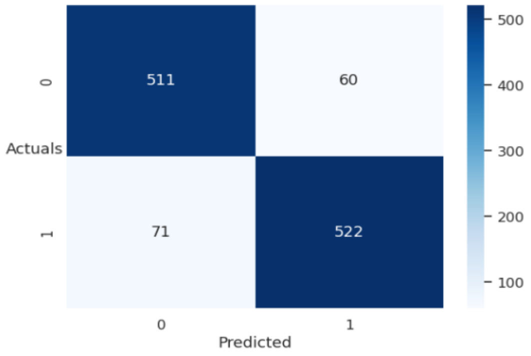
```
The classification report for Bagging test set is
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97       261
         1.0       0.97      0.95      0.96       239

    accuracy                           0.97       500
   macro avg       0.97      0.97      0.97       500
weighted avg       0.97      0.97      0.97       500
The AUC score for Bagging test set is: 0.994
```

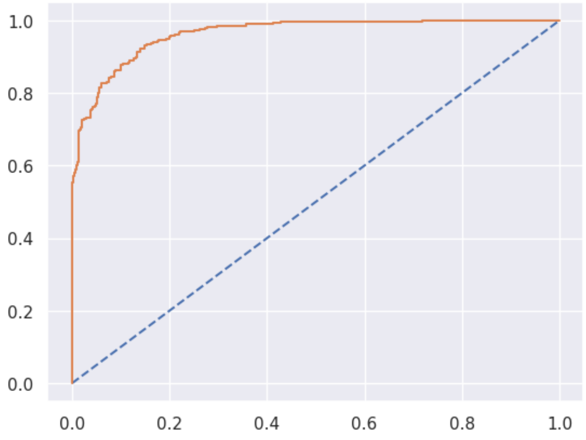




AdaBoosting: On Train Set

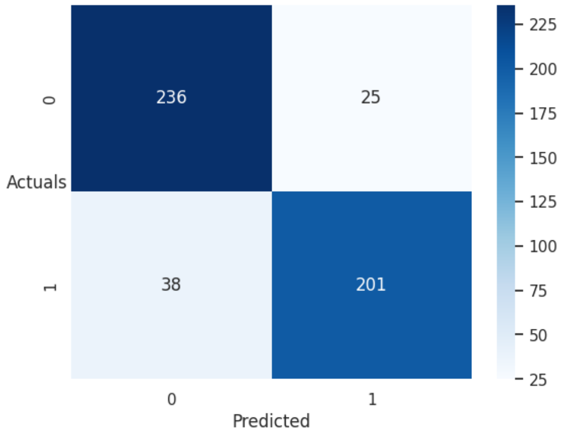
```
The classification report for Adaboosting training set is
              precision    recall  f1-score   support

         0.0       0.88      0.89      0.89       571
         1.0       0.90      0.88      0.89       593

    accuracy                           0.89      1164
   macro avg       0.89      0.89      0.89      1164
weighted avg       0.89      0.89      0.89      1164

The AUC score for AdaBoosting training set is: 0.964
```

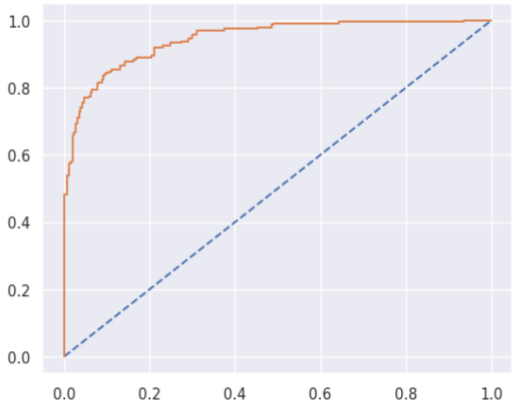




AdaBoosting on Test Set:

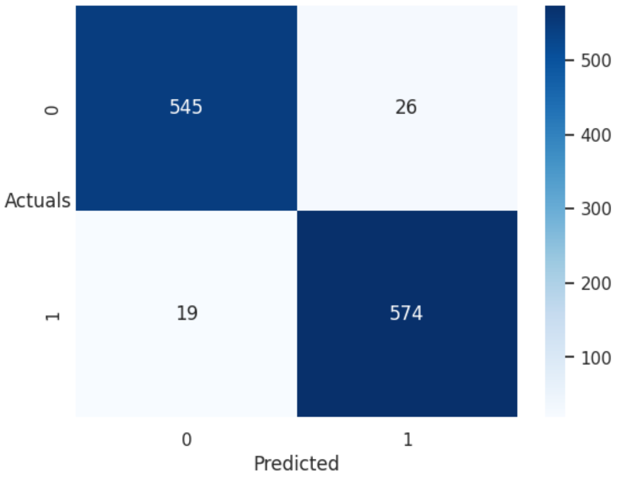
```
The classification report for Adaboosting test set is
              precision    recall  f1-score   support

         0.0       0.86      0.90      0.88       261
         1.0       0.89      0.84      0.86       239

    accuracy                           0.87       500
   macro avg       0.88      0.87      0.87       500
weighted avg       0.87      0.87      0.87       500
The AUC score for AdaBoosting test set is: 0.945
```

