DIGITAL MARKETING USING SOCIAL NETWORK DATA

PGP-DSBA Online November'21
30 Oct 2022

Abstract

Profiling users for targeted digital marketing based on their social media behavior

Venkadasubramanian Jayakumar
Venkadasubramanian_j@outlook.com

Table of Contents

List of Figures

List of Tables

1. Problem Understanding:

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.
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.

Defining problem statement

Create 2 different models for each type of device customer use to access social networking platforms to learn the customer behaviour on travel related pages and their recent travel check-ins and also their influence on others. Model should be created for two type of device, one is for Laptop another one is for Mobile (anything which is not laptop shall be considered as mobile device)

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.

Understanding business/social opportunity

- **Business Opportunity:**

  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

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:**

  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.

Variable Description

Variable Name	Type	Description
UserID :	int64	Unique ID of user
Buy_ticket :	object	Buy ticket in next month
Yearly_avg_view_on_travel_page :	float64	Average yearly views on any travel related page by user
preferred_device :	object	Through which device user preferred to do login
total_likes_on_outstation_ checkin_given :	float64	Total number of likes given by a user on out of station checkings in last year
yearly_avg_Outstation_checkins :	object	Average number of out of station check-in done by user
member_in_family :	object	Total number of relationship mentioned by user in the account
preferred_location_type :	object	Preferred type of the location for travelling of user
Yearly_avg_comment_on_ travel_page :	float64	Average yearly comments on any travel related page by user
total_likes_on_outofstation_ checkin_received	int64	Total number of likes received by a user on out of station checkings in last year
week_since_last_outstation_checkin :	int64	Number of weeks since last out of station check-in update by user
following_company_page :	object	Weather the customer is following company page (Yes or No)
montly_avg_comment_on_ company_page :	int64	Average monthly comments on company page by user
working_flag :	object	Weather the customer is working or not
travelling_network_rating :	int64	Does user have close friends who also like travelling. 1 is highs and 4 is lowest
Adult_flag :	int64	Weather the customer is adult or not
Daily_Avg_mins_spend_on_ traveling_page :	int64	Average time spend on the company page by user on daily basis

Table 1 Variable Information

2. Data Report

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.

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

UserID	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	member_in_family	preferred_location_type	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin	following_company_page	montly_avg_comment_on_company_page	working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
1000001	Yes	307	iOS and Android	38570	1	2	Financial	94	5993	8	Yes	11	No	1	0	8
1000002	No	367	iOS	9765	1	1	Financial	61	5130	1	No	23	Yes	4	1	10
1000003	Yes	277	iOS and Android	48055	1	2	Other	92	2090	6	Yes	15	No	2	0	7
1000004	No	247	iOS	48720	1	4	Financial	56	2909	1	Yes	11	No	3	0	8
1000005	No	202	iOS and Android	20685	1	1	Medical	40	3468	9	No	12	No	4	1	6

Table 2 Sample of Given Dataset

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

Missing Values:
Number of Rows with at least one missing values: 1304
Percentage of Rows with at least 1 missing values: 11.09%
Total Missing value proportion in the given data 0.76%

- 5 columns has missing values. However missing values are not more than 5% of each column observations.
- Except 'following_company_page' variable all other variables are of continuous data type.
- 1304 rows has atleast 1 missing value, which is 11% of the total observations.
- And overall there are <1% of the missing values in the dataset.
- 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.
- Converting 'Adult_flag', 'week_since_last_outstation_checkin' and 'travelling_network_rating' into 'object' datatype as it is a categorical variable.

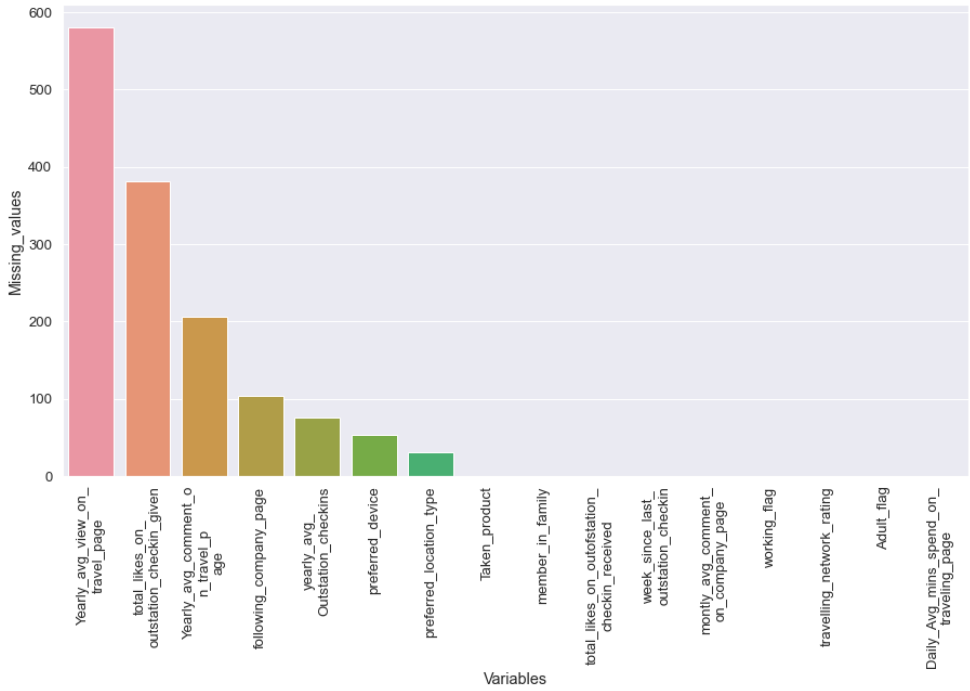


Figure 1 Number of Missing Values

Summary Statistics

Continuous Numerical Variable

	count	mean	std	min	25%	50%	75%	max
Yearly_avg_view_on_travel_page	11179	280.83	68.183	35	232	271	324	464
total_likes_on_outstation_checkin_given	11379	28170	14385	3570	16380	28076	40525	252430
yearly_avg_Outstation_checkins	11685	8.22	8.67	1	1	4	14	29
Yearly_avg_comment_on_travel_page	11554	74.79	24.02	3	57	75	92	815
total_likes_on_ _checkin_received	11760	6531	4706	1009	2940	4948	8393	20065
montly_avg_comment_on_company_page	11760	28.662	48.66	11	17	22	27	500
Daily_Avg_mins_spend_on_traveling_page	11760	13.817	9.07	0	8	12	18	270

Table 3 Summary Statistics - Numerical Variables

- 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 convereted as categorical column.

Categorical Variable

	count	unique	top	freq
Taken_product	11760	2	No	9864
preferred_device	11707	2	Mobile	10599

member_in_family	11760	6	3	4576
preferred_location_type	11729	14	Beach	2424
week_since_last_outstation_checkin	11760	12	1	3070
following_company_page	11657	2	No	8360
working_flag	11760	2	No	9952
travelling_network_rating	11760	4	3	3672
Adult_flag	11760	2	1	6712

Table 4 Summary Statistics - Categorical Variables

- There are more number of customer 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.

Understanding of attributes (variable info, renaming if required)

Value count of "Taken_product"

No	9864
Yes	1896

Value count of "preferred_device"

Tab	4172
iOS and Android	4134
Laptop	1108
iOS	1095
Mobile	600
Android	315
Android OS	145
ANDROID	134
Other	2
Others	2

Value count of "yearly_avg_Outstation_checkins"

1	4543
2	844
10	682
9	340
7	336
3	336
8	320
5	261
4	256
16	255

Value count of "member_in_family"

3	4561
4	3184
2	2256
1	1349
5	384
Three	15
10	11

Value count of "preferred_location_type"

Beach	2424
Financial	2409
Historical site	1856
Medical	1845
Other	643
Big Cities	636
Social media	633
Trekking	528
Entertainment	516
Hill Stations	108
Tour Travel	60
Tour and Travel	47
Game	12
OTT	7
Movie	5

Value count of "week_since_last_outstation_checkin"

6	236
11	229
24	223
29	215
23	215
18	208
15	206
26	199
20	199
25	198
28	180
19	176
14	167
17	160
12	159
22	152
13	150
21	143
27	96
*	1

Value count of "working_flag"

No	9952
Yes	1808

Value count of "Adult_flag"

0	5048
1	4768
2	1264
3	680

1	3070
3	1766
2	1700
4	1118
0	1032
5	728
6	654
7	594
9	472
8	428
10	138
11	60

Value count of "following_company_page"

No	8355
Yes	3285
1	12
0	5

Value count of "travelling_network_rating"

3	3672
4	3456
2	2424
1	2208

- There are few redenduncies and improper data present in the few columns.
- **preferred_device** : Has three different 'Androids', two different 'others' present.
- **yearly_avg_Outstation_checkins**: There is a Wildcharacter (*)
- **member_in_family**: Has 'three' as string.
- **following_company_page**: Has both 'yes / no' and 0/1

Data Wrangling

- **Convert the Preferred device categories into two, only with Laptop and anything other than Laptop shall be converted as Mobile devices**

Laptop	1108
Tab	4172
iOS and Android	4134
iOS	1095
Mobile	600
Android	315
Android OS	145
ANDROID	134
Other	2
Others	2

Mobile

- Mobile 10599
- Laptop 1108

- 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 and Travel' and 'Tour and Travel into one category'.
- Replace '1' to 'No' and '0' to 'Yes' based on its proportion.
- 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.

3. Exploratory Data Analysis

Number of Numerical columns: 7
Number of Categorical Columns: 9

Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

Distribution of Numerical variables:

- ***Yearly average views on Travel pages*** is normally distributed with most number of views are around 250 to 300 views.
- Other than that all the variables are right skewed with outliers.
- ***total_likes_on_outstation_checkins_given***, ***yearly_avg_comment_on_travel_page***, ***montly_avg_comment_on_company_page*** and ***Daily_avg_mins_spend_on_traveling_page*** has very few extreme values at the upper band thus all are right skewed.
- Majority user has the ***yearly average outstation check-ins*** below 5.
- Likewise, ***monthly average comments on company pages*** is also less than 50 for most of the users.
- ***Likes given on outstation check*** in has even distribution of users across different bucket of likes.
- ***yearly_avg_view_on_travel_page*** has very few outliers and normally distributed.
- ***total_likes_on_outstation_checkin_given*** also evenly distributed with only very few outliers at the upper limit which are significantly higher than the other values.
- ***yearly_avg_comment_on_travel_page*** is also has only few outliers which can be treated.
- monthly_avg_comment_on_company_page and Daily_Avg_mins_spend_on_travelling_page has whole group of outliers at the upper range.

Distribution of Categorical Variables:

- User's most preferred device is mobile devices (Tab, iOS, Android)
- Most of the users have a medium size family with 3 members.
- User's most preferred location type is Financial, which means that many may went for a business trip, which then followed by other (undisclosed) and Medical & Game. The least preferred location type would be hill station.

- Many of the users have very recent outstation check-in, within in 1 week or less. Indicates that many of the given users had a recent travel.
- Given data has more number of users who are not following the company's social medial page.
- Surprisingly we have larger number of users with no-work flag, which needs to be further analysed with preferred location type, adult flag and Product taken to understand better who are out targeted customers.
- Majority of the users with travelling network rating as 3 and 4 which indicates that many of the users doesn't have network of travellers.

