Prediction whether the customer is going to adopt the tourism package based on a social media campaign.

Social Media_Tourism_ Project

Final Report Submission

Submitted By- Gunjar Fuley
Batch- PGPDSBA Online Nov_A 2020
Email- gforgunjaar@gmail.com
Phone- 9938126651

GO-GO AIR

Effectiveness of Social Media Campaign for higher revenue through an increase in sales of tickets

SOCIAL MEDIA & Ad Campaigns

According to Wikipedia, **Social media** are interactive technologies that allow the creation or sharing/exchange of information, ideas, interests, and other forms of expression via virtual communities and networks citation. Example Facebook, Twitter, Instagram etc.

A social media campaign is a coordinated marketing designed to reinforce information or sentiments —about a product, service, or overall brand—through at least one social media platform.

Images used for representation

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

Problem of decrease in revenue due to decline in sales of ticket-

Go-Go AIR is a multinational aviation organization headquartered in Mumbai, known for its world class services. Due to huge customer obsession, the organization believes in continues learning and improvement. The management is keen on understanding and removing the flaws across all the departments. Based on continuous feedback, it was understood that the name of the brand Go-Go AIR is degrading significantly due to the frequent and aggressive marketing campaigns. Traditionally, the marketing and sales functions relied on reaching out to the potential customers through the conventional method of cold calling. But Team Go-Go AIR has realized that this practice isn't relevant anymore. In order to replace it, various strategies were suggested to the management. Finally, the esteemed Marketing and Sales Department came up with the idea of reaching out to the masses using social media marketing campaigns. It was eventually approved by management. Hence Go-Go Air decided to collaborate with a social media platform for Ad campaigns.

Need of the study-

The social media ad campaigns attract a huge cost per customer acquisition. Therefore, a pilot project was conducted where the social media ad was displayed to the audience and the data was collected. Based on the data, the management aims to understand digital and social behavior of existing customers. They instructed the Marketing Information System (MIS) team, to come up with a model which will predict that customer will buy the ticket or not.

Understanding business opportunity

The aim of this activity is to achieve an increase in sales revenue by at least 30% in the upcoming financial year through the social media ad campaigns and cost cutting by ending tele calling processes by 50%.

Eventually the MIS team agreed to build a model using machine learning algorithms. They were also asked to check the performance of model on the real data & then recommend deployment. Based on the prediction, the social media ad shall be displayed on the targeted customers.

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

Visual understanding of data

After uploading the data, it was understood that the number of data points or the rows were 11760 and number of features or variables were 17.

This gives the understanding that this data has information of 11760 customers.

In the figure below we can see initial 10 data points. We can see that there are Null (NaN) values in the data.

	0	1	2	3	4	5	6	7	8	9
UserID	1000001	1000002	1000003	1000004	1000005	1000006	1000007	1000008	1000009	1000010
Taken_product	Yes	No	Yes	No	No	No	No	No	No	No
Yearly_avg_view_on_travel_page	307.0	367.0	277.0	247.0	202.0	240.0	NaN	225.0	285.0	270.0
preferred_device	iOS and Android	iOS	iOS and Android	iOS	iOS and Android	iOS	iOS and Android	iOS and Android	iOS	iOS and Android
total_likes_on_outstation_checkin_given	38570.0	9765.0	48055.0	48720.0	20685.0	35175.0	46340.0	NaN	7560.0	45465.0
yearly_avg_Outstation_checkins	1	1	1	1	1	1	1	24	23	27
member_in_family	2	1	2	4	1	2	Three	1	3	3
preferred_location_type	Financial	Financial	Other	Financial	Medical	Financial	Medical	Financial	Financial	NaN
Yearly_avg_comment_on_travel_page	94.0	61.0	92.0	56.0	40.0	79.0	81.0	67.0	44.0	94.0
total_likes_on_outofstation_checkin_received	5993	5130	2090	2909	3468	3068	2670	2693	9526	5237
week_since_last_outstation_checkin	8	1	6	1	9	0	4	1	0	6
following_company_page	Yes	No	Yes	Yes	No	No	Yes	No	No	No
montly_avg_comment_on_company_page	11	23	15	11	12	13	20	22	21	13
working_flag	No	Yes	No	No	No	No	Yes	Yes	Yes	No
travelling_network_rating	1	4	2	3	4	3	1	2	2	2
Adult_flag	0	1	0	0	1	0	3	1	0	2
Daily_Avg_mins_spend_on_traveling_page	8	10	7	8	6	8	12	1	10	17

From the below table we can understand the nature of data whether it is categorical or numeric in nature.

S. No.	Feature Name	Data Type
1	UserID	
2	Taken_product	Categorical
3	Yearly_avg_view_on_travel_page	Numeric
4	preferred_device	Categorical
5	total_likes_on_outstation_checkin_given	Numeric
6	yearly_avg_Outstation_checkins	Numeric
7	member_in_family	Numeric
8	preferred_location_type	Categorical

9	Yearly_avg_comment_on_travel_page	Numeric
10	total_likes_on_outofstation_checkin_received	Numeric
11	week_since_last_outstation_checkin	Numeric
12	following_company_page	Categorical
13	montly_avg_comment_on_company_page	Numeric
14	working_flag	Categorical
15	travelling_network_rating	Categorical
16	Adult_flag	Categorical
17	Daily_Avg_mins_spend_on_traveling_page	Numeric

Also, below are the descriptive details of the data. From count we can understand that there are several null values in numeric variables. Here, we won't be able to infere much about data.

In variable 'Adult_flag', we can see that the maximum value is 3. This is not valid as it's categorical in nature. It can yes or no, either 0 or 1.

	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

Non - Visual understanding of data

The following is the data dictionary provided for the social media ad campaign database. Along with the description we have mentioned if the remaining is required for any column or not.

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

In the below table, we can see that UserID, total_likes_on_outofstation_checkin_received, week_since_last_outstation_checkin, montly_avg_comment_on_company_page, travelling_network_rating, Adult_flag, Daily_Avg_mins_spend_on_traveling_page has variable type ‘integer’. The variables ‘Yearly_avg_view_on_travel_page’, ‘total_likes_on_outstation_checkin_given’, ‘Yearly_avg_comment_on_travel_page’ has data type ‘float’. The remaining features have data type ‘object’.

We can also see that there are several Null values in the features. We shall treat them in the NULL value treatment ahead.

```
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)
```

From the data we can also understand that yearly average view on travel page is 280 for a user.

The percentage of Users buying ticket is 16.12. The percentage of Users not buying ticket is 83.88.

```
No     9864
Yes    1896
Name: Taken_product, dtype: int64
```

Current Business Environment of Digital Marketing

According to the bazaarvoice.com, The average Conversion Rate for the brands that perform in the top 20% of the Instagram accounts we analyzed is 0.3%. It's significantly higher for brands with 10K-50K followers (2.2%) followed by brands with <10K followers (1.8% conversion), implying brands with smaller following have a higher Conversion Rate.

According to wordstream.com, the average conversion rate across Facebook Ads is 9.21%. As you can see, the average advertiser in this industry converts clicks into meaningful actions at a rate of over 14%!

Univariate Analysis

All the unique values and the frequency of the occurrence of any data point in the entire dataset were done.

Below are the findings:

1. UserID
   Here we have all 11760 unique user IDs. This does not require any treatment.

2. Taken_product
   The number of people who has opted to take the product are ‘yes’ and people not buying the product are labelled as ‘no’.
3. Yearly_avg_view_on_travel_page
   It is yearly average view of every user on the social media page.
4. Preferred Device
   In the column ‘preferred_device’, the attributes ‘Andriod’ and ‘ANDRIOD’ are same but the only difference is of case lower and upper. The attributes ‘Other’ and ‘Others’ are also same therefore they were transformed.
5. total_likes_on_outstation_checkin_given
   It is likes on the outstation check ins whenever done by every user.
6. yearly_avg_Outstation_checkins
   We found that one of the value in column ‘yearly_avg_Outstation_checkins’, is ‘*’. We shall impute it with the mode because the datatype is ‘object’. After the imputation, for further analysis the datatype of variable was changed to ‘float’.
7. member_in_family
   In the column ‘member_in_family’, one of the data point is ‘Three’ instead of ‘3’. So, ‘Three’ was replaced to ‘3’.
8. Preferred_location_type
   In the column ‘preferred_location_type’, the attributes ‘Tour Travel’ and ‘Tour and Travel’ are same. Therefore, we have merged ‘Tour Travel’ into ‘Travel and Tour’.
9. Yearly_avg_comment_on_travel_page
   It is the count of the comments made by any user on the social media post by the company.
10. total_likes_on_outofstation_checkin_received
    It gives the likes received on the personal profile of the user for any out of station check in.
11. week_since_last_outstation_checkin
    It gives the number of weeks since the last outstation check in done by the users.
12. following_company_page
    In the column ‘following_company_page’, some data points are labelled ‘1’ and ‘0’. Here we have assumed ‘No’ as ‘0’ and ‘Yes’ as ‘1’. Both the 1s and 0s were transformed into yes and no respectively.
13. montly_avg_comment_on_company_page
    This variables gives the average number of comments done on company page on a monthly basis.
14. working_flag
    The number of users working are 1808 and the non-working users are 9952.

15. travelling_network_rating
    The variable ‘travelling_network_rating’ is categorical variable but by default it is ‘int64’. Therefore the data type of the variable was changed to ‘category’.
16. Adult_flag
    In the column ‘Adult_Flag’, the variable is categorical but by default it is ‘int64’. Therefore the data type of the variable was changed to ‘object’.
17. Daily_Avg_mins_spend_on_traveling_page
    This variables tells about the daily average minutes spend by the user on the travelling page.