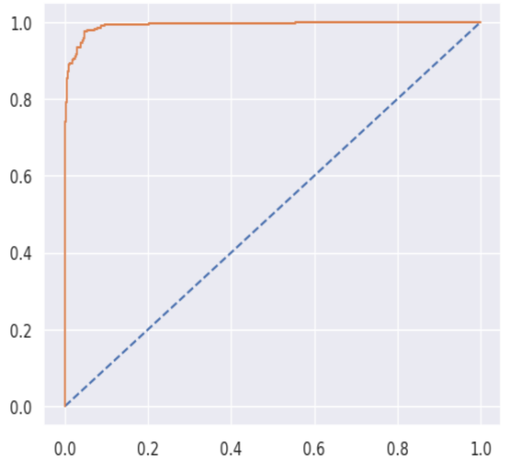




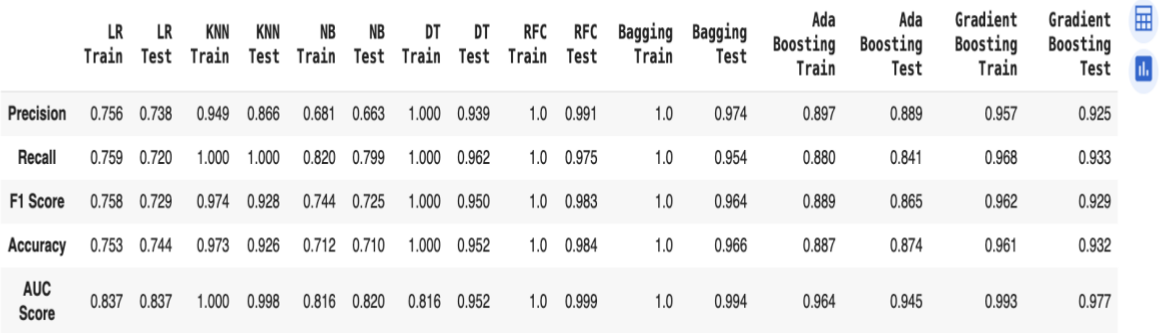
Gradient Boosting:

```
he classification report for Gradientboosting training set is
              precision    recall  f1-score   support

         0.0       0.97      0.95      0.96       571
         1.0       0.96      0.97      0.96       593

    accuracy                           0.96      1164
   macro avg       0.96      0.96      0.96      1164
weighted avg       0.96      0.96      0.96      1164
The AUC score for GradientBoosting training set is: 0.993
```





Model Comparison: Laptop.



	LR Train	LR Test	KNN Train	KNN Test	NB Train	NB Test	DT Train	DT Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.756	0.738	0.949	0.866	0.681	0.663	1.000	0.939	1.0	0.991	1.0	0.974	0.897	0.889	0.957	0.925
Recall	0.759	0.720	1.000	1.000	0.820	0.799	1.000	0.962	1.0	0.975	1.0	0.954	0.880	0.841	0.968	0.933
F1 Score	0.758	0.729	0.974	0.928	0.744	0.725	1.000	0.950	1.0	0.983	1.0	0.964	0.889	0.865	0.962	0.929
Accuracy	0.753	0.744	0.973	0.926	0.712	0.710	1.000	0.952	1.0	0.984	1.0	0.966	0.887	0.874	0.961	0.932
AUC Score	0.837	0.837	1.000	0.998	0.816	0.820	0.816	0.952	1.0	0.999	1.0	0.994	0.964	0.945	0.993	0.977

In order to perform these models, the data was cleaned and unwanted variables were removed. This was followed by treatment of the imbalance in the data using SMOTE.

After that the data was scaled using the standard scalar as there are variables in 1000’s, 100’s etc. With that, train test split was performed in which the data was divided in the ratio of 70:30 where 70% constitutes the training set.