Numerical Variable Distribution

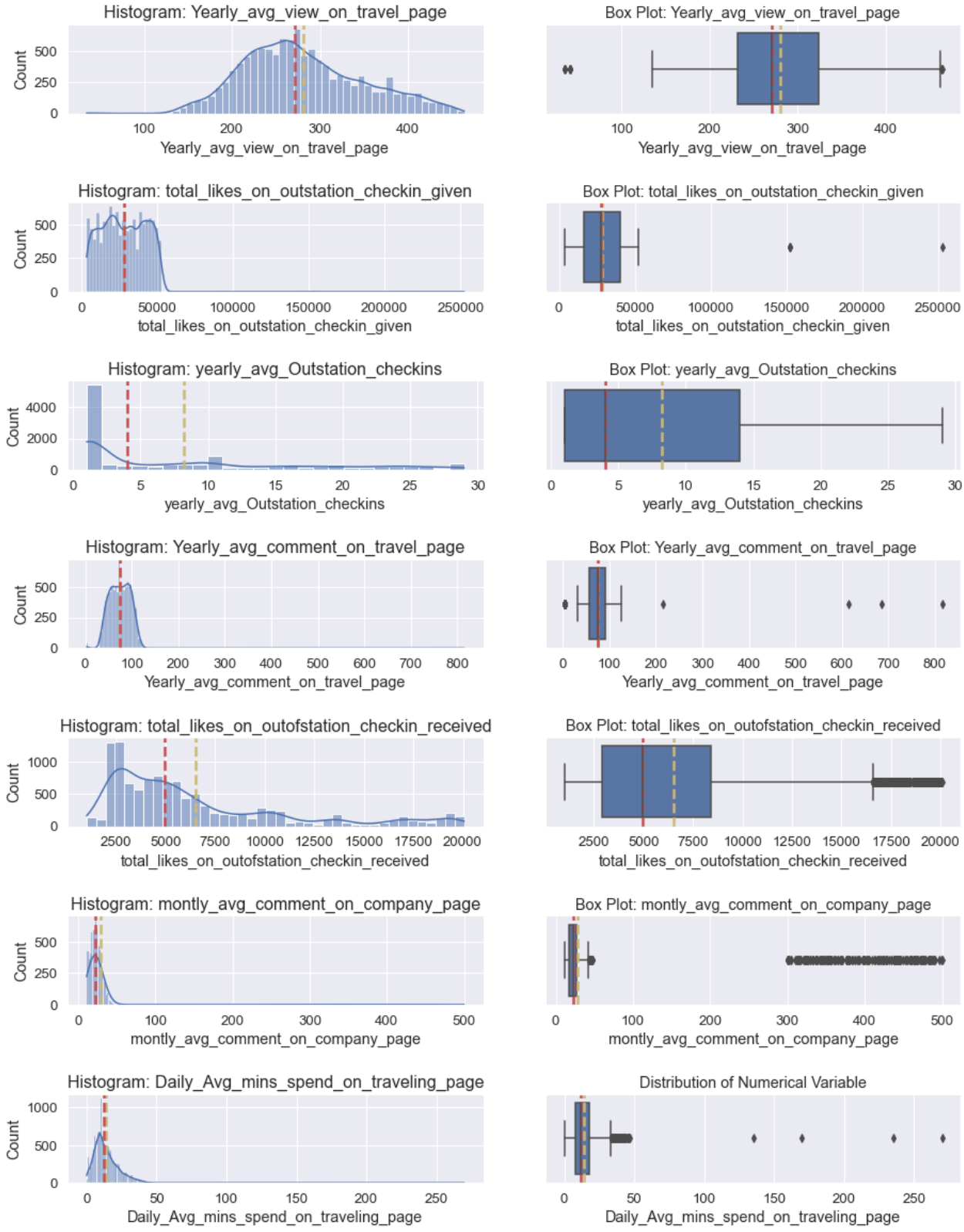


Figure 2 Distribution of Numerical Variables - Hist and Boxplot

Categorical Variable Distribution

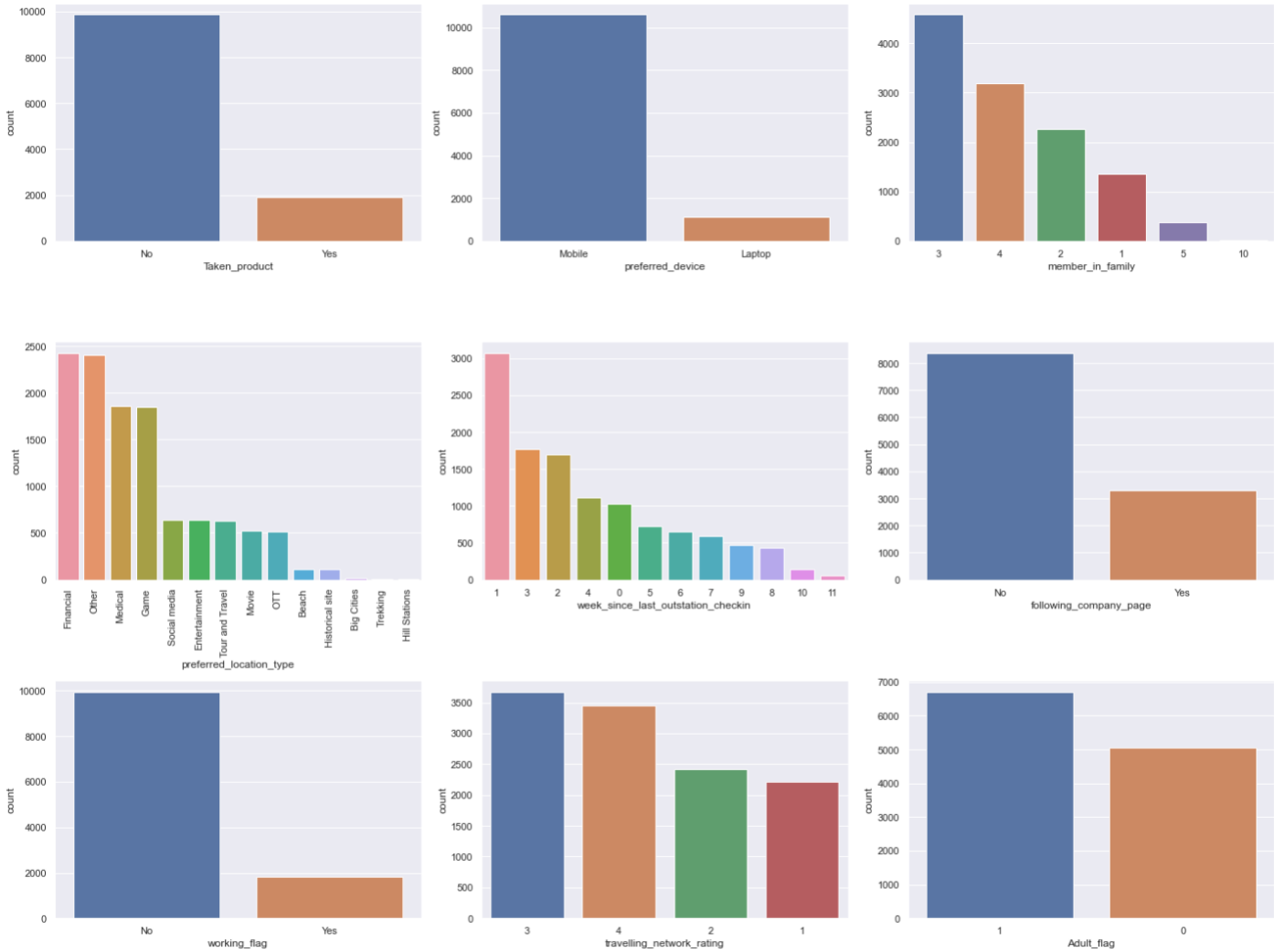


Figure 3 Distribution of Categroical Variables - Bar plot

Bivariate analysis (relationship between different variables , correlations)

Relationship of Target variable with Categorical Variables:

- ***Preferred_device*** Mobile or Laptop has a significance in target variable As the Laptop user has 25% of change in buying the product compared to 15% of Mobile users.
- Number of family member doesn't give any pattern on predicting the target variable.
- The recency of the last outstation check-ins also doesn't provide any significance.
- People who ***follow company's page*** has 40% chance of buying the product compared to only <10% chance on those who don't follow companies page.
- Few of the ***preferred location*** like – ***Beach***, ***Entertainment***, ***Hill Stations*** has higher proportion of Product Taken.