Understanding the categorical features of data

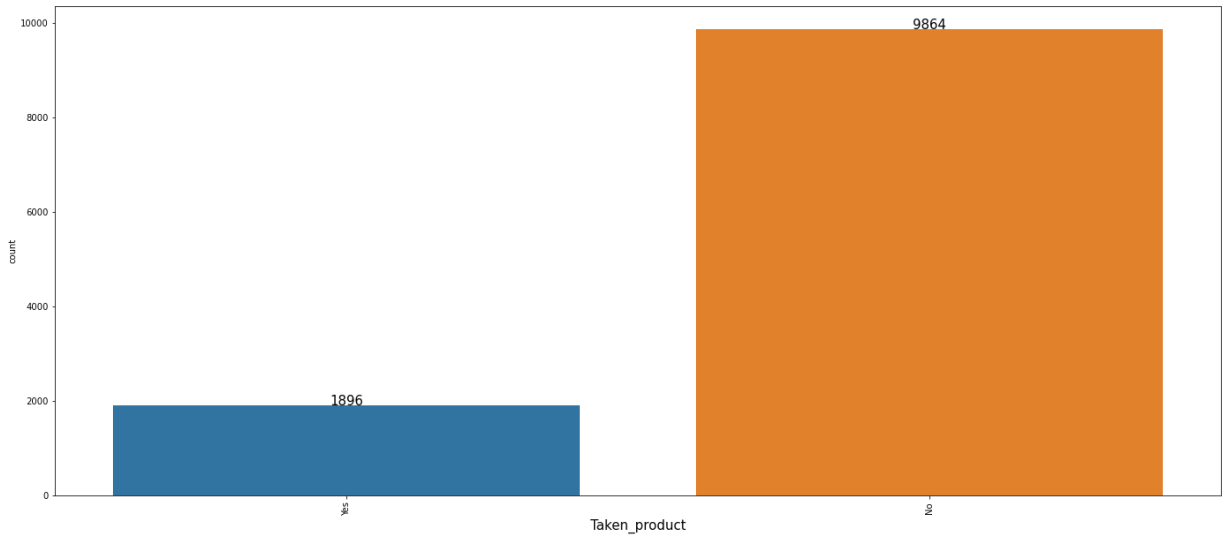

Fig 1

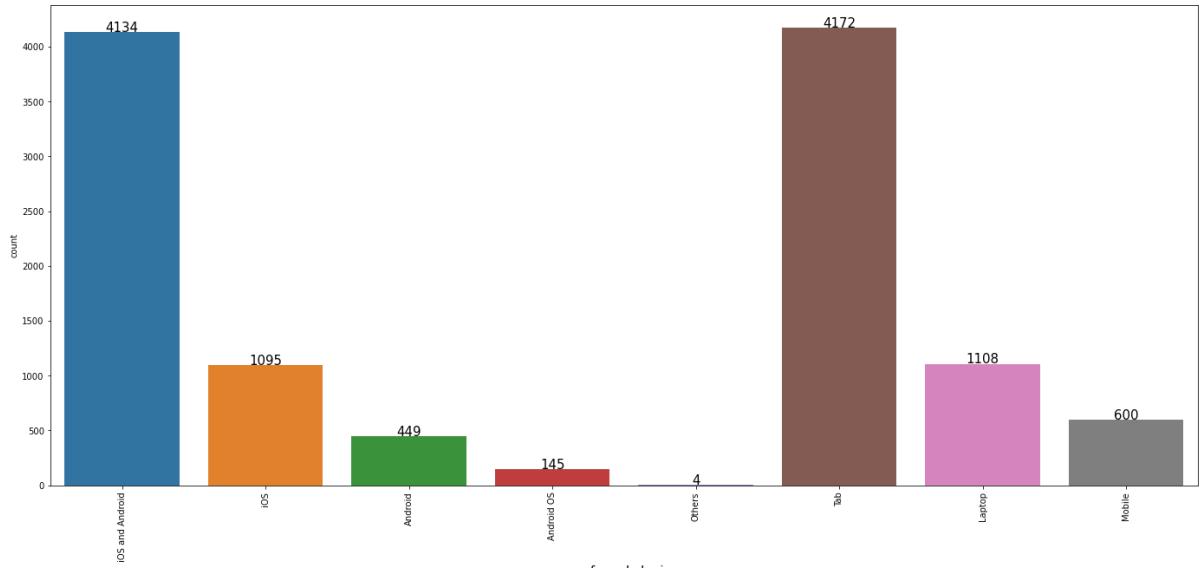
4134
1095
449
145
4
4172
1108
600
count
iOS and Android
iOS
Android
Android OS
Others
Tab
Laptop
Mobile

Fig 2

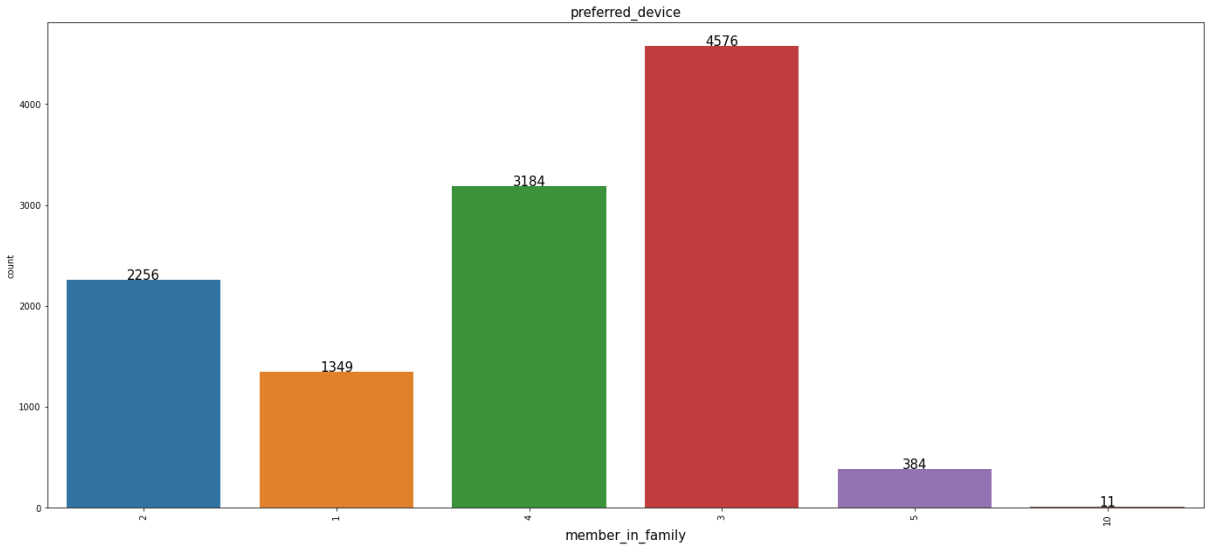
preferred_device
2256
1349
3184
4576
384
11
count
member_in_family

Fig 3

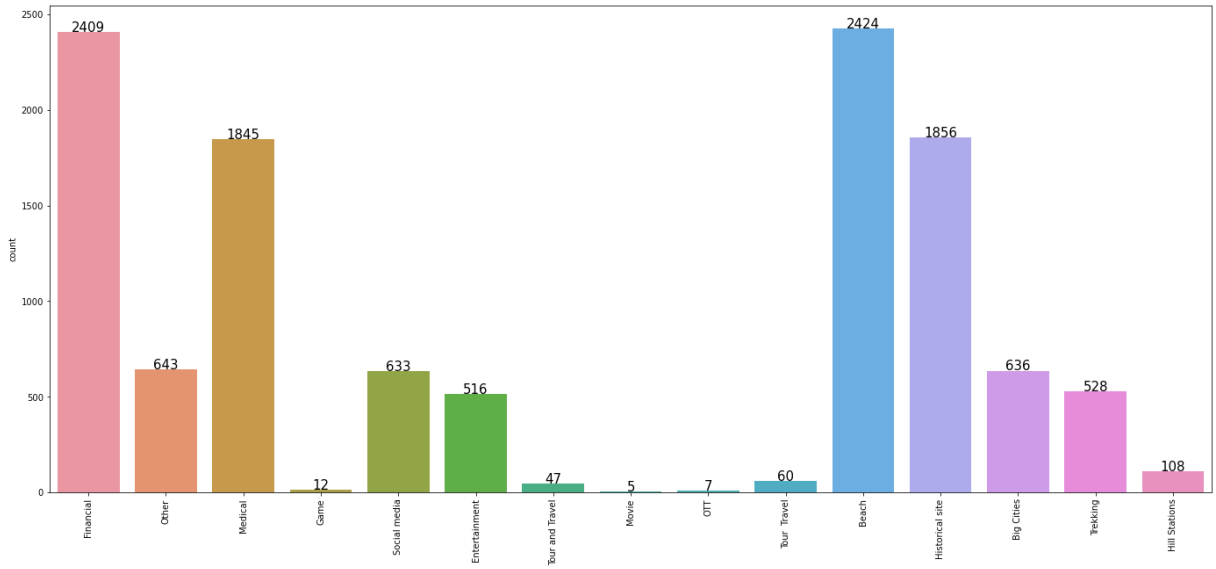
count
2500
2000
1500
1000
500
0
2409
643
1845
12
633
516
47
5
7
60
2424
1856
636
528
108
Financial
Other
Medical
Game
Social media
Entertainment
Tour and Travel
Movie
OTT
Tour Travel
Beach
Historical site
Big Cities
Trekking
Hill Stations

Fig 4

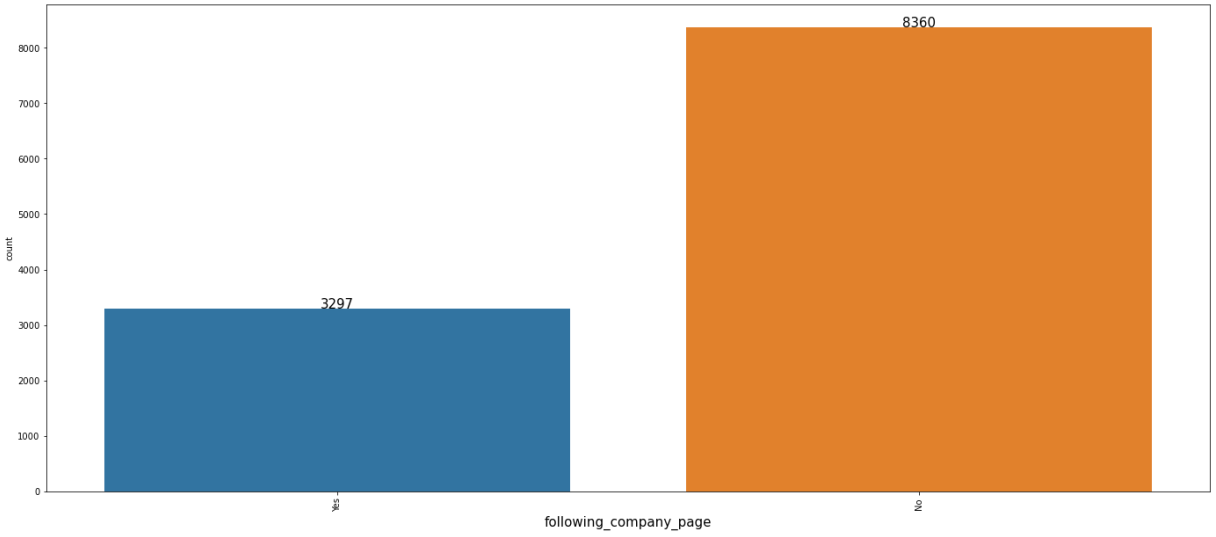
count
8000
7000
6000
5000
4000
3000
2000
1000
0
3297
8360
Yes
No
following_company_page

