Capstone Project

Project Notes -1

Contents:

Problem Statement.

Business Objective:

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

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

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

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

Introduction of the business problem:

Introduction of the business problem

Defining problem statement:

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

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

Need of the study/project:

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

Understanding business/social opportunity:

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

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

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

Data Report:

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

Data includes:

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

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

Understanding of attributes (variable info, renaming if required)

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

Exploratory data analysis:

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

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

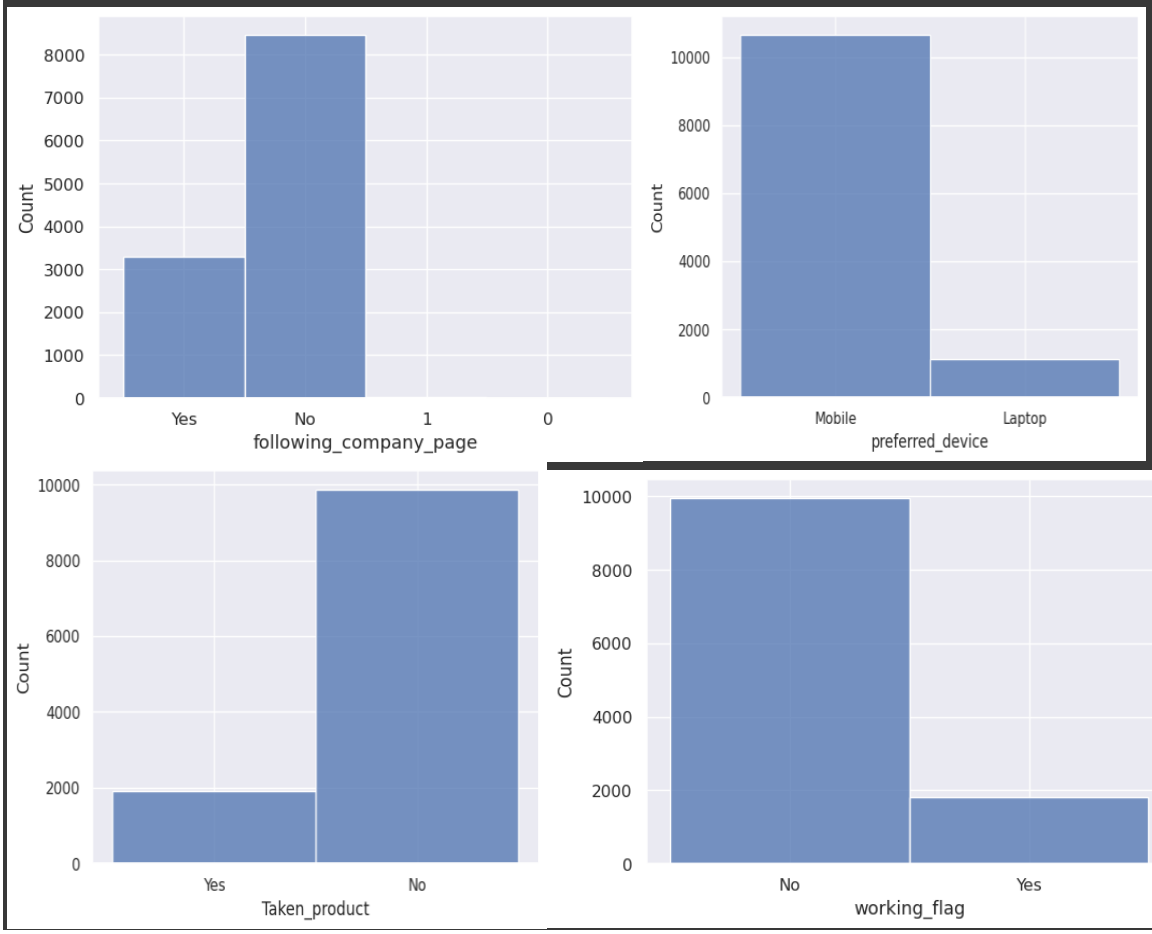
3. Missing values are treated with median and mode for both numerical and categorical columns respectively.

