

Capstone Project- Social media Tourism

PGP_DSBA_Aug_2021

Vishal Rathod

Contents

1-Introduction -Brief introduction about the problem statement and the need of solving it.....	3-4
2-EDA and Business Implication -Uni-variate / Bi-variate / multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?.....	5-13
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)	14-15
4-Model Building -How was the model validated? Just accuracy, or anything else too?.....	16-22
5-Final interpretation / recommendation -Detailed recommendations for the management/client based on the analysis done.....	22-24
Appendix	25

List of Fig-

2.1 Taken_product.....	5
2.2 Preferred device.....	5
2.3 Preferred location type.....	6
2.4 Following company page.....	6
2.5 Working flag.....	7
2.8 Week since last outstation check-in.....	10
2.9 Distribution and box plots.....	11-12
2.11 Heat map.....	13
2.12 Pair plot.....	13
3.1 Box plot with outliers.....	14
3.2 Box plot after removal of outliers.....	15

List of Tables-

2.6-2.8 Bivariate analysis.....	8-10
4.1 Base line Random Forest Model.....	17
4.2 SMOTE Oversampling for Imbalanced Dataset using Random Forest.....	17

4.3 Logistic Regression for dataset.....	18
4.4 Discriminant Analysis LDA	18
4.5 Naive Bayes.....	18
4.6 Model Scores.....	19
4.7 Grid search for Logistic model.....	20
4.8 Bagging using hyperparameters.....	20
4.9 Bagging without grid search.....	21
4.10 Gradient Boosting for laptop.....	21
4.11 Model scores.....	22

1 Introduction

Instead of reaching out to each customer, an aviation company wants to take a more focused approach. With the help of data science, they want to understand the customer's social media behavior and patterns towards buying a travel package. To do so, they collaborated with social media platforms to procure the data.

With this attempt we will evaluate the online and social activity of the target audience in order to deliver digital advertisements to users who are most likely to purchase the tourism package.

We analyzed the data systematically by performing various steps like EDA, Data cleaning and Preprocessing, Model building and finally giving our insights and recommendations on our findings.

Below is a chart that describes various variables present in the data set-

Variable Description

Variable	Description
UserID	Unique ID of user
Buy_ticket	Buy ticket in next month
Yearly_avg_view_on_travel_page	Average yearly views on any travel related page by user
preferred_device	Through which device user preferred to do login
total_likes_on_outstation_checkin_given	Total number of likes given by a user on out of station checkings in last year
yearly_avg_Outstation_checkins	Average number of out of station check-in done by user
member_in_family	Total number of relationship mentioned by user in the account
preferred_location_type	Preferred type of the location for travelling of user
Yearly_avg_comment_on_travel_page	Average yearly comments on any travel related page by user
total_likes_on_outofstation_checkin_received	Total number of likes received by a user on out of station checkings in last year
week_since_last_outstation_checkin	Number of weeks since last out of station check-in update by user
following_company_page	Weather the customer is following company page (Yes or No)
monthly_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

There are 11760 rows and 17 columns in this dataset.

We have 3 float dtypes, 7 int64 dtypes and 7 object dtypes.

There were no duplicate values found in the data.

Many variables have missing values and they would be treated accordingly.

	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
monthy_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

The above table is the description of the data. We can see that the mean and the median values for all these continues type variables have a major difference in them and hence we can assume that these variables are not balanced.

Understanding of attributes (variable info)

While checking for value counts in each variable we found that the column member_in_family had (Three) as one of its values, and hence it was picked by python as an object dtype. Similarly, the variable yearly_avg_Outstation_checkins also had (*) as a value and was picked as an object dtype by the system.

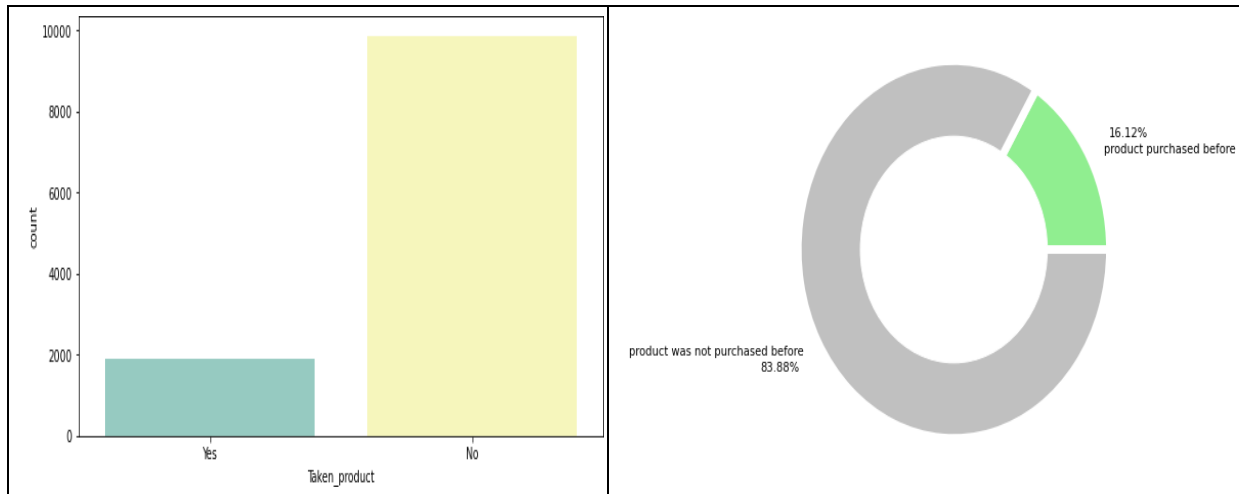
Further we also found that the column [preferred_deivces] had two values which mean the same (other and others) and the column [preferred_ocation_type] also had similar issue where (tour travel and tour and travel) mean the same and hence these values were combined together to one.

2-EDA and Business Implication-

We went ahead made count plots for the categorical variables to understand the frequency of their values and derive some insights from it. This was done before imputing the null values though.