Fig 5

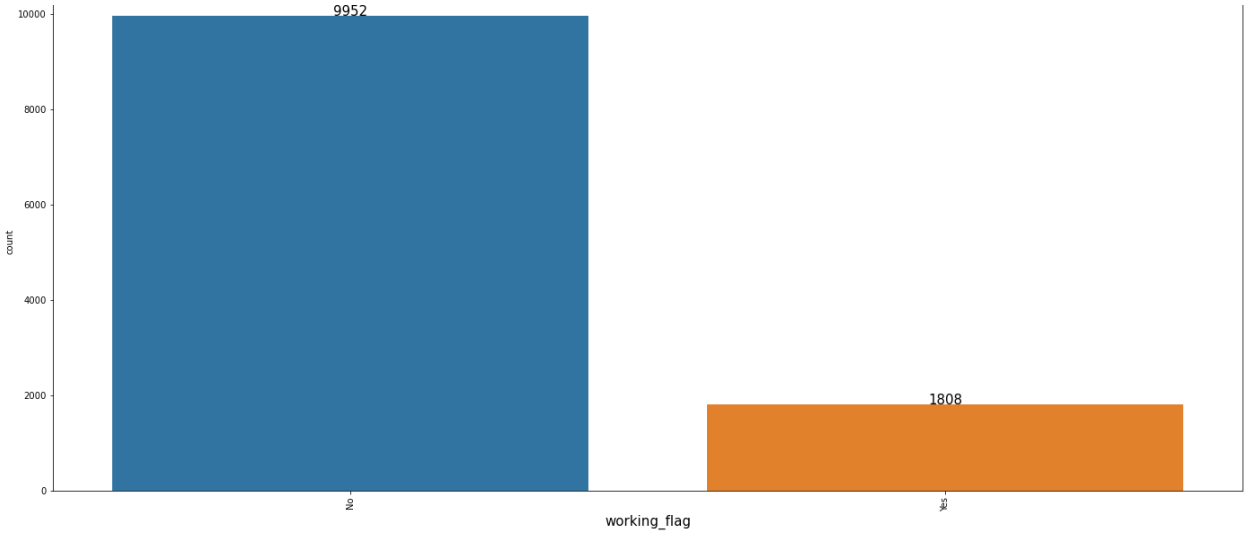
9952
1808
10000
8000
6000
4000
2000
0
count
No
Yes
working_flag

Fig 6

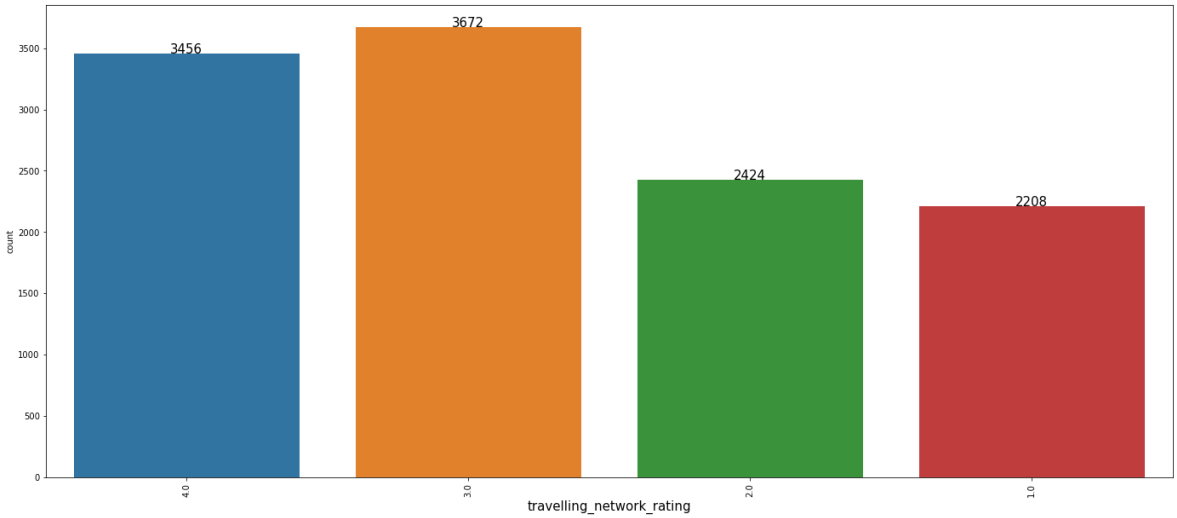

Fig 7

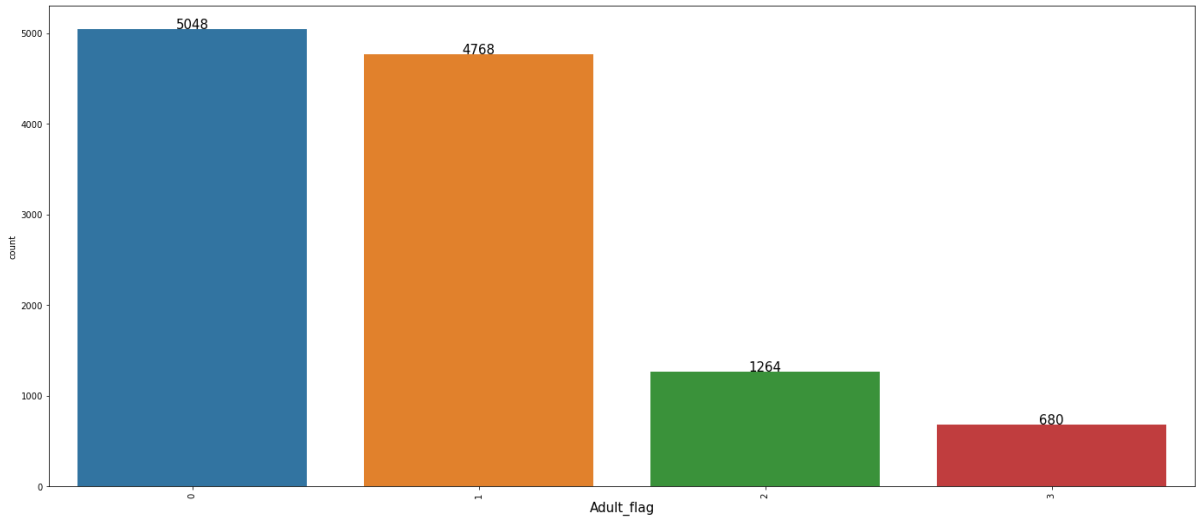

Fig 8

- The above the figures i.e. Fig 1, Fig 2, Fig 3, Fig 4, Fig 5, Fig 6, Fig 7, Fig 8 and Fig 9 belong to the categorical variables.
- In Fig 1, we can see from product taken ‘yes’ that the number of people who bought the product is 9864 and the number of people not going to buy is 1896
- In Fig 2, we understood that the number of prospects is less on Laptop and more on mobiles or tablets. Among non-Laptop devices more number belongs to tab. From operating system (OS), we cannot identify the device because both Android and iOS works well on mobiles as well as tablets.
- From Fig 3, it is visible that most number of families has number of family members as 3. It is followed by 4 members per family.

- From Fig 4, we can understand that the most of the prospects are interested in visiting a beach. It is followed by financial destinations and historical sites respectively. Social media campaign, if aligned with photos related to beach may attract higher traffic.
- From Fig 5, it can be noticed that the number of customers are not following the company page. The number of followers is 3297 where as customers not following are 8390. During, social media campaign videos there should be a reminder given to the customers to follow the page. If they are really interested then they get latest updates, promotions, discounts and other offers launched by the company. This will definitely increase the sale in the travel ticket.
- In Fig 6, we get the number of working customers. We can see that the working people are 9952 and non- working are 1808.
- From Fig 7, the ratings can be understood. The customers have rated 3 stars out 4 in most number of cases. The total of 3 & 4 ratings is 7128. However, the number of customers moderately liking or not giving good rating is also significant.
- Fig 8 belongs to the Adult_flag, the data has some anomalies because of which we can see 4 categories. Actually, categories should be two only i.e. Adult or Not Adult.

a) Bivariate Analysis

We shall understand the bivariate analysis of categorical variables through count plots.

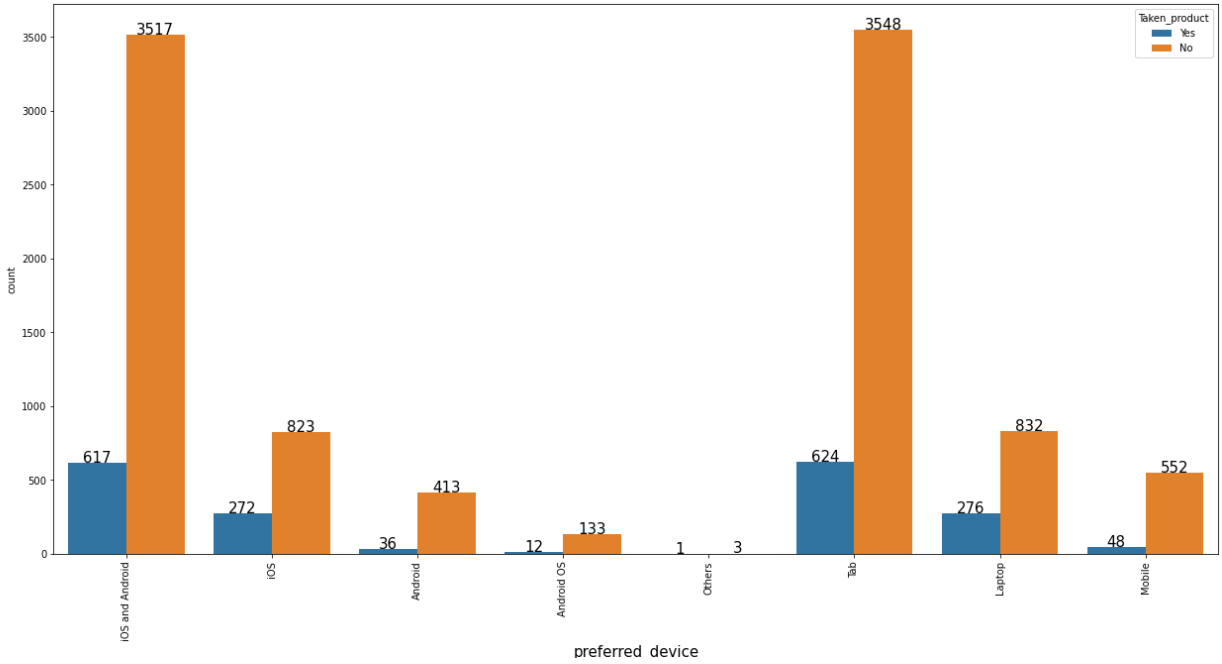

Fig 17

In Fig 17, we can see that most number of customers buying the tickets belongs to 'Tab' and 'iOS and Android'. This means that most of the people buying the ticket are using less of laptop to access the social media campaign. This gives us the understanding that more campaigns should be done on mobile devices compared to

laptops.