1. Logistic Regression/NB models performed lowest with the accuracy of 74.2% and 65.8% on train set/ 70.5% and 63.2% on test set. Accuracy too is lower on both the models.

2. Decision Tree, Random Forest models have provided highest accuracy 0f 100% on Training set & test set 95% & 99% respectively.

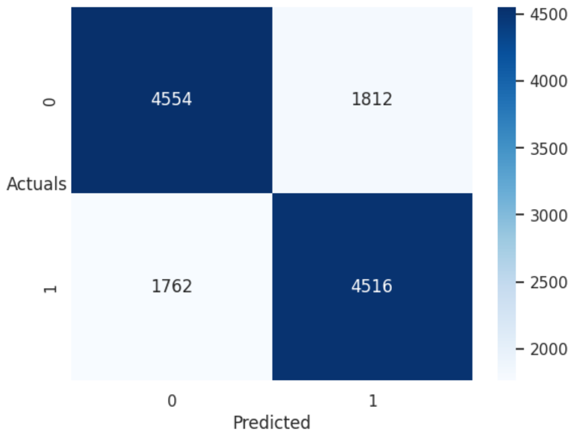
3. The AUC score of both test and train set for K- Nearest Neighbours model and Random Forest model is perfect i.e., 100%. Although, both the models are very good the recall of Random Forest model (98.3%) is slightly lower than the K-Nearest Neighbours model (100%).

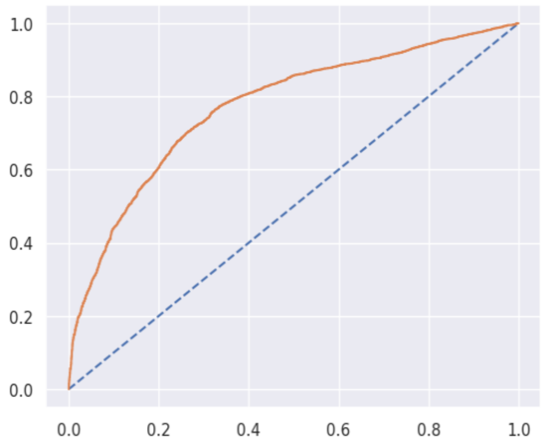
4. DT model achieved the accuracy 100% on training set and 95% on test set followed by RFC.

Model building on Mobile

Logistic Regression on Train Set:

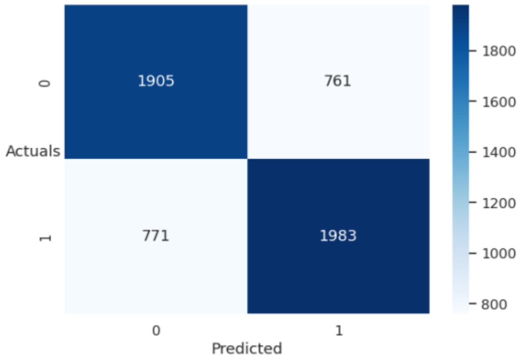
```
The classification report for Logistic Regression training set is
              precision    recall  f1-score   support

         0.0       0.72      0.72      0.72      6366
         1.0       0.71      0.72      0.72      6278

    accuracy                           0.72     12644
   macro avg       0.72      0.72      0.72     12644
weighted avg       0.72      0.72      0.72     12644

The AUC score for Logistic Regression training set is: 0.773
```

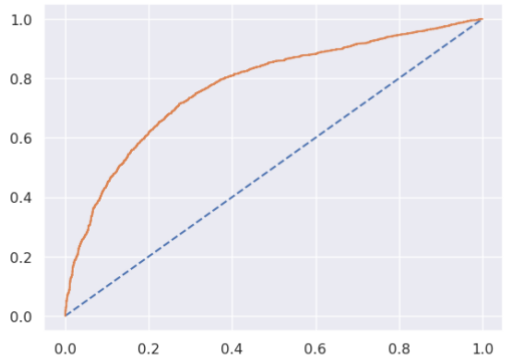




Logistic Regression on Test Set:

```
The classification report for Logistic Regression training set is
              precision    recall  f1-score   support

         0.0       0.71      0.71      0.71      2666
         1.0       0.72      0.72      0.72      2754

    accuracy                           0.72      5420
   macro avg       0.72      0.72      0.72      5420
weighted avg       0.72      0.72      0.72      5420

The AUC score for Logistic Regression test set is: 0.774
```

