SOCIAL MEDIA TOURISM PROJECT

FINAL REPORT SUBMISSION

Vibhav Jaiswal

18/12/2022

Contents

List of Graphs

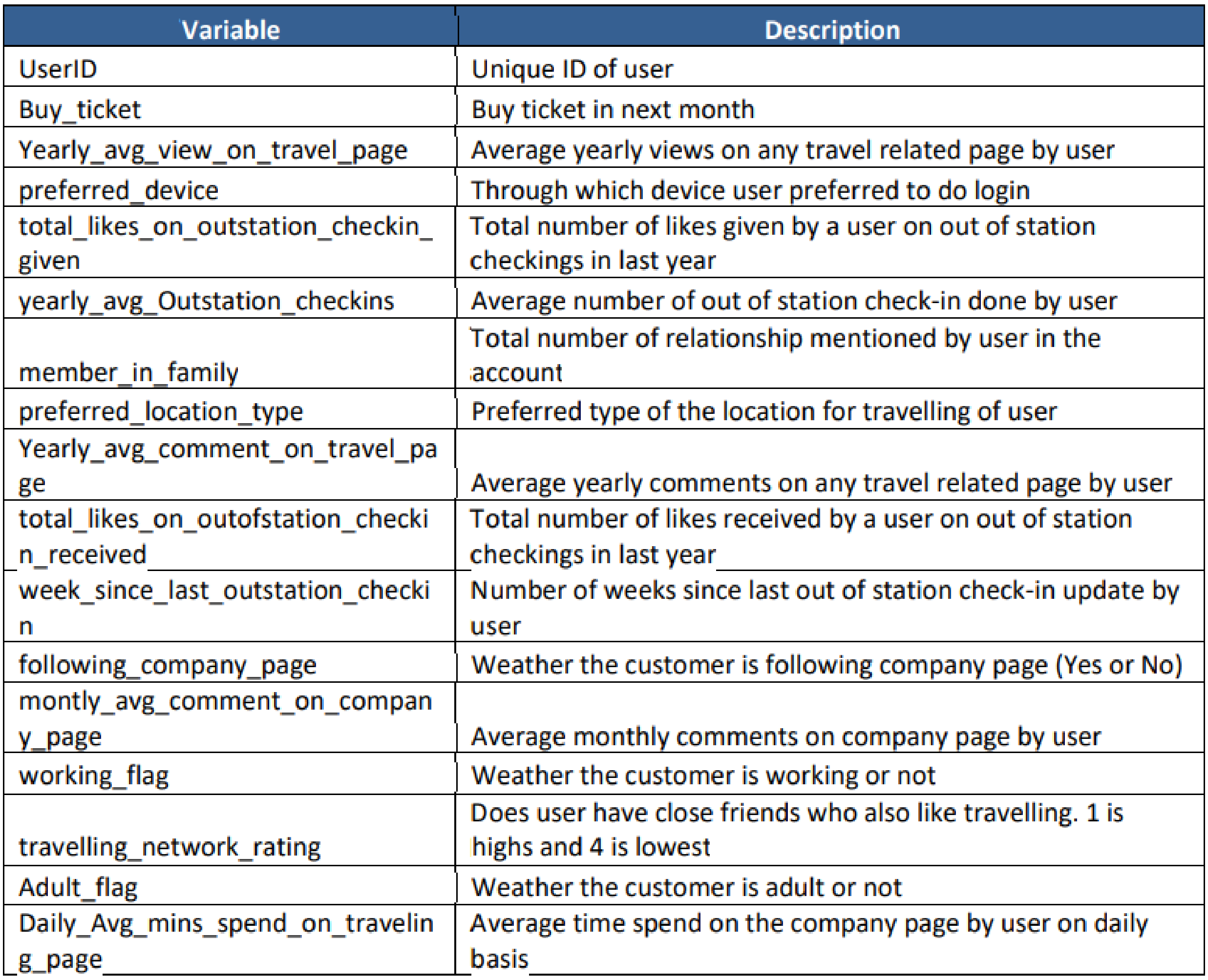
List of Tables

Problem

Problem Statement: 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.

Variable Description



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

1. Introduction - What did you wish to achieve while doing the project?

Defining problem statement

Digital marketing is a coordinated marketing effort to reinforce or assist with a business goal using one or more social media platforms. Campaigns differ from everyday social media efforts because of their increased focus, targeting and measurability. We have an aviation company who is trying to make use of a digital marketing platform to come up with a plan for the targeted customers. They have collaborated with a social networking platform to understand the behaviour of the customers. The purchasing behaviour is device specific – Laptop and Mobile (Models are run separately for each device type). The models used should be accurate as the advertisements for digital platforms are expensive. Now my goal for this project is to predict whether the customer is going to adopt the tourism package based on the social media campaign. Social media lets you reach out to a greater population with information about your business. If you can use social media well, you can improve awareness of your brand, increase the number of visitors to your website, and earn more money. Without it, you may not be able to do well in your business as we are currently living in the tech era.

Need of the study/project

Nowadays people digital footprint is increasing everyday as they tend to get to the cultural of social network. Thus, targeting a customer via digital marketing is much more beneficial to a company rather than a physical advertisement, which might reach larger group of people but we cannot be sure of how much of those people are going to be a potential customer. To find out the customer's interest on travel, which could help the aviation company to pitch their product rightly on time to the right customer, by accessing the customer digital and social behaviour via social networking platforms. And provide digital ad only the customer who has higher propensity of planning a travel in the near future.

This project will help the aviation company get a clearer picture about their customers and help them according to their expectation. Targeted advertisement not only helps the company’s revenue grow but also bring satisfaction to the customers as their individual need is met. Once the advertisements are digitized, it also helps us eradicate the mundane task of calling each and every customer to check if they want to adopt a tour package. If the company has a clear idea on their customers, they can come up with a strategy to improve business. Various tour packages can be framed and sent to the specific group of people.

Understanding business/social opportunity

- **Business Opportunity:**

The business opportunity here is that it helps the company give attention to targeted audiences. This will in turn attract more customers to buy the product. Eventually, this will lead to an increase in revenue. Once the company revenue increases, the company size will also grow and the company will expand. As we can target only the potential customer who has a good chance of buying the product, ROI on Marketing spends could be higher. Also, there is a reduction of tele calling which this translates into 5 less spends on call centres and more control over the marketing spends. This will help business concentrate more only on interested customers and increase the customer retention rate.

- **Social Opportunity:**

A company with good growth will have various employment opportunities. This can be considered the social opportunity of this project. Not only job generation, from the customer's perspective, the company can give discounts on the tour packages or tickets and it can deliver the products on time to the customer without any delay. This will improve the customer experience which will again increase the revenue. As we avoid calling everyone out there to pitch the product, even if they have very little or no chance of buying the product in near future, we could save customer's valuable time and frustration. Thus, the company can have better business opportunities as well as better social opportunities.

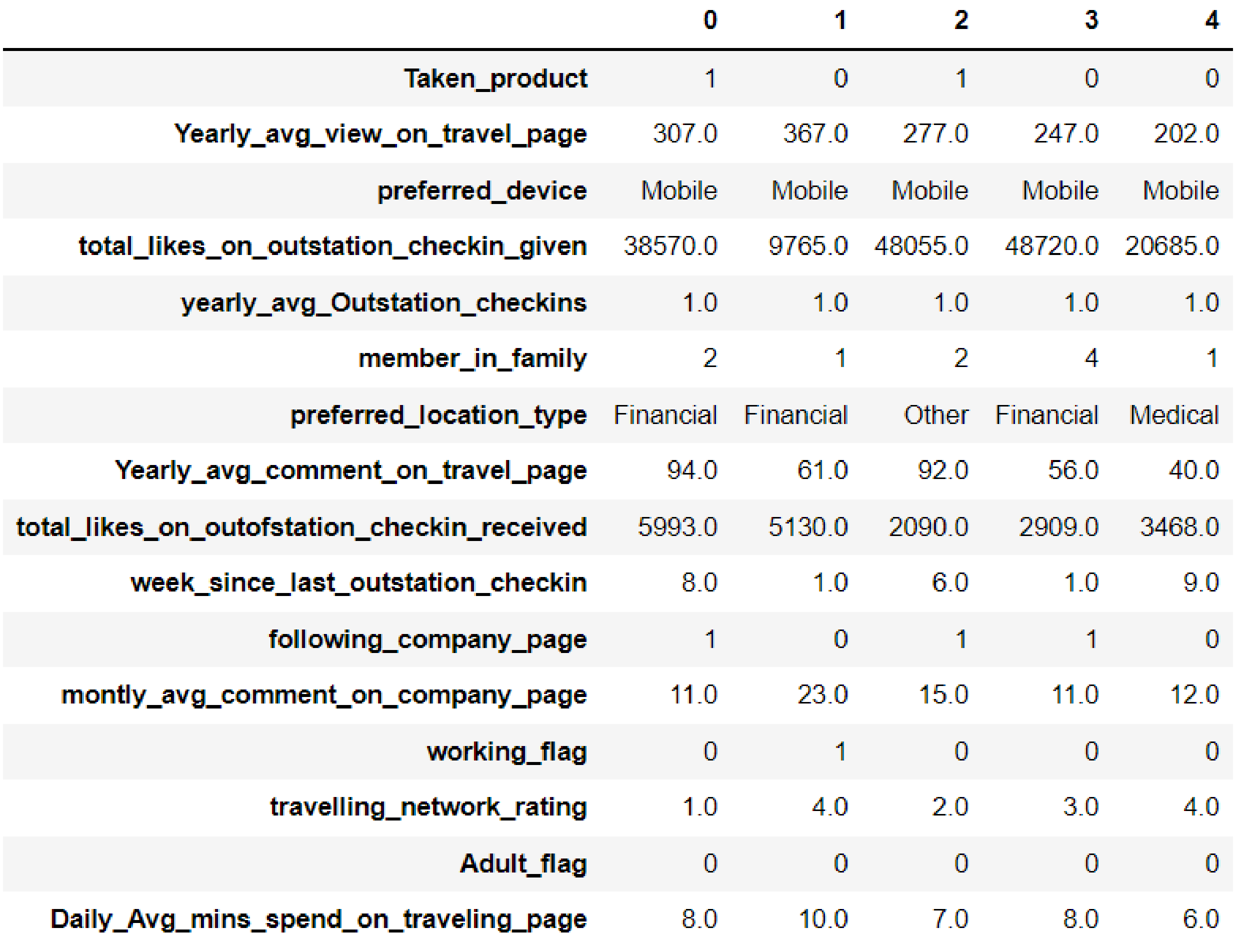
2. EDA - Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.

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

The data has been collected by the third party, here in this case it's the social networking site. Digital and Social behaviour of 11,760 unique customers has been collected in regards to their interest on travel. The data consist of their,

- Likes, comments and reviews on travel related pages.
- Outstation Check-ins, their frequency, likes and interaction with other's check-ins.
- Personal info such as their family, work status, whether they are adults, average time spent on travel related pages.
- Finally, the target columns states whether each customer has brought a ticket for their next trip from the aviation company.
- There are 3 variables *("Yearly_avg_view_on_travel_page", "yearly_avg_Outstation_checkins" and "Yearly_avg_comment_on_travel_page")* that have data points taken for a year.
- There is 1 variable (*"montly_avg_comment_on_company_page")* that has the data points taken for a month.
- There is some data taken on a daily basis and some on a weekly basis. Then the average of these data is taken and added as a part of the dataset.

Visual Inspection of Data (rows, columns, descriptive details)



	0	1	2	3	4
Taken_product	1	0	1	0	0
Yearly_avg_view_on_travel_page	307.0	367.0	277.0	247.0	202.0
preferred_device	Mobile	Mobile	Mobile	Mobile	Mobile
total_likes_on_outstation_checkin_given	38570.0	9765.0	48055.0	48720.0	20685.0
yearly_avg_Outstation_checkins	1.0	1.0	1.0	1.0	1.0
member_in_family	2	1	2	4	1
preferred_location_type	Financial	Financial	Other	Financial	Medical
Yearly_avg_comment_on_travel_page	94.0	61.0	92.0	56.0	40.0
total_likes_on_outofstation_checkin_received	5993.0	5130.0	2090.0	2909.0	3468.0
week_since_last_outstation_checkin	8.0	1.0	6.0	1.0	9.0
following_company_page	1	0	1	1	0
montly_avg_comment_on_company_page	11.0	23.0	15.0	11.0	12.0
working_flag	0	1	0	0	0
travelling_network_rating	1.0	4.0	2.0	3.0	4.0
Adult_flag	0	0	0	0	0
Daily_Avg_mins_spend_on_traveling_page	8.0	10.0	7.0	8.0	6.0

Table 1

The number of Observation (rows): 11760

The number of Variables (columns): 17

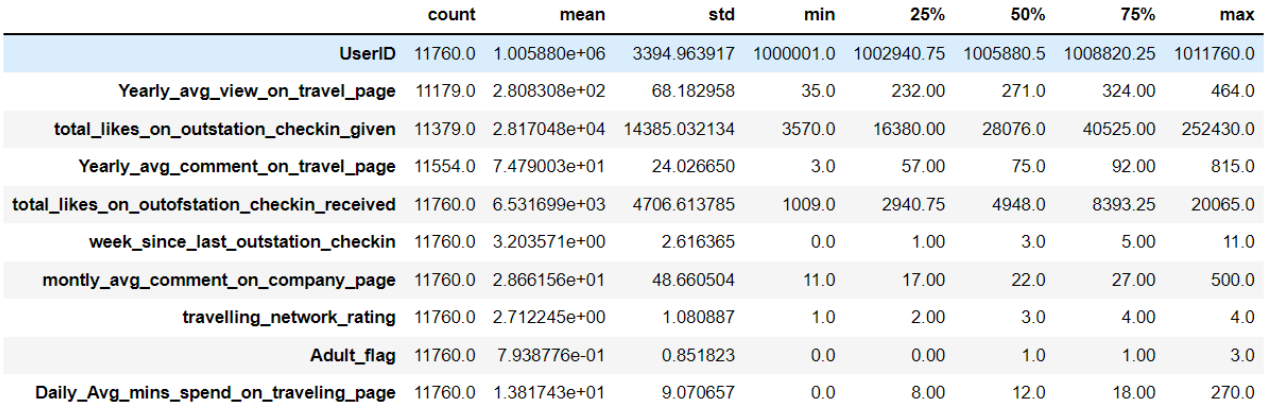
Out of which 7 are object data type and the remaining variables are of integer and float.

Duplicated Observations:

The number of duplicate observations: 0

Summary Statistics

Continuous Numerical Variable



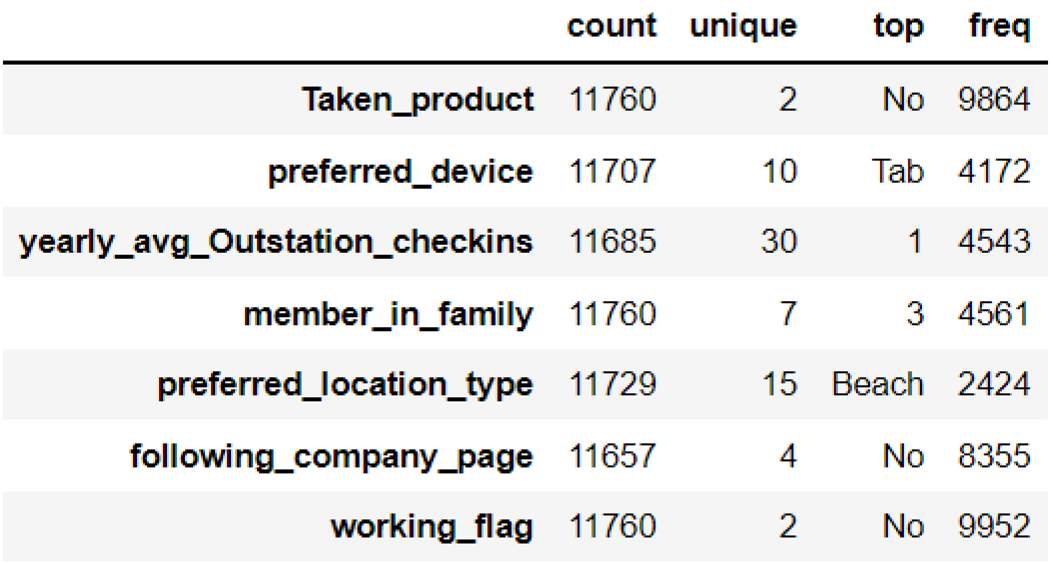
	count	mean	std	min	25%	50%	75%	max
UserID	11760.0	1.005880e+06	3394.963917	1000001.0	1002940.75	1005880.5	1008820.25	1011760.0
Yearly_avg_view_on_travel_page	11179.0	2.808308e+02	68.182958	35.0	232.00	271.0	324.00	464.0
total_likes_on_outstation_checkin_given	11379.0	2.817048e+04	14385.032134	3570.0	16380.00	28076.0	40525.00	252430.0
Yearly_avg_comment_on_travel_page	11554.0	7.479003e+01	24.026650	3.0	57.00	75.0	92.00	815.0
total_likes_on_outofstation_checkin_received	11760.0	6.531699e+03	4706.613785	1009.0	2940.75	4948.0	8393.25	20065.0
week_since_last_outstation_checkin	11760.0	3.203571e+00	2.616365	0.0	1.00	3.0	5.00	11.0
montly_avg_comment_on_company_page	11760.0	2.866156e+01	48.660504	11.0	17.00	22.0	27.00	500.0
travelling_network_rating	11760.0	2.712245e+00	1.080887	1.0	2.00	3.0	4.00	4.0
Adult_flag	11760.0	7.938776e-01	0.851823	0.0	0.00	1.0	1.00	3.0
Daily_Avg_mins_spend_on_traveling_page	11760.0	1.381743e+01	9.070657	0.0	8.00	12.0	18.00	270.0

Table 2

- Few of the variable has larger difference in mean and median (50%) thus only few variables has outliers.

- *'travelling_network_rating'* and *'Adult_flag'* seems to be categorical column. But listed as numerical data which has to be converted as categorical column.

Categorical Variable



	count	unique	top	freq
Taken_product	11760	2	No	9864
preferred_device	11707	10	Tab	4172
yearly_avg_Outstation_checkins	11685	30	1	4543
member_in_family	11760	7	3	4561
preferred_location_type	11729	15	Beach	2424
following_company_page	11657	4	No	8355
working_flag	11760	2	No	9952

Table 3

- There are a greater number of customers who are not taken the products that customer who took the product as per target variable *'Taken_product'*

- Majority of people's preferred device is *'Tab'.*

- Most preferred location shall be *'Beach'.*

- Majority of the family has 3 members as per given dataset.