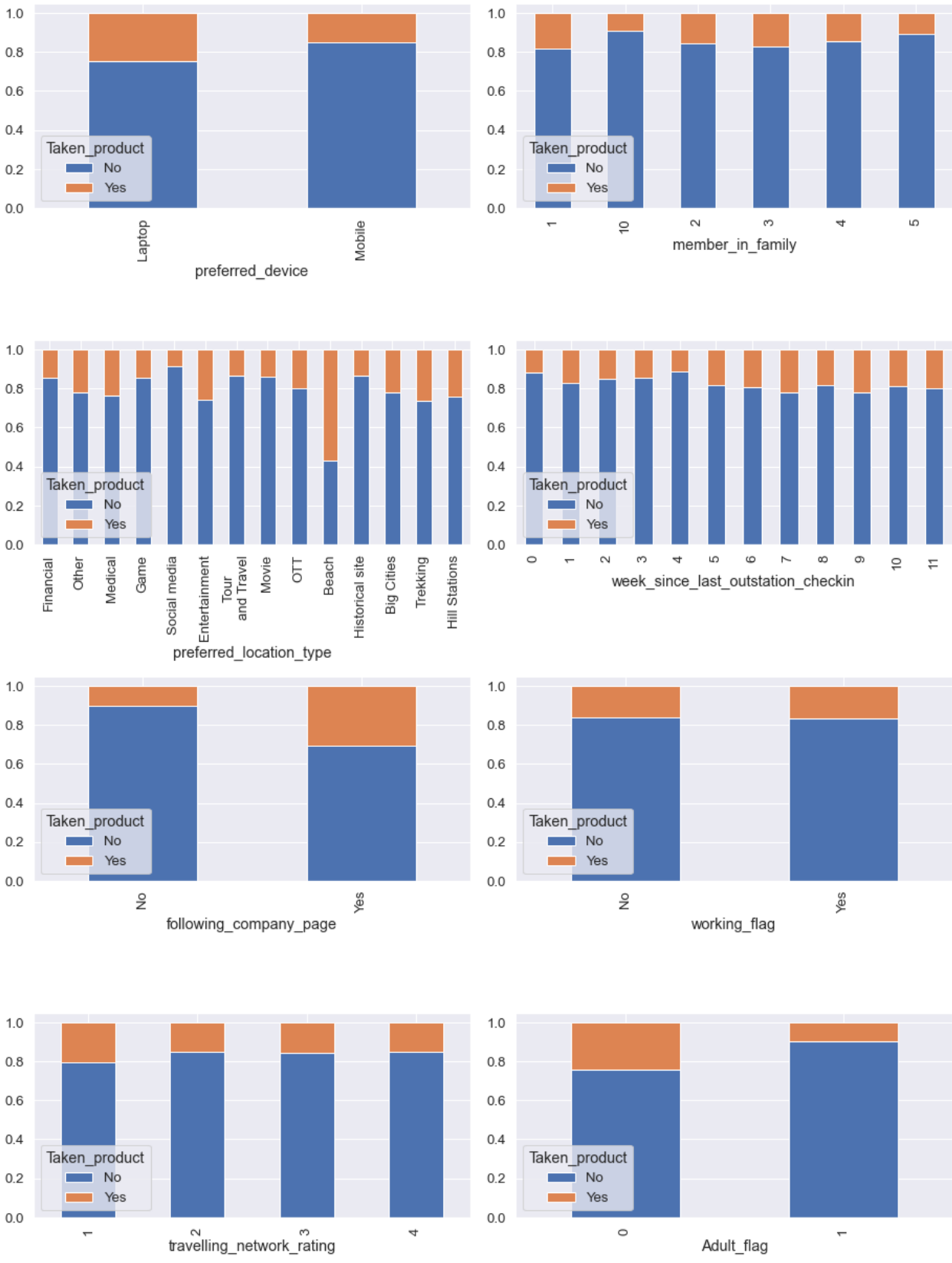
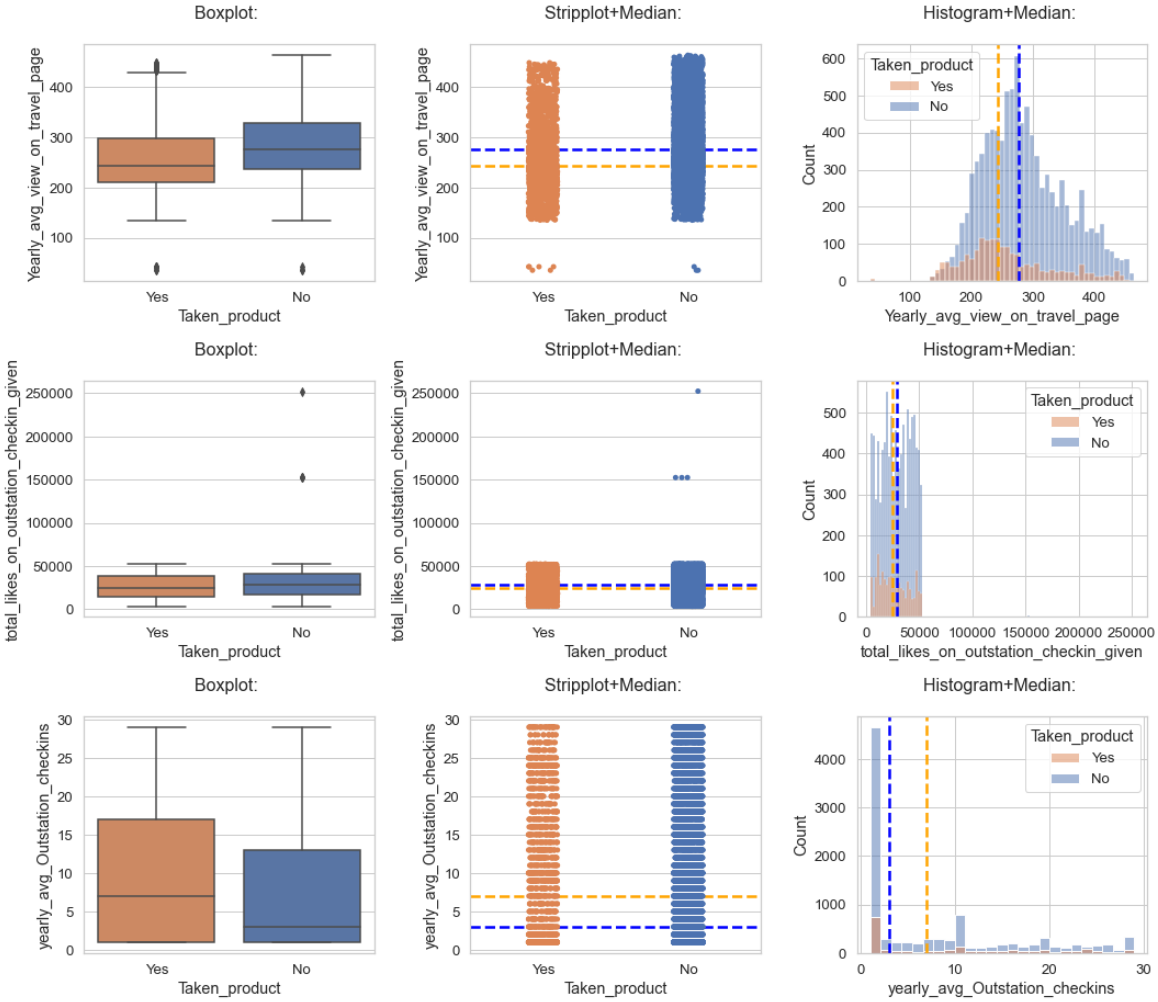


Figure 4 Relationship of Target variable with Categorical Variables

Multivariate Analysis

Relationship of Target variable with Numerical Variables



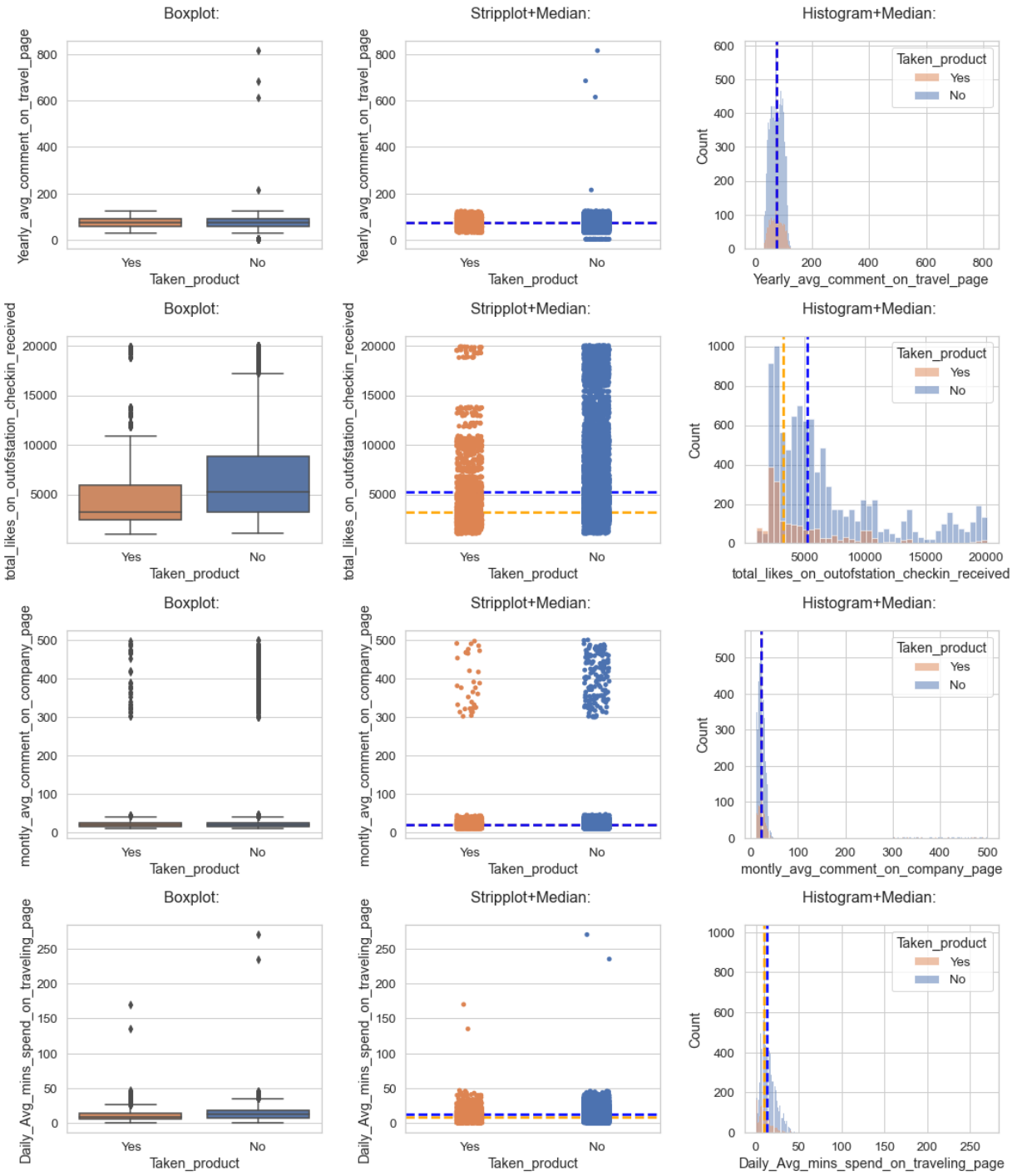


Figure 5 Relationship of Target variable with Numerical Variables

Insights:

- ***Yearly_avg_view_on_travel_page***, ***yearly_avg_outstation_checkins***, ***total_likes_on_outstation_checkin_***received makes a significant difference between the two classes of target variable.
- However, same is not the case with other continuous variables. They are very good at separating the classes.

Influence of Avg Outstation Check-in and Likes received on Target Variable

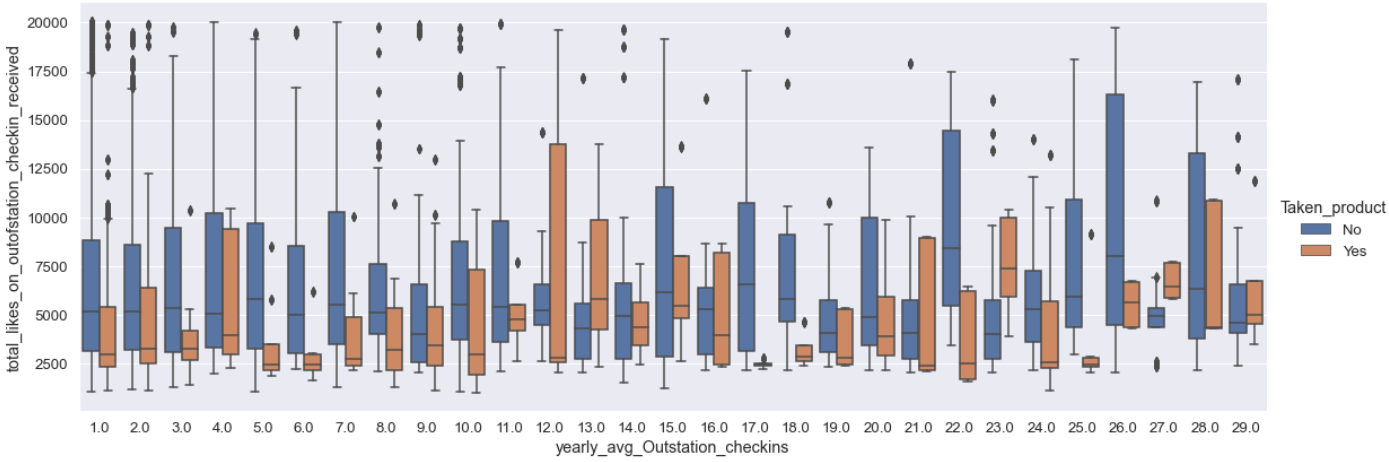


Figure 6 Influence of Avg Outstation Check-in and Likes received on Target Variable

Daily_Avg_mins vs Avg_comments – Following_company_page [Yes or No]

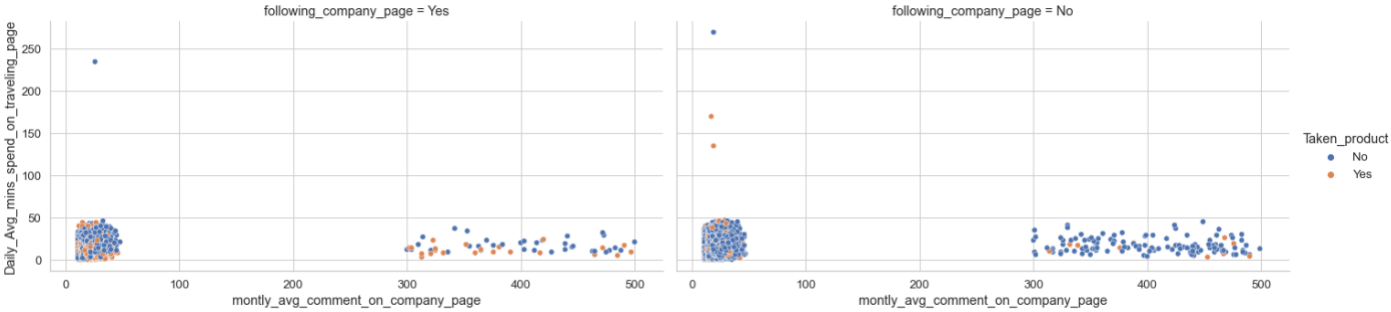


Figure 7 Daily_Avg_mins vs Avg_comments – Following_company

- Highest median of Product taken is at the user group with higher number of ***yearly_avg_Outstation_checkins.***
- Users who interact more (***comments on company page***) and also ***follows company page*** has higher density on ***Taking the product***.