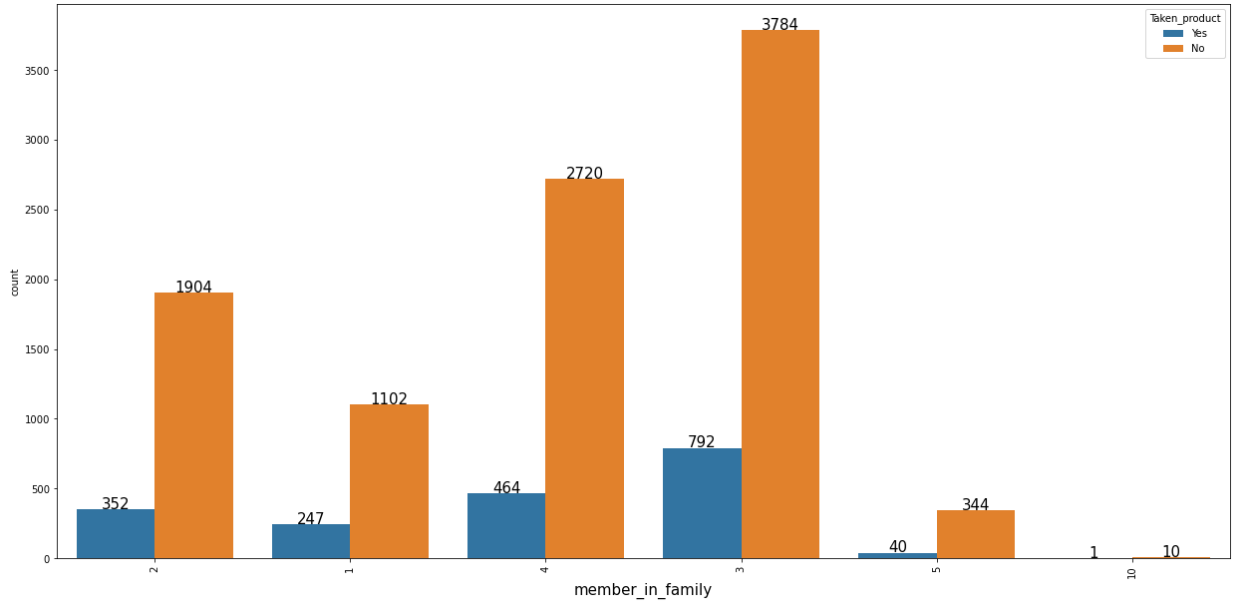

Fig 18

From Fig 18 we can understand that families where number of members are 3 are more likely to buy the ticket. It is followed by number of family member 4 and then 2. We can also understand that where only a member is there has less occurrence of buying. Also, where number of members is large has less occurrence of buying ticket.

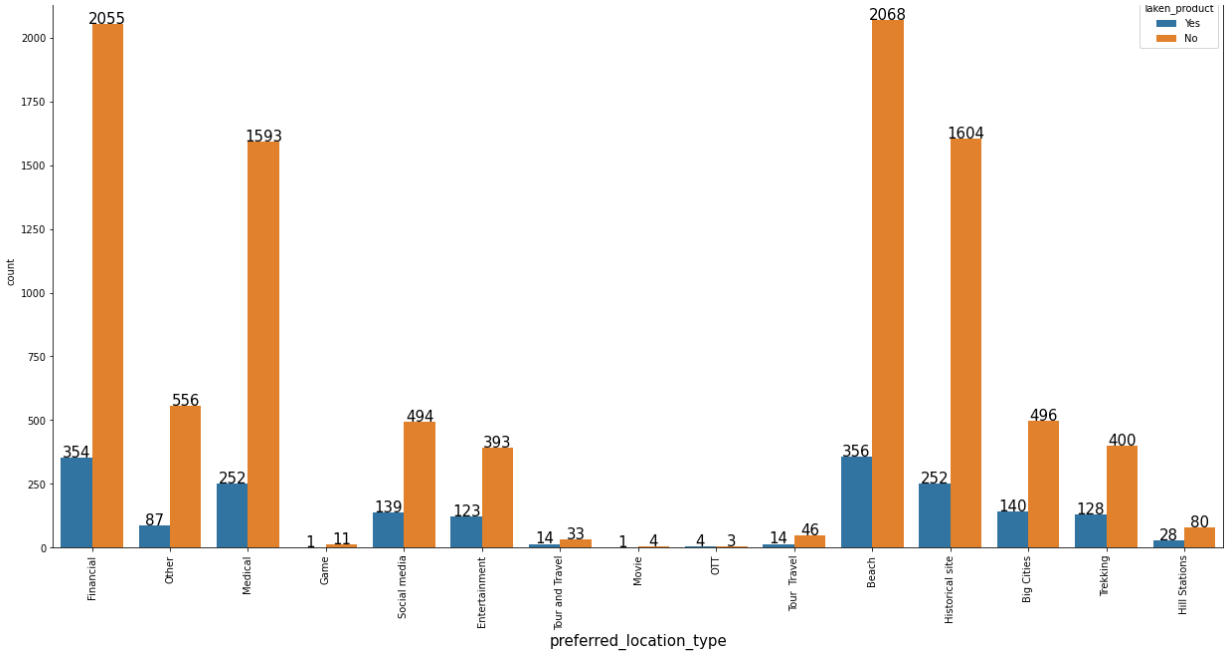

Fig 19

From Fig 19, we can understand that the most favorite destination for people buying the ticket is beach and financial places. The number of customers opting in both the places is almost equal. It is followed by ‘Medical’ and ‘Historical’ places.

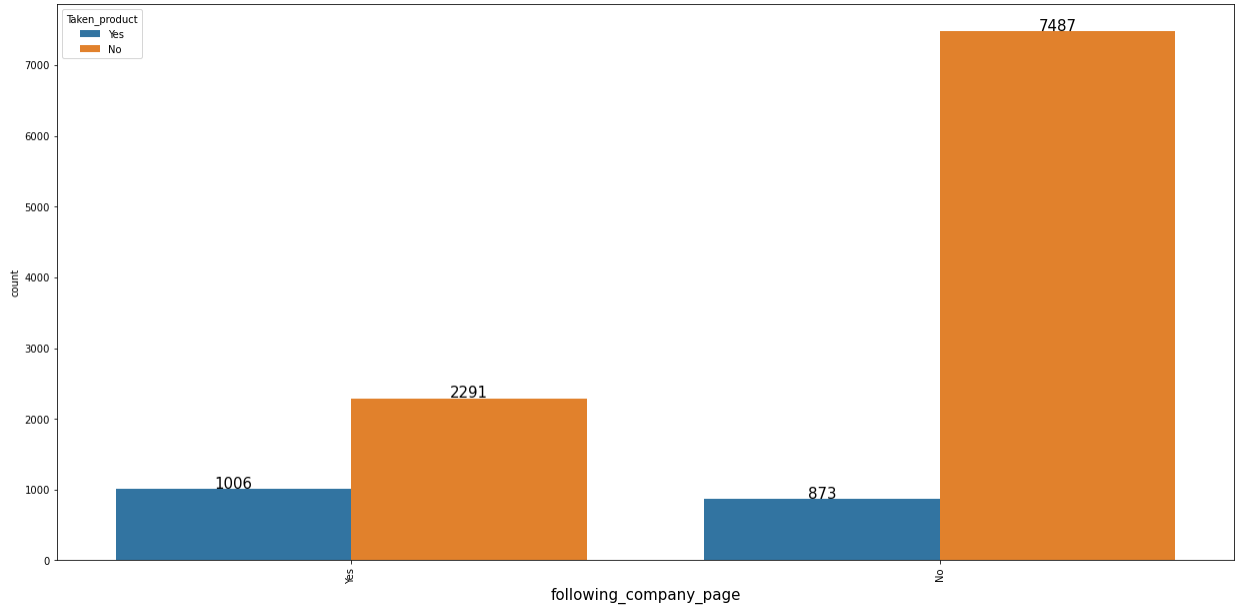

Fig 20

From Fig 20, we can understand that the audience who follow the social media page has more taken the product more than those who do not follow the page.

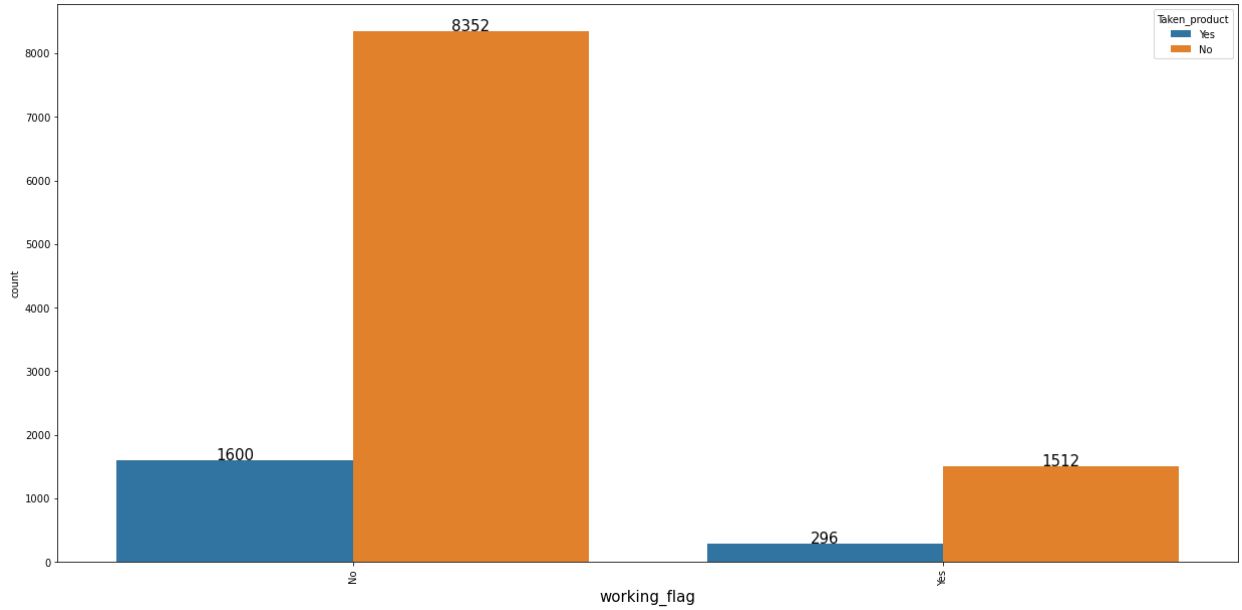

Fig 21

Here from Fig 21, we can understand that working people has very high chances of going to buy the ticket as compared to the non-working audience.