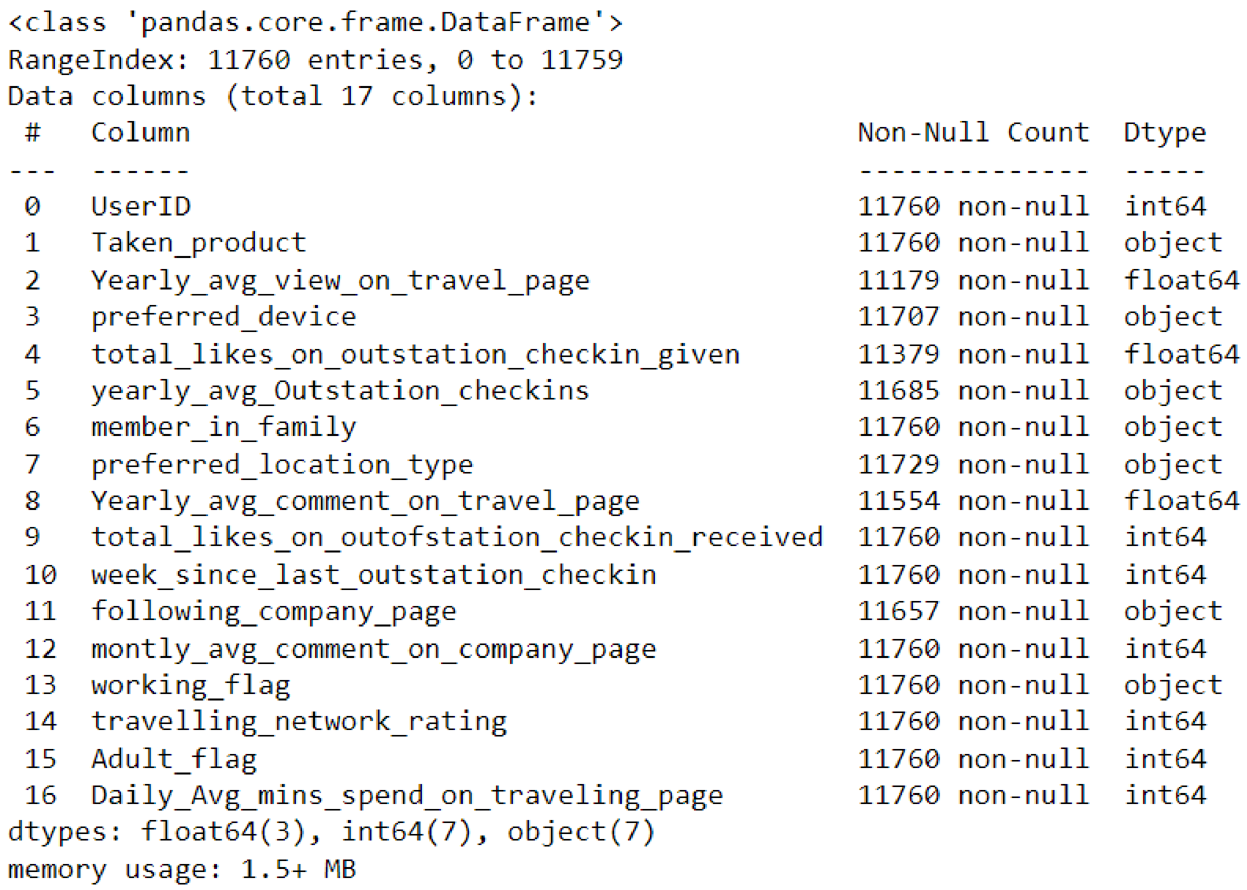
- With the given data we can see it consist of more non-working people.
- Most of the users are with the outstation check-ins within 1 week.

Data Wrangling

- Convert the Preferred device categories into two, only with Laptop and anything other than Laptop shall be converted as Mobile devices
- Replace the '*' with the most frequent value (mode) in the '*Yearly_avg_Outstation_checkins*' and convert them into float datatype.
- Replace 'Three' with the numerical value '3' in "*Member_in_family*" column.
- Combine 'Tour Travel' and 'Tour and Travel' into one category.
- The variable 'Adult_flag' must have only binary value, whether the user is an adult: Yes or No. However, it has 2 and 3 as their values. Values can be imputed as such 0 is not adult and anything other than that shall be adult.