Insights On Correlation Between Independent variables:

- We cannot see any strong correlation between the variables.
- ***Dail_avg_mins_spend_on_traveling_page*** and ***total_likes_on_outofstation_checkin_received*** has moderate positive correlation suggest that person who spends more time on traveling page has higher likes for their check-ins.
- Interestingly ***monthly_avg_comment_on_company_page*** and ***yearly_average_view_on_travel_page*** is negatively correlated
- Also, ***yearly_avg_outstation_checkins*** and ***total_likes_on_outofstation_checkin_received*** is negatively correlated. Thus, the number of check-ins are high tendency to receive likes is low.

- However, the negative correlation are very weak correlation and doesn't show any strong relationship.

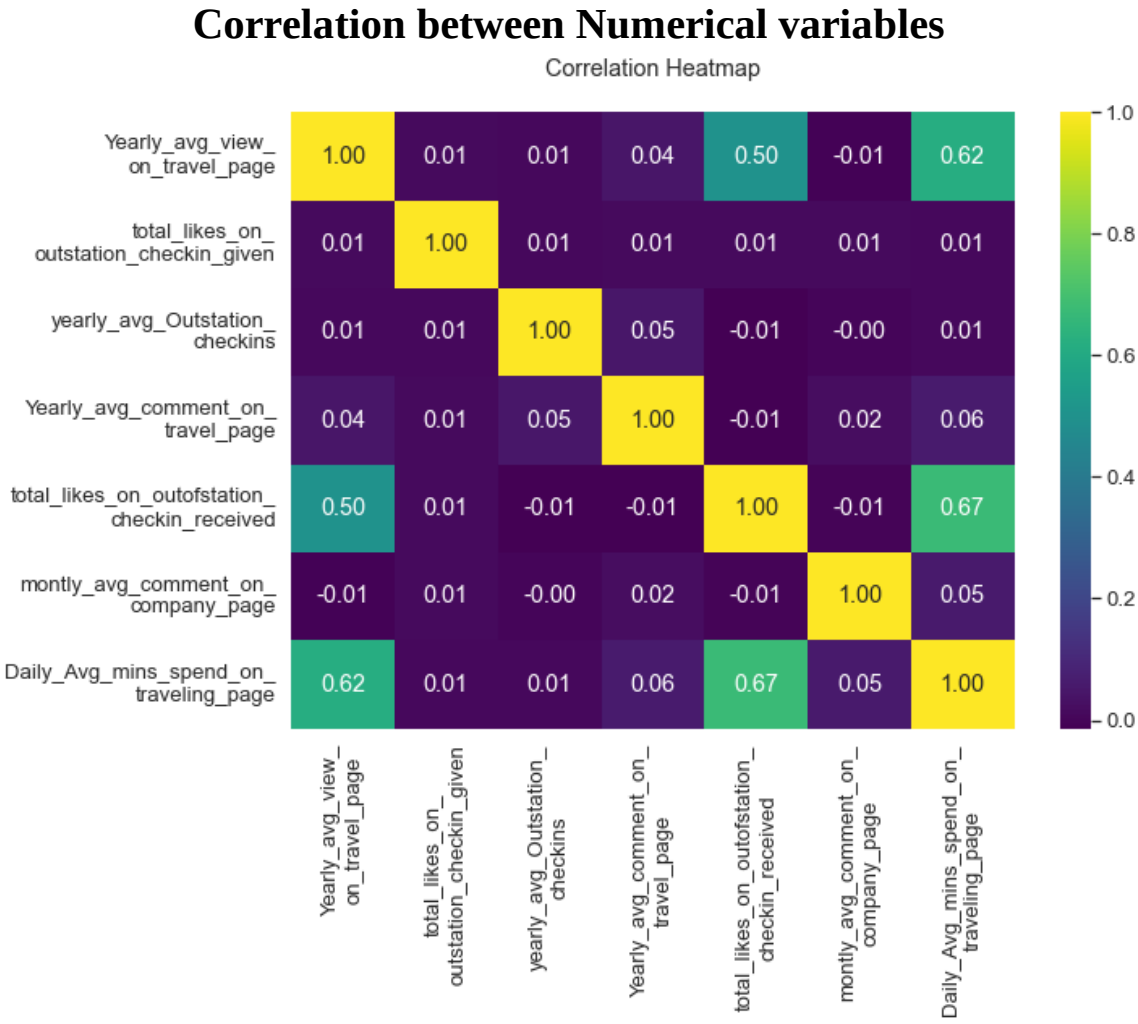


Figure 8 Correlation between Numerical variables

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)

Missing values are treated based on mode for categorical variables and median for continuous variables. And imputing any null values in the data set.

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.

Boxplot after Outliers removed:

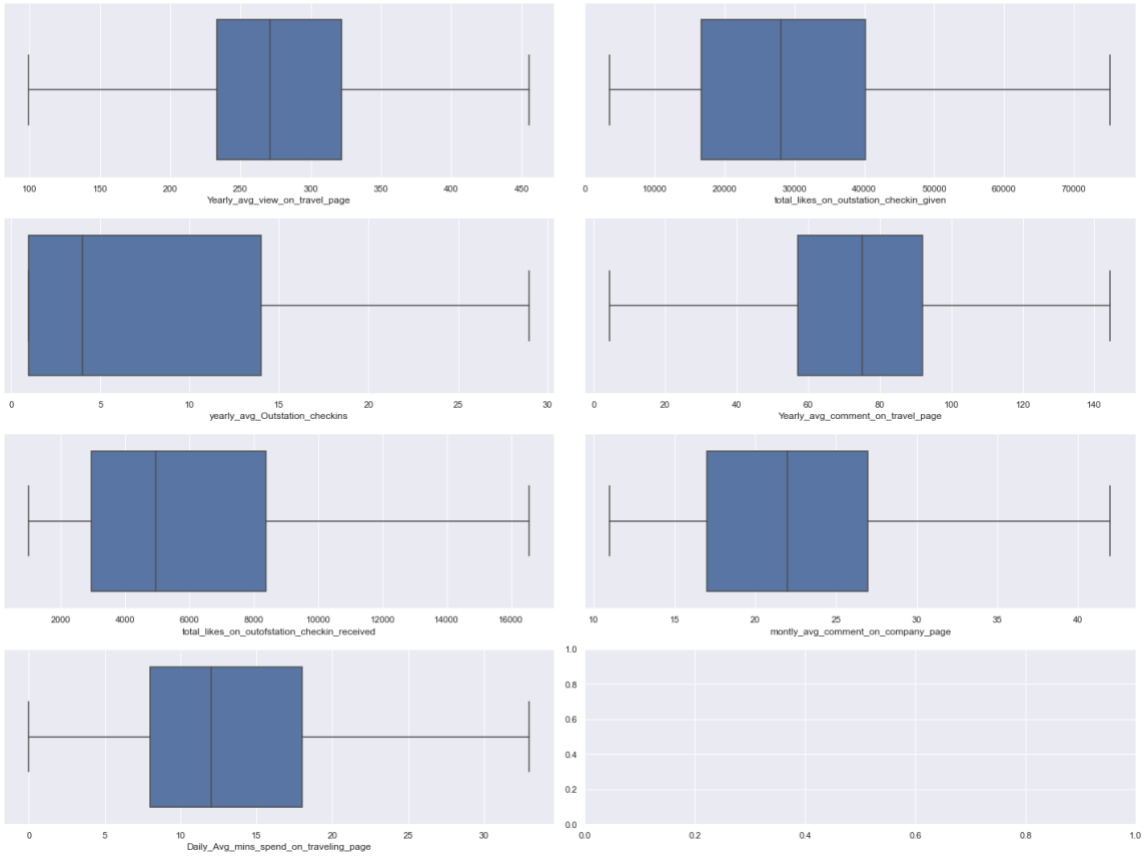


Figure 9 Box plot without outliers

Variable transformation (if applicable)

It's necessary to convert all the variable with numerical values as Machine learning algorithms will work only with the numerical values. Thus it is important to map all the categorical variable into their appropriate numerical values.

'preferred_location_type':

It has different categories of customer preferred locations which is not of any order and doesn't have any priority. Thus we shall encode it will dummy variables and drop the first column to avoid anu multicollinearity.

'Taken_product' , 'following_company_page' , 'working_flag':

These variables only has binary values which shall be converted to 0s and 1s mapped to 'No' and 'Yes' respectively.

And all other object data type variables are converted to integer except 'preferred_device' which shall be used to bifurcate the data set into based on the user preferred device to create two different models.

Feature Engineering

Addition of new variables (if required)

New variable 'Popularity' based on average number of likes per each outstation check ins can be created in order to find the most influential users by dividing the values into 4 buckets

- 0-500 likes per check-in as 1 (Ordinary User),
- 501-5000 likes per check-in as 2 (Moderate User),
- 50001-12500 likes per check-in as 3 (Popular user),

- '>12500 as 4 (Influencer).

Based on this user's popularity we can check whether their popularity status influence in the converting them as a customer to taken the company's product.

New variable 'Popularity Score" shall be created as such it give the average number of likes per outstation check-ins posted by the user.