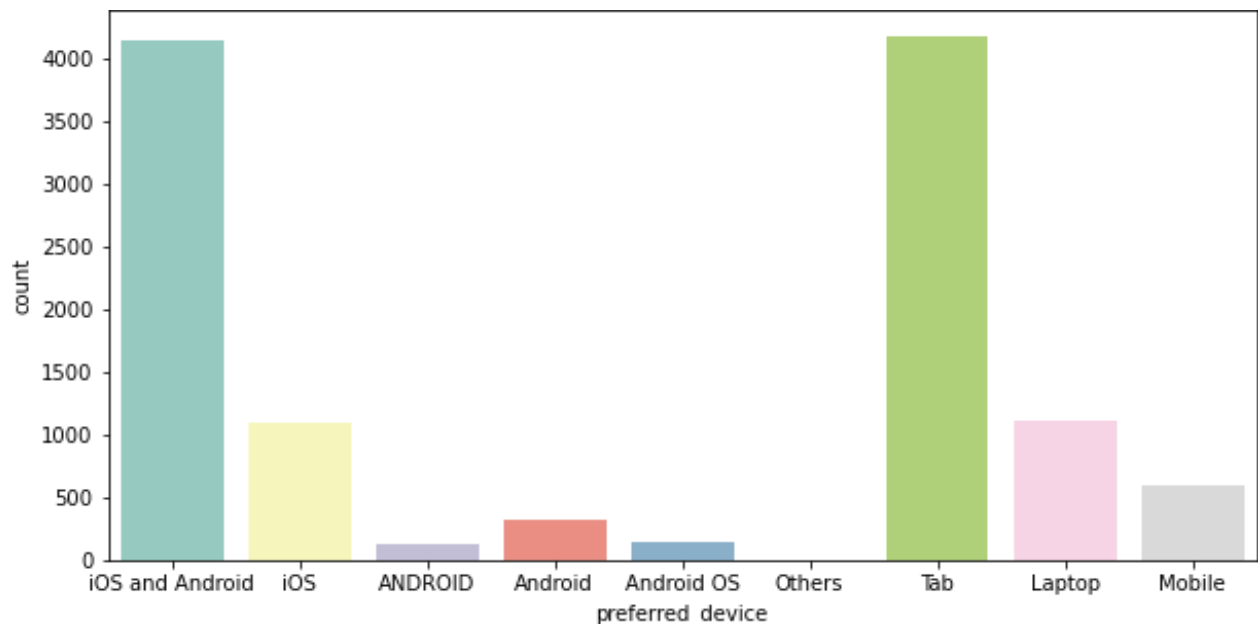
Univariate analysis-

2.1.Taken_product-



We can see that the out of the total number of users only 16% have taken the product and remaining 83% have not.

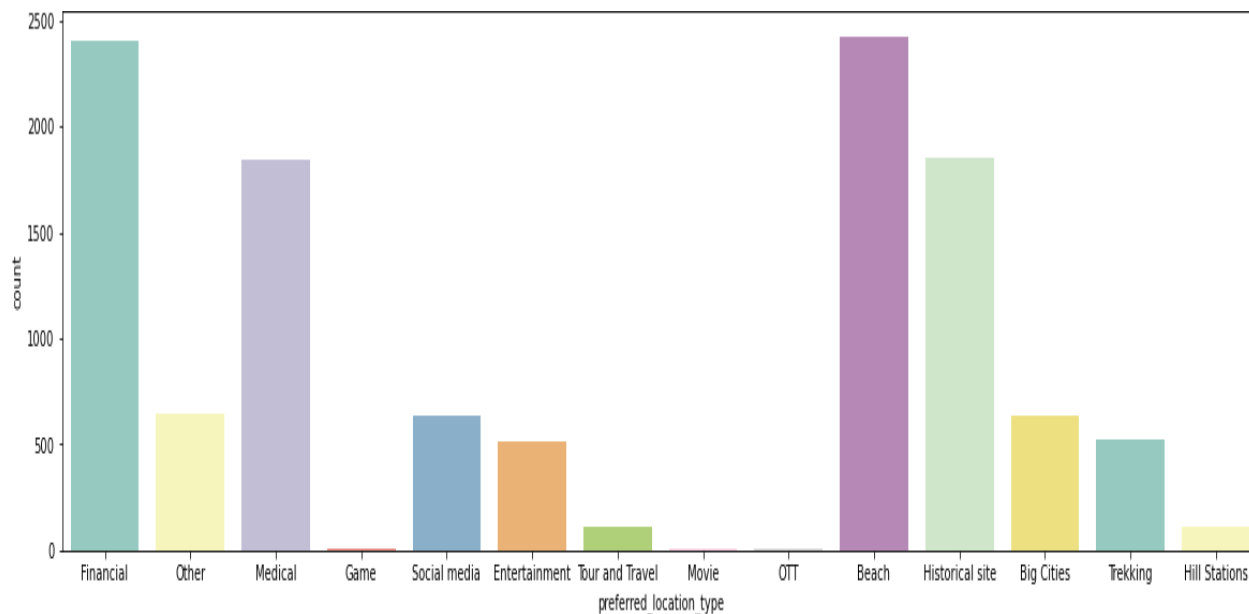
2.2.Preferred_device



From the above plot we can see that people tend to use their mobile phones or a hand-held device

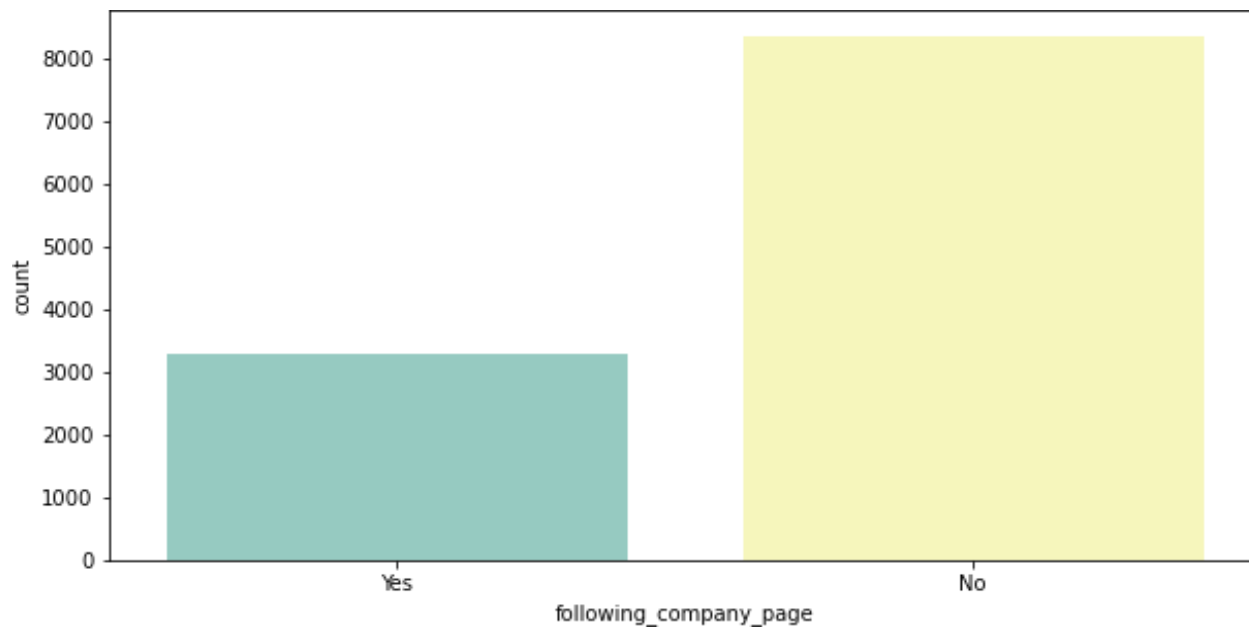
rather than using a laptop.