Understanding of attributes

The image below gives the basic information of the data set. There are 17 variables, out of which 3 variables are of type float, 7 variables are of type int and 7 variables are of type object. The data given is for 11760 individuals. There are null values that require processing.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11760 entries, 0 to 11759
Data columns (total 17 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   UserID                                 11760 non-null  int64  
 1   Taken_product                          11760 non-null  object 
 2   Yearly_avg_view_on_travel_page         11179 non-null  float64
 3   preferred_device                       11707 non-null  object 
 4   total_likes_on_outstation_checkin_given  11379 non-null  float64
 5   yearly_avg_Outstation_checkins         11685 non-null  object 
 6   member_in_family                       11760 non-null  object 
 7   preferred_location_type                11729 non-null  object 
 8   Yearly_avg_comment_on_travel_page      11554 non-null  float64
 9   total_likes_on_outofstation_checkin_received  11760 non-null  int64  
 10  week_since_last_outstation_checkin     11760 non-null  int64  
 11  following_company_page                 11657 non-null  object 
 12  montly_avg_comment_on_company_page     11760 non-null  int64  
 13  working_flag                           11760 non-null  object 
 14  travelling_network_rating              11760 non-null  int64  
 15  Adult_flag                             11760 non-null  int64  
 16  Daily_Avg_mins_spend_on_traveling_page  11760 non-null  int64  
dtypes: float64(3), int64(7), object(7)
memory usage: 1.5+ MB
```

Table 4

As mentioned in problem statement, anything which is not a laptop can be considered as mobile phone usage. Hence renaming of other types to mobile is done under variable '*preferred_device'.*

UNIVARIATE ANALYSIS

Univariate analysis is the simplest form of analysing data. "Uni" means "one", so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression) and its major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

(i) Continuous Variables

Skewness

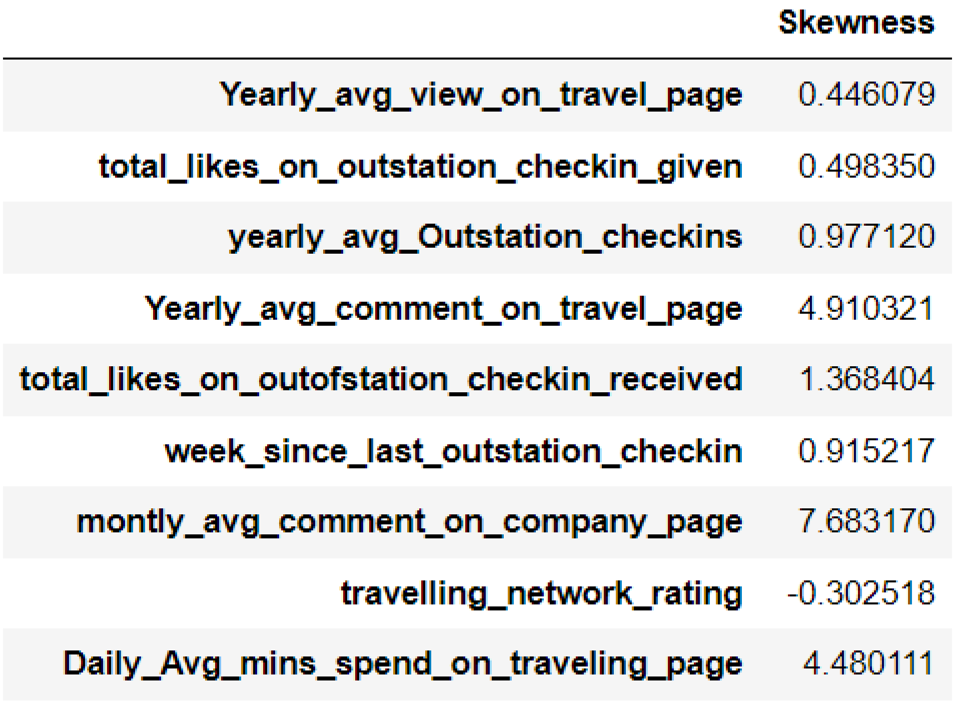
Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value.

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

Table 5

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.



	Skewness
Yearly_avg_view_on_travel_page	0.446079
total_likes_on_outstation_checkin_given	0.498350
yearly_avg_Outstation_checkins	0.977120
Yearly_avg_comment_on_travel_page	4.910321
total_likes_on_outofstation_checkin_received	1.368404
week_since_last_outstation_checkin	0.915217
montly_avg_comment_on_company_page	7.683170
travelling_network_rating	-0.302518
Daily_Avg_mins_spend_on_traveling_page	4.480111

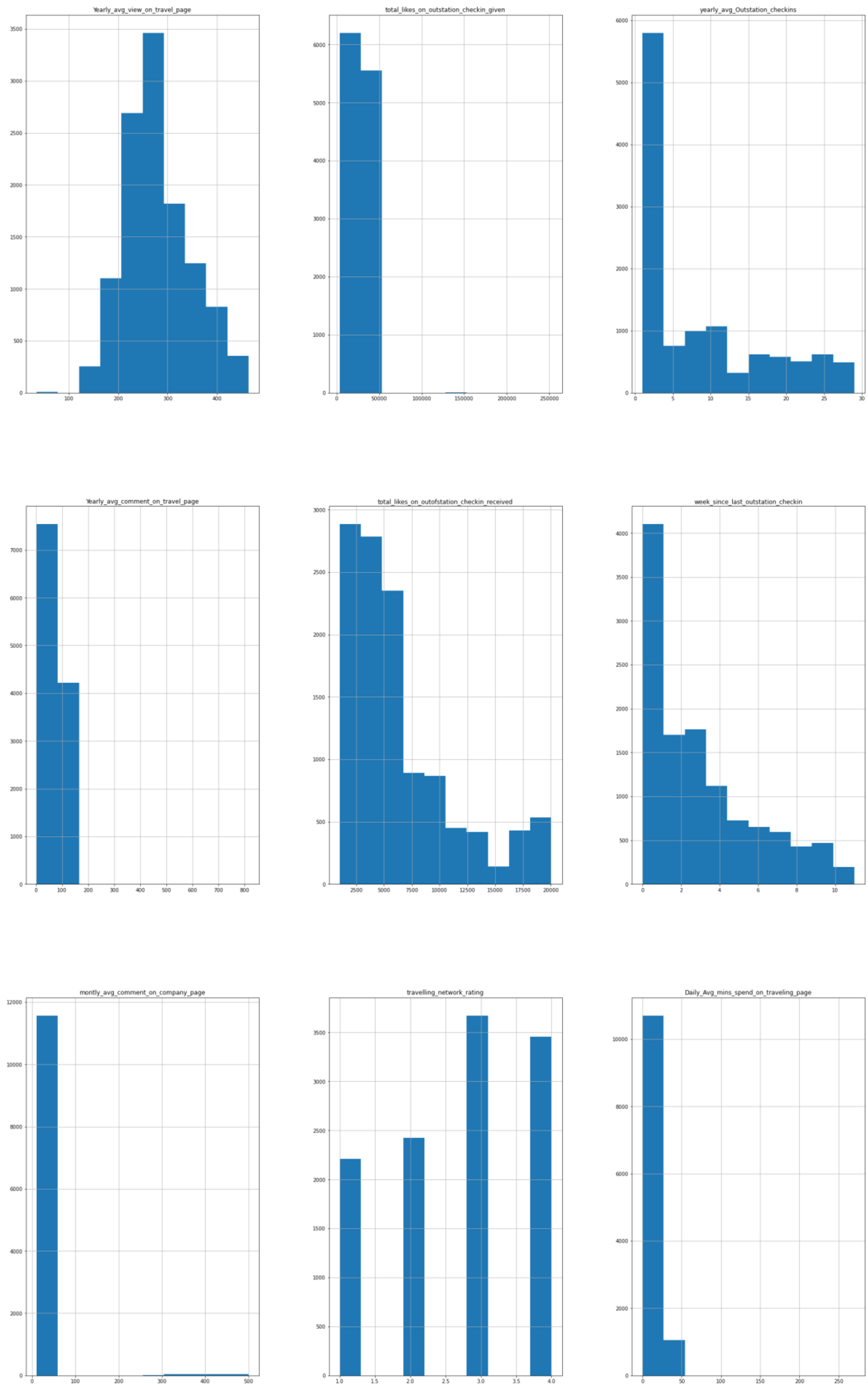
Table 6

Right skewed variables
Yearly_avg_view_on_travel_page
Total_likes_on_outstation_checkin_given
Yearly_avg_Outstation_checkins
Yearly_avg_comment_on_travel_page
Total_likes_on_outofstation_checkin_received
Week_since_last_outstation_checkin
Montly_avg_comment_on_company_page
Daily_Avg_mins_spend_on_traveling_page
Left skewed variables
travelling_network_rating

Table 7

Histogram

The **histograms** are used for **numerical variables** to perform univariate analysis. It is clear from the graph (Graph 1) that all the numerical variables are skewed.

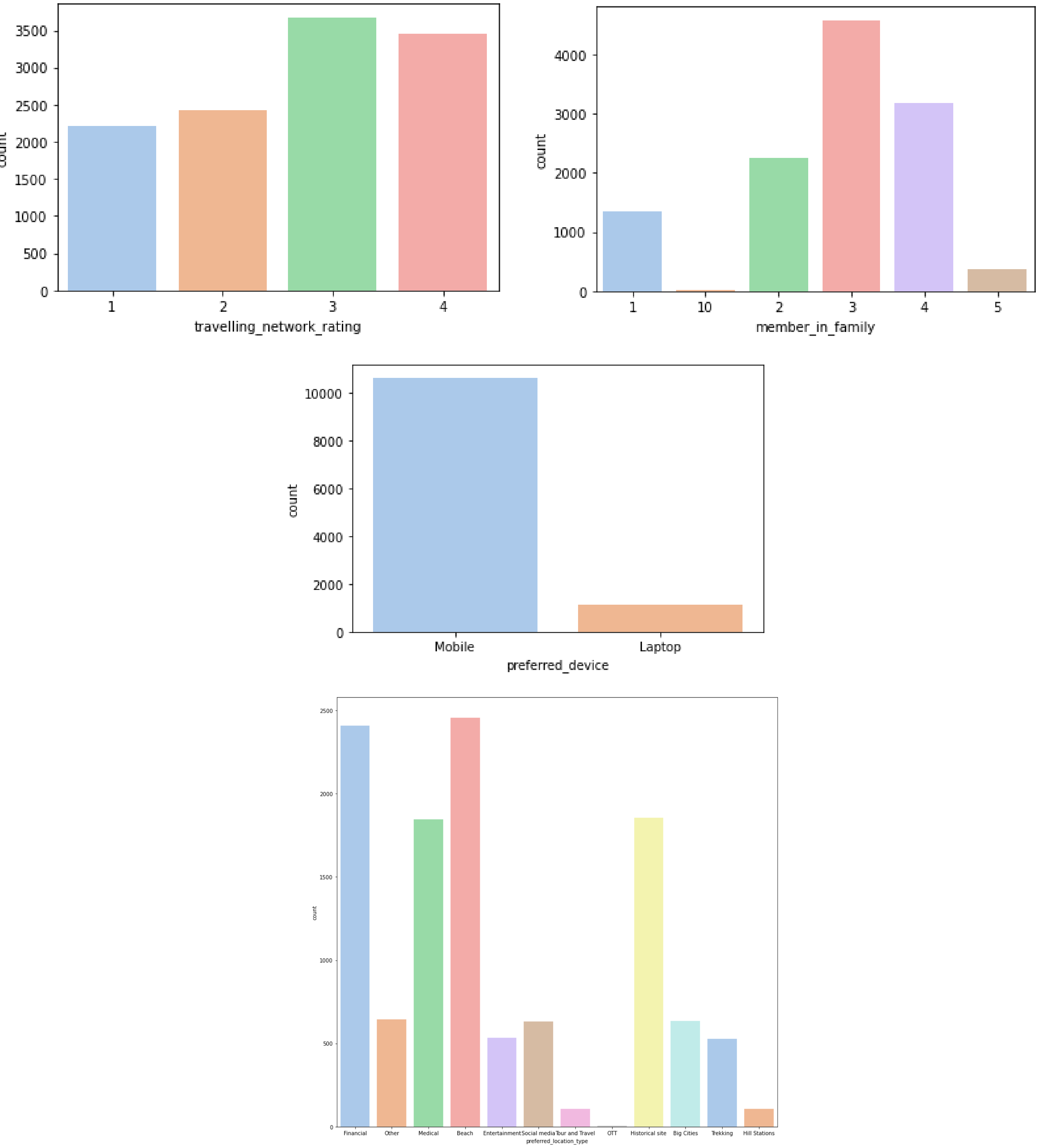

Yearly_avg_view_on_travel_page
0
500
1000
1500
2000
2500
3000
3500
100
200
300
400
total_likes_on_outstation_checkin_given
0
1000
2000
3000
4000
5000
6000
0
50000
100000
150000
200000
250000
yearly_avg_Outstation_checkins
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25
30
Yearly_avg_comment_on_travel_page
0
1000
2000
3000
4000
5000
6000
7000
0
100
200
300
400
500
600
700
800
total_likes_on_outofstation_checkin_received
0
500
1000
1500
2000
2500
3000
2500
5000
7500
10000
12500
15000
17500
20000
week_since_last_outstation_checkin
0
500
1000
1500
2000
2500
3000
3500
4000
0
2
4
6
8
10
montly_avg_comment_on_company_page
0
2000
4000
6000
8000
10000
12000
0
100
200
300
400
500
travelling_network_rating
0
500
1000
1500
2000
2500
3000
3500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Daily_Avg_mins_spend_on_traveling_page
0
2000
4000
6000
8000
10000
0
50
100
150
200
250

Graph 1

(ii) Categorical Variables

Count Plot

The **count plots** are used for **categorical variables** to perform univariate analysis. It gives the count of each category in a particular variable.



Graph 2

From above plot, we can understand that **travelling_network_rating** 3 is highest indicating that majority of user friends don’t like travelling.

member_in_family plot indicates that majority of users has 3 members in their family.

preferred_device for majority of the users is Mobile.

preferred_location_type for majority of users is beach and financial.

BIVARIATE AND MULTIVARIATE ANALYSIS

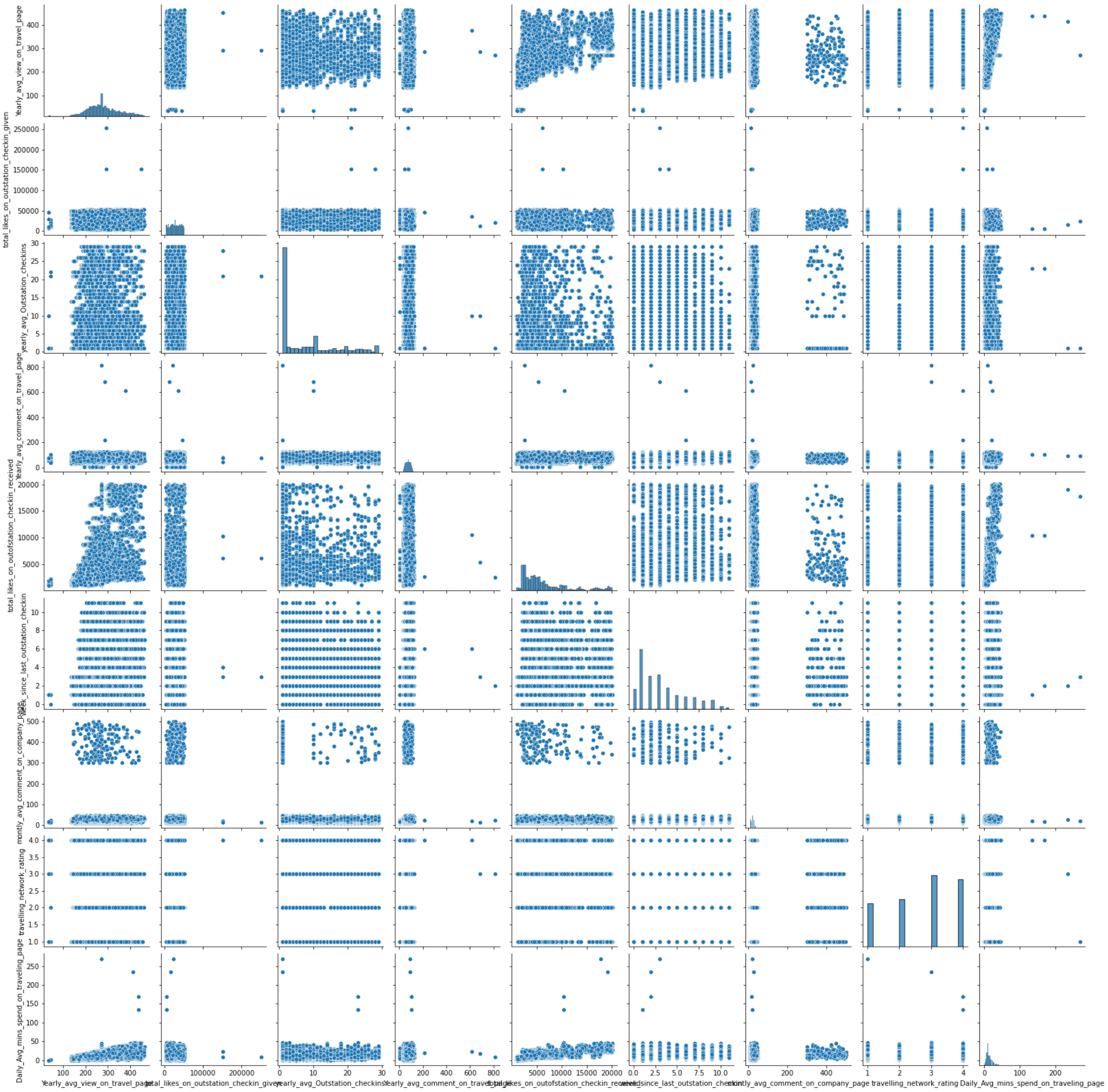
Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association.

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

(i) Continuous variables

Pair Plot and Heat Maps

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column.



Graph 3

From the above pair plot, we can understand that, there is very minimal correlation between features.

Also, skewness exists for almost all features.

The heat map can also be used to check the association between two numeric variables. All the boxes with a value higher than 0.8 are highly correlated. But in the given data set none of the variables have a value 0.8 or more. The heat map for all the numerical variables is below.

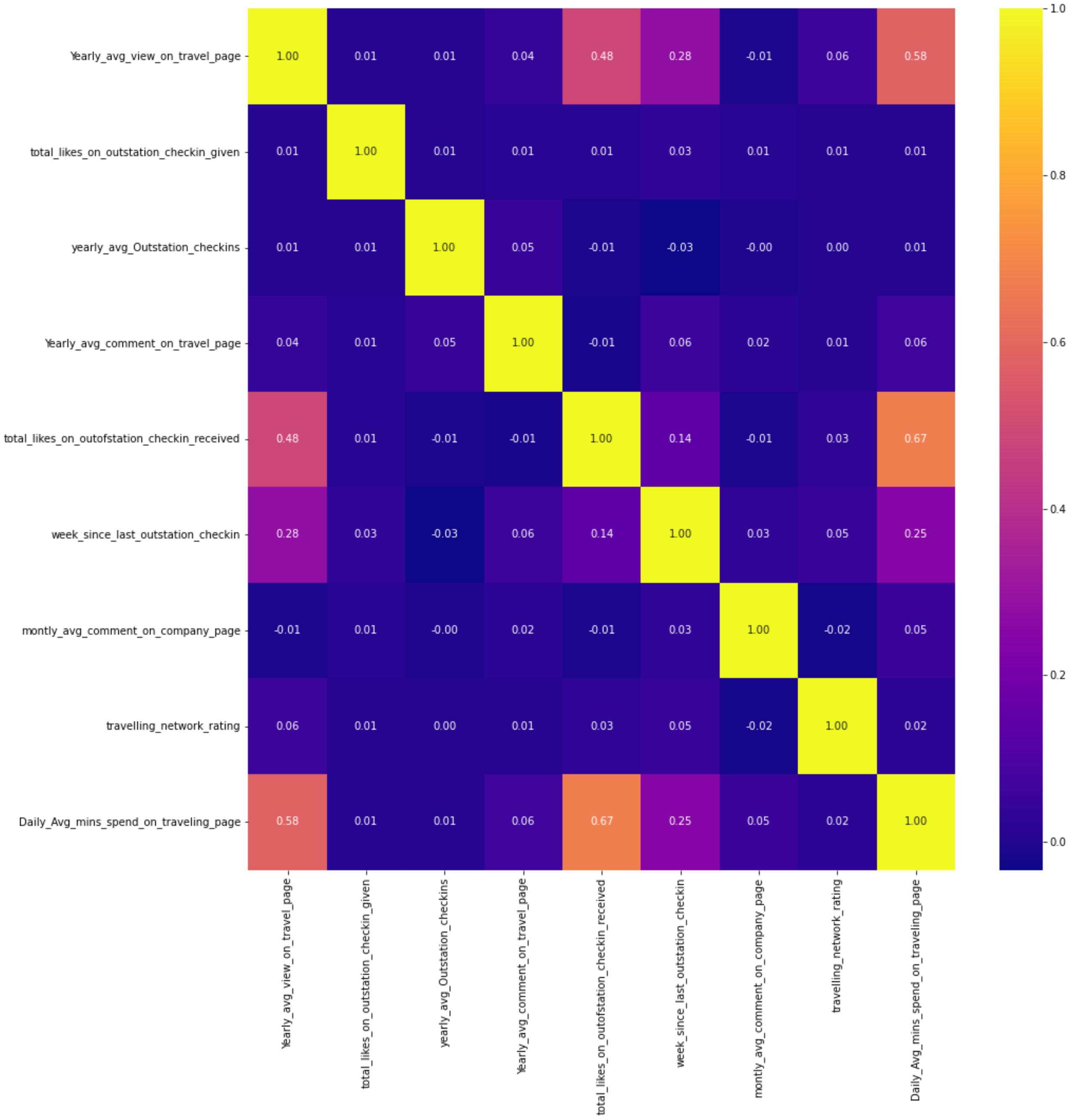
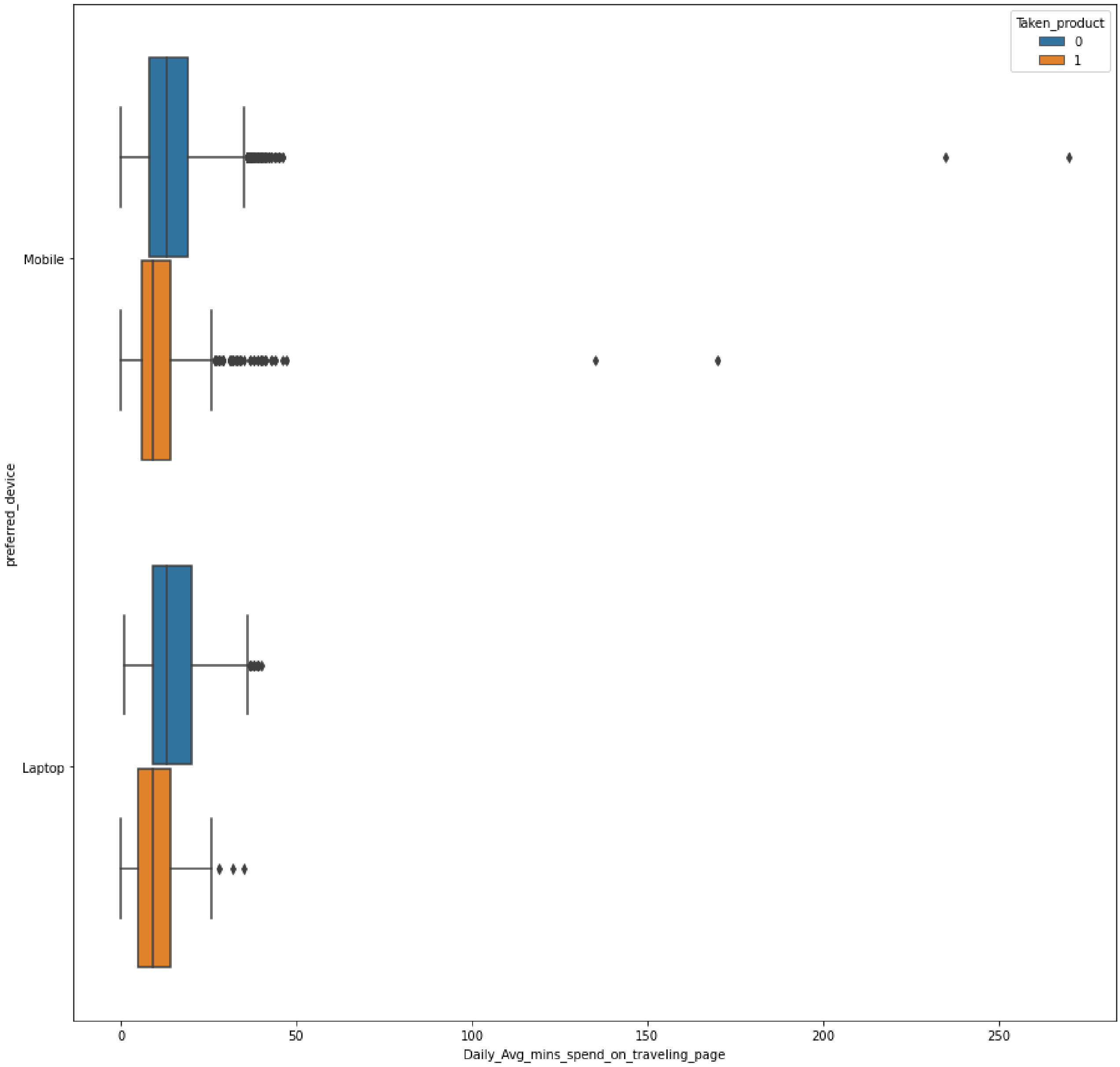


Table 8

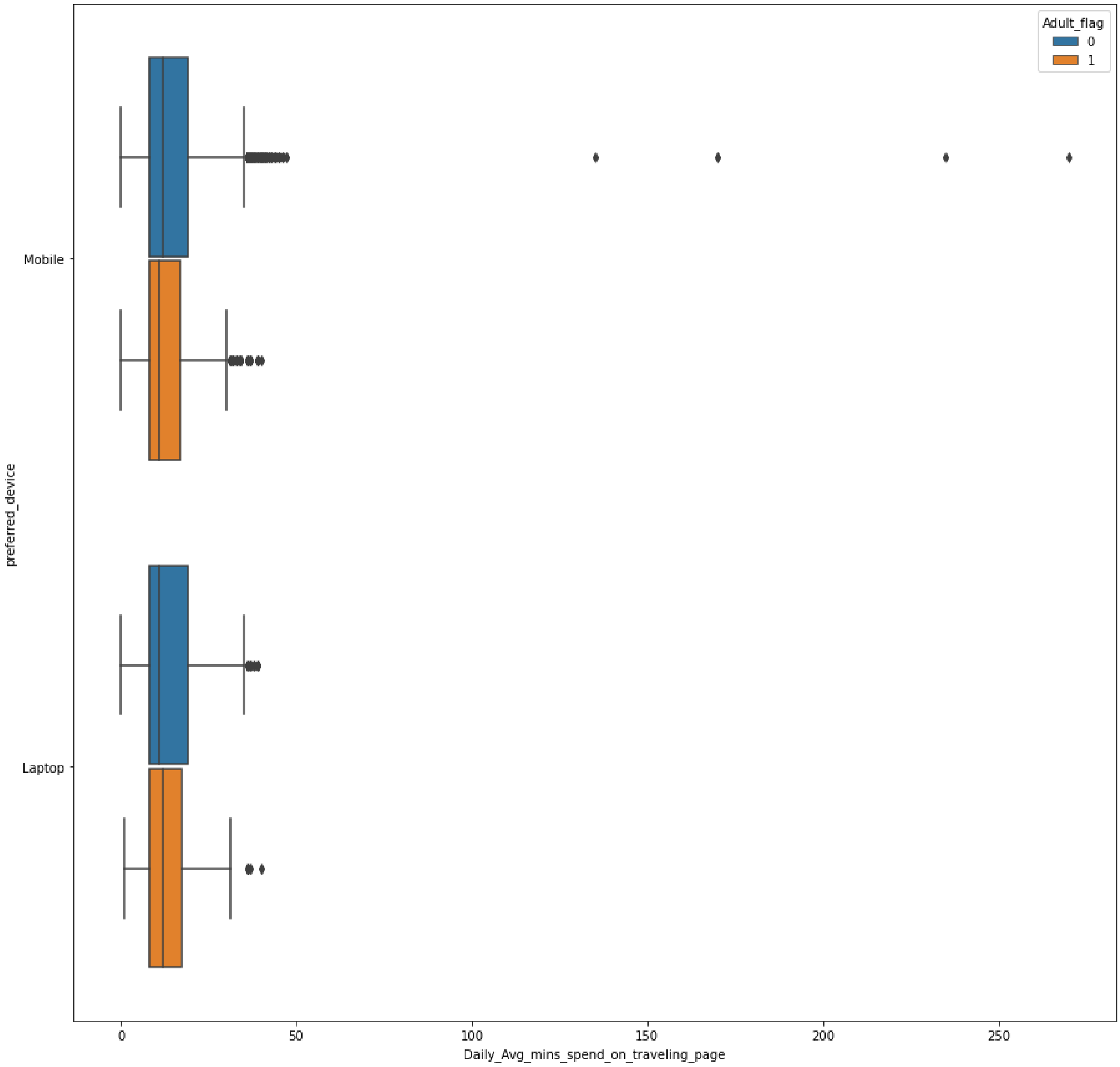
(ii) Categorical variables

Box plot

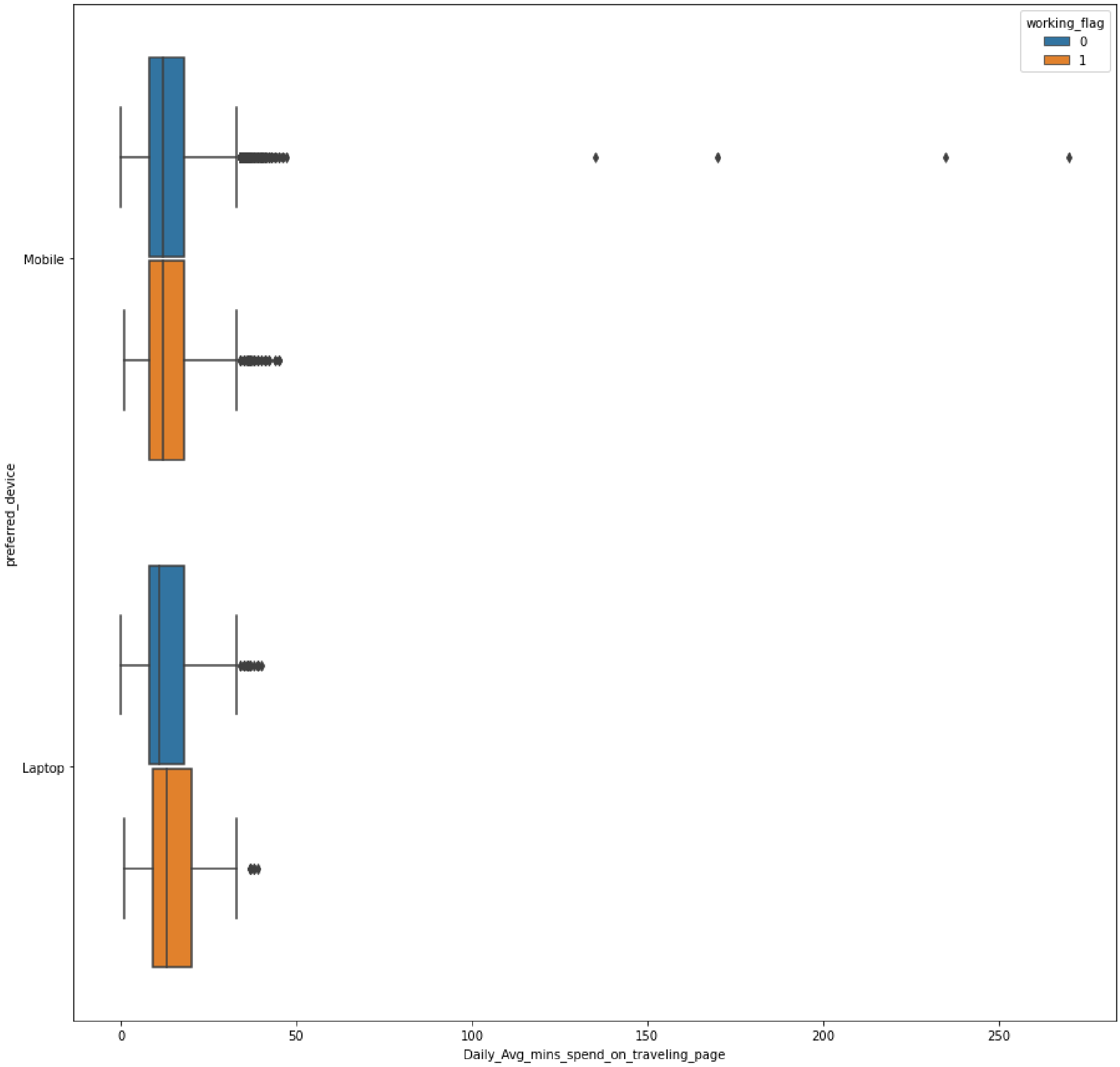
A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.


Taken_product
0
1
Mobile
Laptop
preferred_device
0
50
100
150
200
250
Daily_Avg_mins_spend_on_traveling_page

Graph 4


Adult_flag
0
1
Mobile
Laptop
preferred_device
0
50
100
150
200
250
Daily_Avg_mins_spend_on_traveling_page

Graph 5



Graph 6

Findings from the graphs –

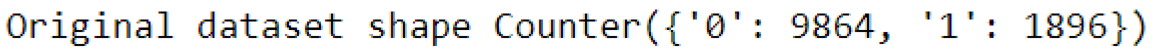
● The device mobile and laptop almost have the same median when it comes to users taking the product.

● For adults, both the devices are used majorly.

● There are almost equal number of Mobile users who have taken the product and those who have not taken the product.

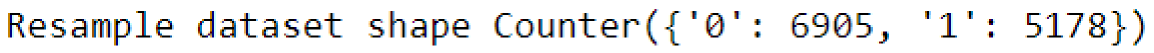
Data Unbalanced

Data is unbalanced as number of counts for ‘1’ is *1896* as compared to ‘0’ which has *9864* counts.