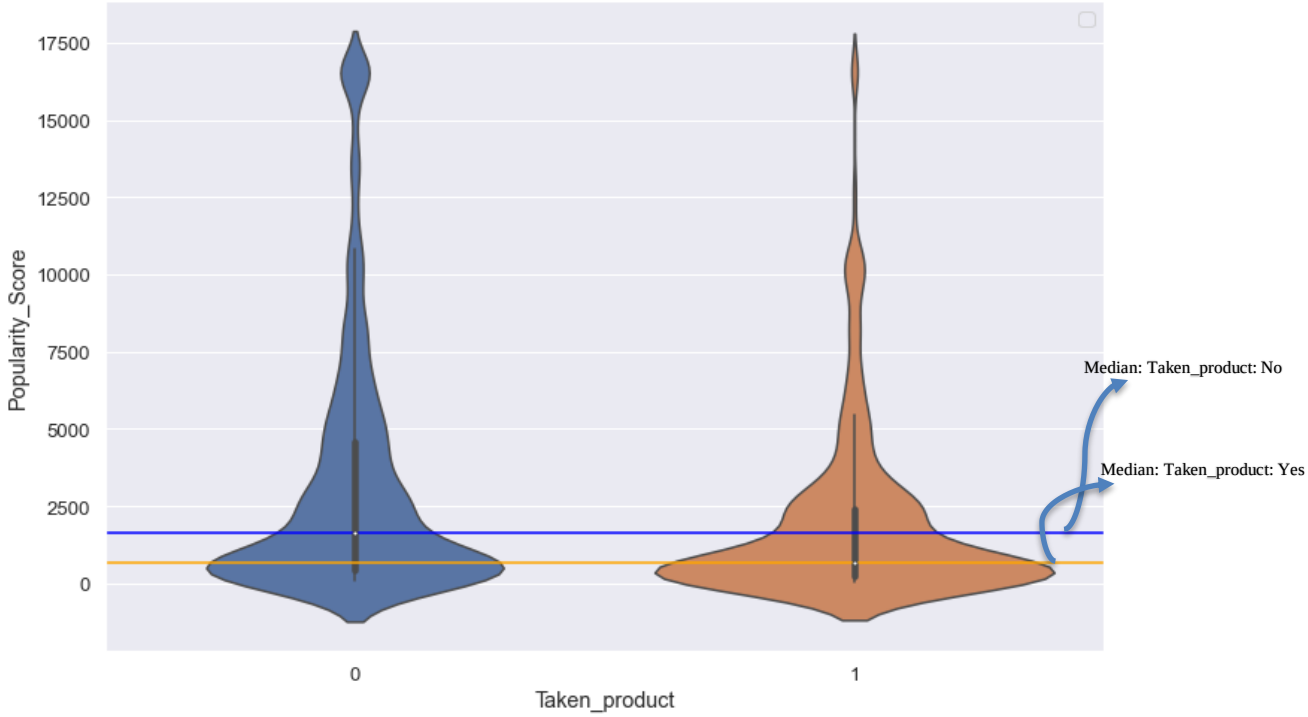


Figure 10 Boxplot: New variable vs Target Variable

Thus the popularity score (average like per outstation check-in) of a user influence in predicting the target variable as the median of class '0' is much higher than median of class '1'.

New variable 'Traveller' based on yearly average number of outstation check-ins can be created in order to find the most influential users by dividing the values into 3 buckets

- <=2 average check-ins per year as 1 (Not a traveller),
- 3-10 average check-ins per year as 2 (Moderate traveller),
- 11-29 average check-ins per year as 3 (Frequent traveller)

Based on this user's travel frequency we can check whether their frequent travel and popularity status influence in the converting them as a customer to taken the company's product.

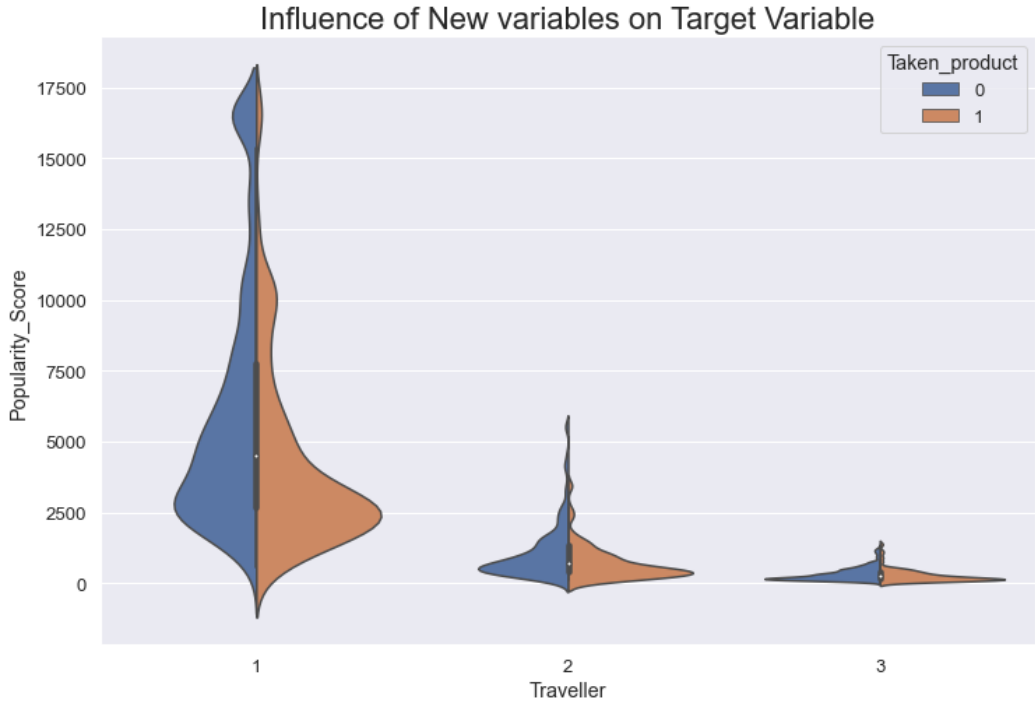


Figure 11 Boxplot - New variable 'Traveller' vs Popularity Score

Figure 12 Boxplot - New variable 'Traveller' vs Popularity Score

Testing the Significance of the New Variables:

ANOVA – "Popularity Scores"

Hypothesis testing for relationship between Popularity scores and the Target variable.

- H0(null hypothesis) : There is no relationship
- H1(alternative hypothesis) : There is a relationship

	df	sum_sq	mean_sq	F	PR(>F)
C(Taken_product):C(Traveller)	**5**	**8.4332E+10**	**1.6866E+10**	**1944.37232**	**0.0**
Residual	**11754**	**1.0196E+11**	**8674422.17**		

As the p-value is <0.05 we can reject the null hypothesis. And conclude there will a relationship between the two variable

Chi-square Test

Hypothesis Test on the newl created variable 'Popularity' and 'Traveller'

Let do a chi-square test on these to newly created features with the target variable for its relation.

- H0 (null hypothesis) : No relationship exist between the Predictor and Target variable
- H1 (alternative hypothesis) : There is a relationship between the Predictor and Target variable

As the p-value is <0.05 for both chi-squared test, we can reject the null-hypothesis, thus variable 'Popularity' and 'Traveller' has significance in predicting the 'Taken_product'

Traveller	Popularity	0	1	Proportion of 1
1	1	0	0	0%
	2	2432	544	22%
	3	1614	178	11%
	4	600	20	3%

2	1	730	239	33%
	2	1863	207	11%
	3	32	0	0%
	4	0	0	0%
3	1	2072	631	30%
	2	521	77	15%
	3	0	0	0%
	4	0	0	0%

Table 5 % of Taken_product against new variables

Thus by grouping the users based on the Travel frequency based on the check-ins show significance in predicting the target of 'Taken_product'. As the user with **higher travel frequency and less popularity** has more proportion of opting for the company's product.

4. Business insights from EDA

Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Before checking the balance of the data, let bifurcate the dataset into two based on the User preferred device - 'Laptop' and 'Mobile'.

Checking for class imbalance

Laptop data		Mobile Data	
0	: 0.750903	0	: 0.847916
1	: 0.249097	1	: 0.152084

As the Target class:1 of the Dependent variable is more than 10% of the given data in both of the data set there is no need for any data balancing.

Business Insights:

- 'Popularity' of the user give significance on the Target variable however, it is opposite that user who is highly popular tend to take the company's product. Thus company has higher scope of approaching these customers as they have move influence on people in the social media since their average likes per check-in is high.
- 'Traveller' frequency of the user also has significance in predication the class '0' or '1' of target variable. Also, user who is a frequent traveller tends to take the company products. Even then we still have many customer to convert into the class '1' by making them to prefer the company's product.

Model Selection:

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, 27) Size of Target Variable (Y_laptop): (1108,)	Size of Predictor variables (X_laptop): (10652, 27) Size of Target variables (Y_laptop): (10652,)

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, 27) Size of X_test for laptop: (333, 27) Size of y_train for laptop: (775,) Size of y_test for laptop: (333,)	Size of X_train for mobile: (7456, 27) Size of X_test for mobile: (3196, 27) Size of y_train for mobile: (7456,) Size of y_test for mobile: (3196,)

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):

- Tree based models (Decision Tree, Random Forest)
- Ensemble Model (XGBoost Classifier)
- Logistic Regression
- Artificial Neural Network.

Choice of model Evaluation Metrics:

For the first iteration, let us build the models with default parameters. Depending upon the performance of the model, whether it overfit or underfit, we shall do the hyperparameter tuning to achieve a better results on both train and test data set for the necessary parameters.

Here, we shall consider **Recall** of the positive class as one of the important metric in evaluating the model performance. Higher Recall shows that the model is good at predicting all the customer who are likely to buy the product by reducing the False Positives.

Models for Laptop:

Let us check for the Variance Inflation factor to ensure that there is not multicollinearity between the predictor variables before training the model. Turns out that the Predictor variables doesn't has any multicollinearity.

	variables	VIF
25	Popularity	3.4059656
11	Daily_Avg_mins_spend_on_traveling_page	3.0519928
26	Traveller	2.7705468
4	total_likes_on_outofstation_checkin_received	2.7694149

0	Yearly_avg_view_on_travel_page	1.8824151
7	montly_avg_comment_on_company_page	1.4372816
8	working_flag	1.3381731
17	preferred_location_type_Historical site	1.2318844
5	week_since_last_outstation_checkin	1.1954274
12	preferred_location_type_Big Cities	1.1638623
24	preferred_location_type_Trekking	1.1331262
2	member_in_family	1.1207941
21	preferred_location_type_Other	1.1128924
16	preferred_location_type_Hill Stations	1.0571269
9	travelling_network_rating	1.0567969
1	total_likes_on_outstation_checkin_given	1.054339
3	Yearly_avg_comment_on_travel_page	1.0525087
10	Adult_flag	1.0487914
6	following_company_page	1.0437911
13	preferred_location_type_Entertainment	
14	preferred_location_type_Financial	
15	preferred_location_type_Game	
18	preferred_location_type_Medical	
19	preferred_location_type_Movie	
20	preferred_location_type_OTT	
22	preferred_location_type_Social media	
23	preferred_location_type_Tour and Travel	

Table 6 VIF - Laptop data

Performance metrics for Different models

Confusion Matrix

Laptop Train Data

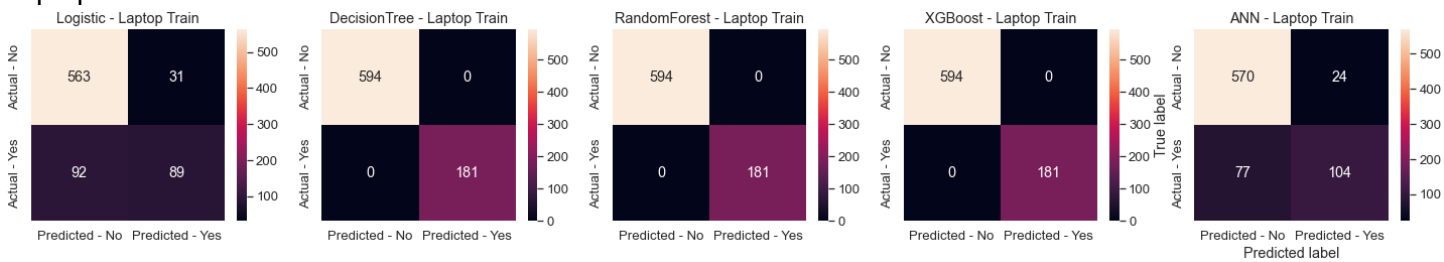


Figure 13 Confusion Matrix - Laptop Train (Default Model)