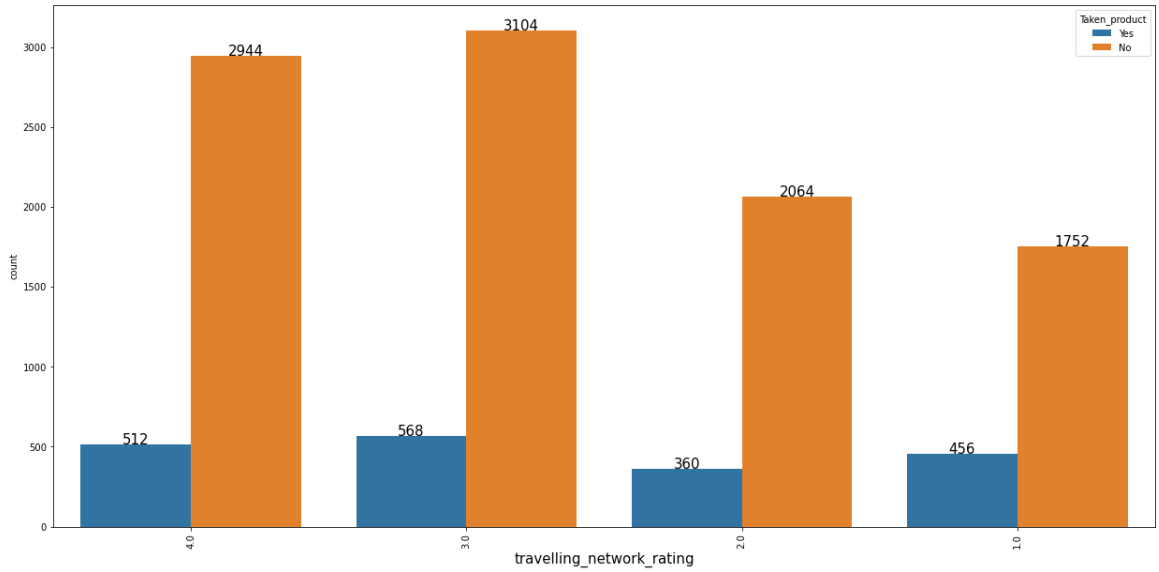

Fig 22

From Fig 22, we can understand that the rating has influenced the buying of ticket. The people who have rated 3 and 4 stars have taken product more as compared to the people giving 1 and 2 rating.

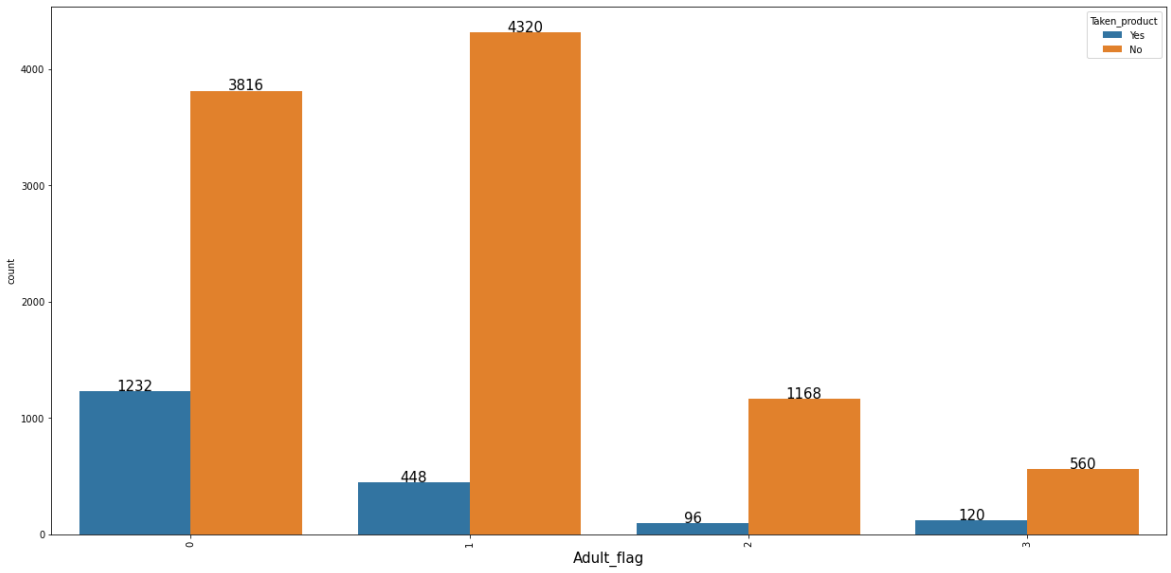

Fig 23

From Fig 23, we can understand that the people who are not adults have opted product more than the people who are adults.

Now, we shall understand the bivariate analysis of numerical variables through the boxplots.

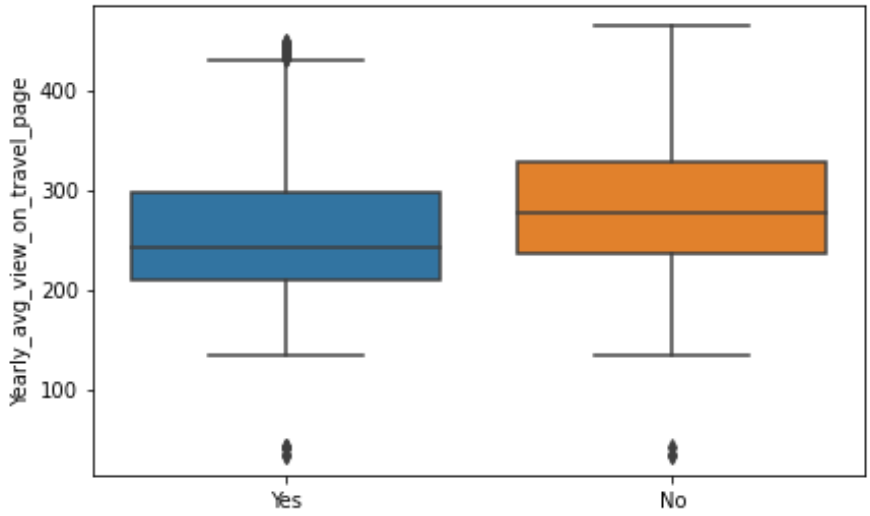

Fig 24

From Fig 24 we can understand the people who have spent significant time on the social media page haven't bought the ticket. The people who have taken the product has less view time average on social media page.

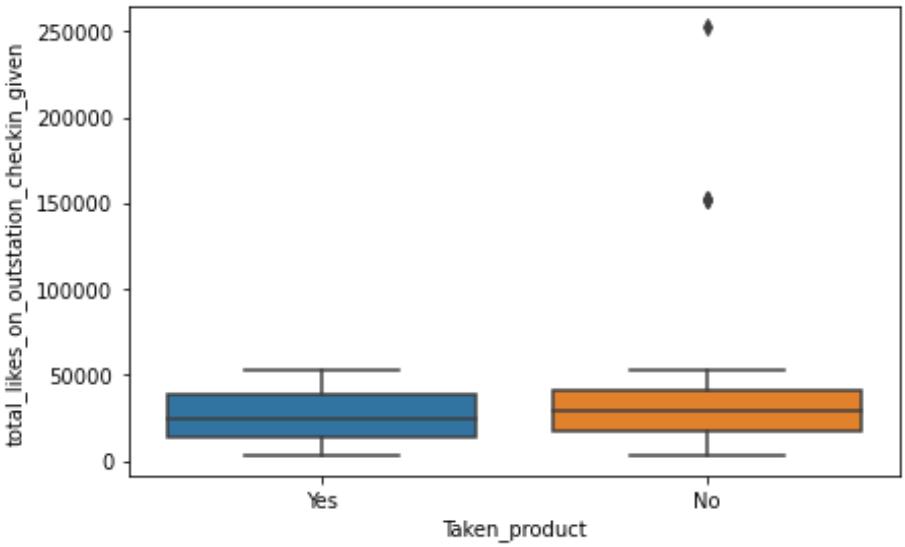

Fig 25

From the Fig 25 we can understand that the people liking the social media has more chances of buying the product.

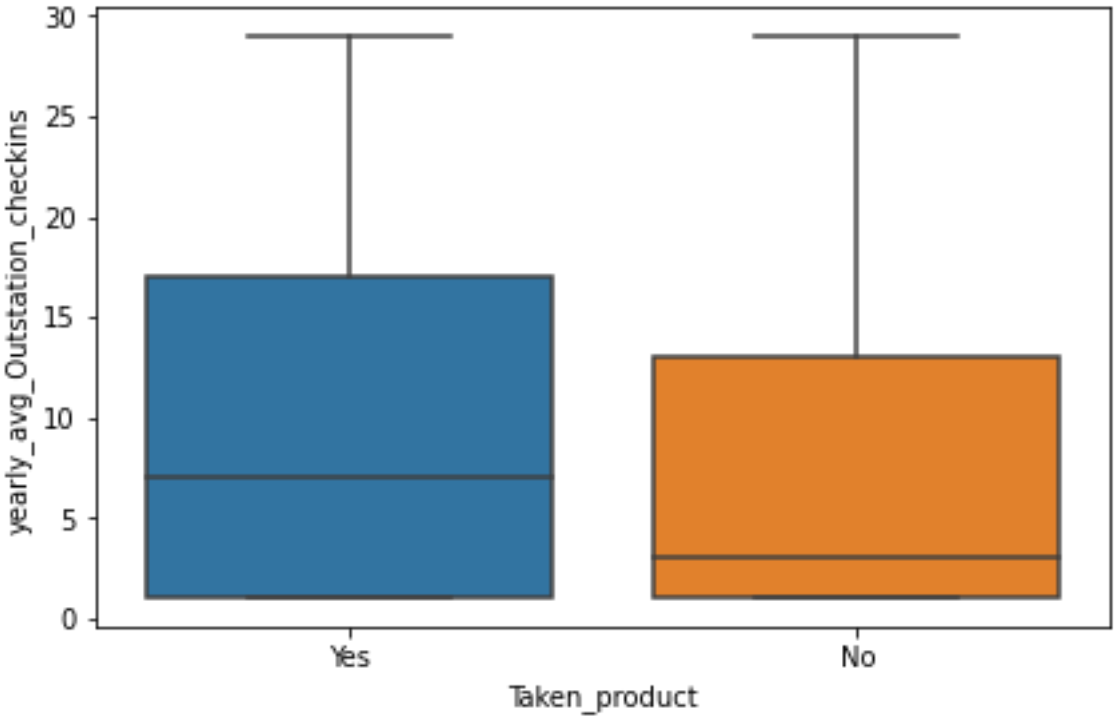

Fig 26

From Fig 26 we can understand that the people often going to outstation trips have more opted to take the product as compared to those who go out very less.