2.3.Preferred location type-



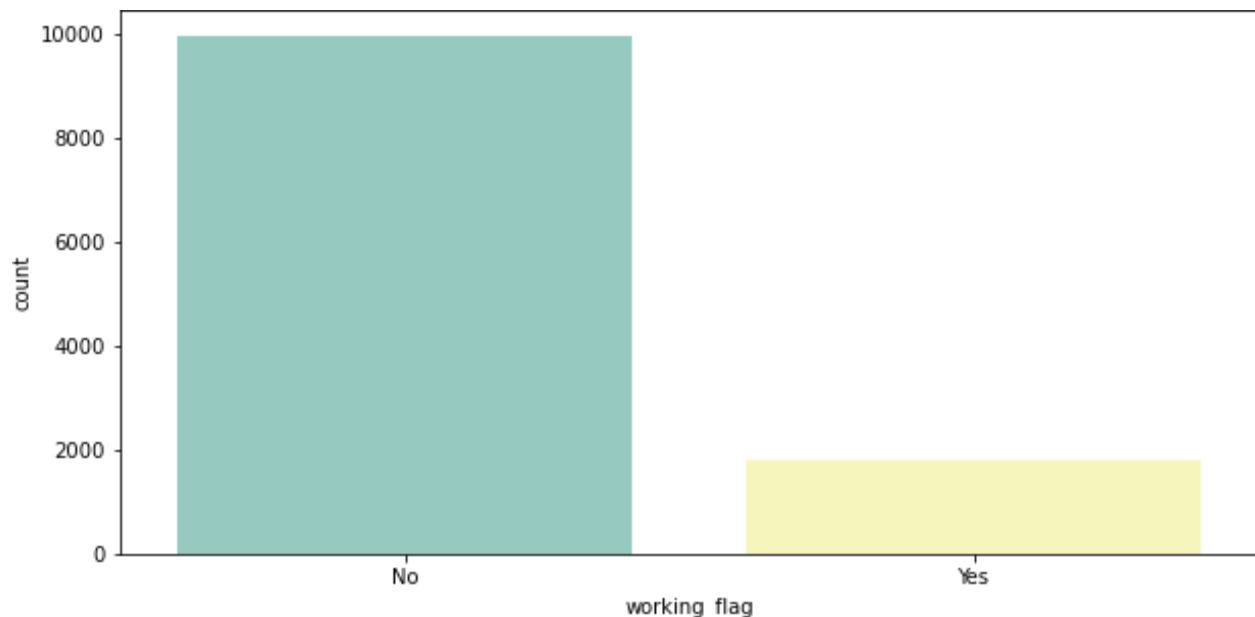
The above plot shows us the purpose of travelling and it is clear that most travelled destination is Beach and for financial purposes. Medical and historical site falls next in the order of most travelled.

2.4.Following company page-



The above plot shows how many users actually follow the company page and we can clearly see that not too many of them do so.

2.5. Working flag-



The above plot depicts the count of working and non-working users. We can see that non-working class is significantly higher than the working class.

Business Implications-

The above plots give us valuable insights understanding the distribution of sub-categories within each variable.

The plot for (taken_product) tells us that only 16% of the total users have purchased the product and the remaining 83% have not. So, it is extremely important to understand why such a big difference has occurred and the right strategy needs to be implemented for the same.

The next plot for the variable (preferred device) gives us an understanding of what kind of device the users like to use while visiting various social media channels for their travel plans. This insight will help to strategize the correct marketing plan to focus and deploy the right advertisements for the right device.

Preferred location type plot can give us an understanding where do users like to travel so that the right kind of advertisement can be made focusing on the most liked destinations.

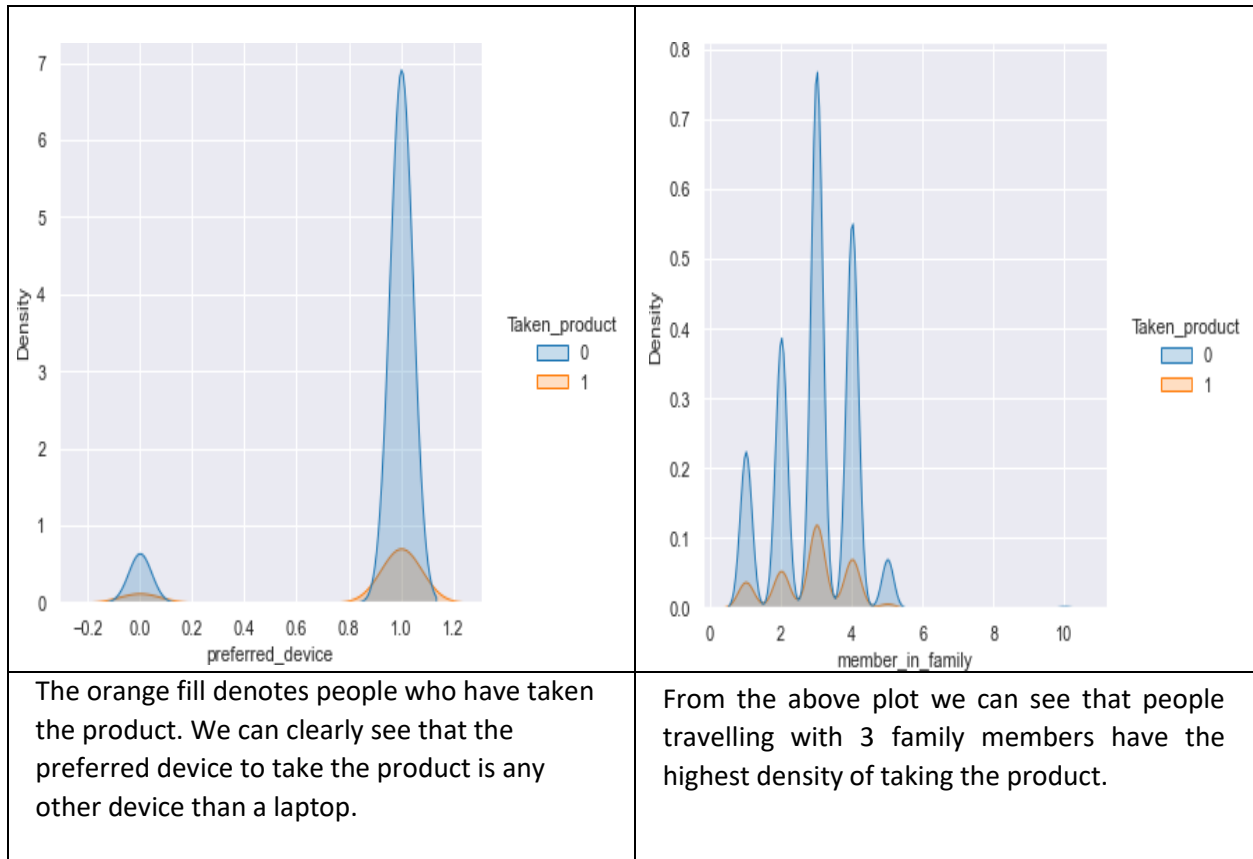
The following page variable helps us to understand how many users are actually following the company page. This can help the company to derive a strategy and target the user to lure them into followers.

Working flag plot shows how the company can target 2 classes of working individuals and can derive a strategy or offer packages accordingly.