Laptop Test Data

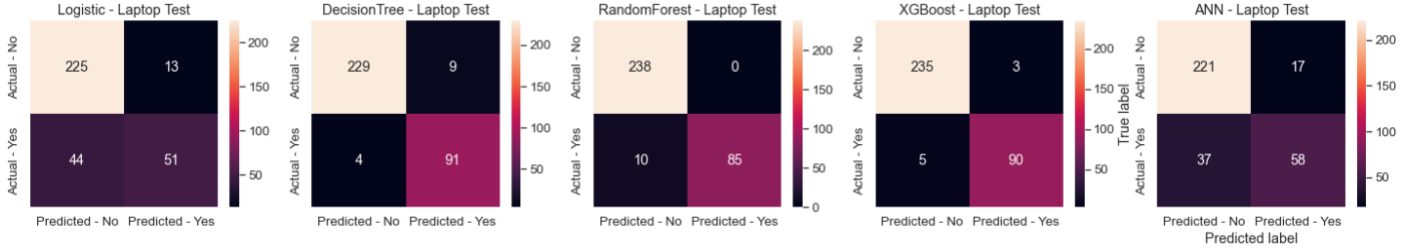


Figure 14 Confusion Matrix - Laptop Test (Default Model)

Classification Report:

Laptop – Training Data:

Model	Accuracy	Precision	Recall	F1_Score
Logistic	0.8412903	0.7416667	0.49171271	0.59136213
DecisionTree	1	1	1	1
RandomForest	1	1	1	1
XGBoost	1	1	1	1
ANN	0.84	0.7355372	0.49171271	0.58940397

Table 7 Model Metrics (Default Model) - Laptop Train

Laptop – Testing data:

Model	Accuracy	Precision	Recall	F1_Score
Logistic	0.81981982	0.75362319	0.54736842	0.63414634
Decision Tree	0.957957958	0.9009901	0.95789474	0.92857143
Random Forest	0.972972973	1	0.90526316	0.95027624
XGBoost	0.975975976	0.96774194	0.94736842	0.95744681
ANN	0.828828829	0.77941176	0.55789474	0.65030675

Table 8 Model Metrics (Default Model) - Laptop Test

Insights for the default Model:

- Random Forest and XGBoost has the **least number of False positives (Type II error)** in the Positive Class (1)
- Logistic regression is the least performing model followed by ANN.
- However, out of 5 model 3 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.

Hyperparameter Tuning – Laptop

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

Artificial Neural Network:

- ⇒ Hidden layer sizes – Describes number of neurons and hidden layer to be used for the neural network. Higher the value more complex the network is and consumes more resource & time.
- ⇒ Tolerance – Error value between each step of learning rate, smaller the value more complex the model is and take more time to converge.
- ⇒ Max iter – Maximum number of iteration unless the convergence is achieve.

Performance Metric after Hyperparameter Tuning – Laptop

After multiple iteration with different parameter values for the hyperparameter, below are the performance metrics of the tuned models.

	Tuned_Logistic	Tuned_DecisionTree	Tune_RandomForest	Tuned_XGBoost	Tuned_ANN
Accuracy_train	0.850322581	0.854193548	0.974193548	0.976774194	0.861935484
Accuracy_test	0.822822823	0.870870871	0.936936937	0.927927928	0.843843844
Precision_train	0.821782178	0.788135593	1	0.982248521	0.94047619
Precision_test	0.846153846	0.817073171	0.94047619	0.927710843	0.905660377
Recall_train	0.458563536	0.513812155	0.889502762	0.917127072	0.436464088
Recall_test	0.463157895	0.705263158	0.831578947	0.810526316	0.505263158
F1_Score_train	0.588652482	0.622073579	0.941520468	0.948571429	0.596226415
F1_score_test	0.598639456	0.757062147	0.882681564	0.865168539	0.648648649
auc_train	0.834030917	0.903965995	0.999469836	0.995433153	0.853302826
auc_test	0.869305617	0.913157895	0.9910659	0.970809376	0.876780186

Table 9 Performance Metrics After Tuning - Laptop

After fine tuning them model to avoid overfitting we can see that the XGBoost performance better that other model in terms of evaluation metrics. Also, I has a good consist performance between train and test data. Thus we will choose this our model of choice.

Models for Mobile:

Let us check for the Variance Inflation factor to ensure that there is not multicollinearity between the predictor variables before training the model. Turns out that the Predictor variables doesn't has any multicollinearity.

	variables	VIF
25	Popularity	5.50098241
4	total_likes_on_outofstation_checkin_received	3.78975948
26	Traveller	3.70547595
11	Daily_Avg_mins_spend_on_traveling_page	2.93896956
14	preferred_location_type_Financial	2.0869098
18	preferred_location_type_Medical	1.90753934
7	montly_avg_comment_on_company_page	1.87133861
0	Yearly_avg_view_on_travel_page	1.76413381
17	preferred_location_type_Historical site	1.52650348
22	preferred_location_type_Social media	1.36097897
8	working_flag	1.33635345
13	preferred_location_type_Entertainment	1.30685761
21	preferred_location_type_Other	1.26505283
12	preferred_location_type_Big Cities	1.21758589
24	preferred_location_type_Trekking	1.17748846
5	week_since_last_outstation_checkin	1.14220113
3	Yearly_avg_comment_on_travel_page	1.07267896
23	preferred_location_type_Tour and Travel	1.07142366
2	member_in_family	1.05593287
16	preferred_location_type_Hill Stations	1.04070819
10	Adult_flag	1.02239669
9	travelling_network_rating	1.01808443
1	total_likes_on_outstation_checkin_given	1.01355177
6	following_company_page	1.01041198
15	preferred_location_type_Game	1.00888468
20	preferred_location_type_OTT	1.00695421
19	preferred_location_type_Movie	1.00495187

Table 10 VIF - Mobile Data

Performance metrics for models

Confusion Matrix

Mobile Train Data

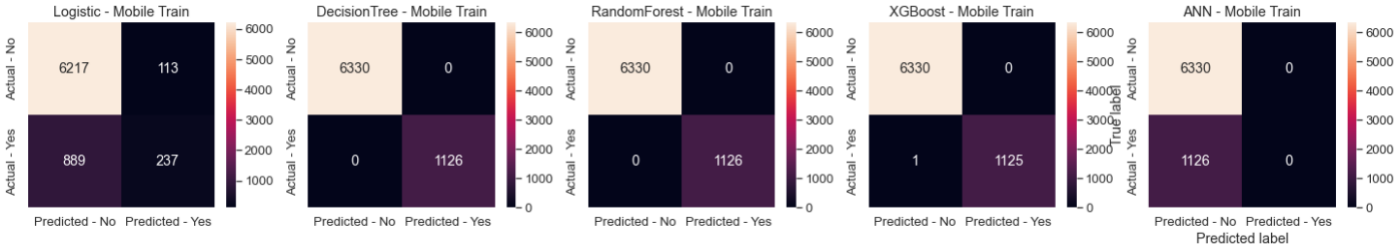


Figure 15 Confusion Matrix (Default)- Mobile Train Data

Mobile Test Data

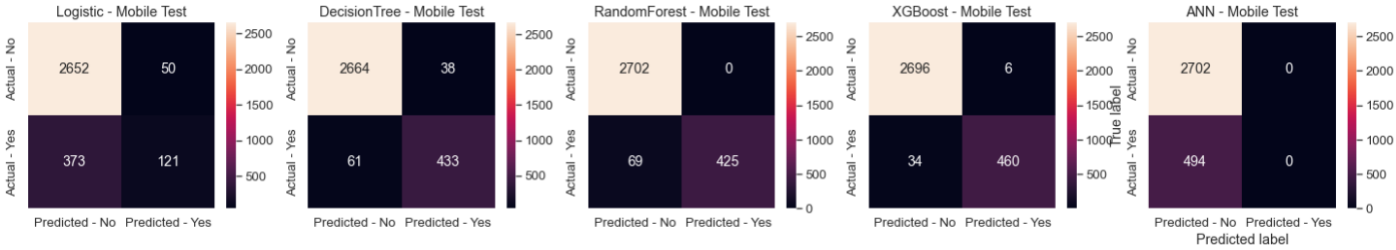


Figure 16 Confusion Matrix (Default)- Mobile Test Data

Classification Report:

Mobile – Training Data:

Model	Accuracy	Precision	Recall	F1_Score
Logistic	0.86668455	0.68539326	0.21669627	0.32928475
Decision Tree	1	1	1	1
RandomForest	1	1	1	1
XGBoost	1	1	1	1
ANN	0.86574571	0.71331058	0.18561279	0.29457364

Table 11 Classification Metrics (Default) - Mobile Train

Mobile – Testing data:

Model	Accuracy	Precision	Recall	F1_Score
Logistic	0.86827284	0.71098266	0.24898785	0.36881559
Decision Tree	0.97121402	0.92405063	0.88663968	0.90495868
RandomForest	0.97465582	1	0.83603239	0.9106946
XGBoost	0.98529412	0.99120879	0.91295547	0.95047418
ANN	0.86827284	0.76258993	0.2145749	0.33491311

Table 12 Classification Metrics (Default) - Mobile Test

Insights for the default Model:

- Random Forest and XGBoost has the **least number of False positives (Type II error)** in the Positive Class (1)
- Logistic regression is the least performing model followed by ANN.
- However, out of 5 model 3 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.

Hyperparameter Tuning – Laptop

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

Performance Metric after Hyperparameter Tuning – Laptop

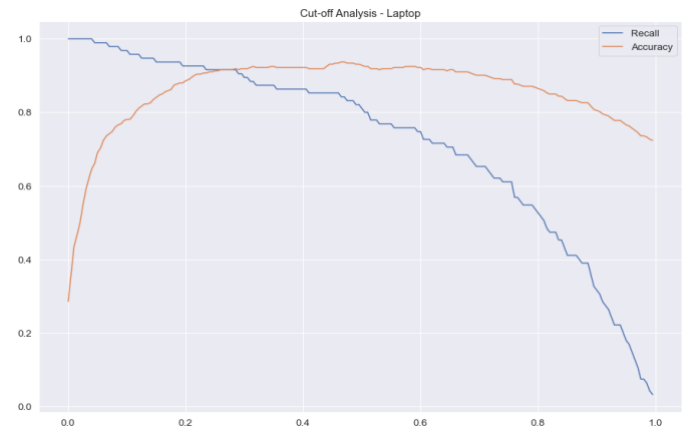
	Tuned_Logistic	Tuned_DecisionTree	Tune_RandomForest	Tuned_XGBoost	Tuned_ANN
Accuracy_train	0.86681867	0.880633047	0.959093348	0.993562232	0.874463519
Accuracy_test	0.868898623	0.876720901	0.936170213	0.969962453	0.871714643
Precision_train	0.684210526	0.754310345	0.997575758	0.999074074	0.735148515
Precision_test	0.714285714	0.745098039	0.996575342	0.976076555	0.712121212
Recall_train	0.219360568	0.310834813	0.730905861	0.958259325	0.263765542
Recall_test	0.253036437	0.307692308	0.589068826	0.825910931	0.285425101
F1_Score_train	0.332212508	0.440251572	0.843669913	0.97824116	0.388235294
F1_score_test	0.373692078	0.435530086	0.740458015	0.894736842	0.407514451
auc_train	0.795761535	0.809203755	0.997804444	0.999647707	0.850689715
auc_test	0.802187314	0.805368718	0.984598303	0.991391142	0.837028052

Table 13 Classification Metrics (Tuned) - Mobile data

Cut Off Analysis

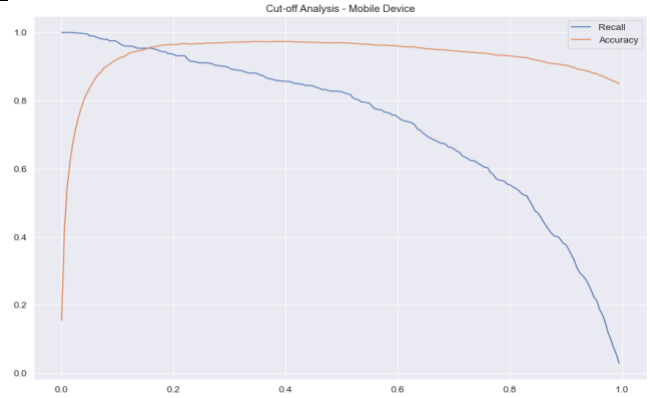
To find the optimal threshold of probability above which the prediction is consider as positive class of prediction. We shall find such optimal threshold to achieve the maximum accuracy and Recall trade-off.