```
Original dataset shape Counter({'0': 9864, '1': 1896})
```

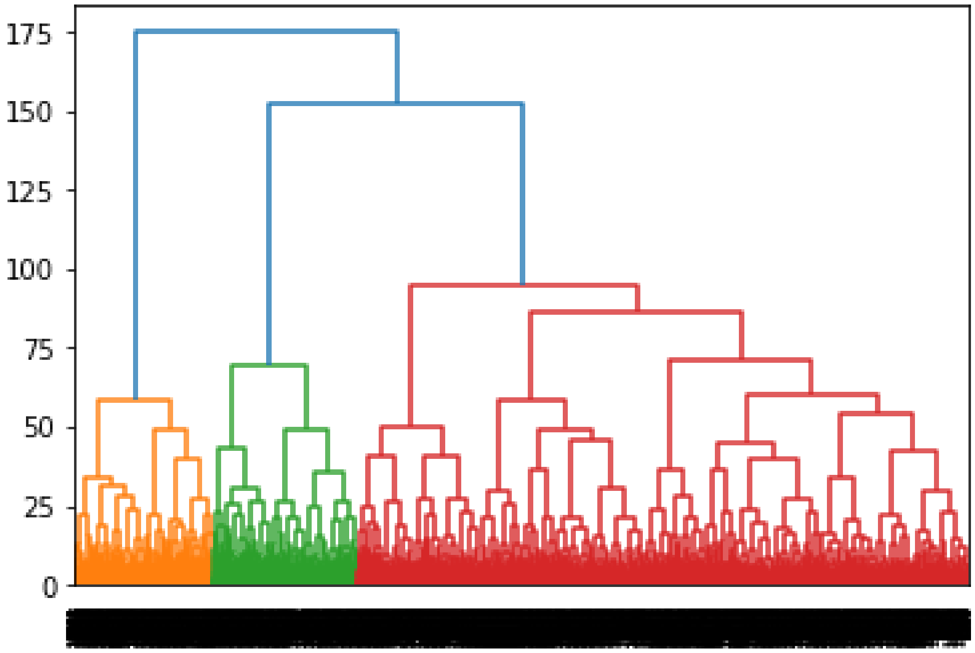
Hence to make it balanced, *SMOTE* is applied. Data set is significantly balanced on using *SMOTE* as number of counts for ‘1’ is increased to *5178* as compared to ‘0’ which has *6905* counts.



```
Resample dataset shape Counter({'0': 6905, '1': 5178})
```

Hierarchical Clustering (HC):

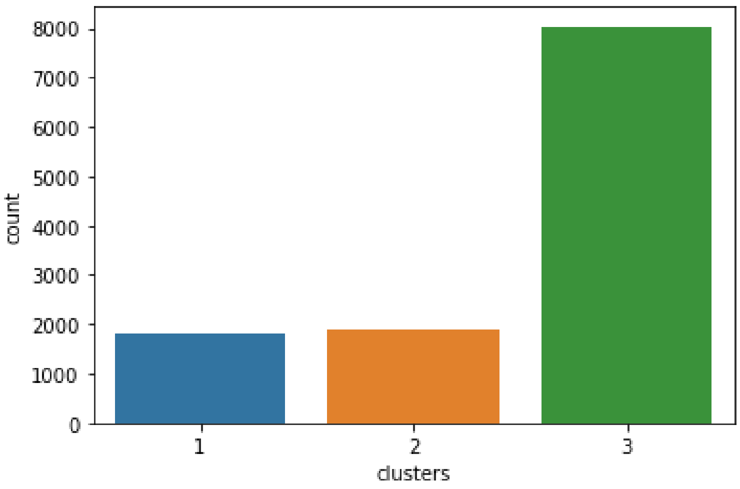
HC was performed on the scaled data using ward linkage.



Graph 7

From the above dendrogram, we noted that data is segregated into 3 clusters.

Further from countplot and value_counts() function, we get no. of data points in each clusters.

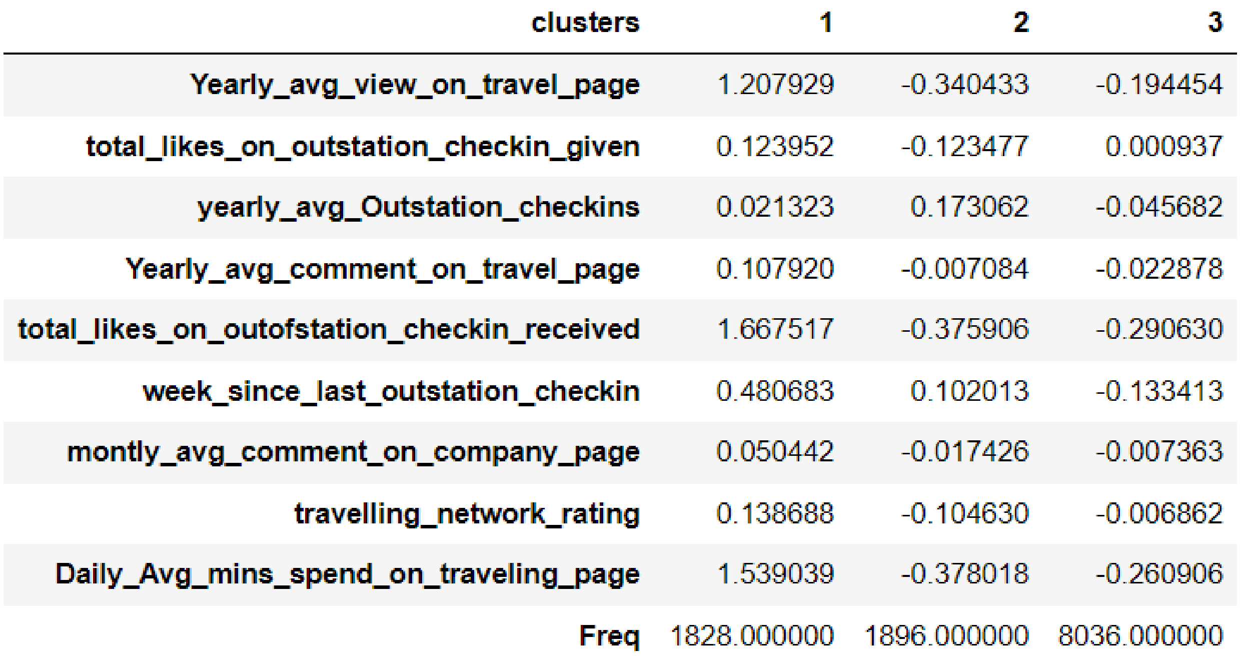


```
1     1828
2     1896
3     8036
Name: clusters, dtype: int64
```

Graph 8

From above details it is clear that cluster 3 has highest number of data points followed by cluster 2 and cluster 1.

Means for each of the numerical features corresponding to different clusters is shown below.



clusters	1	2	3
Yearly_avg_view_on_travel_page	1.207929	-0.340433	-0.194454
total_likes_on_outstation_checkin_given	0.123952	-0.123477	0.000937
yearly_avg_Outstation_checkins	0.021323	0.173062	-0.045682
Yearly_avg_comment_on_travel_page	0.107920	-0.007084	-0.022878
total_likes_on_outofstation_checkin_received	1.667517	-0.375906	-0.290630
week_since_last_outstation_checkin	0.480683	0.102013	-0.133413
montly_avg_comment_on_company_page	0.050442	-0.017426	-0.007363
travelling_network_rating	0.138688	-0.104630	-0.006862
Daily_Avg_mins_spend_on_traveling_page	1.539039	-0.378018	-0.260906
Freq	1828.000000	1896.000000	8036.000000

Table 9

Cluster 3 has maximum number of data points with positive mean only for *total_likes_on_outstation_checkin_given*. Hence majority of the users has given positive likes on out of station checking's in last year.

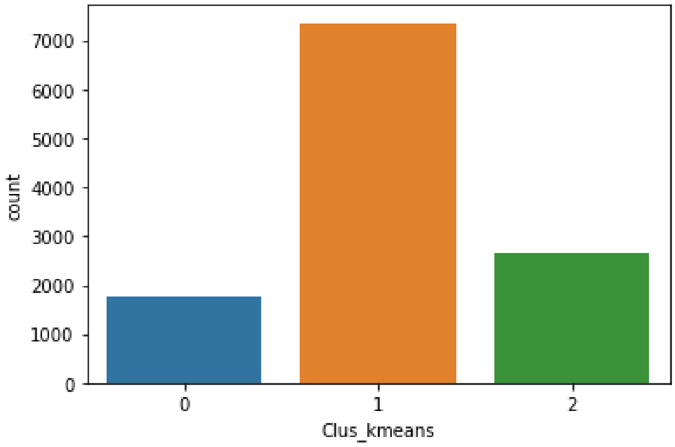
montly_avg_comment_on_company_page, travelling_network_rating, yearly_avg_Outstation_checkins, Yearly_avg_comment_on_travel_page means are negative but very close to zero.

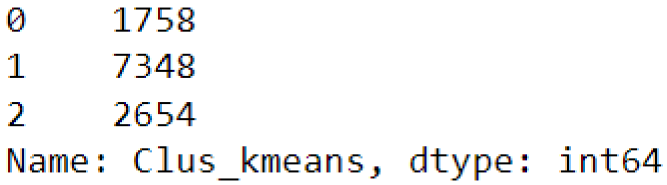
However, *Yearly_avg_view_on_travel_page, total_likes_on_outofstation_checkin_received, week_since_last_outstation_checkin, Daily_Avg_mins_spend_on_traveling_page* means are significantly negative and needs attention to turn them positive.

K Means Clustering (KC):

For KC, KMeans and WSS are performed to get appropriate number of clusters and same is mapped to data points (refer Python Notebook file).

Below is the countplot and value_counts() for formed clusters.



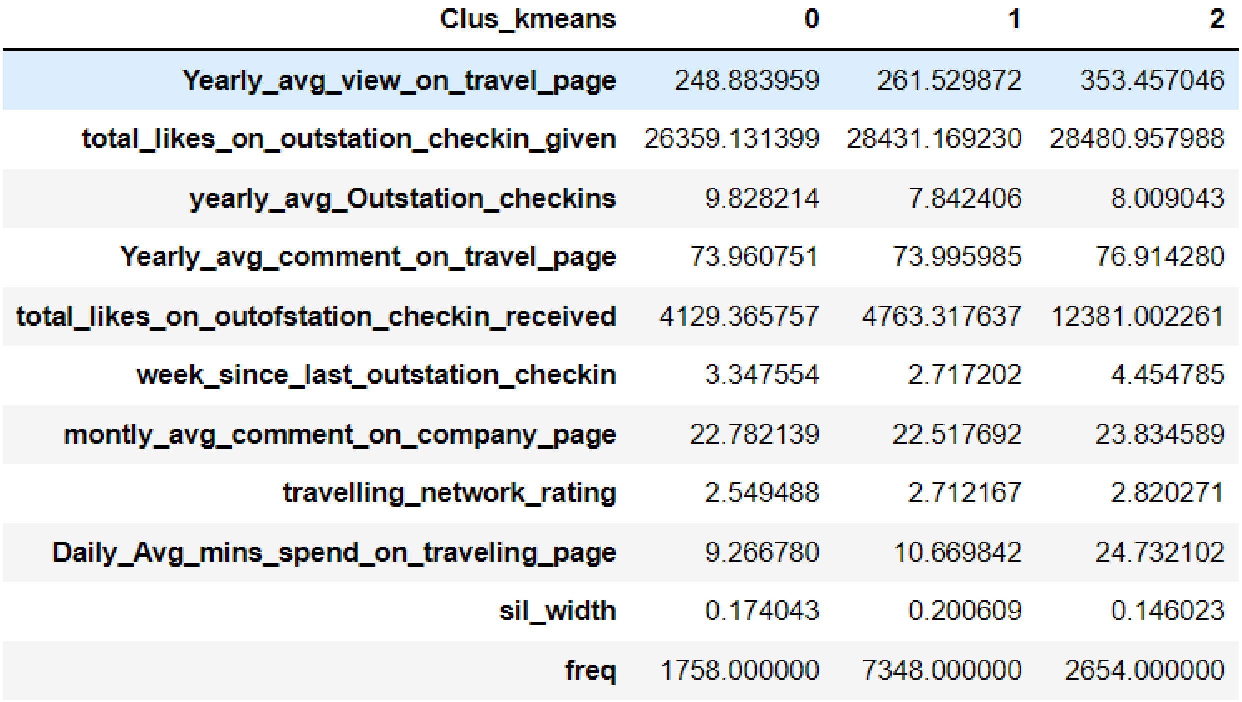


```
0    1758
1    7348
2    2654
Name: Clus_kmeans, dtype: int64
```

Graph 9

From above details it is clear that cluster 1 has highest number of data points followed by cluster 2 and cluster 0.

Means for each of the numerical features corresponding to different clusters is shown below.



Clus_kmeans	0	1	2
Yearly_avg_view_on_travel_page	248.883959	261.529872	353.457046
total_likes_on_outstation_checkin_given	26359.131399	28431.169230	28480.957988
yearly_avg_Outstation_checkins	9.828214	7.842406	8.009043
Yearly_avg_comment_on_travel_page	73.960751	73.995985	76.914280
total_likes_on_outofstation_checkin_received	4129.365757	4763.317637	12381.002261
week_since_last_outstation_checkin	3.347554	2.717202	4.454785
montly_avg_comment_on_company_page	22.782139	22.517692	23.834589
travelling_network_rating	2.549488	2.712167	2.820271
Daily_Avg_mins_spend_on_traveling_page	9.266780	10.669842	24.732102
sil_width	0.174043	0.200609	0.146023
freq	1758.000000	7348.000000	2654.000000

Table 10

From above table we can understand that Cluster 2 has highest means for all features followed by Cluster 1 and Cluster 0

Since Cluster 1 has maximum number of data points, there is scope of improvement in features to achieve highest means.

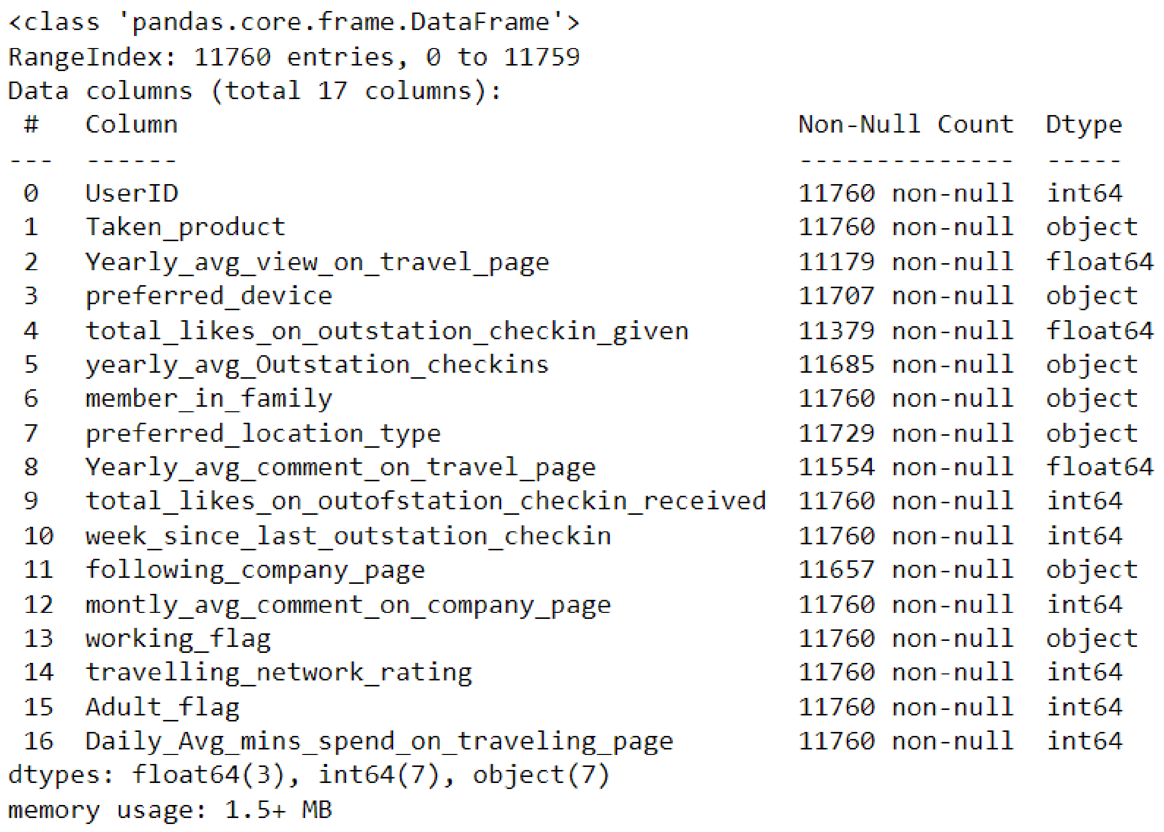
3. Data Cleaning and Pre-processing - Approach used for identifying and treating missing values and outlier treatment (and why) - Need for variable transformation (if any) - Variables removed or added and why (if any)

Removal of unwanted variables (if applicable)

As of now there is no need of dropping any variables. Only variable we can get rid of is of UserID columns as it doesn't provide any information about the class of target variable.

Missing Value treatment (if applicable)

There were a lot of missing variables in the given data set. It can be seen in the below image where the unhighlighted variables have null values. The missing values were treated with imputation. It is not advisable to drop the missing values as there were only a few row values missing and they can be imputed.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11760 entries, 0 to 11759
Data columns (total 17 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   UserID                                  11760 non-null  int64  
 1   Taken_product                           11760 non-null  object 
 2   Yearly_avg_view_on_travel_page          11179 non-null  float64
 3   preferred_device                        11707 non-null  object 
 4   total_likes_on_outstation_checkin_given 11379 non-null  float64
 5   yearly_avg_Outstation_checkins          11685 non-null  object 
 6   member_in_family                        11760 non-null  object 
 7   preferred_location_type                 11729 non-null  object 
 8   Yearly_avg_comment_on_travel_page       11554 non-null  float64
 9   total_likes_on_outofstation_checkin_received 11760 non-null  int64  
 10  week_since_last_outstation_checkin      11760 non-null  int64  
 11  following_company_page                  11657 non-null  object 
 12  montly_avg_comment_on_company_page      11760 non-null  int64  
 13  working_flag                            11760 non-null  object 
 14  travelling_network_rating               11760 non-null  int64  
 15  Adult_flag                              11760 non-null  int64  
 16  Daily_Avg_mins_spend_on_traveling_page  11760 non-null  int64  
dtypes: float64(3), int64(7), object(7)
memory usage: 1.5+ MB
```

Table 11

The variables *Yearly_avg_view_on_travel_page, preferred_device, total_likes_on_outstation_checkin_given, yearly_avg_Outstation_checkins, preferred_location_type, Yearly_avg_comment_on_travel_page and following_company_page* had to be treated for missing values. The object type variables were imputed with mode function and the numerical variables were imputed with median function. As the mode and median are resistant to outliers, these were chosen for imputation.