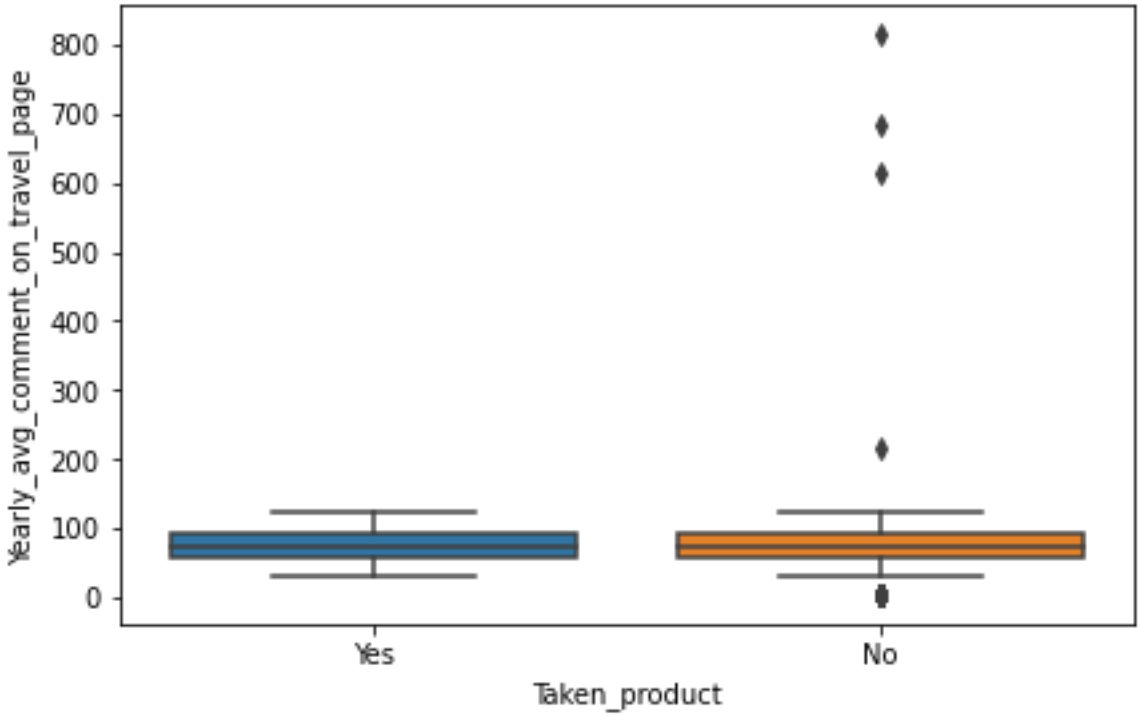

Fig 27

From Fig 27 we can understand that the people commenting on the social media page does not take significant effect on buying the ticket.

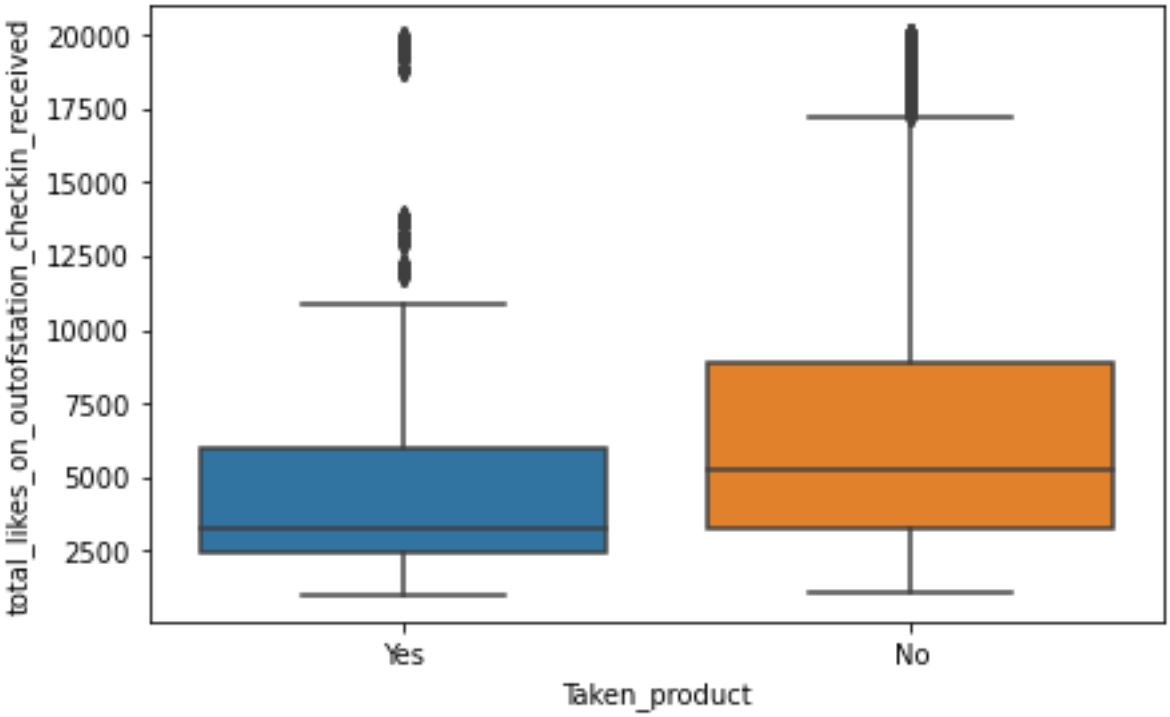

Fig 28

From Fig 28, we can understand that the likes on outstation check-ins does not have significant effect on buying pattern of a customer.

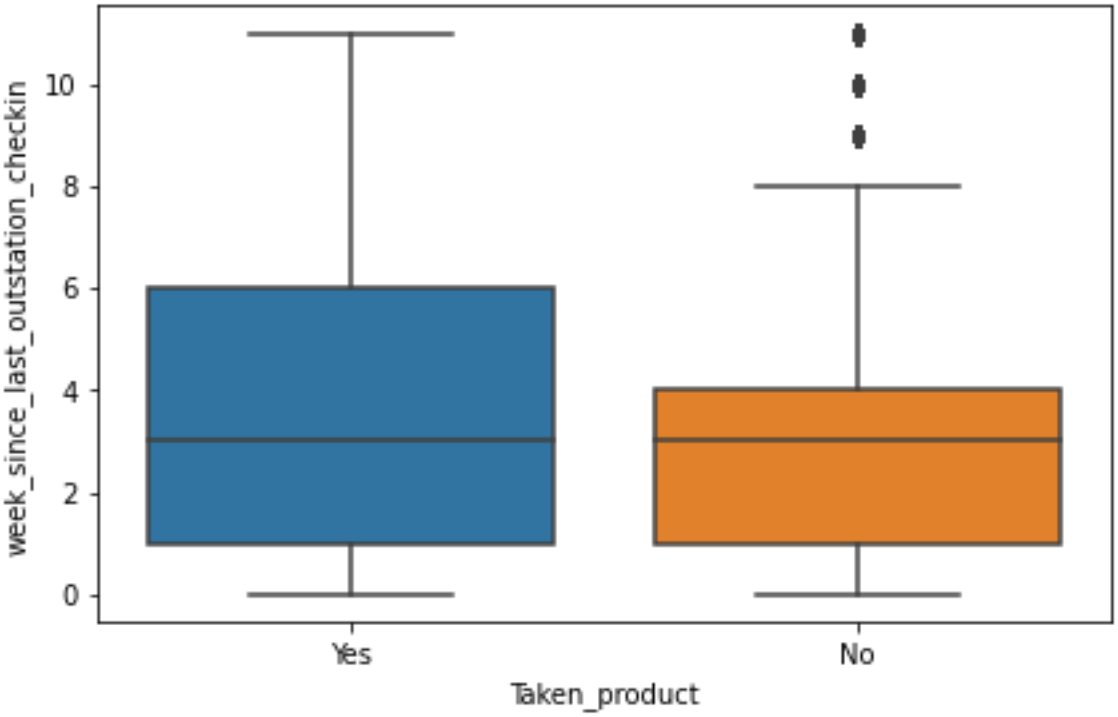

Fig 29

From Fig 29, we can understand that the more weeks since last outstation checking has more opted for going to buy the ticket.

Correlation Matrix

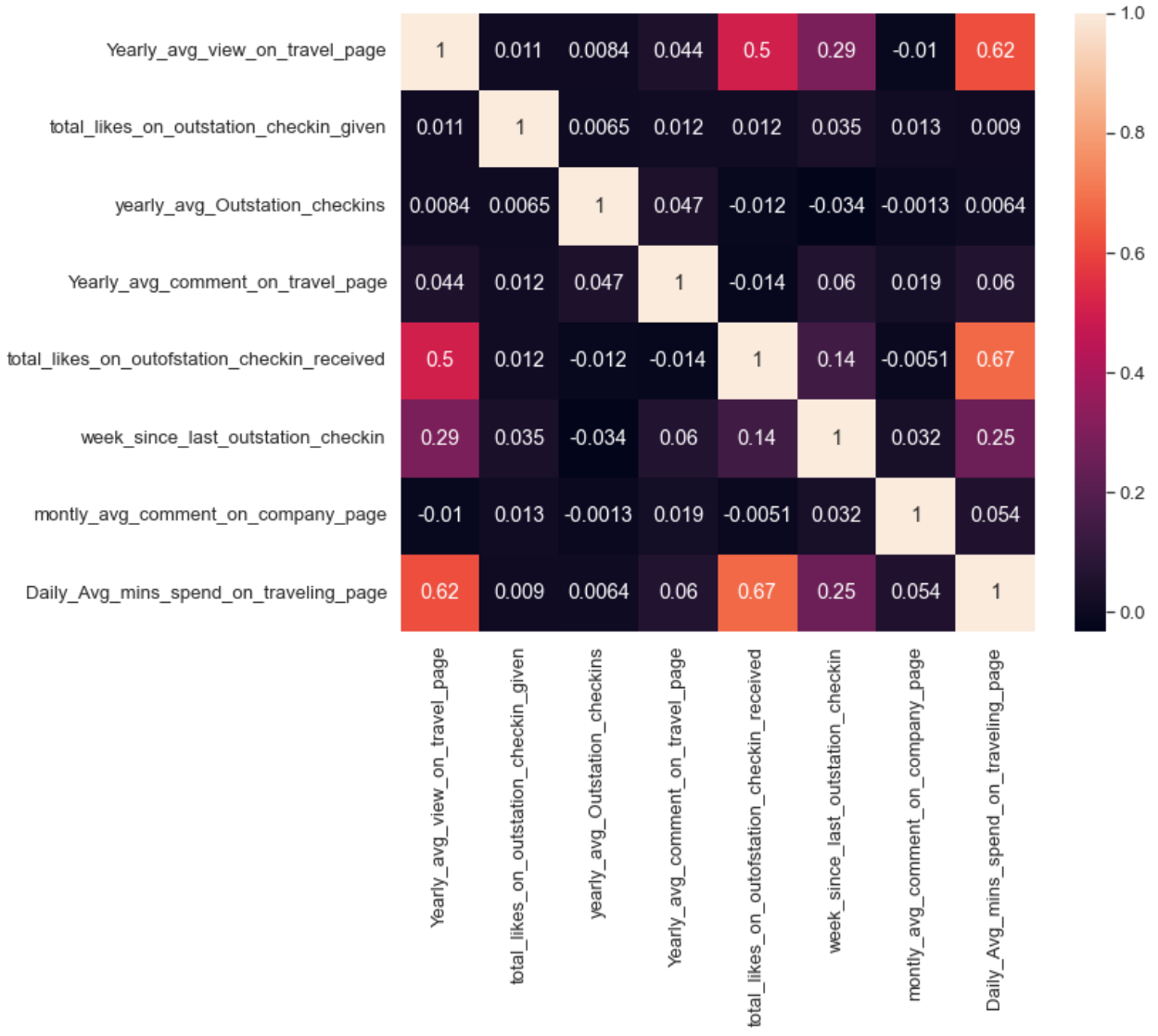

Fig 30

- From Fig 30 we can understand that there is high correlation of 0.67 between “Daily average minutes spend on travelling page” and “total likes on outstation checkin received”
- There is moderate correlation of 0.62 between “Daily average minutes spend on travelling page” and “yearly average view on travel page”