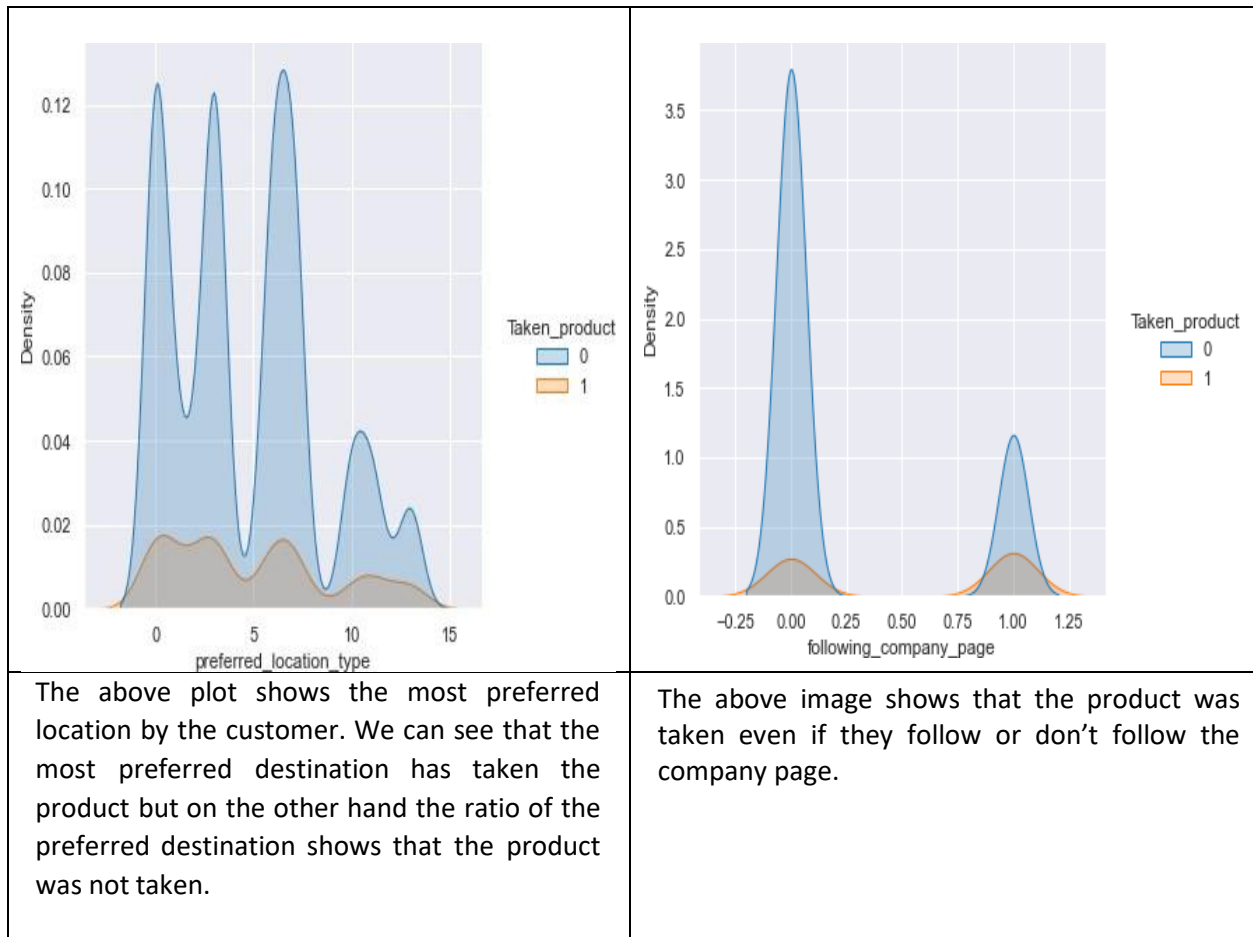
Bivariate Analysis-

These are the distribution and plots after we encoded the variables. This was done to understand relationship between our target variable [taken_product] and various columns.

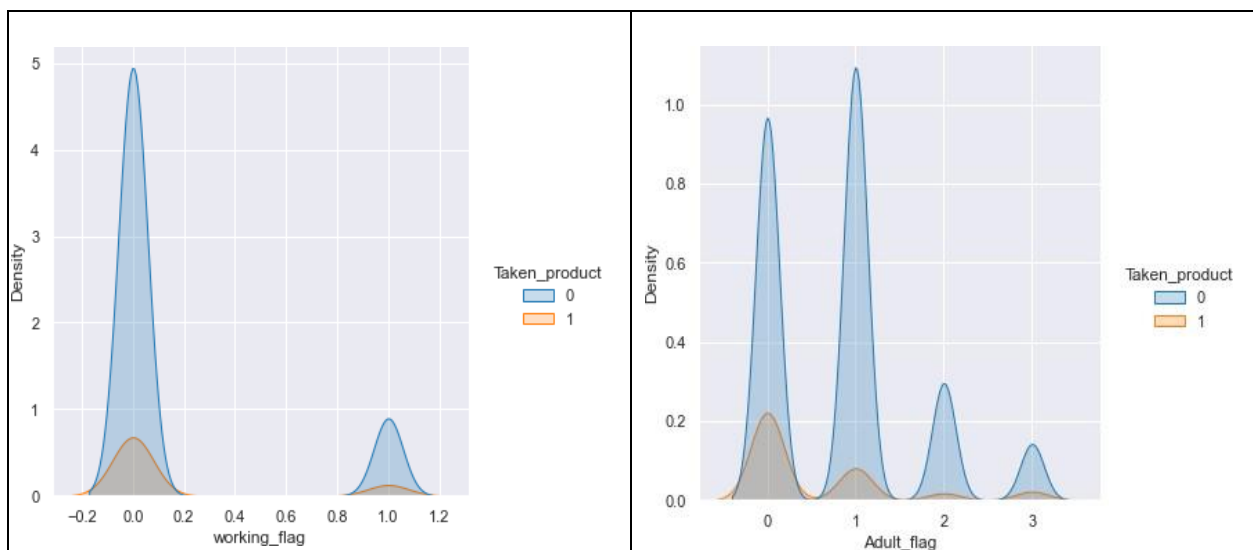
2.6.



2.7.



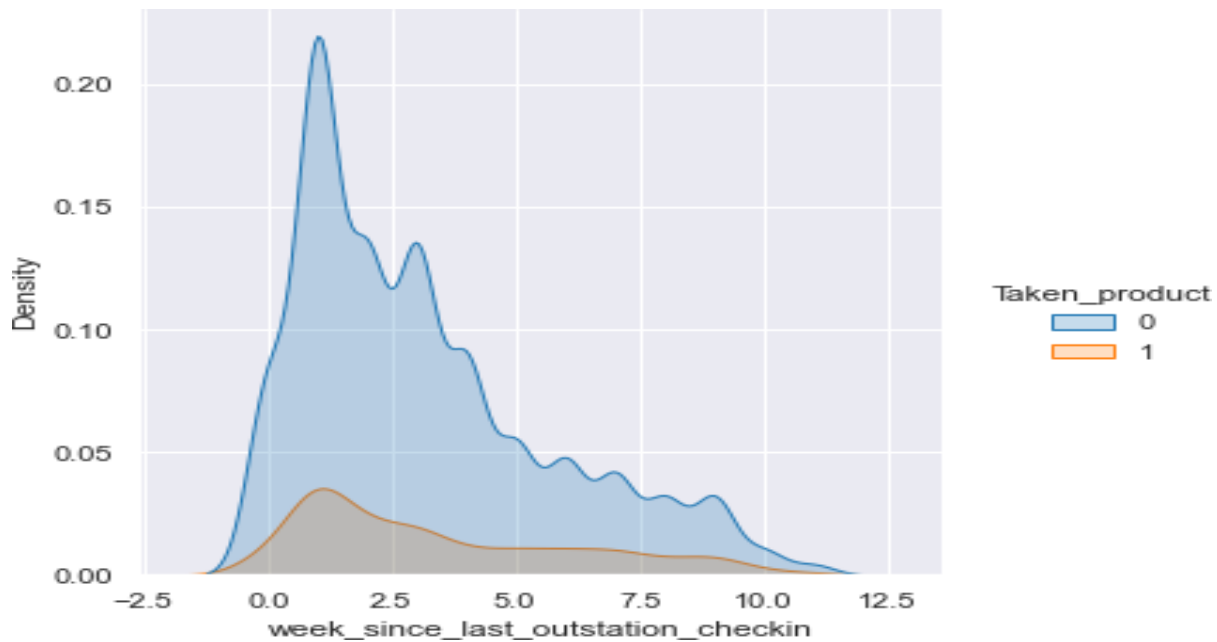
2.8.