Outlier Treatment (if required)

As many of the continuous variables has outliers and extreme values which shall be removed as many of the Machine learning algorithm such as Logistic Regression are sensitive to outliers.

Any values above 1.5 x IQR from Q3 shall be floored to that limit, likewise any values below 1.5 x IQR from Q1 shall be capped to that lower limit. IQR shall be calculated as difference between Q3 and Q1.

There are outliers in the below mentioned variables. This is evident from the box plots.

Yearly_avg_view_on_travel_page ,

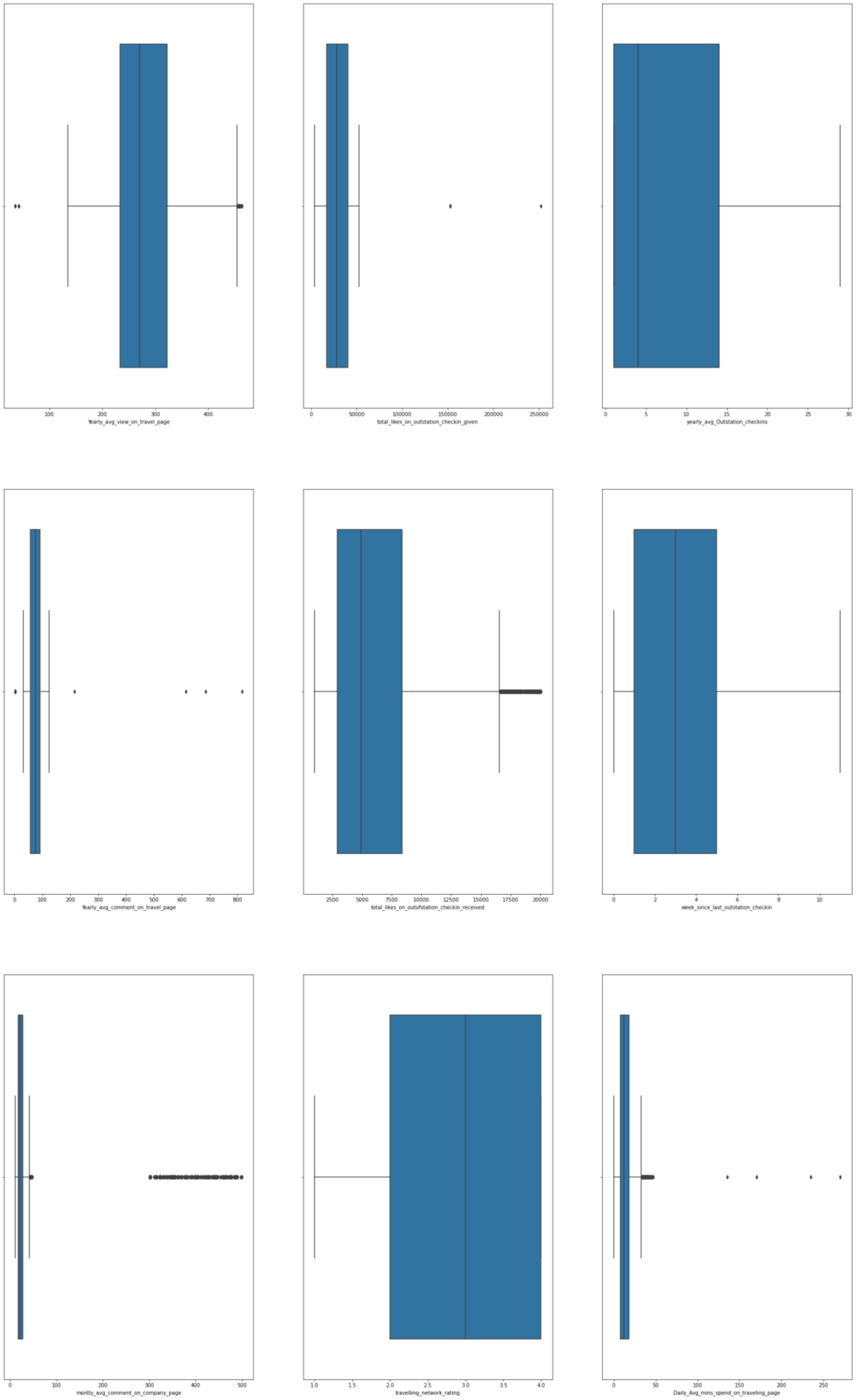
Total_likes_on_outstation_checkin_given

Yearly_avg_comment_on_travel_page

Total_likes_on_outofstation_checkin_received

Montly_avg_comment_on_company_page

Daily_Avg_mins_spend_on_traveling_page


100
200
300
400
Yearly_avg_view_on_travel_page
0
50000
100000
150000
200000
250000
total_likes_on_outstation_checkin_given
0
5
10
15
20
25
30
yearly_avg_Outstation_checkins
0
100
200
300
400
500
600
700
800
Yearly_avg_comment_on_travel_page
2500
5000
7500
10000
12500
15000
17500
20000
total_likes_on_outofstation_checkin_received
0
2
4
6
8
10
week_since_last_outstation_checkin
0
100
200
300
400
500
montly_avg_comment_on_company_page
1.0
1.5
2.0
2.5
3.0
3.5
4.0
travelling_network_rating
0
50
100
150
200
250
Daily_Avg_mins_spend_on_traveling_page

Graph 10

The outlier in the data set is treated using the IQR method.

Inter quartile range (IQR) method –

Each dataset can be divided into quartiles. The first quartile point indicates that 25% of the data points are below that value whereas the second quartile is considered as the median point of the dataset.

The inter quartile method finds the outliers on numerical datasets by following the procedure below,

Find the first quartile, Q1.

Find the third quartile, Q3.

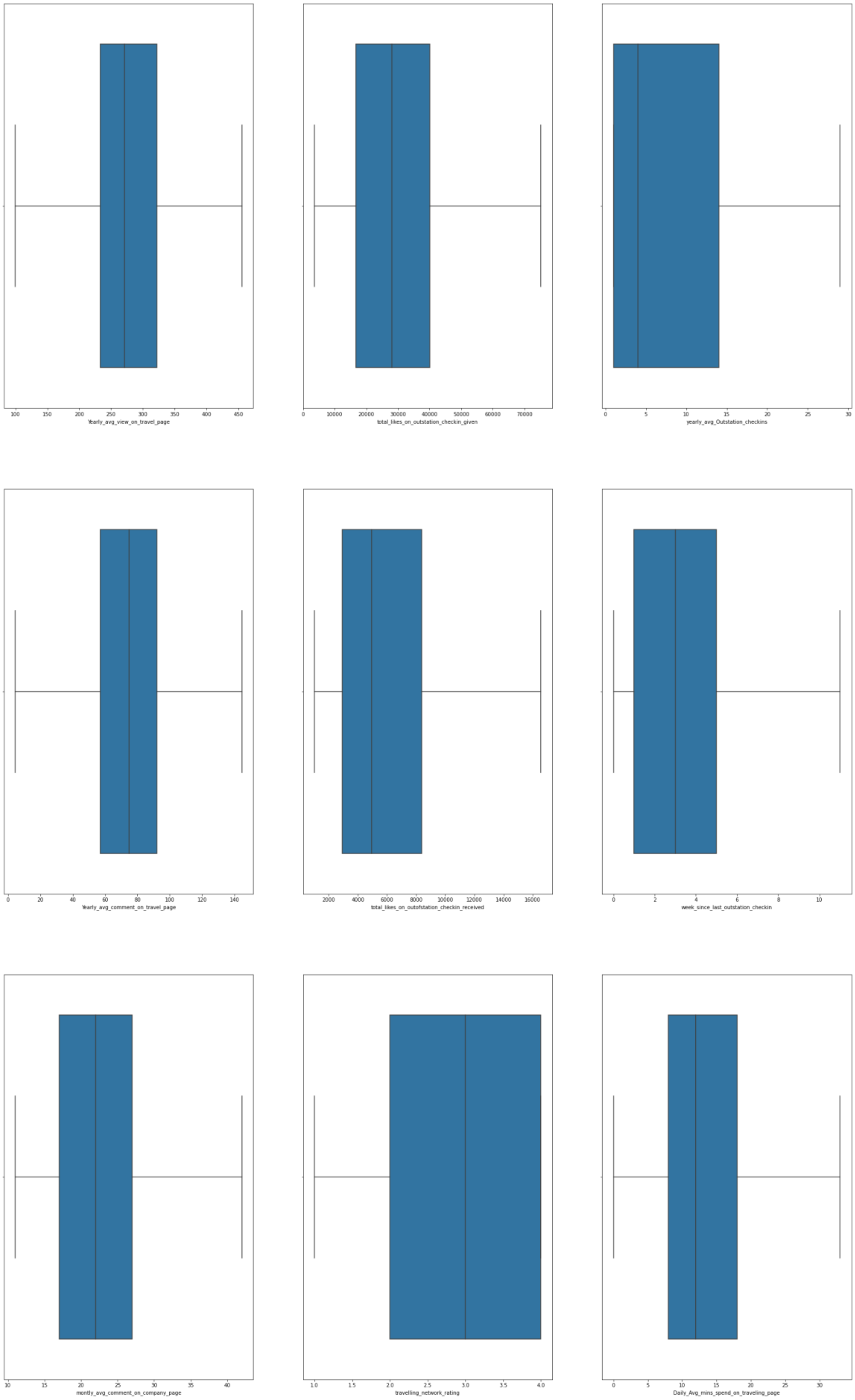
Calculate the IQR. IQR= Q3-Q1.

Define the normal data range with lower limit as Q1–1.5*IQR and upper limit as Q3+1.5*IQR.

Any data point outside this range is considered an outlier and should be removed for further analysis.

The concept of quartiles and IQR can best be visualized from the boxplot. It has the minimum and maximum point defined as Q1–1.5*IQR and Q3+1.5*IQR respectively.

Any point outside this range is outlier.


100 150 200 250 300 350 400 450
Yearly_avg_view_on_travel_page
0 10000 20000 30000 40000 50000 60000 70000
total_likes_on_outstation_checkin_given
0 5 10 15 20 25 30
yearly_avg_Outstation_checkins
0 20 40 60 80 100 120 140
Yearly_avg_comment_on_travel_page
2000 4000 6000 8000 10000 12000 14000 16000
total_likes_on_outofstation_checkin_received
0 2 4 6 8 10
week_since_last_outstation_checkin
10 15 20 25 30 35 40
montly_avg_comment_on_company_page
1.0 1.5 2.0 2.5 3.0 3.5 4.0
travelling_network_rating
0 5 10 15 20 25 30
Daily_Avg_mins_spend_on_traveling_page

Graph 11

Variable transformation (if applicable)

There are multiple transformations done in the data set.

The variable “*member_in_family*” had both numerical and string values. (i.e) it had 3 and three as the value in the rows. That has been changed to a numerical value.

The binary valued data has been transformed to 0 and 1 where 0 being No and 1 being Yes.

The variable “*Adult_flag*” has been transformed to a binary data, where 0 & 1 being 0 and 2 & 3 being 1. This change was done assuming the data is talking about whether the user is an adult or not. In that case, 2 & 3 will not have any meaning to it.

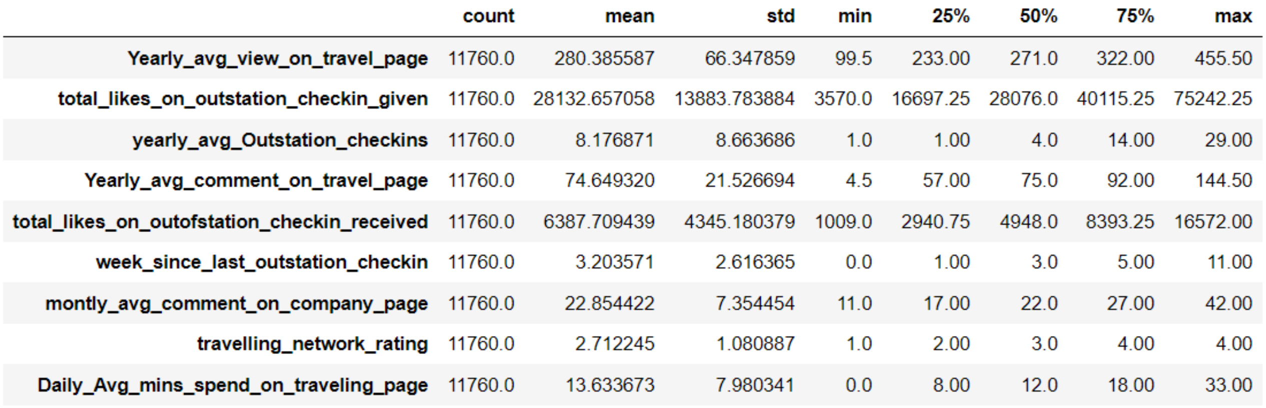
The variable “*preferred_location_type*” has many redundant choices that has been reduced.

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set “*total_likes_on_outstation_checkin_given*” is having values in thousands and “*yearly_avg_Outstation_checkins*” is just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the formula (x-mean)/standard deviation. We will be doing this only for the numerical variables.

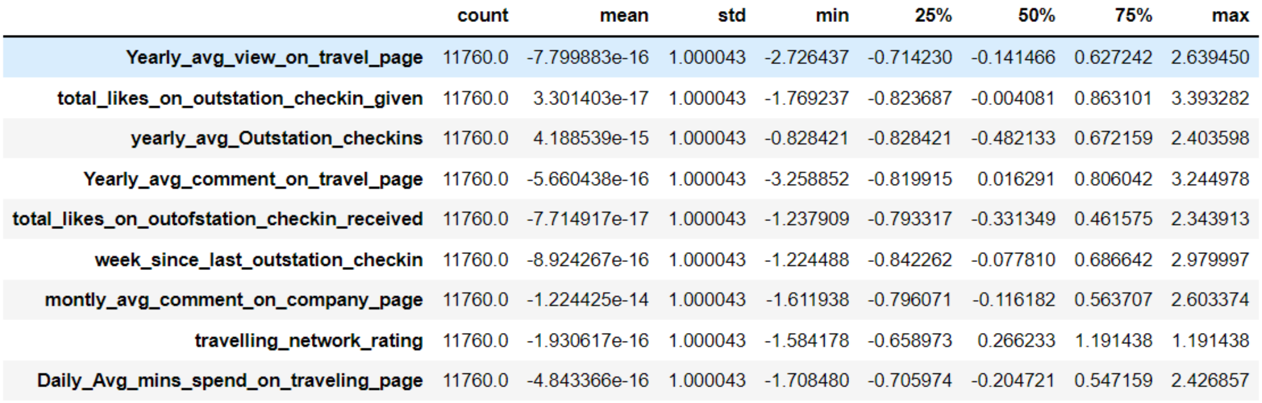
Before Scaling:



	count	mean	std	min	25%	50%	75%	max
Yearly_avg_view_on_travel_page	11760.0	280.385587	66.347859	99.5	233.00	271.0	322.00	455.50
total_likes_on_outstation_checkin_given	11760.0	28132.657058	13883.783884	3570.0	16697.25	28076.0	40115.25	75242.25
yearly_avg_Outstation_checkins	11760.0	8.176871	8.663686	1.0	1.00	4.0	14.00	29.00
Yearly_avg_comment_on_travel_page	11760.0	74.649320	21.526694	4.5	57.00	75.0	92.00	144.50
total_likes_on_outofstation_checkin_received	11760.0	6387.709439	4345.180379	1009.0	2940.75	4948.0	8393.25	16572.00
week_since_last_outstation_checkin	11760.0	3.203571	2.616365	0.0	1.00	3.0	5.00	11.00
montly_avg_comment_on_company_page	11760.0	22.854422	7.354454	11.0	17.00	22.0	27.00	42.00
travelling_network_rating	11760.0	2.712245	1.080887	1.0	2.00	3.0	4.00	4.00
Daily_Avg_mins_spend_on_traveling_page	11760.0	13.633673	7.980341	0.0	8.00	12.0	18.00	33.00

Table 11

After Scaling:



	count	mean	std	min	25%	50%	75%	max
Yearly_avg_view_on_travel_page	11760.0	-7.799883e-16	1.000043	-2.726437	-0.714230	-0.141466	0.627242	2.639450
total_likes_on_outstation_checkin_given	11760.0	3.301403e-17	1.000043	-1.769237	-0.823687	-0.004081	0.863101	3.393282
yearly_avg_Outstation_checkins	11760.0	4.188539e-15	1.000043	-0.828421	-0.828421	-0.482133	0.672159	2.403598
Yearly_avg_comment_on_travel_page	11760.0	-5.660438e-16	1.000043	-3.258852	-0.819915	0.016291	0.806042	3.244978
total_likes_on_outofstation_checkin_received	11760.0	-7.714917e-17	1.000043	-1.237909	-0.793317	-0.331349	0.461575	2.343913
week_since_last_outstation_checkin	11760.0	-8.924267e-16	1.000043	-1.224488	-0.842262	-0.077810	0.686642	2.979997
montly_avg_comment_on_company_page	11760.0	-1.224425e-14	1.000043	-1.611938	-0.796071	-0.116182	0.563707	2.603374
travelling_network_rating	11760.0	-1.930617e-16	1.000043	-1.584178	-0.658973	0.266233	1.191438	1.191438
Daily_Avg_mins_spend_on_traveling_page	11760.0	-4.843366e-16	1.000043	-1.708480	-0.705974	-0.204721	0.547159	2.426857

Table 12

4. Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.