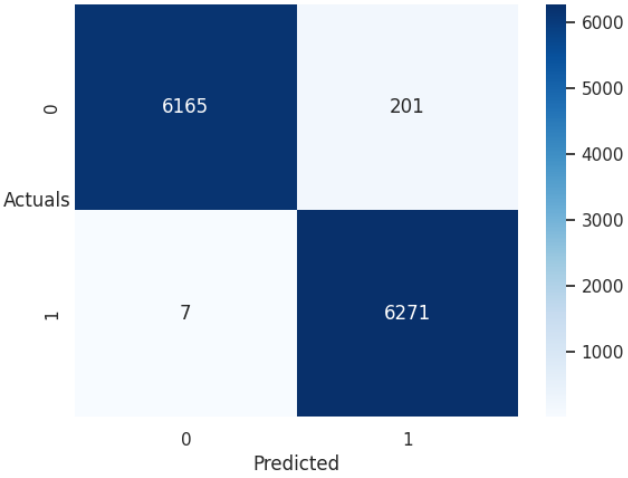


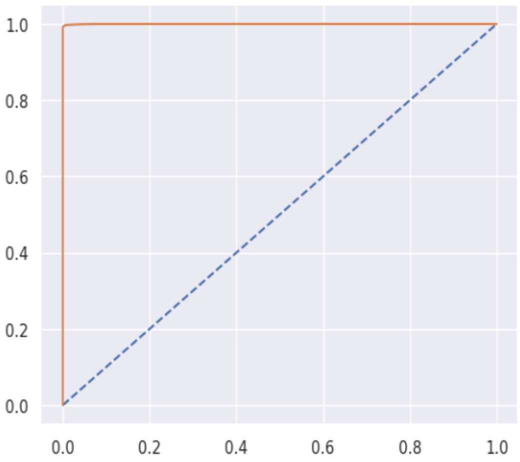


KNN Model:

Making prediction on Train Set: KNN

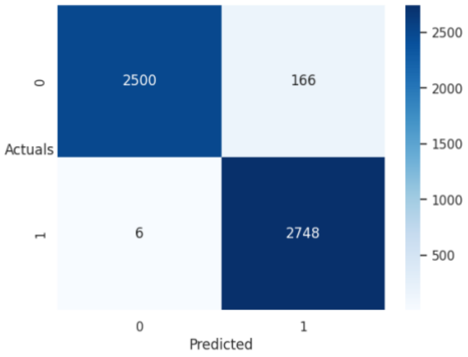
```
The classification report for KNN set is
              precision    recall  f1-score   support

         0.0       1.00      0.97      0.98      6366
         1.0       0.97      1.00      0.98      6278

    accuracy                           0.98     12644
   macro avg       0.98      0.98      0.98     12644
weighted avg       0.98      0.98      0.98     12644

The AUC score for KNN training set is: 1.000
```

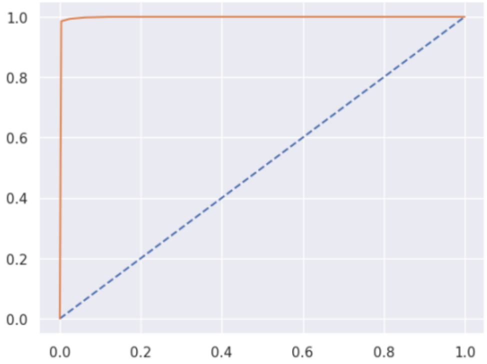




Making prediction on Test Set: KNN

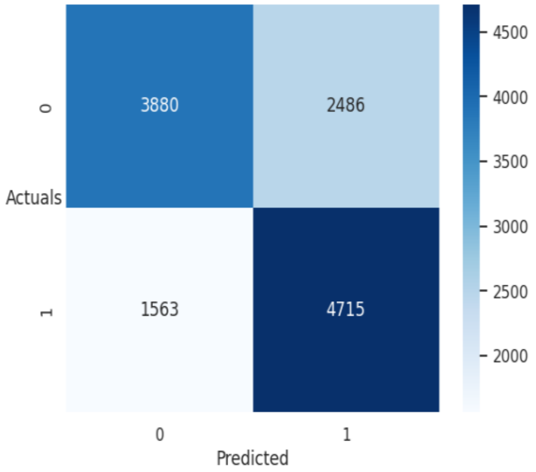
```
The classification report for KNN set is
              precision    recall  f1-score   support

         0.0       1.00      0.94      0.97      2666
         1.0       0.94      1.00      0.97      2754

    accuracy                           0.97      5420
   macro avg       0.97      0.97      0.97      5420
weighted avg       0.97      0.97      0.97      5420
The AUC score for KNN test set is: 0.998
```