4. Outliers will be treated as per IQR.

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

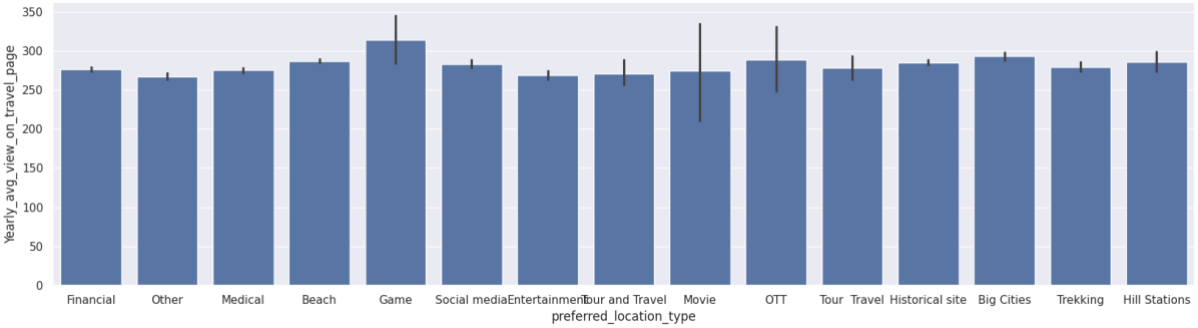
Univariate analysis.

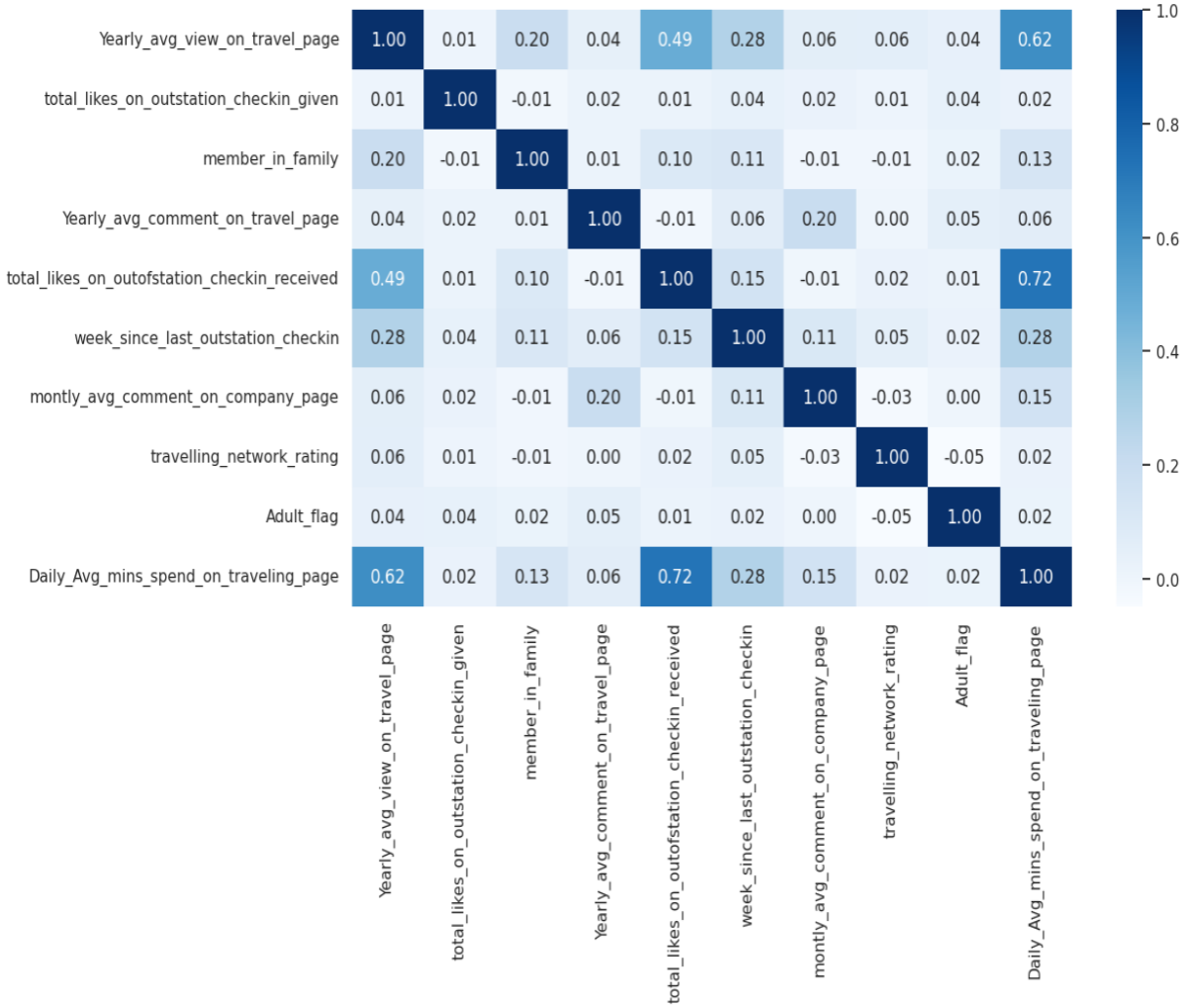
Some of the plots for reference:

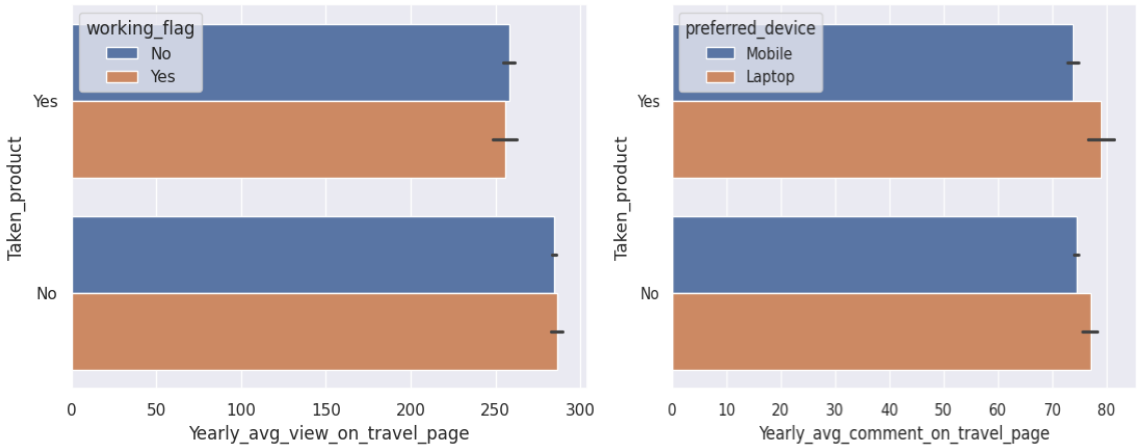


- With the above analysis it shows that more people uses mobile and less using laptops
- Product has been taken by less people.
- Most of the people visiting sites seems to be not working.
- Also, less people following the company.

Bivariate Analysis:







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

Business insights from EDA:

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

Model building and interpretation.

Checking top 5 rows again to verify the attributes after processing.

```
df.head()
```

	UserID	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	m
0	1000001	1	307.0	1	38570.0	1	
1	1000002	0	367.0	1	9765.0	1	
2	1000003	1	277.0	1	48055.0	1	
3	1000004	0	247.0	1	48720.0	1	
4	1000005	0	202.0	1	20685.0	1	

Split the data into training and test set in 70:30 ratio

Checking both the train & test data again.

```
[388] display(Train.head())

      print(Train.shape)
```

	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	member_
7187	0	217.0	1	42600.0	1	
1892	1	271.0	1	17115.0	1	
4269	0	219.0	1	12365.0	1	
6746	0	311.0	0	41450.0	2	
3790	0	242.0	1	30210.0	1	

(8232, 17)

```
display(Test.head())

print(Test.shape)
```

	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	member_
8032	0	250.0	1	41449.0	1	
8356	1	375.0	0	51637.0	28	
7760	0	228.0	1	11103.0	2	
4974	0	339.0	1	33433.0	1	
761	1	270.0	1	23100.0	15	

(3528, 17)

Model 1 with all data

OLS Regression Results

Dep. Variable:	Taken_product	R-squared:	0.154
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	99.71
Date:	Tue, 06 Feb 2024	Prob (F-statistic):	3.67e-284
Time:	15:59:06	Log-Likelihood:	-2769.2
No. Observations:	8232	AIC:	5570.
Df Residuals:	8216	BIC:	5683.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5298	0.031	16.837	0.000	0.468	0.592
Yearly_avg_view_on_travel_page	-0.0006	7.45e-05	-7.993	0.000	-0.001	-0.000
preferred_device	-0.0998	0.013	-7.669	0.000	-0.125	-0.074
total_likes_on_outstation_checkin_given	-1.426e-06	2.69e-07	-5.295	0.000	-1.95e-06	-8.98e-07
yearly_avg_Outstation_checkins	0.0034	0.000	7.899	0.000	0.003	0.004
member_in_family	0.0051	0.004	1.365	0.172	-0.002	0.012
preferred_location_type	0.0014	0.001	1.044	0.296	-0.001	0.004
Yearly_avg_comment_on_travel_page	0.0001	0.000	0.696	0.486	-0.000	0.000
total_likes_on_outofstation_checkin_received	-6.911e-06	1.27e-06	-5.438	0.000	-9.4e-06	-4.42e-06
week_since_last_outstation_checkin	0.0176	0.002	11.475	0.000	0.015	0.021
following_company_page	0.1973	0.008	23.615	0.000	0.181	0.214
montly_avg_comment_on_company_page	-0.0013	0.001	-2.193	0.028	-0.003	-0.000
working_flag	0.0054	0.012	0.471	0.637	-0.017	0.028
travelling_network_rating	-0.0215	0.004	-6.132	0.000	-0.028	-0.015
Adult_flag	-0.0997	0.006	-17.335	0.000	-0.111	-0.088
Daily_Avg_mins_spend_on_traveling_page	-0.0029	0.001	-3.612	0.000	-0.004	-0.001