Laptop Device



Cut_off	Accuracy	Recall
0.465	**0.93693693**	**0.84210526**
0.47	**0.93693693**	**0.84210526**
0.455	**0.93393393**	**0.85263157**
0.46	**0.93393393**	**0.85263157**
0.475	**0.93393393**	**0.83157894**

Mobile Device



Cut_off	Accuracy	Recall
0.345	0.97340425	0.88056680
0.385	0.97340425	0.85829959
0.39	0.97340425	0.85829959
0.35	0.97309136	0.87854251
0.395	0.97309136	0.85627530

⇒ Let us consider 0.4 for Laptop and 0.3 for mobile as a cut-off and predict the class as 1 when it is above the given threshold.

Feature Importance

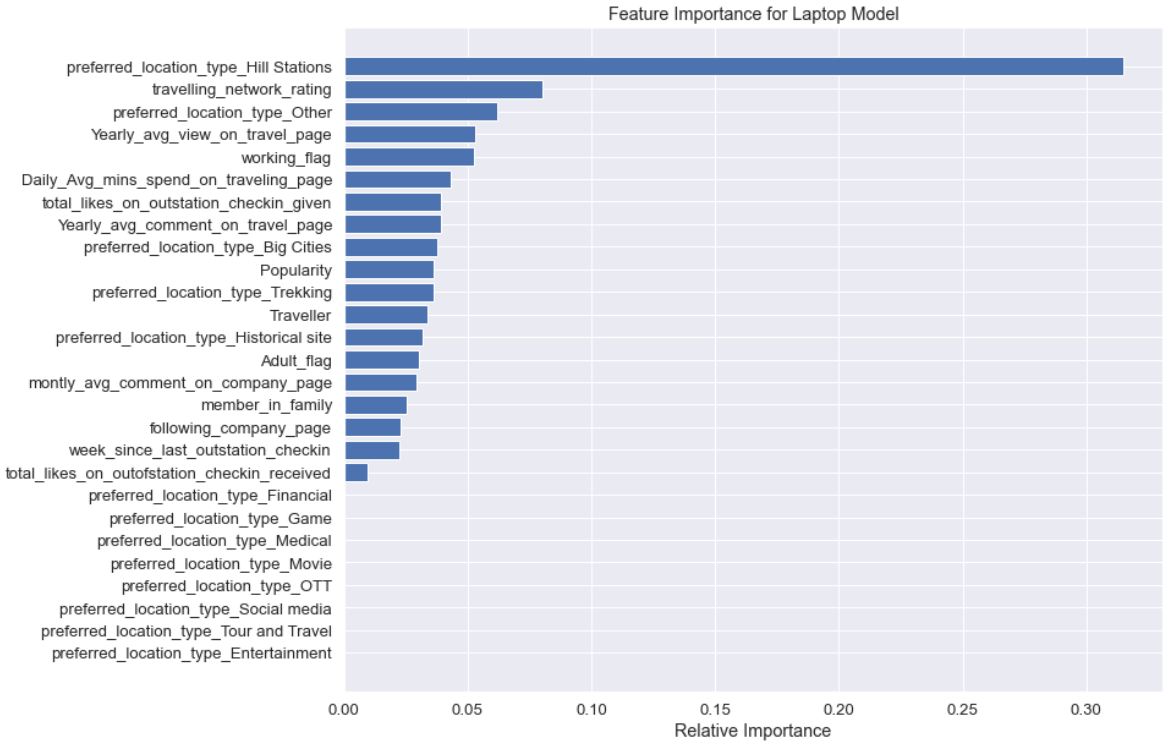


Figure 17 Feature Importance - Laptop

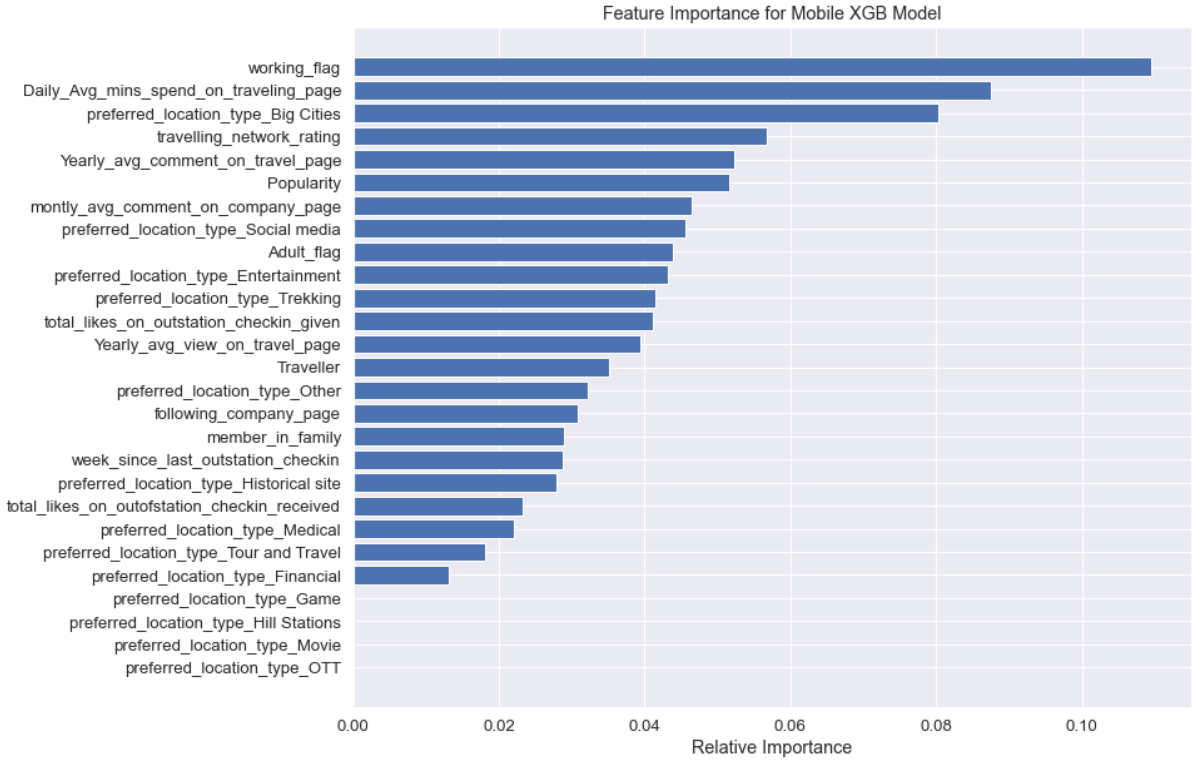


Figure 18 Feature Importance - Mobile

- With the Importance feature of both the models from XGBoost we shall consider the top 10 feature to find the insights and which feature as good significance on the Target variable.