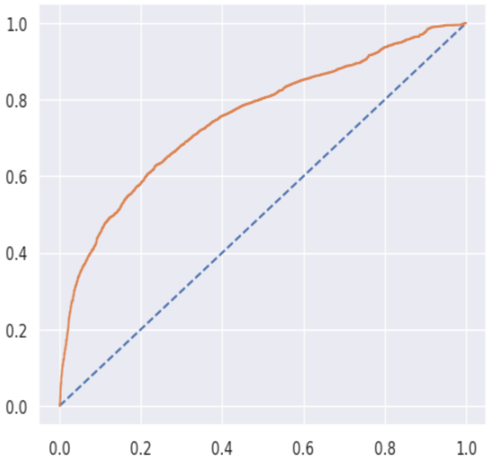




Naive Bayes Model:Prediction on Train set

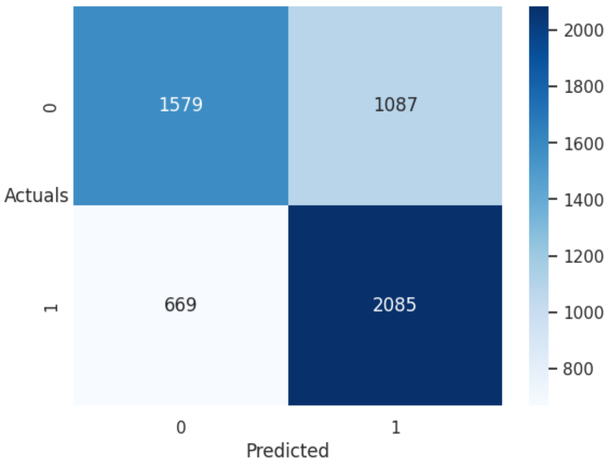
```
The classification report for Naive Bayes Model set is
              precision    recall  f1-score   support

         0.0       0.71      0.61      0.66      6366
         1.0       0.65      0.75      0.70      6278

    accuracy                           0.68     12644
   macro avg       0.68      0.68      0.68     12644
weighted avg       0.68      0.68      0.68     12644
```

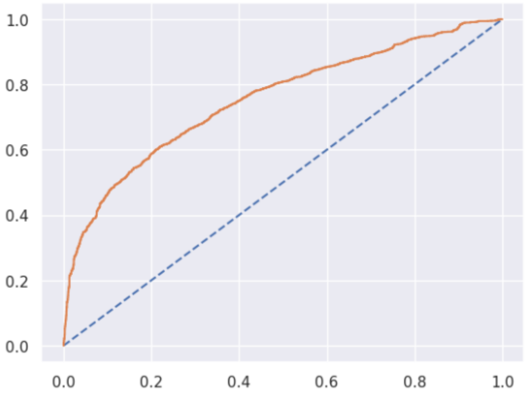




Prediction on Test set

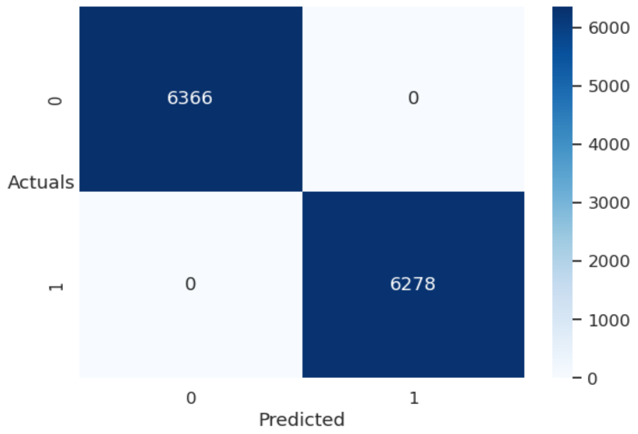
```
The classification report for Naive Bayes Model set is
              precision    recall  f1-score   support

         0.0       0.70      0.59      0.64      2666
         1.0       0.66      0.76      0.70      2754

    accuracy                           0.68      5420
   macro avg       0.68      0.67      0.67      5420
weighted avg       0.68      0.68      0.67      5420
The AUC score for Naive Bayes test set is: 0.754
```