Omnibus:	1877.427	Durbin-Watson:	2.020
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3448.153
Skew:	1.474	Prob(JB):	0.00
Kurtosis:	4.165	Cond. No.	2.74e+05

Model did not performed too well. will check and VIF and see with dropping high VIF attributes making any difference.

VIF calculation for 'Daily_Avg_mins_spend_on_traveling_page' & 'total_likes_on_outofstation_checkin_received' is on a higher side will drop them and find if we getting a better model.

Model 2 with deleted attributes:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Taken_product   R-squared:                       0.154
Model:                            OLS   Adj. R-squared:                  0.152
Method:                 Least Squares   F-statistic:                     99.71
Date:                Tue, 06 Feb 2024   Prob (F-statistic):          3.67e-284
Time:                        16:04:04   Log-Likelihood:                -2769.2
No. Observations:                8232   AIC:                             5570.
Df Residuals:                    8216   BIC:                             5683.
Df Model:                          15
Covariance Type:            nonrobust
================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                          0.5298      0.031     16.837      0.000       0.468       0.591
Yearly_avg_view_on_travel_page    -0.0006   7.45e-05     -7.993      0.000      -0.001      -0.000
preferred_device                  -0.0998      0.013     -7.669      0.000      -0.125      -0.074
total_likes_on_outstation_checkin_given -1.426e-06   2.69e-07     -5.295      0.000   -1.95e-06   -8.98e-07
yearly_avg_Outstation_checkins     0.0034      0.000      7.899      0.000       0.003       0.004
member_in_family                   0.0051      0.004      1.365      0.172      -0.002       0.012
preferred_location_type            0.0014      0.001      1.044      0.296      -0.001       0.004
Yearly_avg_comment_on_travel_page  0.0001      0.000      0.696      0.486      -0.000       0.000
total_likes_on_outofstation_checkin_received -6.911e-06   1.27e-06     -5.438      0.000   -9.4e-06   -4.42e-06
week_since_last_outstation_checkin 0.0176      0.002     11.475      0.000       0.015       0.021
following_company_page             0.1973      0.008     23.615      0.000       0.181       0.214
montly_avg_comment_on_company_page -0.0013      0.001     -2.193      0.028      -0.003      -0.000
working_flag                       0.0054      0.012      0.471      0.637      -0.017       0.028
travelling_network_rating         -0.0215      0.004     -6.132      0.000      -0.028      -0.015
Adult_flag                        -0.0997      0.006    -17.335      0.000      -0.111      -0.088
Daily_Avg_mins_spend_on_traveling_page -0.0029      0.001     -3.612      0.000      -0.004      -0.001
==============================================================================
Omnibus:                     1877.427   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3448.153
Skew:                           1.474   Prob(JB):                         0.00
Kurtosis:                       4.165   Cond. No.                     2.74e+05
==============================================================================
```

Not significant after dropping high VIF attributes
Will try and drop few more

```
vif_cal(input_data=Train.drop(['Taken_product','total_likes_on_outofstation_checkin_received','Daily_Avg_mins_spend_on_traveling_page','montly_avg_c