- There is low correlation of 0.5 between “yearly average view on travel page” and “total likes on outstation checkin received”

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)

Missing Value Treatment

During the univariate analysis and the bivariate analysis, we had divided the dataset into two parts. One part composed of ‘categorical’ & ‘object’ features. Whereas the other part composed of integer as well as float data types.

In the categorical variables we have used mode for the missing value treatment.

In case of numerical variables, we have used median imputation method for the missing value treatment. Median is the best measure of central tendency to fill in missing values.

Below are the categorical variables after NULL value treatment.

```
Taken_product                0
preferred_device             0
member_in_family             0
preferred_location_type      0
following_company_page       0
working_flag                 0
travelling_network_rating    0
Adult_flag                   0
dtype: int64
```

Below are the numerical variables after NULL value treatment.

```
Yearly_avg_view_on_travel_page                  0
total_likes_on_outstation_checkin_given         0
yearly_avg_Outstation_checkins                  0
Yearly_avg_comment_on_travel_page               0
total_likes_on_outofstation_checkin_received    0
week_since_last_outstation_checkin              0
montly_avg_comment_on_company_page              0
Daily_Avg_mins_spend_on_traveling_page          0
dtype: int64
```

Below is the data before the outlier treatment.

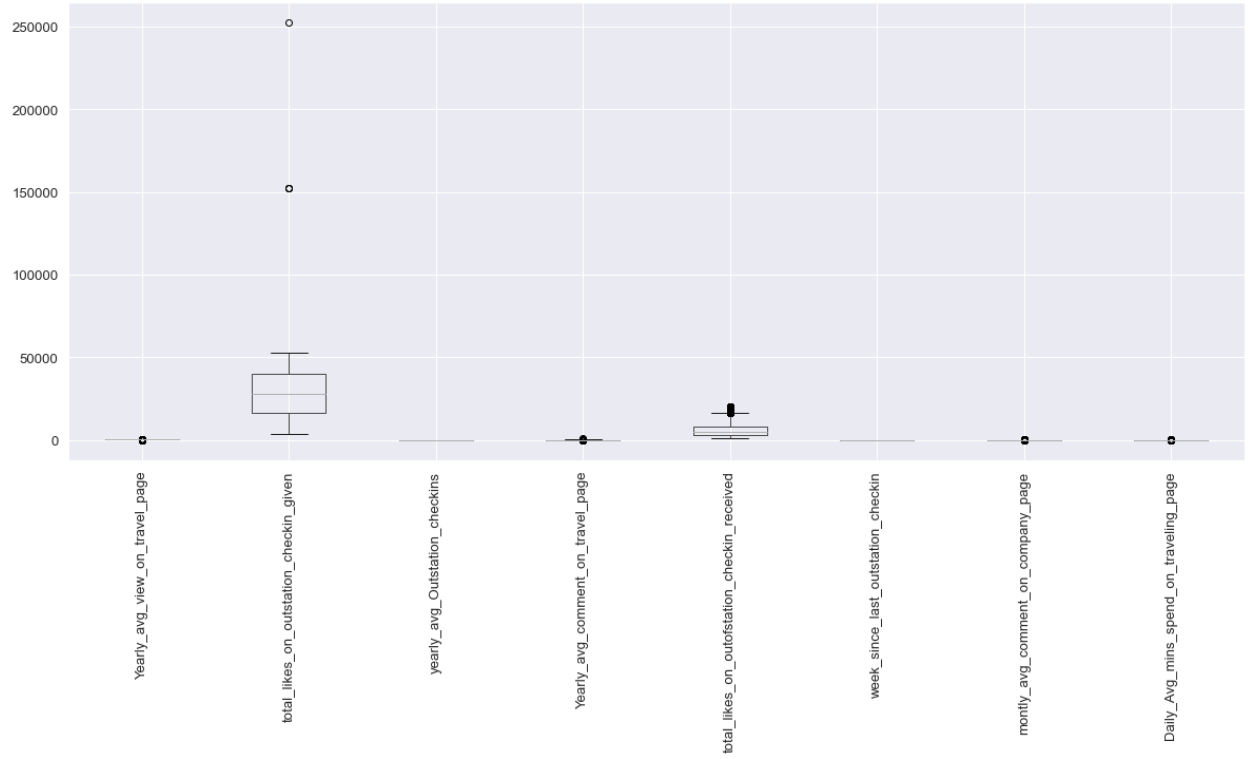

Fig 31

In the above graph Fig 31, we can see that all the variables have outliers except "week since last outstation check-in". We shall be treating the outliers by imputing them with the standard technique of imputing with upper quantile and lower quantile limits. The upper value is calculated by Q3+(1.5 * IQR) & lower value is calculated by Q1-(1.5 * IQR). After imputation the data looks like the following image, Fig 32.

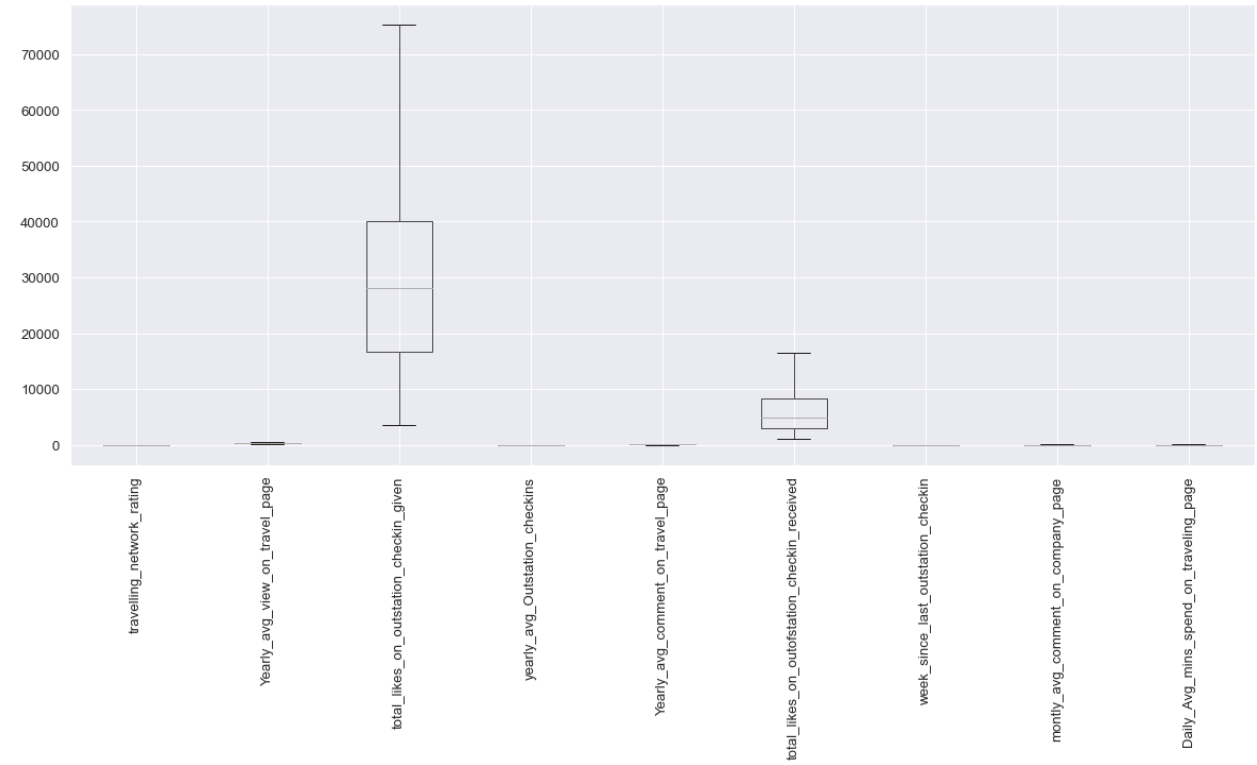

Fig 32

Variable Transformation & Addition of new variable

- In the beginning of the project only we understood the fact that the social media page of the company is viewed by majorly two types of devices. These devices are Laptop and Mobile. The categories like tablet, Android, iOS, Mobile and Others shall fall under category “Mobile”. Remaining data points shall come under category “Laptop”. Therefore, we have to we have done variable transformation of all the variables which are not Laptop into “Mobile”. A new variable was created labelled “Mobile_Or_Laptop”. Subsequently a new variable was created called “Labelled_Mobile_Or_Laptop” where “Laptop” is labelled 0 and “Mobile” was labelled as 1.

- In case of working_flag, the data points with ‘Yes’ are labelled as 1 and ‘No’ are labelled as 0. These changes are incorporated in a new variable called “Labelled_working_flag”.

- In case of Taken_product, the data points with ‘Yes’ are labelled as 1 and ‘No’ are labelled as 0. These changes are incorporated in a new variable called “Labelled_ Taken_product”.

- In case of following_company_page, the data points with ‘Yes’ are labelled as 1 and ‘No’ are labelled as 0. These changes are incorporated in a new variable called “Labelled_ following_company_page”.