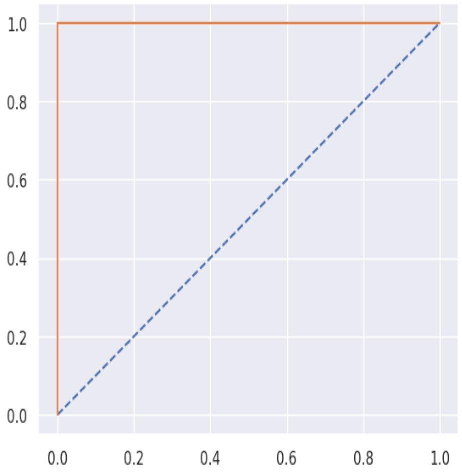




Decision Tree Classifier: prediction on Train Set

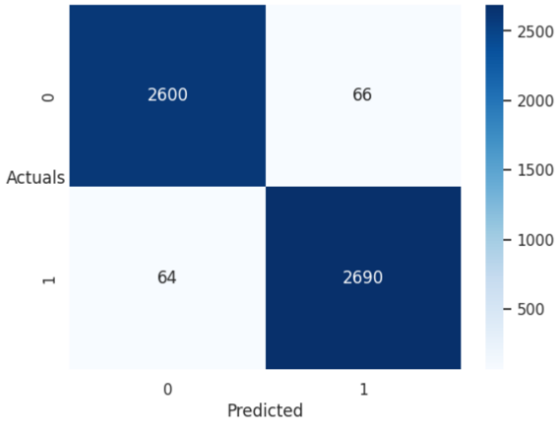
```
The classification report for Decision Tree training set is
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      6366
         1.0       1.00      1.00      1.00      6278

    accuracy                           1.00     12644
   macro avg       1.00      1.00      1.00     12644
weighted avg       1.00      1.00      1.00     12644

The AUC score for Decision Tree training set is: 1.000
```

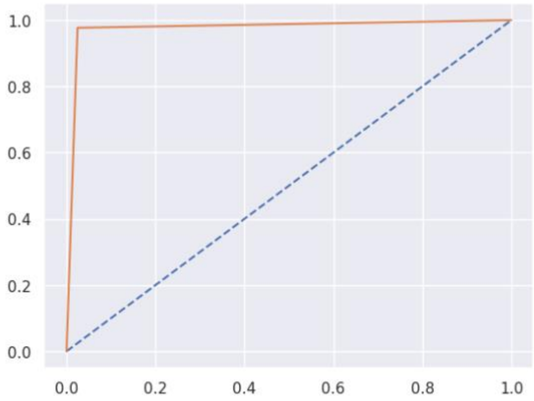




Decision Tree Classifier: prediction on Test Set

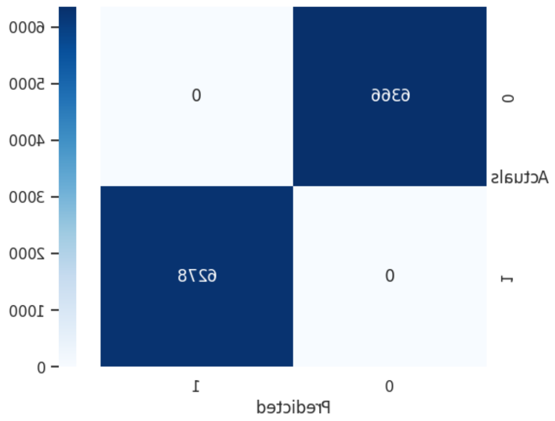
```
The classification report for Decision Tree training set is
              precision    recall  f1-score   support

         0.0       0.98      0.98      0.98      2666
         1.0       0.98      0.98      0.98      2754

    accuracy                           0.98      5420
   macro avg       0.98      0.98      0.98      5420
weighted avg       0.98      0.98      0.98      5420
The AUC score for Decision Tree test set is: 0.976
```

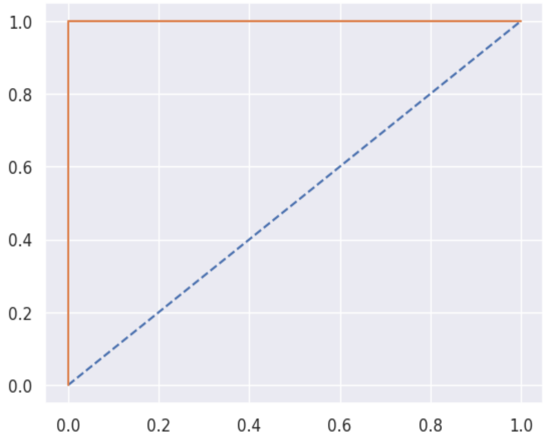




Random Forest Classifier: Predictions on Train Set

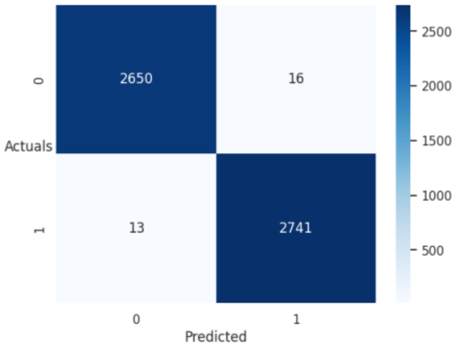
```
The classification report for RFC training set is
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      6366
         1.0       1.00      1.00      1.00      6278

    accuracy                           1.00     12644
   macro avg       1.00      1.00      1.00     12644
weighted avg       1.00      1.00      1.00     12644
The AUC score for RFC training set is: 1.000
```