preferred_device  VIF =  1.02
total_likes_on_outstation_checkin_given  VIF =  1.0
yearly_avg_Outstation_checkins  VIF =  1.01
member_in_family  VIF =  1.02
preferred_location_type  VIF =  inf
Yearly_avg_comment_on_travel_page  VIF =  1.02
week_since_last_outstation_checkin  VIF =  1.02
following_company_page  VIF =  1.0
working_flag  VIF =  1.0
travelling_network_rating  VIF =  1.01
Adult_flag  VIF =  1.01
preferred_location_type  VIF =  inf
<ipython-input-392-b323439b0457>:10: RuntimeWarning: divide by zero encountered in double_scalars
  vif=round(1/(1-rsq),2)
<ipython-input-392-b323439b0457>:10: RuntimeWarning: divide by zero encountered in double_scalars
  vif=round(1/(1-rsq),2)
```

Will Try and run the another model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Taken_product   R-squared:                       0.154
Model:                            OLS   Adj. R-squared:                  0.152
Method:                 Least Squares   F-statistic:                     99.71
Date:                Tue, 06 Feb 2024   Prob (F-statistic):          3.67e-284
Time:                        16:16:33   Log-Likelihood:                -2769.2
No. Observations:                8232   AIC:                             5570.
Df Residuals:                    8216   BIC:                             5683.
Df Model:                          15
Covariance Type:            nonrobust
================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                         0.5298      0.031     16.837      0.000       0.468       0.592
Yearly_avg_view_on_travel_page   -0.0006   7.45e-05     -7.993      0.000      -0.001      -0.000
preferred_device                 -0.0998      0.013     -7.669      0.000      -0.125      -0.074
total_likes_on_outstation_checkin_given -1.426e-06   2.69e-07     -5.295      0.000   -1.95e-06   -8.98e-07
yearly_avg_Outstation_checkins    0.0034      0.000      7.899      0.000       0.003       0.004
member_in_family                  0.0051      0.004      1.365      0.172      -0.002       0.012
preferred_location_type           0.0014      0.001      1.044      0.296      -0.001       0.004
Yearly_avg_comment_on_travel_page 0.0001      0.000      0.696      0.486      -0.000       0.000
total_likes_on_outofstation_checkin_received -6.911e-06   1.27e-06     -5.438      0.000   -9.4e-06   -4.42e-06
week_since_last_outstation_checkin 0.0176      0.002     11.475      0.000       0.015       0.021
following_company_page            0.1973      0.008     23.615      0.000       0.181       0.214
montly_avg_comment_on_company_page -0.0013      0.001     -2.193      0.028      -0.003      -0.000
working_flag                      0.0054      0.012      0.471      0.637      -0.017       0.028
travelling_network_rating        -0.0215      0.004     -6.132      0.000      -0.028      -0.015
Adult_flag                       -0.0997      0.006    -17.335      0.000      -0.111      -0.088
Daily_Avg_mins_spend_on_traveling_page -0.0029      0.001     -3.612      0.000      -0.004      -0.001
==============================================================================
Omnibus:                     1877.427   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3448.153
Skew:                           1.474   Prob(JB):                         0.00
Kurtosis:                       4.165   Cond. No.                     2.74e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.74e+05. This might indicate that there are
```

As noticed, even 3rd model did not given significant performance .

Let us make prediction on all models.

```
[405] ## Prediction on Training and Test set

    y_pred_1_Train = lm1.fittedvalues
    y_pred_2_Train = lm2.fittedvalues
    y_pred_3_Train = lm3.fittedvalues

    y_pred_1_Test = lm1.predict(Test)
    y_pred_2_Test = lm2.predict(Test)
    y_pred_3_Test = lm3.predict(Test)
```

```
[406] ## RMSE Calculation on Training and Test set

    from sklearn.metrics import mean_squared_error

    print('RMSE on Training Set for Model 1:',mean_squared_error(y_true=Train['Taken_product'],y_pred=y_pred_1_Train,squared=False))
    print('RMSE on Training Set for Model 2:',mean_squared_error(y_true=Train['Taken_product'],y_pred=y_pred_2_Train,squared=False))
    print('RMSE on Training Set for Model 3:',mean_squared_error(y_true=Train['Taken_product'],y_pred=y_pred_3_Train,squared=False))

RMSE on Training Set for Model 1: 0.33873085761234994
RMSE on Training Set for Model 2: 0.34239387680391054
RMSE on Training Set for Model 3: 0.34902310672905956
```

```
    print('RMSE on Test Set for Model 1:',mean_squared_error(y_true=Test['Taken_product'],y_pred=y_pred_1_Test,squared=False))
    print('RMSE on Test Set for Model 2:',mean_squared_error(y_true=Test['Taken_product'],y_pred=y_pred_2_Test,squared=False))
    print('RMSE on Test Set for Model 3:',mean_squared_error(y_true=Test['Taken_product'],y_pred=y_pred_3_Test,squared=False))

RMSE on Test Set for Model 1: 0.33606728315168993
RMSE on Test Set for Model 2: 0.34049429953614263
RMSE on Test Set for Model 3: 0.34517381997703955
```

RMSE is getting better for each model after iterating attributes.
it has given slight better result on test but still it is not significant.

Conclusion:
Though the RMSE for 3rd model has improved however it is still insignificant.
We may need to get more data or attributes to decided and suggest to make an Approach to run digital campaigns for the travel company.