The above image shows the number of non-working professionals who have taken the product is much higher the working ones.

The above plot indicates that people who bought the product maximum were not travelling with adult and can be assumed that these set of people travel mostly with their children.

2.8.



Business Implications-

The distribution plot for preferred device clearly shows that the users prefer using a hand held or mobile device to browse various websites for their travel plans. The company can derive a strategy and digital advertisement to cater to these users who like to use mobile devices.

The next plot that shows relation with the taken product variable is member in the family and it can be seen that people usually like to travel in a group of 3 as this group has taken the product more then any other. Hence the company can form strategies and plan out to design packages that attract users who like to travel in a group of 3 people.

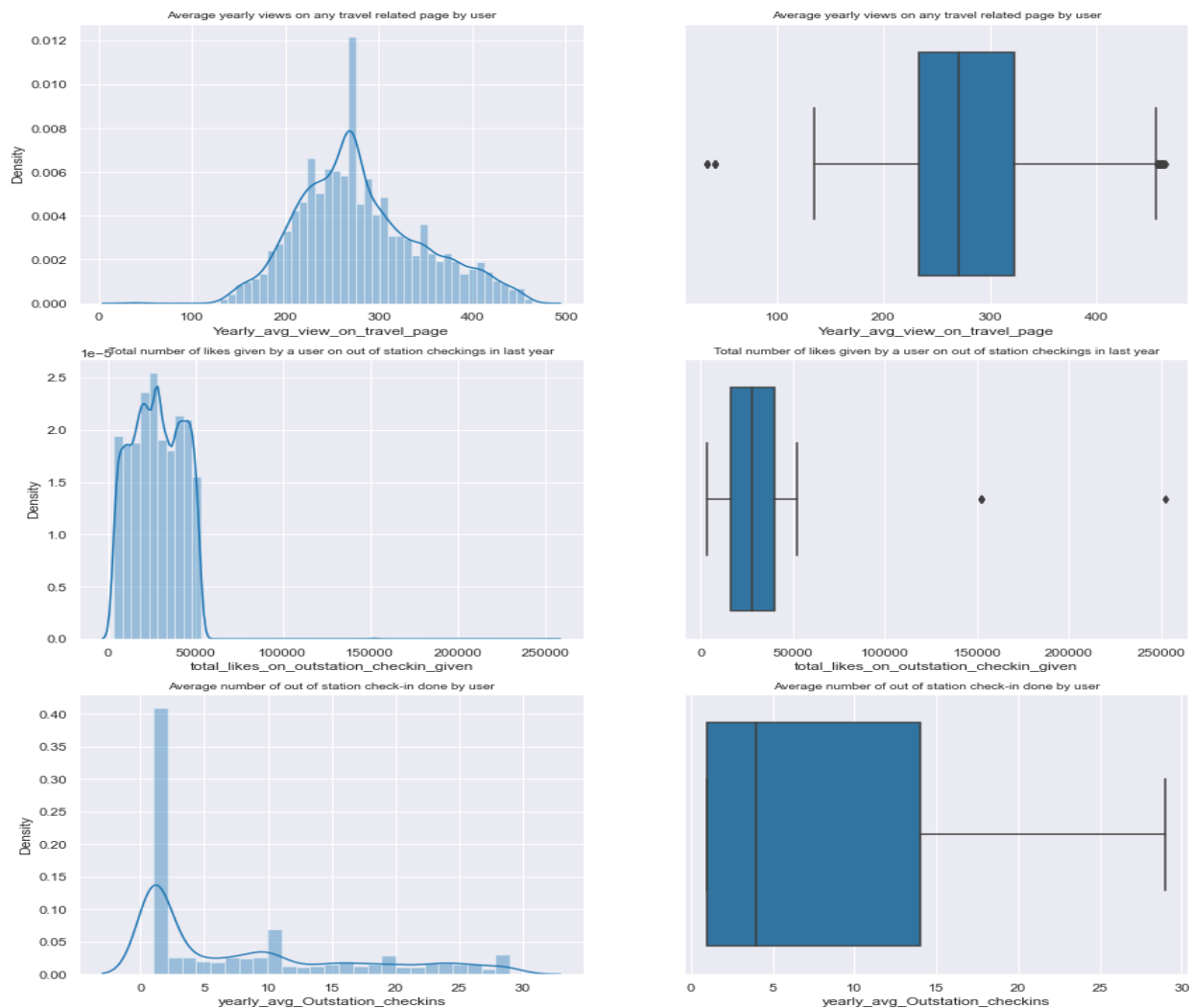
The next plot is most preferred location and we can see that people traveling to a certain destination has bought the product so the company can refine their strategy in creating lucrative and attractive offers to these destinations to lure users to buy their product.

The next plot is following company page and we can see that if following or not following users have still bought the product. This means this variable doesn't carry too value and can the company can concentrate on other important variables to design the right strategy.

The next plot is working flag and we can see that the not working users have bought the product more than the working class. This means the users are either young adults still studying in college or not working and the right strategy could be derived to target this class of users.

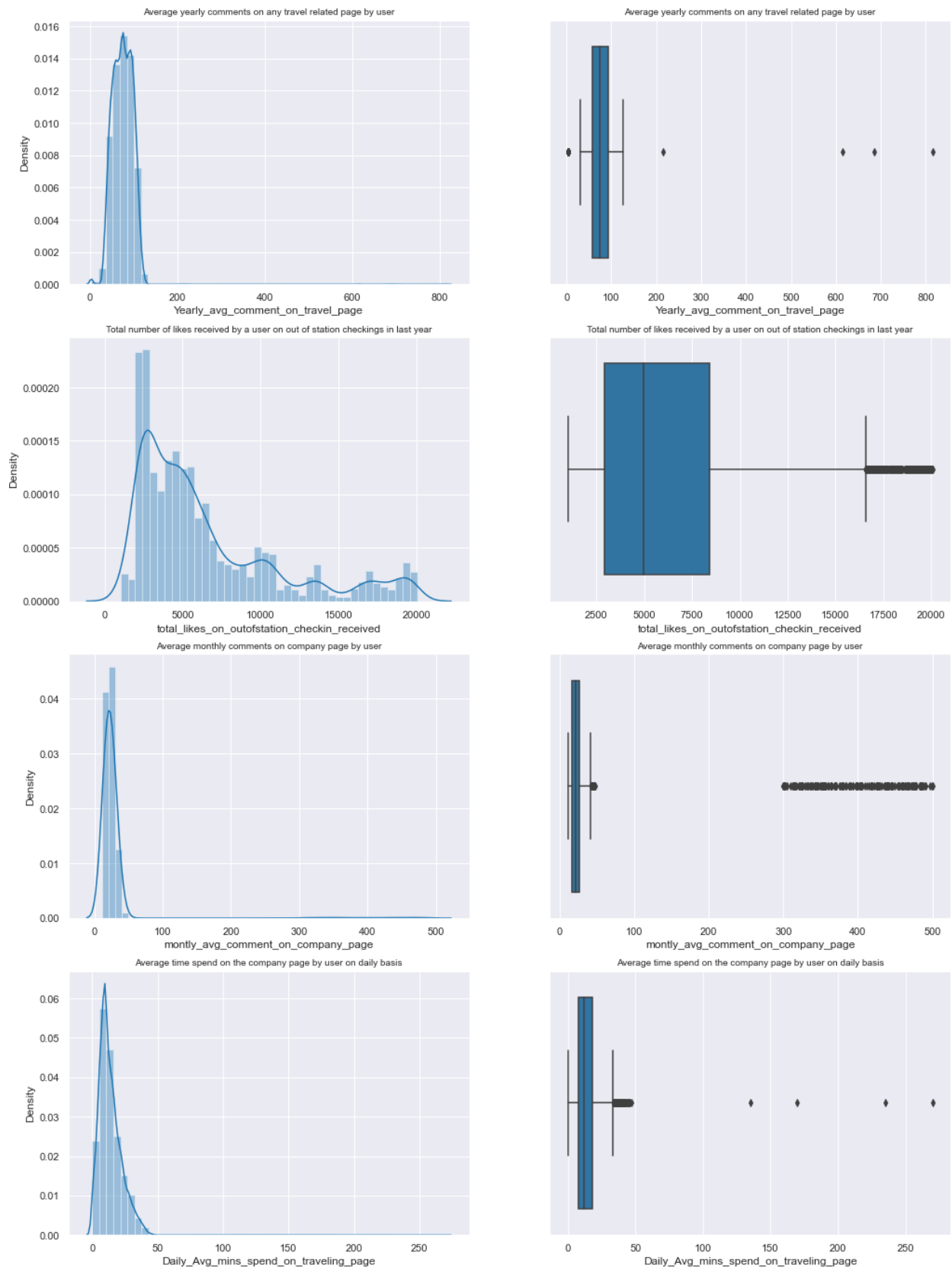
The plot adult flag shows that class 0 has bought the product maximum than others. This means we can assume these users are young people who go on fun holidays and it is paid by their parents. The company can design attractive packages to offer this class and increase the revenue.