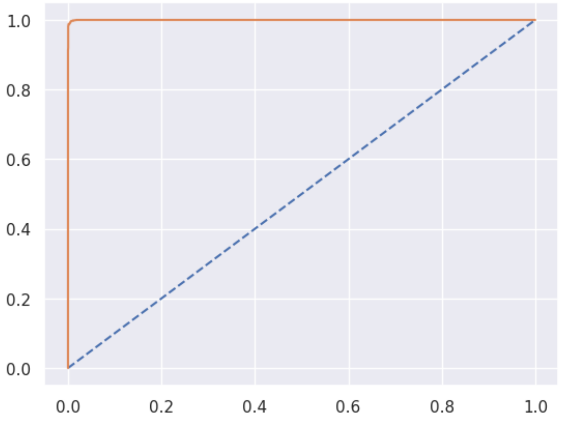




Random Forest Classifier: Predictions on Test Set

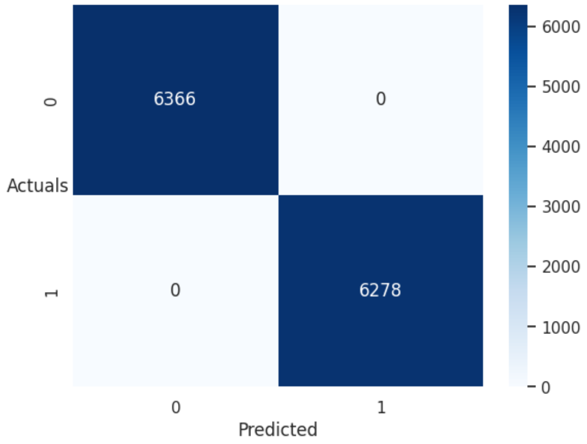
```
The classification report for RFC training set is
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99      2666
         1.0       0.99      1.00      0.99      2754

    accuracy                           0.99      5420
   macro avg       0.99      0.99      0.99      5420
weighted avg       0.99      0.99      0.99      5420
The AUC score for RFC test set is: 1.000
```

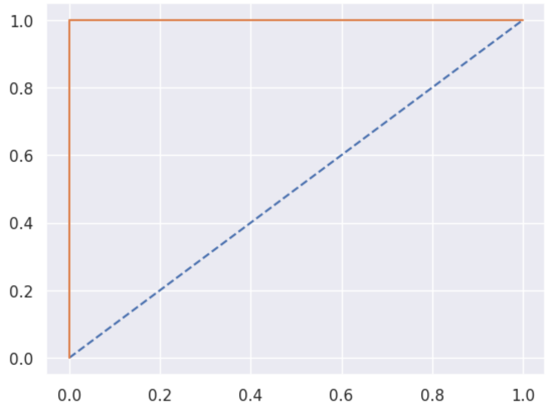




<u>Bagging technique:</u> <u>Train set</u>

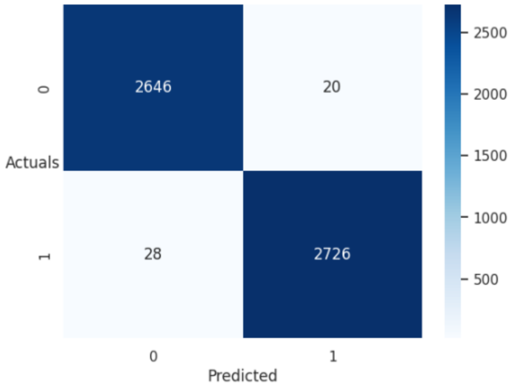
```
The classification report for Bagging training set is
               precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      6366
         1.0       1.00      1.00      1.00      6278

    accuracy                           1.00     12644
   macro avg       1.00      1.00      1.00     12644
weighted avg       1.00      1.00      1.00     12644
The AUC score for Bagging training set is: 1.000
```