Significance on Target variable – Laptop

Numerical Variables

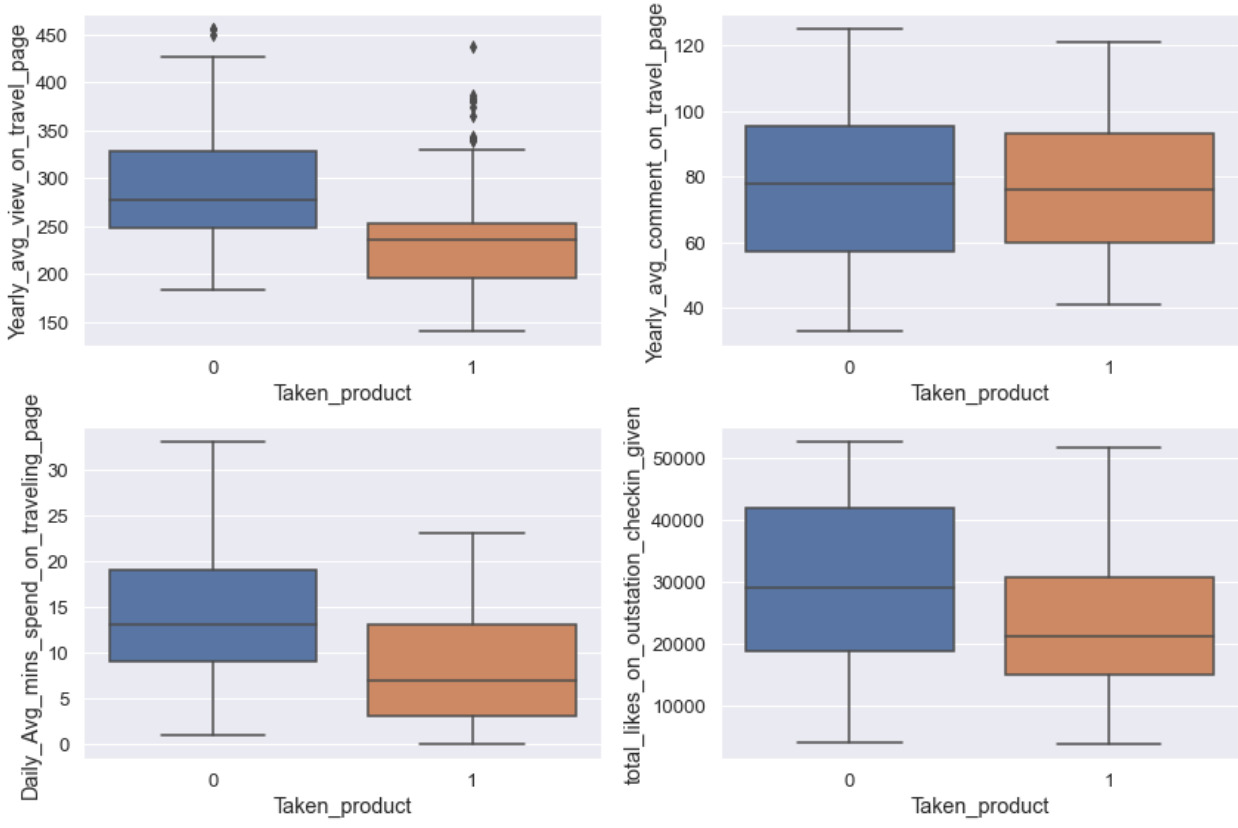


Figure 19 Significance on Target variable -Numerical

Categorical Variables

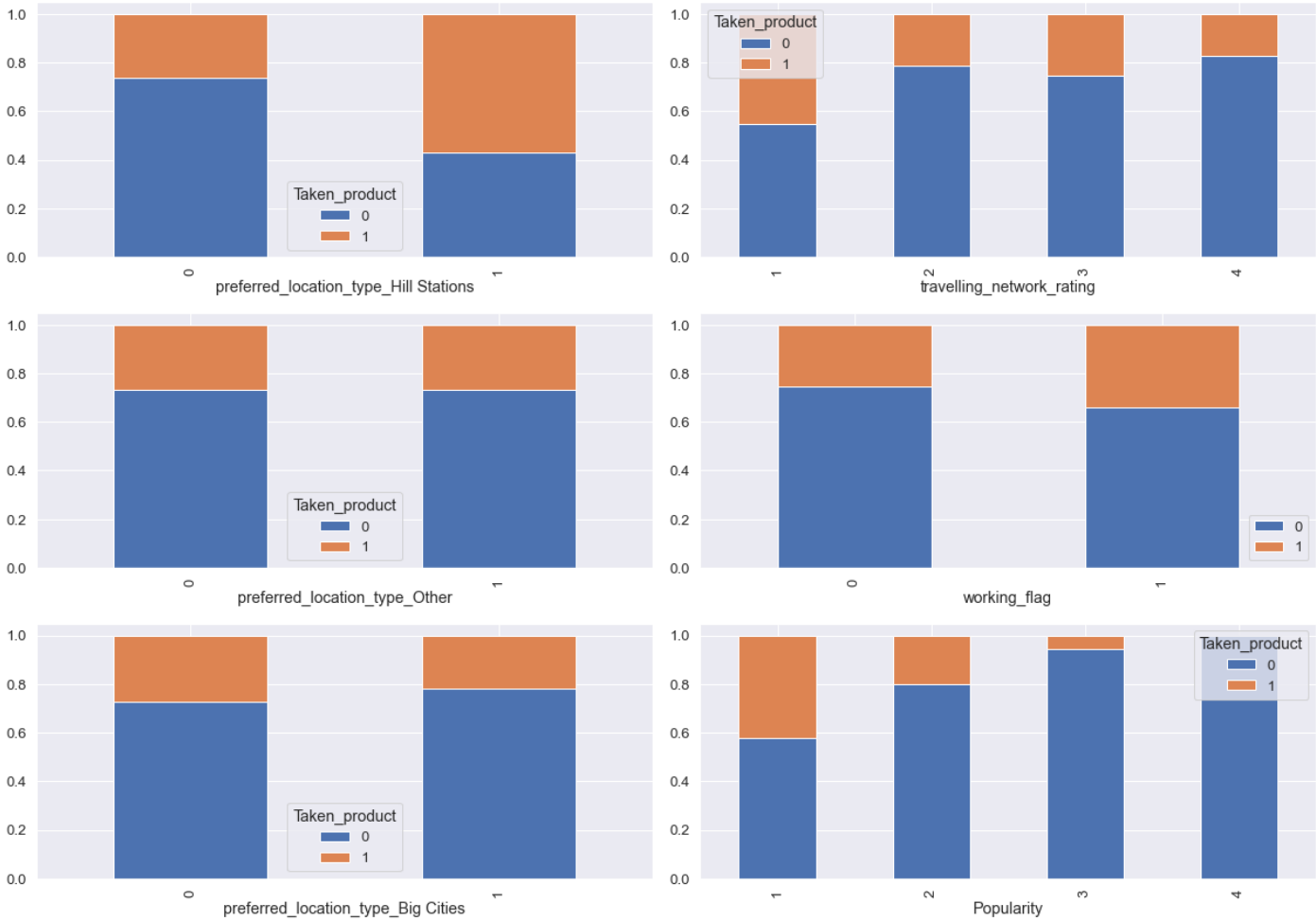


Figure 20 Significance on Target variable -Categorical

Insights for Laptop as Preferred Device:

- User who has their preferred location as "Hill stations" ~30% more chance in converting to prefer the company's product than the users who don't prefer Hill Stations.
- User who are working has ~10% more chance of buying the company's product.
- Out of the four class of Travel rating user who with the least Travel network rating (rating 1) has high ~30% more chance of buying the product than other ratings.
- Users with the least Popularity rating has ~25% higher chance.
- However, Popular user are still not preferring our company's product hence there is a huge opportunity to target such user to convert them into the potential customer.

Significance on Target variable – Mobile

Numerical Variables:

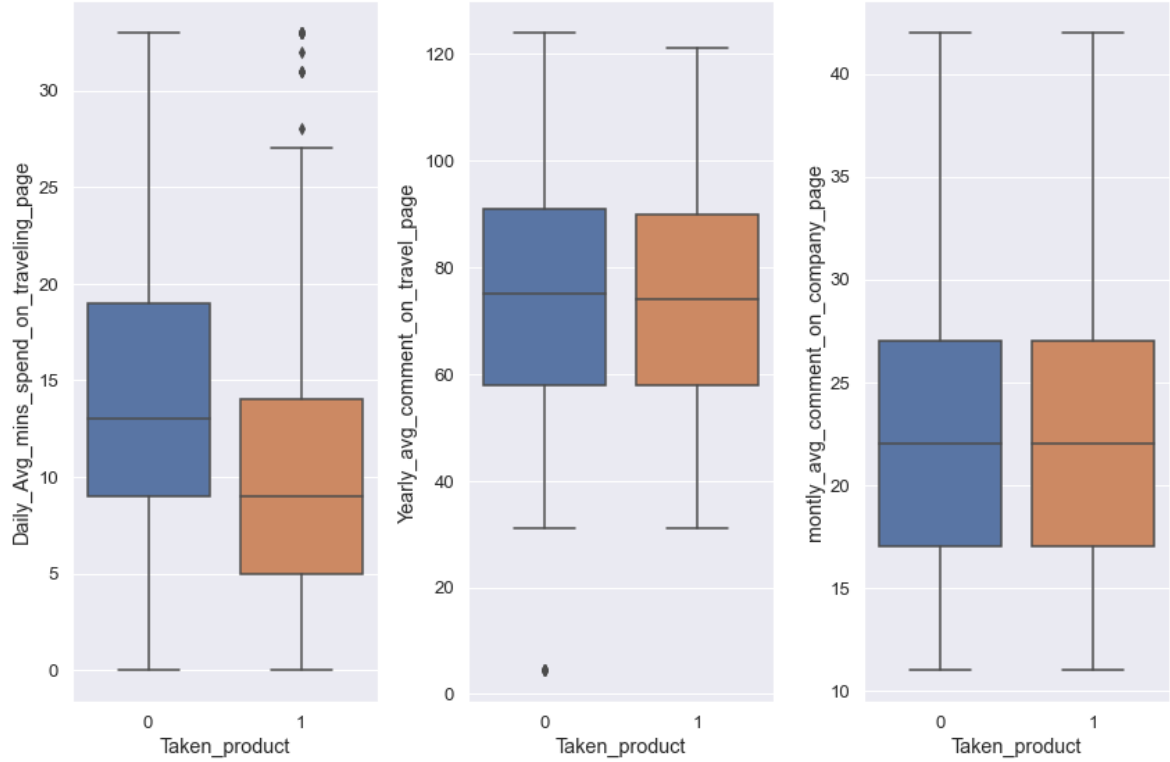


Figure 21 Significance on Target variable -Numerical

Categorical Variables:

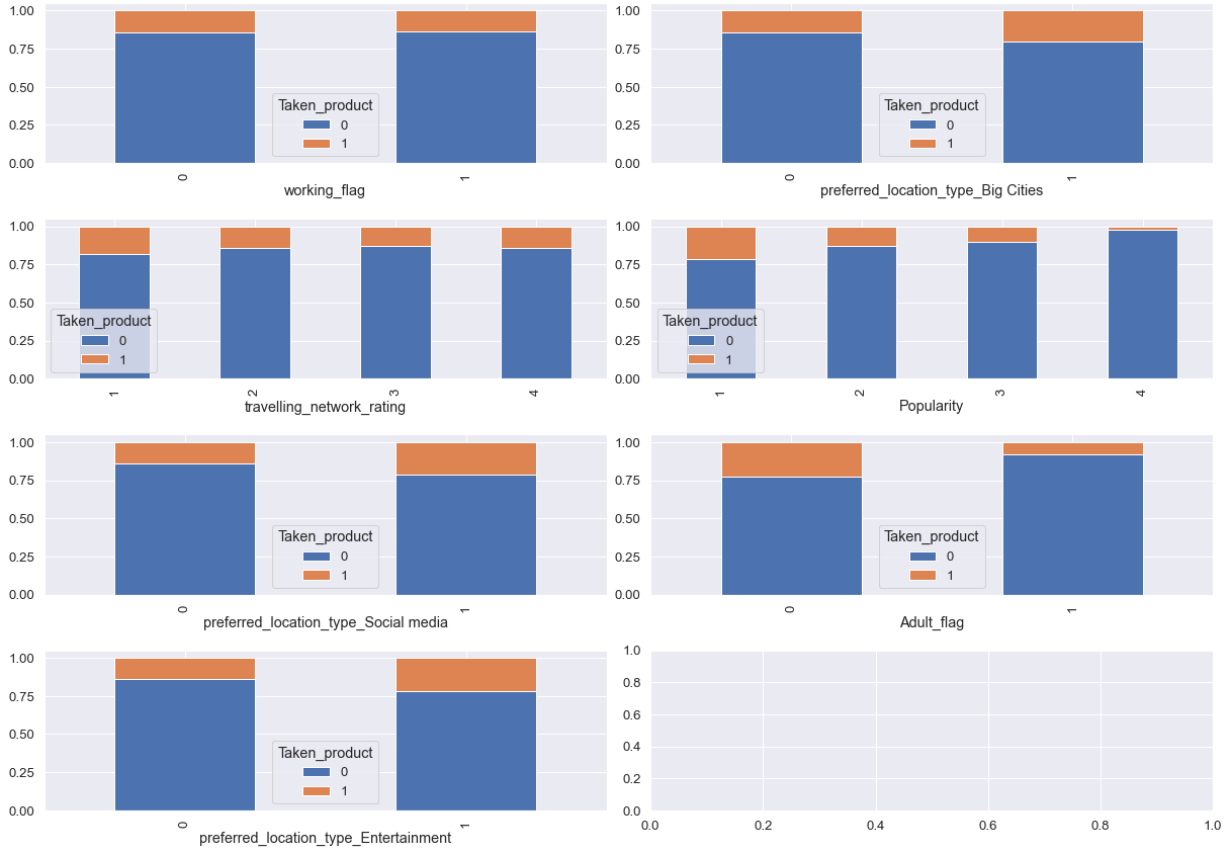


Figure 22 Significance on Target variable -Categorical

Insights for Mobile as Preferred Device:

- Users who are not Adult has ~20% more chance of buying the product.
- User who has Entertainment as their preferred location has a ~15% more chance in converting.
- Users with the least Popularity rating has ~15% higher chance.
- However, Popular user are still not preferring our company's product hence there is a huge opportunity to target such user to convert them into the potential customer.

User Profiling for Targeted Digital Marketing:

After finding the significant variables we shall now cluster the users based on their behaviour on these particularly importance variables, which then help us to target the user group who as better metrics in the importance features we have selected.

By doing K-means Clustering with the optimal cluster of 4 we shall group the users into 4 clusters for both Laptop and Mobile device.

Profiling for Laptop Users:

cluster	Popularity	preferred_location_type_Hill Stations	travelling_network_rating	working_flag	Taken_product
Cluster-1	2	0	1	0	0.0%
Cluster-2	2	0	4	0	0.0%
Cluster-3	2	0	3	1	1.7%
Cluster-4	2	1	3	0	100%

Table 14 Profiling for Laptop Users

Profiling for Mobile Users:

cluster	Popularity	Adult_flag	preferred_location_type_Entertainment	working_flag	% Taken_product
Cluster-1	2	1	0	0	9.4%
Cluster-2	2	0	0	0	2.2%
Cluster-3	2	1	1	0	2.4%
Cluster-4	2	1	0	1	1.5%

Table 15 Profiling for Mobile Users

General Business Insights:

- It is also important to target the customer who are wrongly predicted as converted to improve the sale. But, this shall be give 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.