2.9.



The above and the below plots indicates that the data is not balanced and it is either right skewed or left skewed.

2.10.

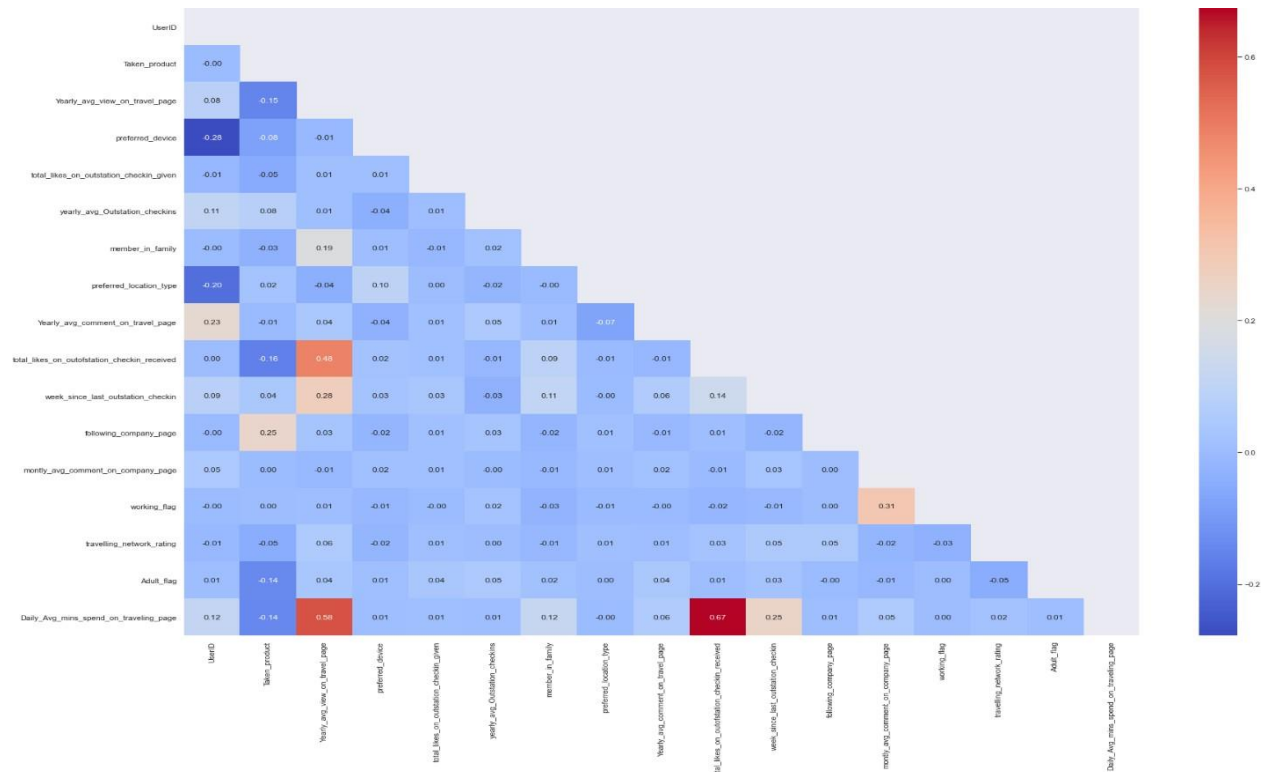


Also, we can clearly see some evidence of outliers in some of the variables.

Multivariate analysis-

Using Pearson's correlation measure we plotted a heat map to find out multicollinearity amongst the variables.

2.11.



We can clearly see that the highlighted boxes in red and orange are the only variables highly correlated to each other. Otherwise, most of them are negatively or almost have 0 correlation with each other.

2.12.



3-Data Cleaning and Pre-processing-

Labelling categorical data-

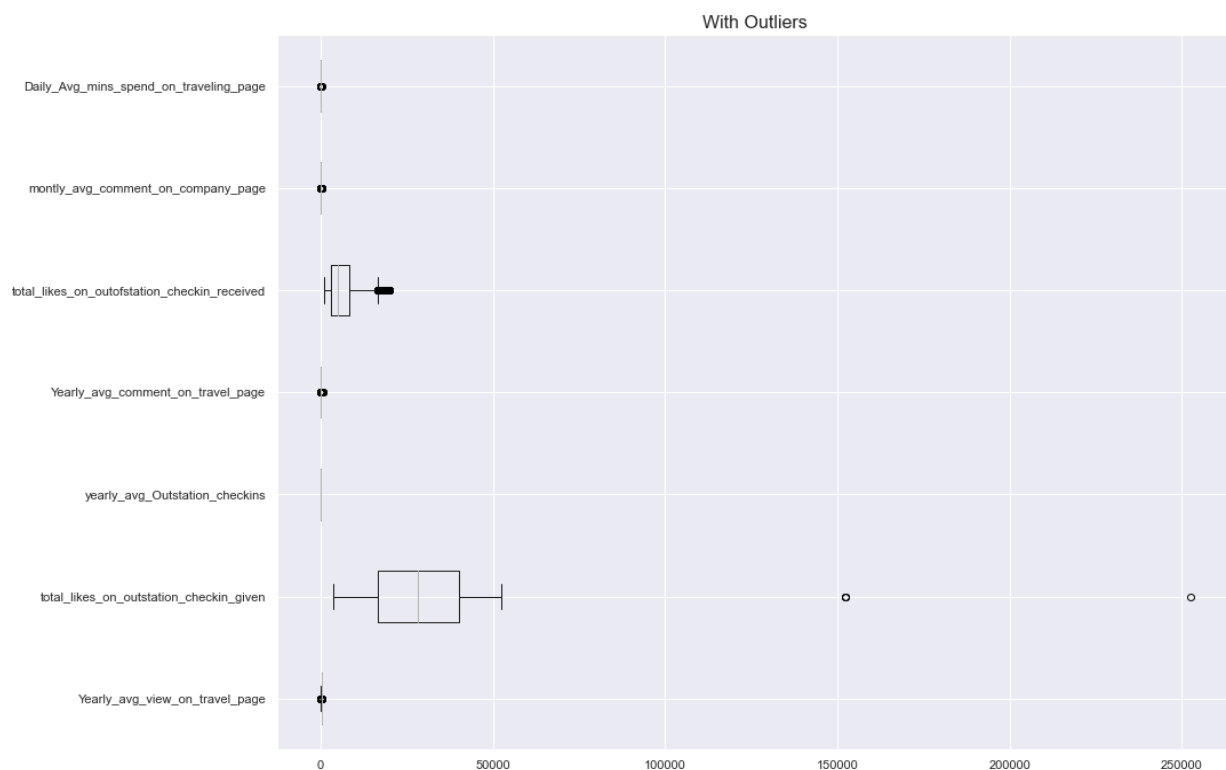
- Labelling people who have made a purchase with the company as '1' and people who have not made a purchase as '0'
- Anything which is not a laptop can be considered as mobile phone usage. Labelling users who are using Laptop as '0' and people who are using anything else as '1'
- Combining the family with 'Three' family members with families with '3' members.
- Labelling all the preferred location with unique values from '0' to '13'
- Labelling users who work as '1' and users who do not work as '0'