- In case of preferred_location_type, the data points were labelled from 14 to 1. 14 is the most preferred location, whereas 1 is the least preferred location. These inferences were drawn from the frequency of occurrence of each destination which is mentioned below for reference. These changes are incorporated in a new variable called “Labelled_ 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

- The variables ‘member_in_family’, ‘yearly_avg_Outstation_checkins’ and ‘Adult_flag’ were converted from categorical variables to float for further analysis.

Removal of unwanted variables

The following variables were removed-

- 'preferred_device'- Because it was converted into ‘Mobile_Or_Laptop’
- 'preferred_location_type'- It was labelled 1 to 14 and new variable ’Labelled_preferred_location_type’ was created
- 'following_company_page'- Converted to 1 & 0 in new labelled column

- 'working_flag'- Converted to 1 & 0 in new labelled column
- 'Mobile_Or_Laptop'- Converted to 1 & 0 in new labelled column
- 'Taken_product'- Converted to 1 & 0 in new labelled column
- The statsmodel technique has been applied to the variables to eliminate the variables which are not contributing. Here after removing the highest p-value in repeated models, the below final variables were obtained. The p-value considered here has to be less than 0.05. Therefore, features were tried and tested manually using backward elimination approach.

Logit Regression Results

Dep. Variable:	Labelled_Taken_product	**No. Observations:**	7879
Model:	Logit	**Df Residuals:**	7867
Method:	MLE	**Df Model:**	11
Date:	Sun, 07 Nov 2021	**Pseudo R-squ.:**	0.1937
Time:	21:53:50	**Log-Likelihood:**	-2805.7
converged:	True	**LL-Null:**	-3479.6
Covariance Type:	nonrobust	**LLR p-value:**	2.164e-282

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.1715	0.254	8.540	0.000	1.673	2.670
travelling_network_rating	-0.2125	0.032	-6.714	0.000	-0.275	-0.150
Adult_flag	-0.6138	0.047	-13.175	0.000	-0.705	-0.522
Yearly_avg_view_on_travel_page	-0.0038	0.001	-5.810	0.000	-0.005	-0.003
total_likes_on_outstation_checkin_given	-1.182e-05	2.45e-06	-4.816	0.000	-1.66e-05	-7.01e-06
yearly_avg_Outstation_checkins	0.0356	0.004	9.216	0.000	0.028	0.043
total_likes_on_outofstation_checkin_received	-9.317e-05	1.37e-05	-6.796	0.000	-0.000	-6.63e-05
week_since_last_outstation_checkin	0.1537	0.013	11.420	0.000	0.127	0.180
Daily_Avg_mins_spend_on_traveling_page	-0.0433	0.007	-5.938	0.000	-0.058	-0.029
Labelled_Mobile_Or_Laptop	-0.7637	0.103	-7.413	0.000	-0.966	-0.562
Labelled_following_company_page	1.5742	0.070	22.357	0.000	1.436	1.712
Labelled_preferred_location_type	-0.1054	0.013	-8.117	0.000	-0.131	-0.080

Fig-33

The variables eliminated in the process were 'Yearly_avg_comment_on_travel_page', 'member_in_family', 'Labelled_working_flag', and 'montly_avg_comment_on_company_page'.

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

This is a classification problem. Therefore after the split, below mentioned Machine Learning Algorithms were applied separately on Laptop & Mobile devices:
Logistic Regression
Linear Discriminant Analysis
Naive Bayes Model
Decision Tree Classifier
Random Forest Classifier
K- Nearest Neighbour
Model tuning (Bagging)
Model tuning (Adaboosting & Gradient Boosting)

Comparing all the models (Laptop)

	LR Train	LR Test	LDA Train	LDA Test	NB Train	NB Test	CART Train	CART Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	KNN Train	KNN Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.739	0.678	0.734	0.667	0.707	0.682	1.0	0.953	1.0	0.987	1.0	0.955	0.890	0.840	0.969	0.851	0.951	0.885
Recall	0.756	0.696	0.770	0.703	0.813	0.784	1.0	0.959	1.0	1.000	1.0	1.000	0.876	0.818	1.000	1.000	0.966	0.939
F1 Score	0.747	0.687	0.752	0.684	0.756	0.730	1.0	0.956	1.0	0.993	1.0	0.977	0.883	0.829	0.984	0.919	0.959	0.911
Accuracy	0.737	0.718	0.739	0.712	0.731	0.742	1.0	0.961	1.0	0.994	1.0	0.979	0.881	0.850	0.983	0.922	0.957	0.919
AUC Score	0.830	0.824	0.829	0.822	0.815	0.823	1.0	0.961	1.0	1.000	1.0	0.999	0.962	0.943	1.000	1.000	0.992	0.943

Fig 34

Inferences for Laptop devices based on Model Building(Fig 34)

A. Logistic Regression
The Logistic Regression has come up with poor F1 Score of 74.7% on the train set and only 68.7% on test set. The precision also is not up to the mark i.e. 73.9% for train and 67.8% for the test datasets.

B. Linear Discriminant Analysis
The LDA has also not performed well. It has given an F1 Score of 75.2% on Train set and 68.4% on the Test set. The precision also isn't up to the mark. Precision is only 73.4% for train set and 66.7 for the test set.

C. Naïve Bayes Model

The Naïve Bayes model has also not shown very poor performance. The F1 Score is 75.6% and 68.4% on train as well as test split datasets. The precision is poor for this model. It is 70.7% & 68.2% for train as well as test sets.

D. Decision Tree Classifier

The Decision Tree Classifier (CART) model has performed reasonably well. The accuracy score for train and test dataset is 100% and 96.1%. The precision is also very good. It is 100% and 95.3% for train and test set respectively. The F1 Score here is 100% and 95.6% on train and test respectively.

E. Random Forest Classifier

The performance of the Random Forest Classifier is the best amongst all the models. The F1 Score here is 100% and 99.3% on train and test respectively. The accuracy is 100% for the train and 99.4% for the test data sets. The precision is also 100% for training dataset and 98.7% testing dataset.

F. K- Nearest Neighbour

This model has performed fairly but not up to the mark. The accuracy is 98.3% for train and 92.2% for the test dataset. The F1 Score here is 98.4% and 91.9% on train and test respectively.

Model Tuning (Bagging & Boosting)

After applying the model tuning technique bagging to the model the performance received was good. F1-Score for train dataset was 100% and 97.7% for the test. In the case of boosting the Gradient boosting gave F1 Score of 95.9% on train and 91.1% on the test datasets.

The final recommendation to business shall be move ahead with Random Forest Classifier, where on the test dataset the False Negatives were 0 and False Positives were only 2.

Comparing all the models(Mobile)

	LR Train	LR Test	LDA Train	LDA Test	NB Train	NB Test	CART Train	CART Test	RFC Train	RFC Test	Bagging Train	Bagging Test	Ada Boosting Train	Ada Boosting Test	KNN Train	KNN Test	Gradient Boosting Train	Gradient Boosting Test
Precision	0.710	0.712	0.708	0.716	0.658	0.662	1.0	0.987	1.0	0.997	1.0	0.994	0.799	0.795	0.984	0.975	0.853	0.847
Recall	0.706	0.733	0.702	0.732	0.739	0.764	1.0	0.985	1.0	0.996	1.0	0.992	0.797	0.811	1.000	0.998	0.822	0.829
F1 Score	0.708	0.723	0.705	0.724	0.696	0.710	1.0	0.986	1.0	0.996	1.0	0.993	0.798	0.803	0.992	0.986	0.837	0.838
Accuracy	0.709	0.717	0.707	0.719	0.678	0.685	1.0	0.986	1.0	0.996	1.0	0.993	0.799	0.800	0.992	0.986	0.840	0.839
AUC Score	0.777	0.783	0.776	0.782	0.757	0.765	1.0	0.986	1.0	1.000	1.0	1.000	0.888	0.886	1.000	0.997	0.930	0.886

Figure- 35

Inferences for Mobile devices based on Model Building(Fig-35)

A. Logistic Regression

The Logistic Regression has come up with poor F1 Score of 70.8% on the train set and only 72.3% on test set. The precision also is not up to the mark i.e. 71.0% for train and 71.2% for the test datasets.

B. Linear Discriminant Analysis

The LDA has also not performed well. It has given an accuracy of 70.7% on Train set and 71.9% on the Test set. The precision also isn't up to the mark. Precision is only 70.8% for train set and 71.6% for the test set. The F1 Score is 70.5% for train and 72.4% for the test.

C. Naïve Bayes Model

The Naïve Bayes model has also not shown very poor performance. The accuracy score is 67.8% and 68.5% on train as well as test split datasets. The precision is poor for this model. It is 65.8% & 66.2% for train as well as test sets. The F1 Score is 69.6% for train and 71.0% for the test.

D. Decision Tree Classifier

The Decision Tree Classifier (CART) model has performed reasonably well. The accuracy score for train and test dataset is 100% and 98.6%. The precision is also very good. It is 100% and 98.7% for train and test set respectively. The F1 Score is 100% for train and 98.6% for the test.

E. Random Forest Classifier

The performance of the Random Forest Classifier is the best amongst all the models. The accuracy is 100% for the train and 99.6% for the test data sets. The precision is also 100% for training dataset and 99.7% testing dataset. The number of False positives are 6 and False Negatives is 8. The F1 Score is 100% for train and 99.6% for the test.

F. K- Nearest Neighbour

This model has performed fairly but not as good as Random Forest Classifier. The accuracy is 99.2% for train and 98.6% for the test dataset. The number of False positives are 47 and False Negatives are 3.

Model Tuning (Bagging & Boosting)

After applying the model tuning technique bagging to the model the performance received was good. F1-Score for train dataset was 100% and 99.3% for the test. In the case of boosting the Gradient boosting gave F1 Score of 83.7% on train and 83.8% on the test datasets.

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

The model was validated using the F1 Score not the accuracy because the data is highly imbalanced as the percentage of Users buying ticket is 16.12. The percentage of Users not buying ticket is 83.88.

Laptop

The final recommendation to business shall be move ahead with Random Forest Classifier, where on the test dataset the False Negatives were 0 and False Positives were only 2.

Mobile

The final recommendation to business shall be move ahead with Random Forest Classifier for mobile devices. The number of False Positives and False Negatives are very high in case of K-Nearest Neighbor.

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

- The budget allocation for campaigns should be in a ratio of 75:25 for Laptop and Mobile respectively considering the traffic and probability of buying
- Social media campaigns, if aligned with photos related to the beach may attract higher traffic
- In social media campaign videos, there should be a reminder given to the customers to follow the page
- By following the social media page the customers will get the latest updates, promotions, discounts, and other offers launched by the company. This will increase the sale of the travel ticket.
- Working people have a high probability of buying the product therefore campaigns should address their concerns
- Based on the model building approach used for both the devices laptop and mobile, we can finally conclude that the Random Forest Classifier shall be the best model for predicting the likeness of a customer for buying the product through the social media campaign by the GO-GO Air company.