1) Model Building and interpretation

Splitting the data set into 'Predictor' and 'Target' variables

Both data set for ‘Laptop’ and ‘Mobile’ devices are split into Predictor and Target variables (X and Y) respectively before further split into train and test data.

Laptop Dataset	**Mobile Dataset**
Size of Predictor variables (X_laptop): (1108, 15) Size of Target Variable (Y_laptop): (1108,)	Size of Predictor variables (X_laptop): (10652, 15) Size of Target variables (Y_laptop): (10652,)

Table 13

Scale the 'predictor' variables as few models are sensitive to scales of different variables
Scale the numerical data into common scale using **StandardScaler** from **Sklearn**.

Divide the data into Test and Train dataset

Let divide each dataset into Train and Test data in order to train the model and validate it for its performance on the unknown data. We shall keep 30% of the data as test data and the remaining 70% as train data.

Laptop Dataset	**Mobile Dataset**
Size of X_train for laptop: (775, 15) Size of X_test for laptop: (333, 15) Size of y_train for laptop: (775,) Size of y_test for laptop: (333,)	Size of X_train for mobile: (7456, 15) Size of X_test for mobile: (3196, 15) Size of y_train for mobile: (7456,) Size of y_test for mobile: (3196,)

Table 14

Choice of Models:

Our choice of models for the particular business problem, which is a classification problem, with the Target variable as binary class (“Yes” – 1 & “No” – 0):

Data frame was split into 2 dataset one for Mobile users and the other for Laptop users and models are built separately for each dataset.

Smote was used to balance the minority class in the target variable.

Since this is classification problem, models such as CART, Random Forest, Logistic Regression, LDA and KNN is used.

These models’ performance is considered better for classification problems.

Cart model shows good results in large datasets.

Random forest being an ensemble technique is expected to perform better.

Logistic regression is simple, fast, portable, and easy to train and perform significantly better if assumptions are met.

KNN is also suitable for classification as it is easy to implement and requires only two inputs i.e., K-value and distance function.

Choice of model Evaluation Metrics:

Precision is the parameter we should evaluate for this approach since the cost of not targeting the appropriate client is higher than the cost of targeting the wrong customer, who may not buy the product. As a result, the organization will lose potential clients, hence I'm contemplating Precision as a desired criteria for this project.

F1 and Recall were also considered as recall is computed as the ratio of Positive samples that were correctly categorized as Positive to the total number of Positive samples. The recall of the model assesses its ability to detect Positive samples. The more positive samples identified, the larger the recall.

Hyperparameter Tuning

Tune the hyperparameter which can prune the trees and avoid overfitting and reduce the complexity of the cost function in non-tree based models.

XGBoost Classifier:

⇒ Number of Estimators (Trees) – Number of Trees used for Ensemble, higher the number of trees gives more robust model and a consistent result.

⇒ Learning Rate – Amount of steps taken by the model function to achieve the minimal optimal point. Lower the number more accurate the results are but take up more resource and time.

⇒ Max Depth – How far the tree can grow, by default is it set to grown until no further possible split which tends to overfit the model. Too high number usually tends to overfit.

⇒ Min Child Weight – Minimum weight required in the node for further split. If weights are lesser than the value given tree stop growing further.

⇒ Subsample – How much of a sample to be considered for each iteration of the model.

Logistic Regression:

⇒ Regularisation – Strength of Regularisation higher the value lesser the effect of Penalty.

⇒ Penalty – Type of Regularisation used to calculate the best fit model by attending the global minima.

⇒ Solver function – Type of Loss function used to calculate the best fit model by attending the global minima

Decision Tree:

⇒ Max Depth – How far the tree can grow, by default is it set to grown until no further possible split which tends to overfit the model. Too high number usually tends to overfit.

⇒ Max Feature – No of features need to be considered for each split. It's good practice to start with half of the features available.

⇒ Min sample leaf – Minimum number of samples required to considered the node as leaf node. Higher the value lesser the model over fits by quickly attaining the leaf node.

⇒ Min sample Split – Minimum number of Samples required to further split the child node. Higher the value lesser the model over fits by pruning the tree.

Random Forest:

⇒ n estimators – Number of Trees used for Ensemble, higher the number of trees gives more robust model and a consistent result.

⇒ Max depth – How far the tree can grow, by default is it set to grown until no further possible split which tends to overfit the model. Too high number usually tends to overfit.

⇒ Max features – No of features need to be considered for each split. It's good practice to start with half of the features available.

⇒ Min sample leaf – Minimum number of samples required to considered the node as leaf node. Higher the value lesser the model over fits by quickly attaining the leaf node.

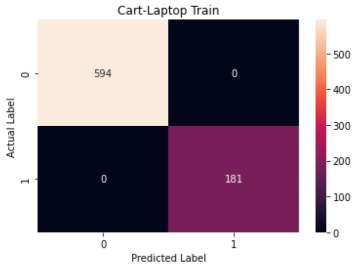
⇒ Min sample Split – Minimum number of Samples required to further split the child node. Higher the value lesser the model over fits by pruning the tree.

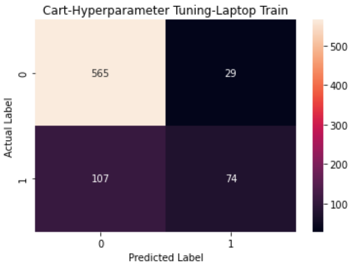
5. Model validation - How was the model validated? Just accuracy, or anything else too?

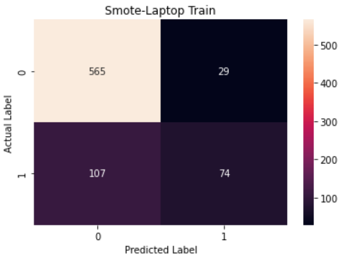
Performance metrics for Different models

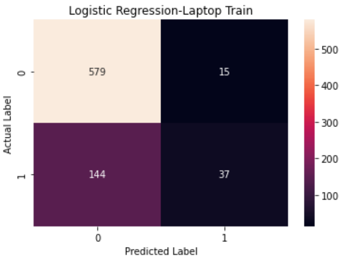
Confusion Matrix

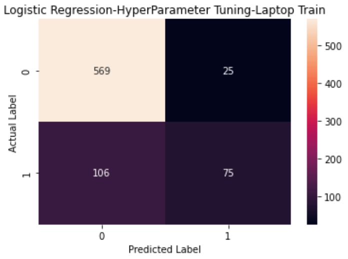
Laptop Train Data

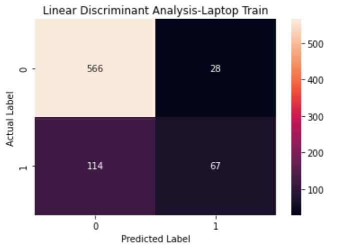


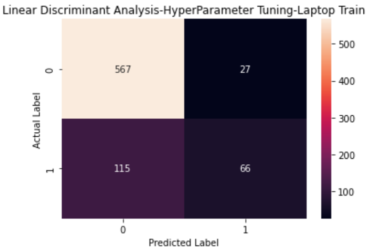


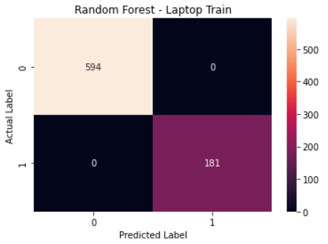


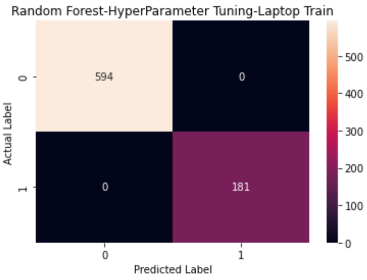


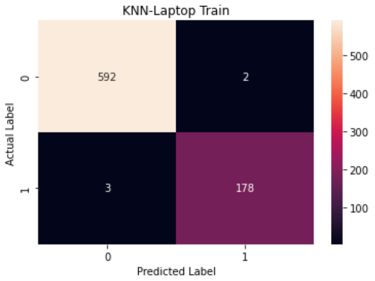


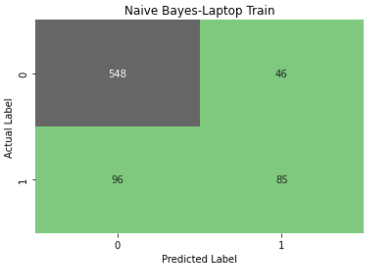


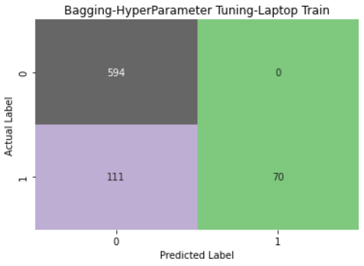


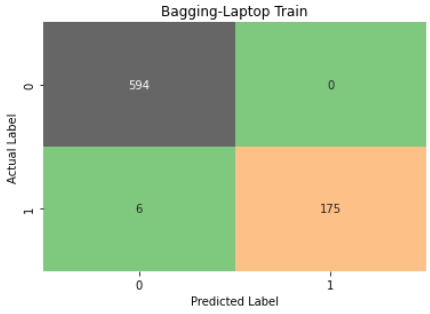


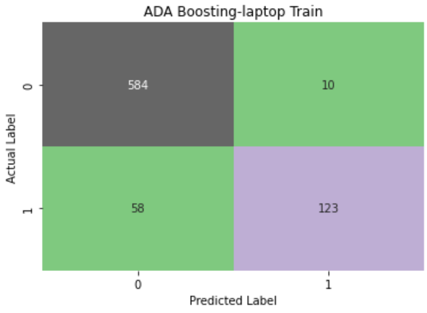


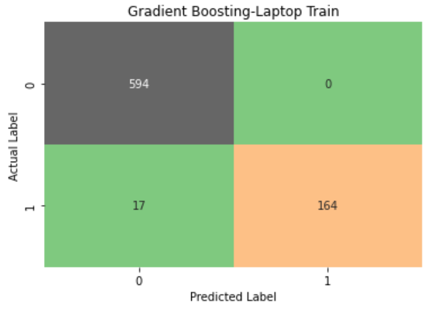












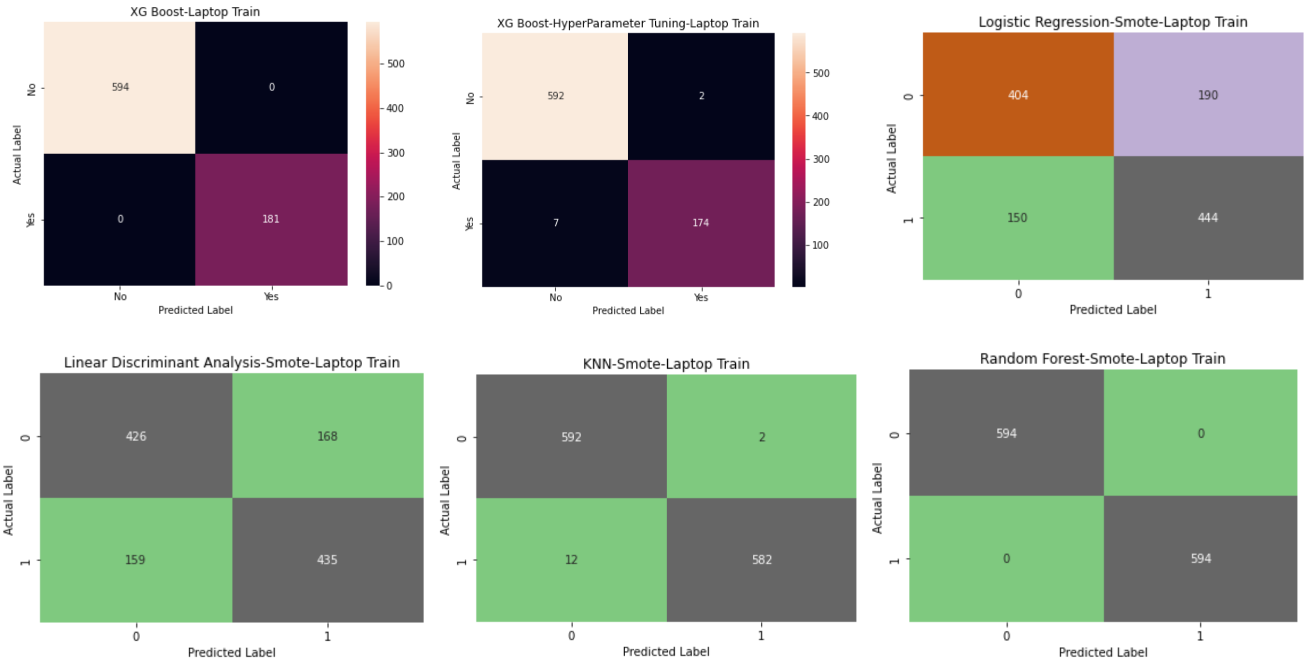
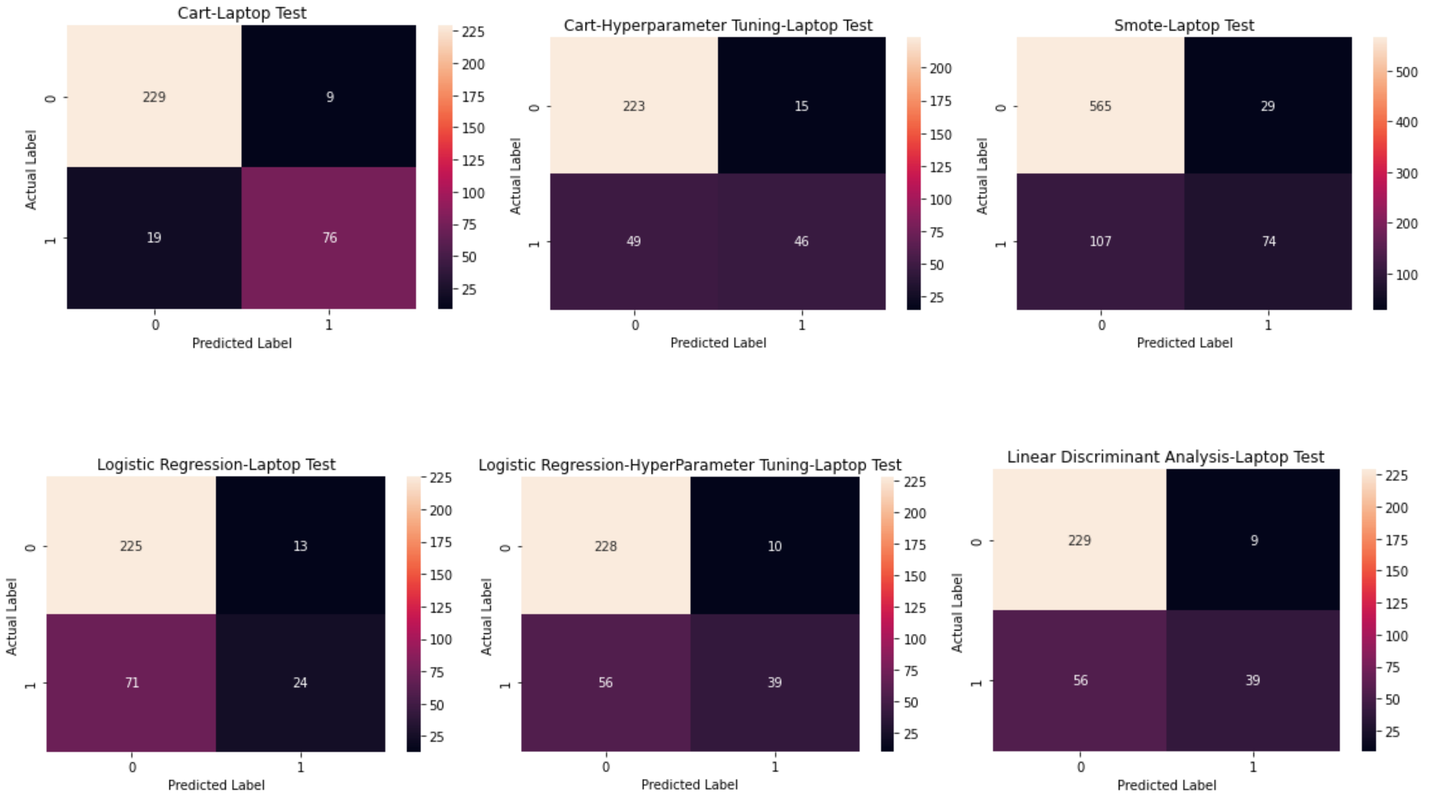
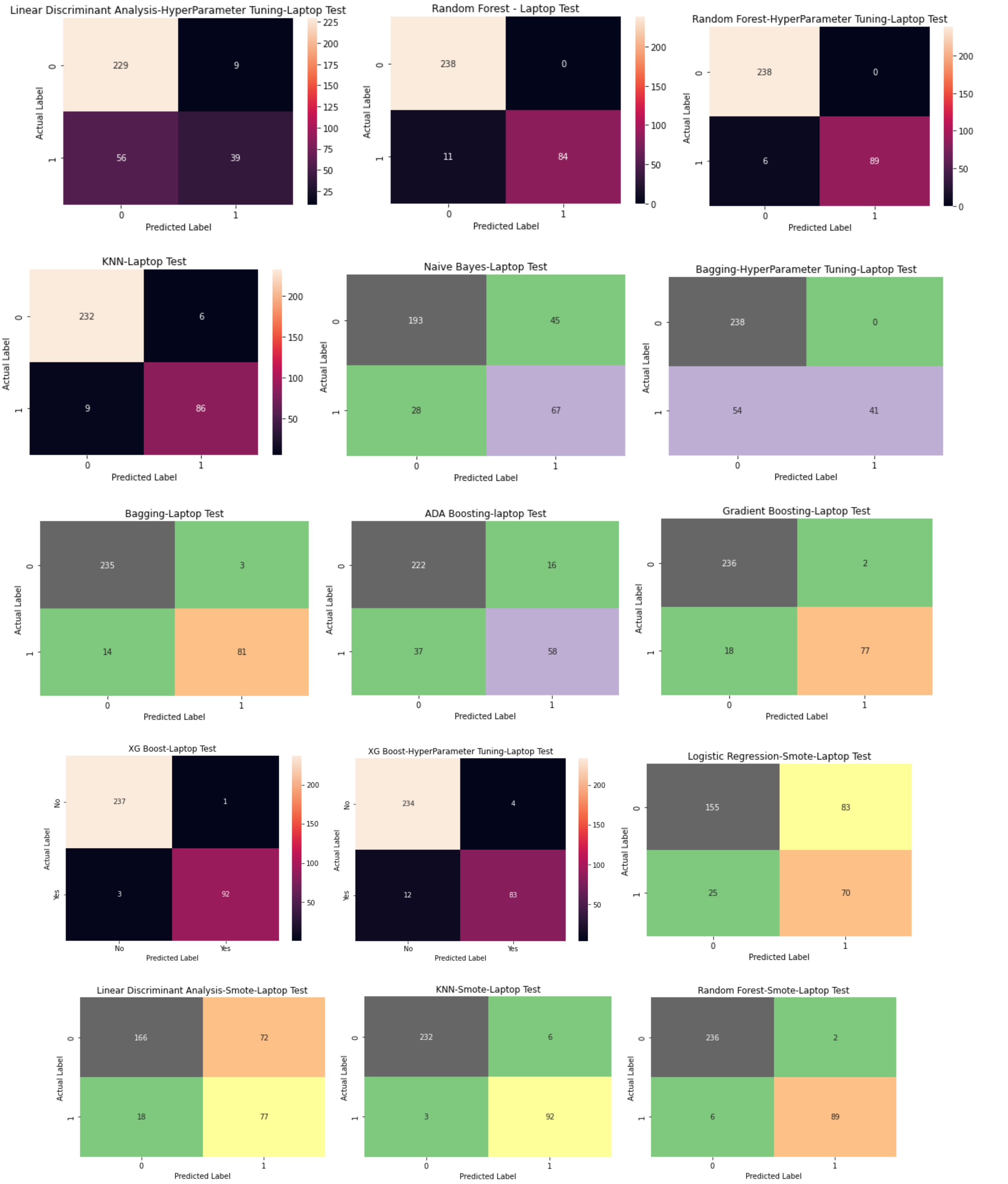


Table 15

Cart, Random Forest and XGBoost has the least number of False positives (***Type II error***) in the Positive Class (1).