Next, we went ahead and changed the dtype to numeric for all the variable so that while imputing null values python doesn't throw an error.

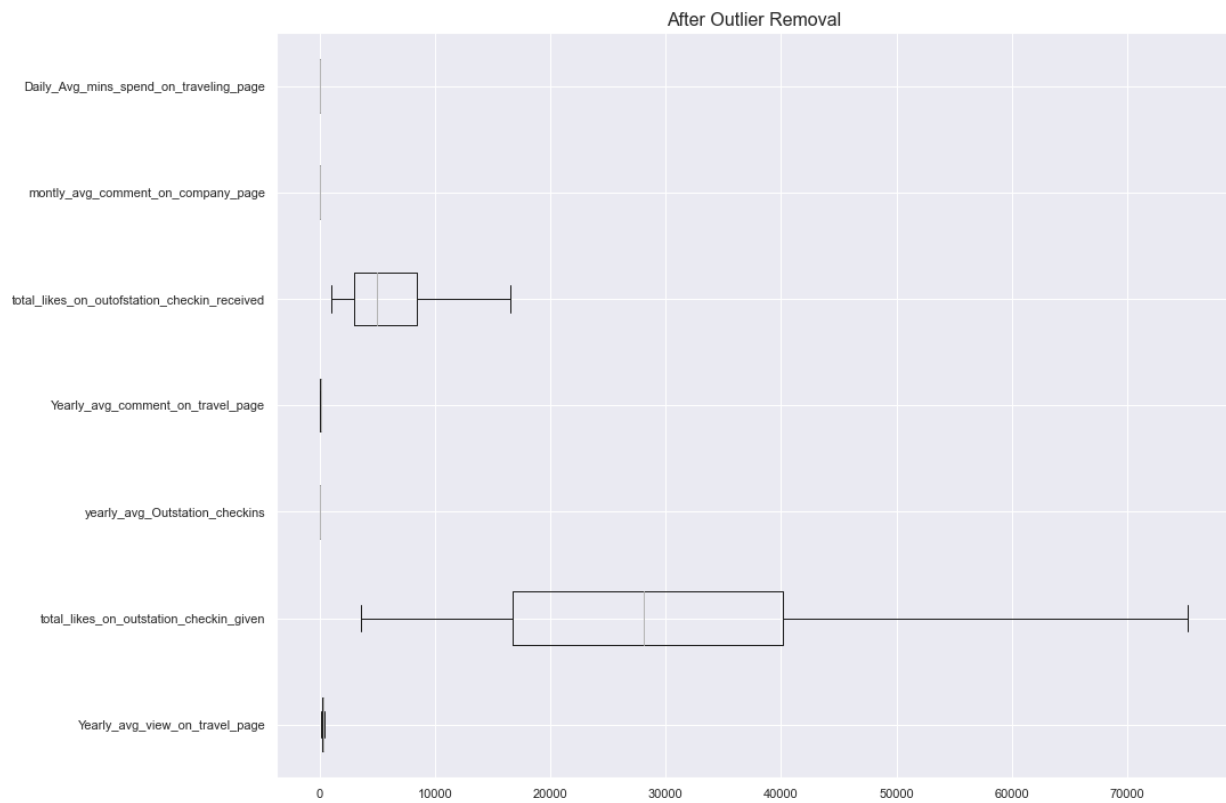
Outlier Treatment-

We took 7 variables over here who have prominent outliers in them and further treated them using the IQR method.

3.1



3.2.



4-Model building-

Feature selection using RFE (Recursive Feature Elimination)-

The technique of RFE was used to find the optimum number of variables for the purpose of model building in order to get good accuracy by deleting or dropping variables which do not contribute or have a low level of relation with the target variable (Taken_product).

10 best variables found with the above technique-

1. Yearly_avg_view_on_travel_page selected=True rank=1
2. total_likes_on_outstation_checkin_given selected=True rank=1
3. yearly_avg_Outstation_checkins selected=True rank=1
4. member_in_family selected=True rank=1
5. Yearly_avg_comment_on_travel_page selected=True rank=1
6. total_likes_on_outofstation_checkin_received selected=True rank=1
7. week_since_last_outstation_checkin selected=True rank=1
8. following_company_page selected=True rank=1
9. travelling_network_rating selected=True rank=1
10. Adult_flag selected=True rank=1

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

Smote was used to balance the minority class in the target variable and it was used to build Random Forest model.

Metric considered for this project-

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

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

List of Models build-

Random Forest
Logistic Regression
LDA Discriminant Analyses
Naïve Bayes
Bagging
ADA boost
Gradient boosting

Analysis of different models-4.1

Baseline Random Forest model									
Laptop users					Mobile users				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	238	0	0.99	1.00	0.99	2702
1	1.00	1.00	1.00	95	1	0.99	0.94	0.96	494
accuracy			1.00	333	accuracy			0.99	3196
macro avg	1.00	1.00	1.00	333	macro avg	0.99	0.97	0.98	3196
weighted avg	1.00	1.00	1.00	333	weighted avg	0.99	0.99	0.99	3196
Test Accuracy					Test Accuracy				
rfcl.score(X_test,y_test)					rfcl.score(X_test,y_test)				
1.0					0.9893617021276596				

4.2

SMOTE Oversampling for Imbalanced Laptop Dataset using Random Forest					SMOTE Oversampling for Imbalanced Mobile Dataset using Random Forest				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	238	0	0.99	1.00	1.00	2702
1	0.99	1.00	0.99	95	1	0.99	0.96	0.98	494
accuracy			1.00	333	accuracy			0.99	3196
macro avg	0.99	1.00	1.00	333	macro avg	0.99	0.98	0.99	3196
weighted avg	1.00	1.00	1.00	333	weighted avg	0.99	0.99	0.99	3196