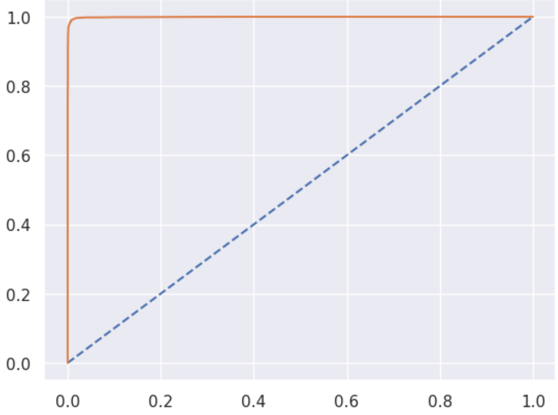




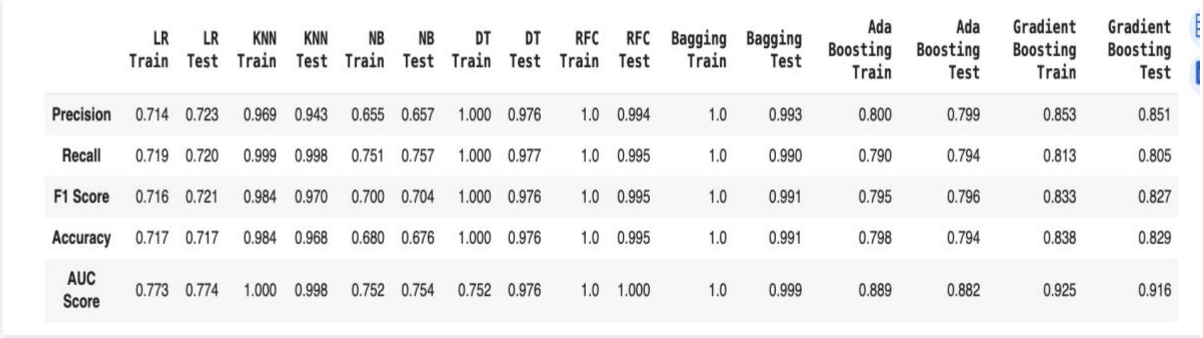
<u>Bagging technique:</u> <u>Test set</u>

```
The classification report for Bagging test set is
               precision    recall  f1-score   support

         0.0       0.99      0.99      0.99      2666
         1.0       0.99      0.99      0.99      2754

    accuracy                           0.99      5420
   macro avg       0.99      0.99      0.99      5420
weighted avg       0.99      0.99      0.99      5420
```





Model Comparison: Mobile.



	LR Train	LR Test	KNN Train	KNN Test	NB Train	NB Test	DT Train	DT Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.714	0.723	0.969	0.943	0.655	0.657	1.000	0.976	1.0	0.994	1.0	0.993	0.800	0.799	0.853	0.851
Recall	0.719	0.720	0.999	0.998	0.751	0.757	1.000	0.977	1.0	0.995	1.0	0.990	0.790	0.794	0.813	0.805
F1 Score	0.716	0.721	0.984	0.970	0.700	0.704	1.000	0.976	1.0	0.995	1.0	0.991	0.795	0.796	0.833	0.827
Accuracy	0.717	0.717	0.984	0.968	0.680	0.676	1.000	0.976	1.0	0.995	1.0	0.991	0.798	0.794	0.838	0.829
AUC Score	0.773	0.774	1.000	0.998	0.752	0.754	0.752	0.976	1.0	1.000	1.0	0.999	0.889	0.882	0.925	0.916

On comparing with the "Mobile" users as the number of rows increases the accuracy of most of the models decreased.

- Logistic Regression with 71% accuracy on both Train & Test & Naive Bayes with 68% on train and 67% on test are the lowest models performed.
- Decision Tree model/ Random Forest models have performed the best with the accuracy of 100% on Train and 97% & 99% respectively on test set.
- Bagging technique did help to perform the mode slightly better but it declined as compared to laptop uses.
- Ada boosting and Gradient boosting performed better on laptop users.

Random Forest is Test set has accuracy of 99.5% but it predicted number of False positives reducing its accuracy in comparison to the K-nearest Neighbours. therefore KNN model stand out as best among.

- KNN model having such high Recall can help the company in identifying Potential customers who can buy product in the future. Also, this will also help for a better reach and have target base Approach accordingly.
- It can help in increasing flow on the company's site, resulting optimizing the click per cost expense for the company.
- As, the higher the number of hits on website increases, more chances of purchasing the product also increases bringing in the surge in revenues for the company.