Laptop Test Data





Linear Discriminant Analysis-HyperParameter Tuning-Laptop Test

Random Forest - Laptop Test

Random Forest-HyperParameter Tuning-Laptop Test

KNN-Laptop Test

Naive Bayes-Laptop Test

Bagging-HyperParameter Tuning-Laptop Test

Bagging-Laptop Test

ADA Boosting-laptop Test

Gradient Boosting-Laptop Test

XG Boost-Laptop Test

XG Boost-HyperParameter Tuning-Laptop Test

Logistic Regression-Smote-Laptop Test

Linear Discriminant Analysis-Smote-Laptop Test

KNN-Smote-Laptop Test

Random Forest-Smote-Laptop Test

Table 16

Random Forest and XGBoost has the least number of False positives (***Type II error***) in the Positive Class (1).

Classification Report:

Laptop – Training Data:

Models	Accuracy	Precision	Recall	F1_Score
Cart	1.00	1.00	1.00	1.00
Cart using pruning	0.82	0.72	0.41	0.52
Smote	0.82	0.72	0.41	0.52
Logistic Regression	0.79	0.71	0.20	0.32
Linear Discriminant Analysis	0.82	0.71	0.37	0.49
Random Forest	1.00	1.00	1.00	1.00
KNN	0.99	0.99	0.98	0.99
Smote – Logistic Regression	0.71	0.70	0.75	0.72
Smote - LDA	0.72	0.72	0.73	0.73
Smote – KNN	0.99	1.00	0.98	0.99
Smote - RF	1.00	1.00	1.00	1.00
Logistic Regression-GridSearch CV	0.83	0.75	0.41	0.53
LDA- GridSearch CV	0.82	0.71	0.36	0.48
Random Forest-GridSearch CV	1.00	1.00	1.00	1.00
Naïve Bayes	0.82	0.65	0.47	0.54
Bagging	0.99	1.00	0.97	0.98
Bagging-GridSearch CV	0.86	1.00	0.39	0.56
ADA Boosting	0.91	0.92	0.68	0.78
Gradient Boosting	0.98	1.00	0.91	0.95
XG Boost	1.00	1.00	1.00	1.00
XG Boost Hyperparameter Tune	0.99	0.99	0.96	0.97

Table 17

Laptop – Testing Data:

Models	Accuracy	Precision	Recall	F1_Score
Cart	0.92	0.89	0.80	0.84
Cart using pruning	0.81	0.75	0.48	0.59
Smote	0.81	0.75	0.48	0.59
Logistic Regression	0.75	0.65	0.25	0.36
Linear Discriminant Analysis	0.80	0.81	0.41	0.55
Random Forest	0.97	1.00	0.88	0.94
KNN	0.95	0.93	0.91	0.92
Smote – Logistic Regression	0.68	0.46	0.74	0.56
Smote - LDA	0.73	0.52	0.81	0.63
Smote – KNN	0.97	0.94	0.97	0.95
Smote - RF	0.98	0.98	0.94	0.96
Logistic Regression-GridSearch CV	0.80	0.80	0.41	0.54
LDA- GridSearch CV	0.80	0.81	0.41	0.55
Random Forest-GridSearch CV	0.98	1.00	0.94	0.97
Naïve Bayes	0.78	0.60	0.71	0.65
Bagging	0.95	0.96	0.85	0.91
Bagging-GridSearch CV	0.84	1.00	0.43	0.60
ADA Boosting	0.84	0.78	0.61	0.69
Gradient Boosting	0.94	0.97	0.81	0.89
XG Boost	0.99	0.99	0.97	0.98
XG Boost Hyperparameter Tune	0.95	0.95	0.87	0.91

Table 18

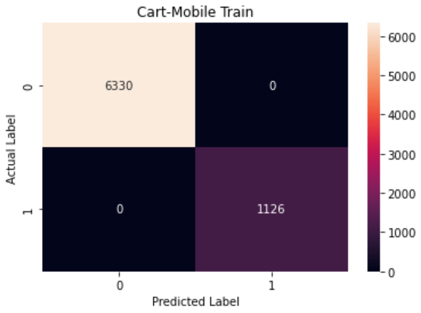
- Logistic regression is the least performing model followed by Linear Discriminant Analysis.
- ***Cart, Random Forest and XG Boost*** are ***overfitting*** which achieve perfect accuracy of 100% in Train data thus the model is not generalised enough.
- Hence, we shall tune the hyperparameters for each model to achieve an optimal performance on both Train and Test data.
- After fine tuning models to avoid overfitting, we can see that the ***XGBoost-Hyper Parameter Tune*** performance better than other models in terms of evaluation metrics. Also, it has consistent performance between train and test data. Thus, we will choose this as our model of choice.

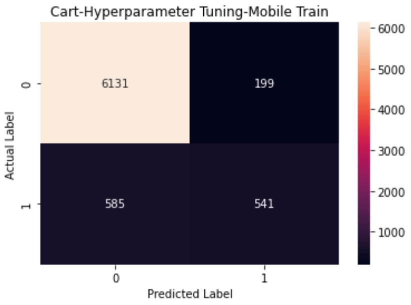
Models for Mobile:

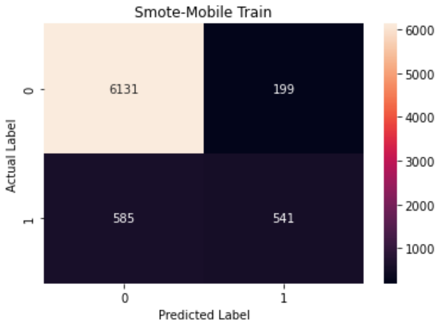
Performance metrics for models

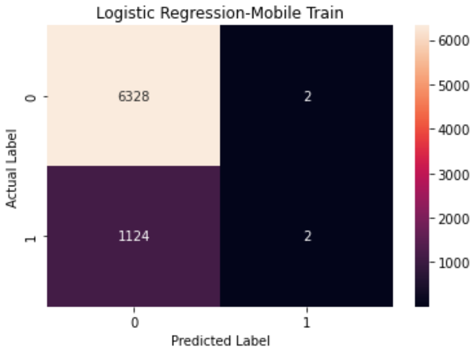
Confusion Matrix

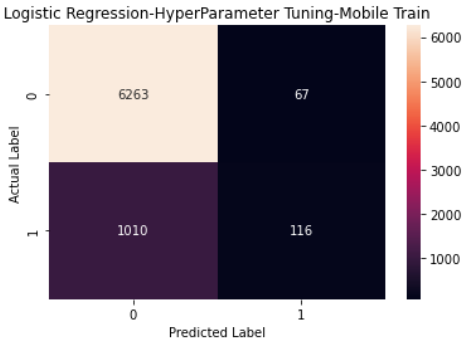
Mobile Train Data

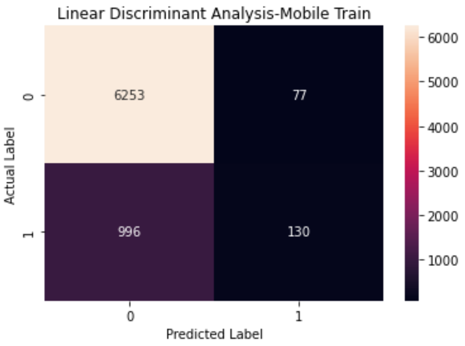

Cart-Mobile Train

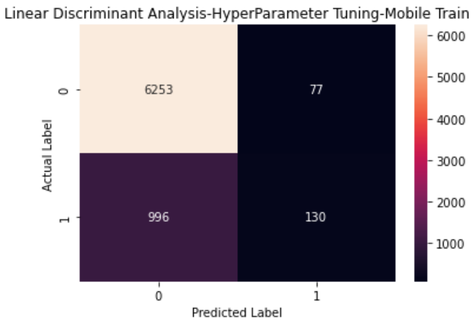

Cart-Hyperparameter Tuning-Mobile Train

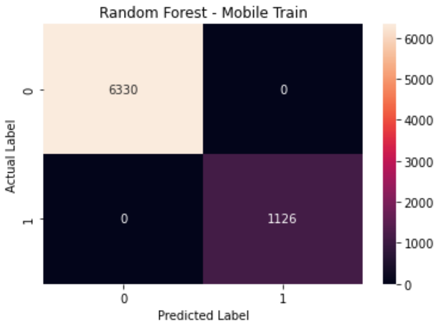

Smote-Mobile Train

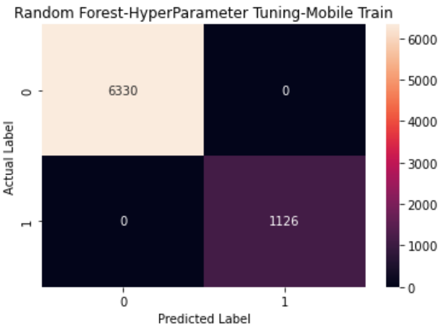

Logistic Regression-Mobile Train


Logistic Regression-HyperParameter Tuning-Mobile Train


Linear Discriminant Analysis-Mobile Train


Linear Discriminant Analysis-HyperParameter Tuning-Mobile Train


Random Forest - Mobile Train


Random Forest-HyperParameter Tuning-Mobile Train

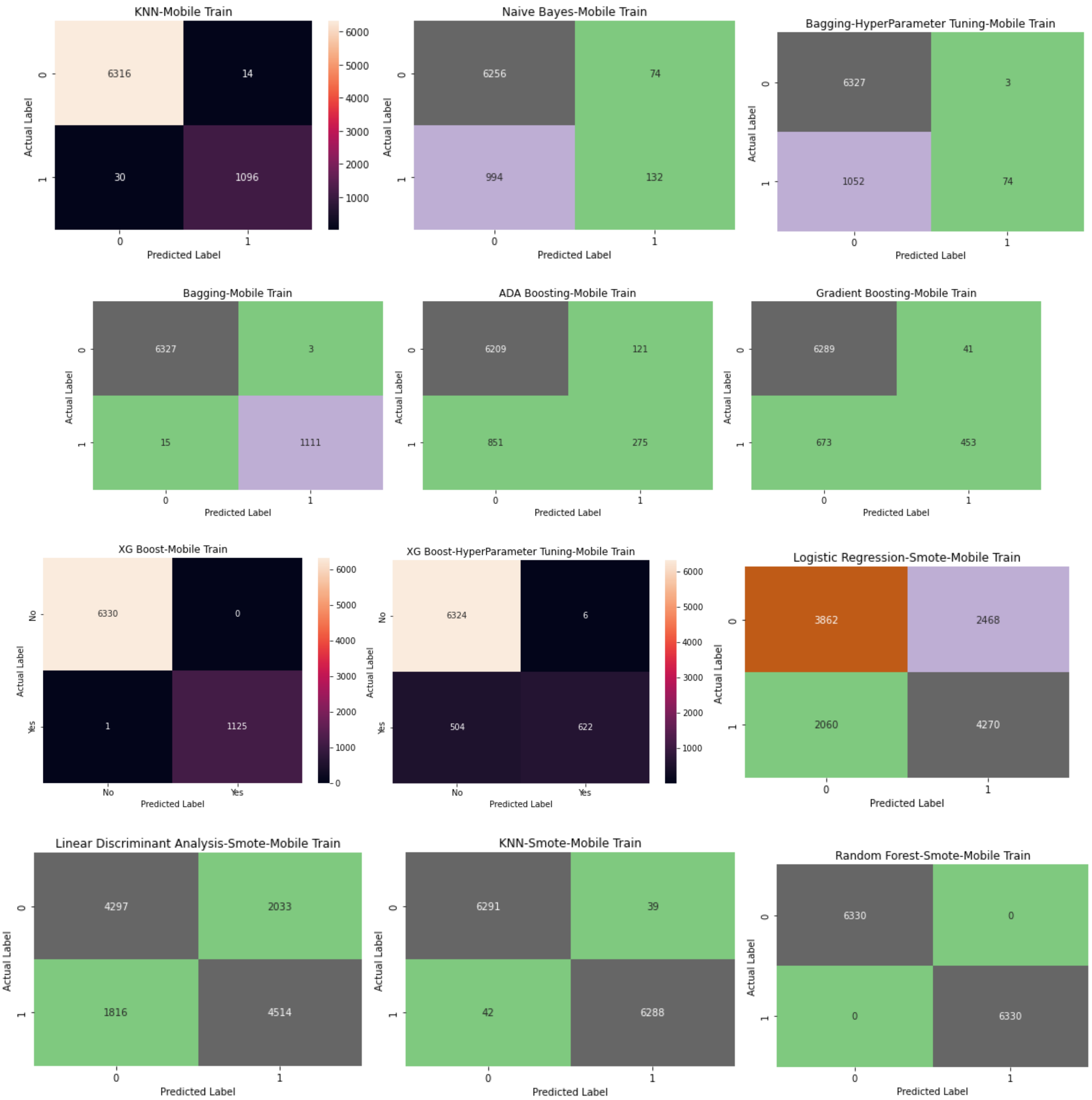
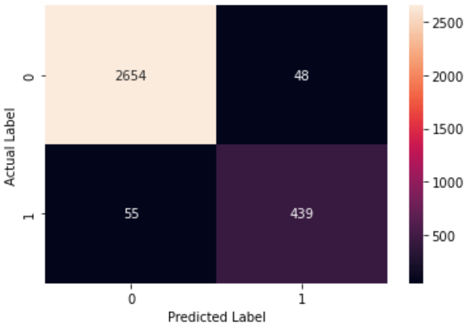


Table 19

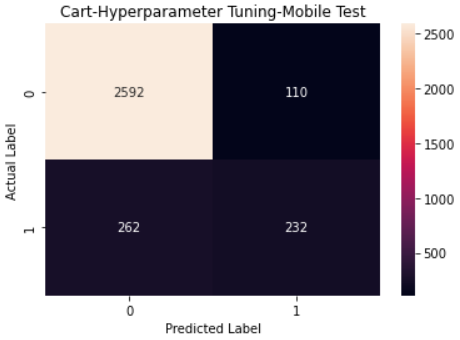
Random Forest and XGBoost has the **least number of False positives (Type II error)** in the Positive Class (1)

Mobile Test Data

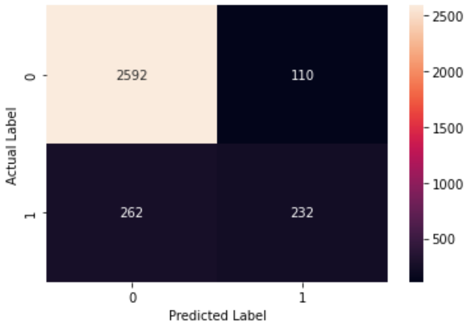
Cart-Mobile Test



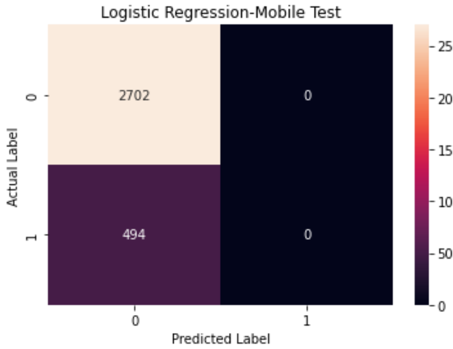
Cart-Hyperparameter Tuning-Mobile Test



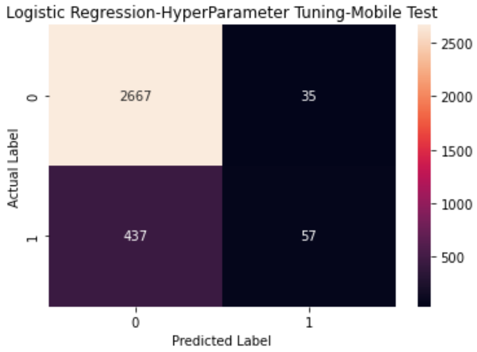
Smote-Mobile Test



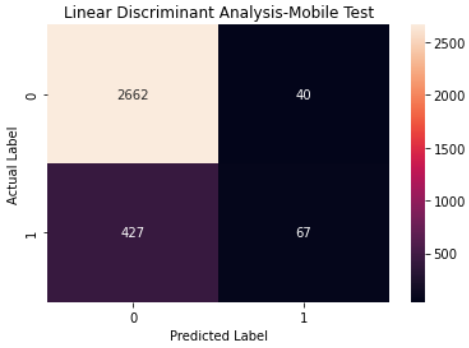
Logistic Regression-Mobile Test



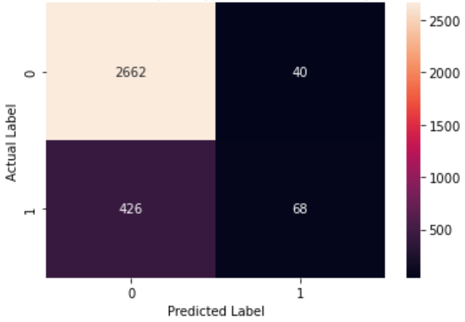
Logistic Regression-HyperParameter Tuning-Mobile Test



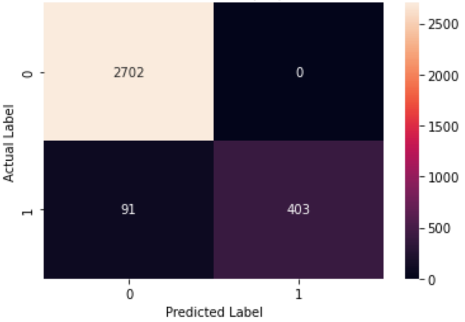
Linear Discriminant Analysis-Mobile Test



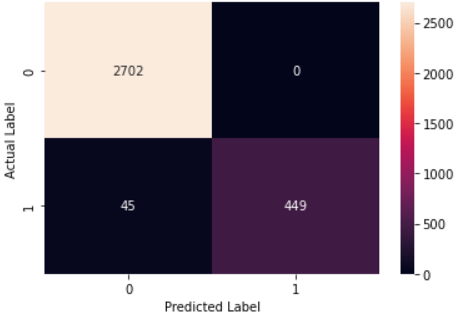
Linear Discriminant Analysis-HyperParameter Tuning-Mobile Test



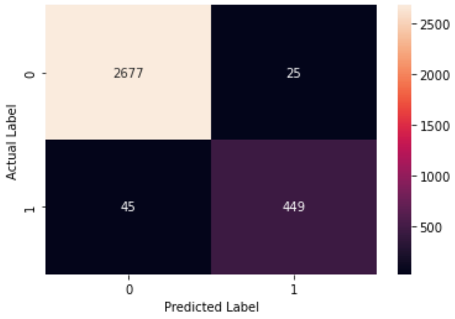
Random Forest - Laptop Test



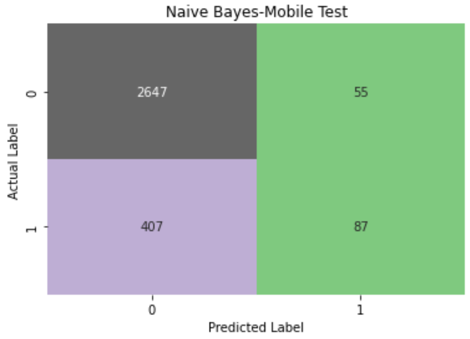
Random Forest-HyperParameter Tuning-Mobile Test



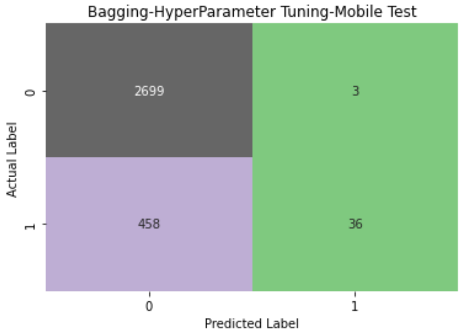
KNN-Mobile Test



Naive Bayes-Mobile Test



Bagging-HyperParameter Tuning-Mobile Test



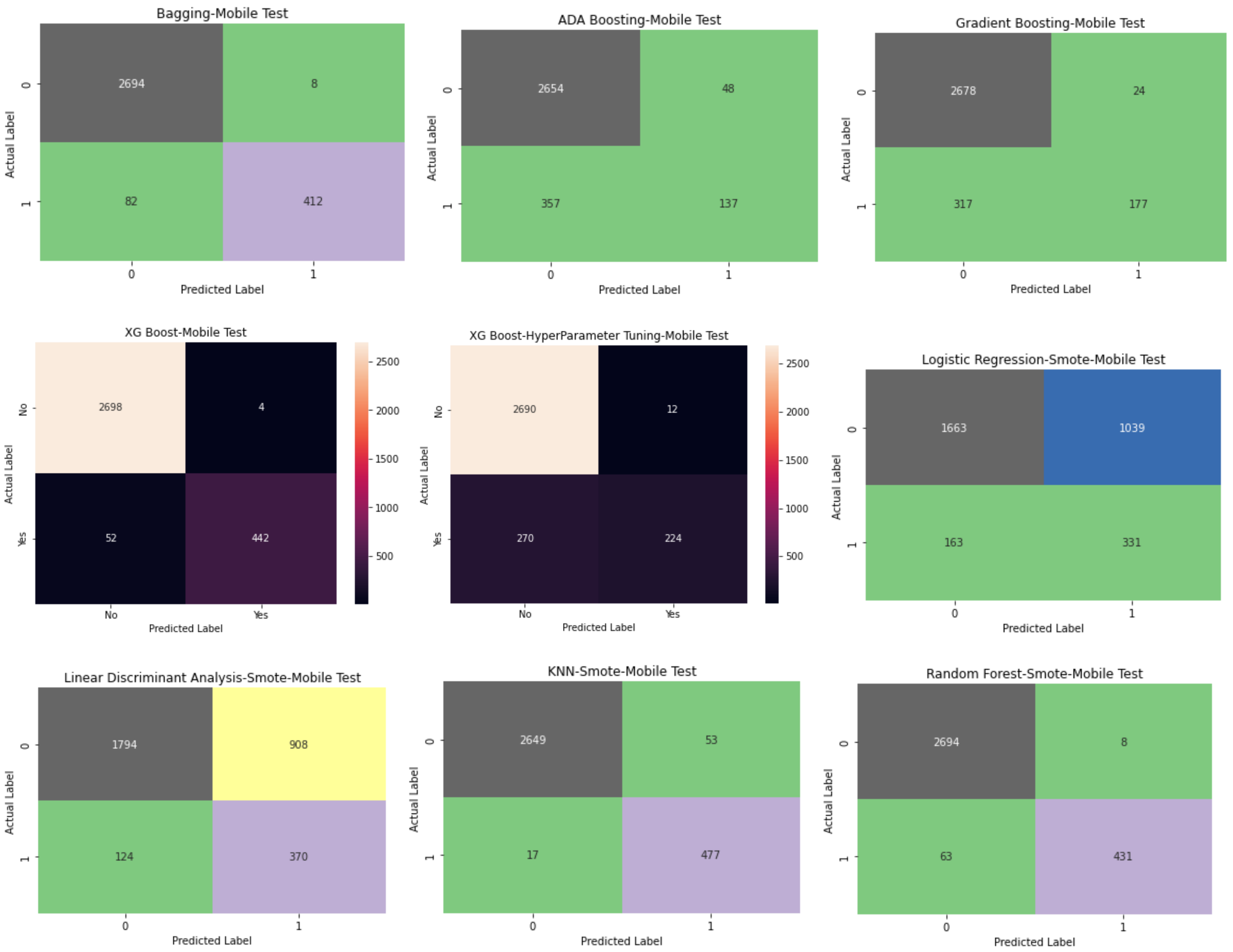


Table 20

Random Forest and XGBoost has the **least number of False positives (Type II error)** in the Positive Class (1)

Classification Report:

Mobile – Training Data:

Models	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Cart	1.00	1.00	1.00	1.00
Cart using pruning	0.89	0.73	0.48	0.58
Smote	0.89	0.73	0.48	0.58
Logistic Regression	0.85	0.50	0.00	0.00
Linear Discriminant Analysis	0.86	0.63	0.12	0.20
Random Forest	1.00	1.00	1.00	1.00
KNN	0.99	0.99	0.97	0.98
Smote – Logistic Regression	0.64	0.63	0.67	0.65
Smote - LDA	0.70	0.69	0.71	0.70
Smote – KNN	0.99	0.99	0.99	0.99

Models	Accuracy	Precision	Recall	F1_Score
Smote - RF	1.00	1.00	1.00	1.00
Logistic Regression-GridSearch CV	0.86	0.63	0.10	0.18
LDA- GridSearch CV	0.86	0.63	0.12	0.20
Random Forest-GridSearch CV	1.00	1.00	1.00	1.00
Naïve Bayes	0.86	0.64	0.12	0.20
Bagging	1.00	1.00	0.99	0.99
Bagging-GridSearch CV	0.86	0.96	0.07	0.12
ADA Boosting	0.87	0.69	0.24	0.36
Gradient Boosting	0.90	0.92	0.40	0.56
XG Boost	1.00	1.00	1.00	1.00
XG Boost Hyperparameter Tune	0.93	0.99	0.55	0.71

Table 21

Mobile – Testing Data:

Models	Accuracy	Precision	Recall	F1_Score
Cart	0.97	0.90	0.89	0.90
Cart using pruning	0.88	0.68	0.47	0.56
Smote	0.88	0.68	0.47	0.56
Logistic Regression	0.85	0.00	0.00	0.00
Linear Discriminant Analysis	0.85	0.63	0.14	0.22
Random Forest	0.97	1.00	0.82	0.90
KNN	0.98	0.95	0.91	0.93
Smote – Logistic Regression	0.62	0.24	0.67	0.36
Smote - LDA	0.68	0.29	0.75	0.42
Smote – KNN	0.98	0.90	0.97	0.93
Smote - RF	0.98	0.98	0.87	0.92
Logistic Regression-GridSearch CV	0.85	0.62	0.12	0.19
LDA- GridSearch CV	0.85	0.63	0.14	0.23
Random Forest-GridSearch CV	0.99	1.00	0.91	0.95
Naïve Bayes	0.86	0.61	0.18	0.27
Bagging	0.97	0.98	0.83	0.90
Bagging-GridSearch CV	0.86	0.92	0.07	0.14
ADA Boosting	0.87	0.74	0.28	0.40
Gradient Boosting	0.89	0.88	0.36	0.51
XG Boost	0.98	0.99	0.89	0.94
XG Boost Hyperparameter Tune	0.91	0.95	0.45	0.61

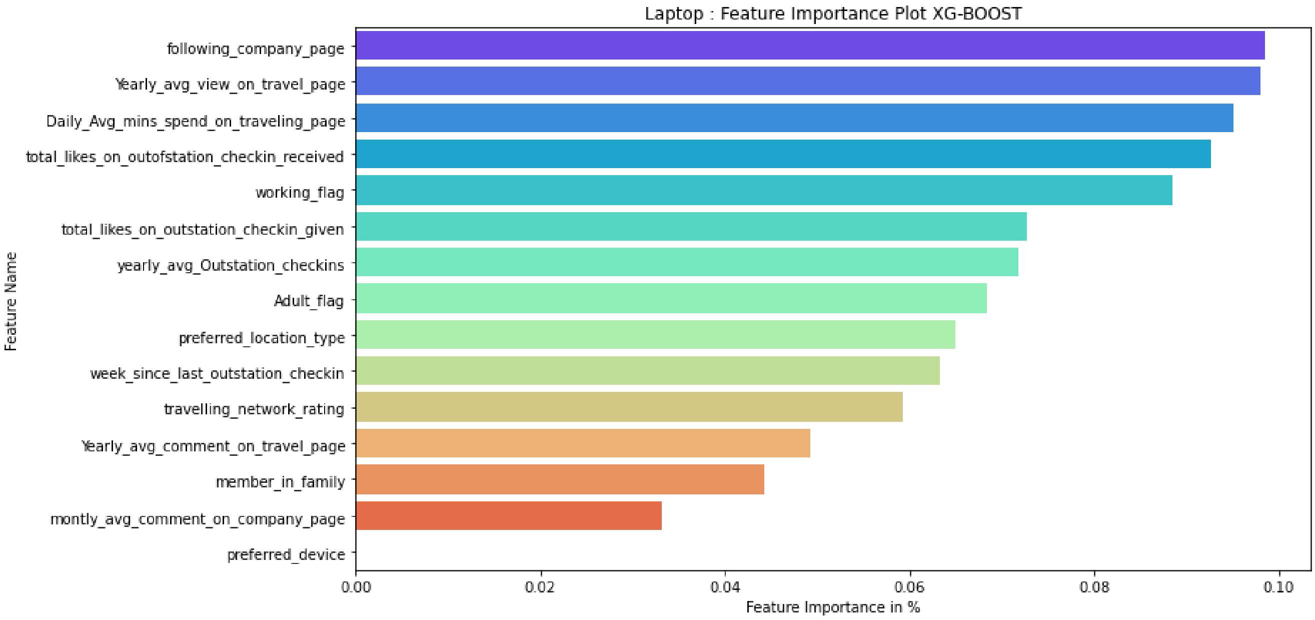
Table 22

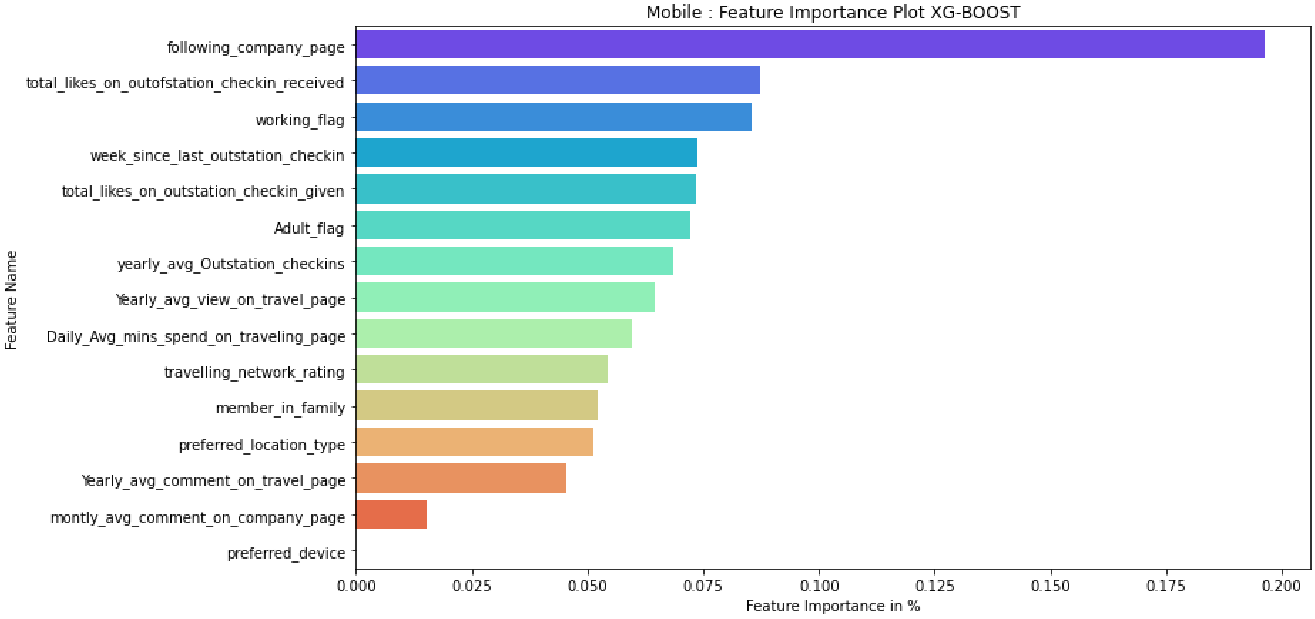
- Logistic regression is the least performing model followed by Linear Discriminant Analysis.
- ***Cart, Random Forest, Bagging and XG Boost*** are ***overfitting*** which achieve perfect accuracy of 100% in Train data thus the model is not generalised enough.
- Hence, we shall tune the hyperparameters for each model to achieve an optimal performance on both Train and Test data.
- After fine tuning model to avoid overfitting, we can see that the ***XGBoost-Hyper Parameter Tune*** performance better than other model in terms of evaluation metrics. Also, it has consistent performance between train and test data. Thus, we will choose this our model of choice

6. Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client.

Analysis:

Feature Importance





Graph 12

- Based on the above feature importance chart, tour plan can be suggested to the users based on the variables identified as important using feature importance.

- The ***following_company_page*** come out to be the first in priority. This means that the customers following company page are much interested in tours / travels and likely to purchase tour package. This can be used to come up with a tour package for those users.

- The ***total_likes_on_outstation_checkin_given*** also suggest customers interest in tour / travels. This means that the user uses social media platforms highly and their interest is on tours. This can be used to come up with a tour package for those users.

- Similarly, ***total_likes_on_outofstation_checkin_received*** variable shows that the users are much interested in travel.

- The variables such as the ***Yearly_avg_view_on_travel_page*** and ***Yearly_avg_comment_on_travel_page*** will give us more insights on what the users like and based on that a special tour package can be introduced.

- The variable ***week_since_last_outstation_checkin*** will also be very useful to assess when these users will go on their next trip.

Recommendations:

- The analysis on the dataset has given the company an understanding how the users behave on social media platform's and if the organization focuses on the feature variables it can turn the revenue game in their favour.

- As digital advertising is costly, the company can take decision on which device it wants to deploy their best advertisement and, in this case, it should be hand held or mobile device which happens to the favourite to browse the net amongst the users.

- The company can make tailored packages for fun destinations like beach for instance and include a clause if taken in a group more than 3 individuals then the user will get a discount.

- The company can include complimentary services for travellers travelling for the purpose of finance or business that could be included with their stay if they opt for the package offered to them.

- Special discounts could be given to travellers travelling for medical purpose. By doing so, they can gain the trust of the customers and in return they can have a loyal customer following.

- For the customers to stay on their page and follow the page the company could use and introduce interactive games which offers the customer discounts or complimentary services or a free stay with an affiliated hotel and so on if they participate and win the games.

- They could ask the customers to write or comment on the company page and get points that can be redeemed in the future to get travel discounts.

- Coupons could be provided to repeat customers for their family and friends for their next travel plan and in return the company can have a new customer base, thus increasing the footfall and revenue.

- The total likes and comments given by the users indicate they are active users in the media platform. The highly active customers should be given high priority as there is a strong chance that they will take the product.

- It is also important to target the customer who are wrongly predicted as converted to improve the sale. But this shall be given second priority only when there is still a budget to spend.

- When the user is from Laptop device, who are lesser popular, good Travel Networking Rating, prefers Hill Stations and also Working has very high propensity to buy ticket for their next trip. Thus, those user clusters can be profiled and Target for digital Marketing.

- When the user is from Mobile device, who are lesser popular, working, Adult and prefers Entertainment has very high propensity to buy ticket for their next trip. Thus, those user clusters can be profiled and Target for digital Marketing.

- We can come up with a travel combo for the users who are the frequent travellers.

- We could request a testimony from the loyal customers to promote the brand.

- Targeted advertisements to a specific destination can be sent to the users based on their interest.

- There can be discounts offered to the customers who did not take the product to attract them.