4.3

Logistic Regression for Laptop dataset					Logistic Regression for Mobile dataset				
0.7537537537537538 [[233 5] [77 18]]					0.8444931163954944 [[2699 3] [494 0]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.75	0.98	0.85	238	0	0.85	1.00	0.92	2702
1	0.78	0.19	0.31	95	1	0.00	0.00	0.00	494
accuracy			0.75	333	accuracy			0.84	3196
macro avg	0.77	0.58	0.58	333	macro avg	0.42	0.50	0.46	3196
weighted avg	0.76	0.75	0.69	333	weighted avg	0.71	0.84	0.77	3196

4.4

Discriminant Analysis LDA for Laptop dataset					Discriminant Analysis LDA for Mobile dataset				
0.7957957957957958 [[227 11] [57 38]]					0.8620150187734669 [[2666 36] [405 89]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.80	0.95	0.87	238	0	0.87	0.99	0.92	2702
1	0.78	0.40	0.53	95	1	0.71	0.18	0.29	494
accuracy			0.80	333	accuracy			0.86	3196
macro avg	0.79	0.68	0.70	333	macro avg	0.79	0.58	0.61	3196
weighted avg	0.79	0.80	0.77	333	weighted avg	0.84	0.86	0.83	3196

4.5

Naive Bayes for laptop dataset					Naive Bayes for Mobile dataset				
0.7897897897897898 [[201 37] [33 62]]					0.8560700876095119 [[2690 12] [448 46]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.84	0.85	238	0	0.86	1.00	0.92	2702
1	0.63	0.65	0.64	95	1	0.79	0.09	0.17	494
accuracy			0.79	333	accuracy			0.86	3196
macro avg	0.74	0.75	0.75	333	macro avg	0.83	0.54	0.54	3196
weighted avg	0.79	0.79	0.79	333	weighted avg	0.85	0.86	0.80	3196

4.6-Model Scores-

Models	Precision-Laptop	Accuracy-Laptop	Precision-Mobile	Accuracy-Mobile
Baseline Random Forest	100%	100%	99%	99%
Smote Using Random Forest	99%	100%	99%	99%
Logistic Regression	78%	75%	0%	84%
Discriminant Analysis-LDA	78%	80%	71%	86%
Naïve Bayes	63%	79%	79%	86%

Models	Recall-Laptop	F1-Laptop	Recall- Mobile	F1-Mobile
Baseline Random Forest	100%	100%	94%	96%
Smote Using Random Forest	100%	99%	96%	98%
Logistic Regression	19%	31%	0%	0%
Discriminant Analysis-LDA	40%	53%	18%	29%
Naïve Bayes	65%	64%	0.09%	17%

Model Tuning- Ensemble Techniques-

Modelling using Grid search for Logistic model for laptop dataset

Best Parameters for the model found as follows

Best: 0.817112 using {'C': 0.01, 'penalty': 'l2', 'solver': 'newton-cg'}

4.7

Grid search for Logistic model for laptop dataset					Grid search for Logistic model for Mobile dataset				
0.7867867867867868 [[236 2] [69 26]]					0.8610763454317898 [[2676 26] [418 76]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.77	0.99	0.87	238	0	0.86	0.99	0.92	2702
1	0.93	0.27	0.42	95	1	0.75	0.15	0.26	494
accuracy			0.79	333	accuracy			0.86	3196
macro avg	0.85	0.63	0.65	333	macro avg	0.80	0.57	0.59	3196
weighted avg	0.82	0.79	0.74	333	weighted avg	0.85	0.86	0.82	3196

Best Parameter-

{'base_estimator__max_depth': 6, 'base_estimator__max_leaf_nodes': 15, 'n_estimators': 40}

4.8

Bagging using hyperparameters for laptop dataset					Bagging using hyperparameters for mobile dataset				
0.8648648648648649 [[238 0] [45 50]]					0.867334167709637 [[2695 7] [417 77]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.84	1.00	0.91	238	0	0.87	1.00	0.93	2702
1	1.00	0.53	0.69	95	1	0.92	0.16	0.27	494
accuracy			0.86	333	accuracy			0.87	3196
macro avg	0.92	0.76	0.80	333	macro avg	0.89	0.58	0.60	3196
weighted avg	0.89	0.86	0.85	333	weighted avg	0.87	0.87	0.82	3196

4.9

Bagging without grid search laptop dataset						Bagging without grid search Mobile dataset					
0.9819819819819819 [[238 0] [6 89]]						0.9831038798498123 [[2692 10] [44 450]]					
	precision	recall	f1-score	support			precision	recall	f1-score	support	
0	0.98	1.00	0.99	238		0	0.98	1.00	0.99	2702	
1	1.00	0.94	0.97	95		1	0.98	0.91	0.94	494	
accuracy			0.98	333		accuracy			0.98	3196	
macro avg	0.99	0.97	0.98	333		macro avg	0.98	0.95	0.97	3196	
weighted avg	0.98	0.98	0.98	333		weighted avg	0.98	0.98	0.98	3196	
ADA Boosting Model for laptop						ADA Boosting Model for mobile					
0.8768768768768769 [[226 12] [29 66]]						0.8704630788485607 [[2637 65] [349 145]]					
	precision	recall	f1-score	support			precision	recall	f1-score	support	
0	0.89	0.95	0.92	238		0	0.88	0.98	0.93	2702	
1	0.85	0.69	0.76	95		1	0.69	0.29	0.41	494	
accuracy			0.88	333		accuracy			0.87	3196	
macro avg	0.87	0.82	0.84	333		macro avg	0.79	0.63	0.67	3196	
weighted avg	0.87	0.88	0.87	333		weighted avg	0.85	0.87	0.85	3196	

4.10

Gradient Boosting for laptop						Gradient Boosting for Mobile					
0.954954954954955 [[235 3] [12 83]]						0.9020650813516896 [[2669 33] [280 214]]					
	precision	recall	f1-score	support			precision	recall	f1-score	support	
0	0.95	0.99	0.97	238		0	0.91	0.99	0.94	2702	
1	0.97	0.87	0.92	95		1	0.87	0.43	0.58	494	
accuracy			0.95	333		accuracy			0.90	3196	
macro avg	0.96	0.93	0.94	333		macro avg	0.89	0.71	0.76	3196	
weighted avg	0.96	0.95	0.95	333		weighted avg	0.90	0.90	0.89	3196	

4.11 -Model Scores-

Models	Precision-Laptop	Accuracy-Laptop	Precision-Mobile	Accuracy-Mobile
Logistic Regression Grid Search	93%	79%	75%	86%
Bagging Using Hyperparameters	100%	86%	92%	87%
Bagging without Grid Search	100%	98%	98%	98%
ADA Boost	85%	88%	69%	87%
Gradient Boost	97%	95%	87%	90%
Models	Recall-Laptop	F1-Laptop	Recall- Mobile	F1-Mobile
Logistic Regression Grid Search	27%	42%	15%	26%
Bagging Using Hyperparameters	53%	69%	15%	27%
Bagging without Grid Search	94%	97%	91%	94%
ADA Boost	69%	76%	29%	41%
Gradient Boost	87%	92%	43%	58%

5-INTERPRETATION OF MOST OPTIMAL MODEL:

Performing visual inspection, stacking, and bagging models based on our model evaluation. Finally, we can integrate all of the results and conclude that Random Forest is the best performing model for both mobile phone users and laptop users, with a 99 percent accuracy rate.

Precision, the required criterion for this problem, is also found to be much higher for Random Forest models, with 99 percent for Laptop users and 99 percent for Mobile phone users.

Precision quantifies the number of positive class predictions that actually belong to the positive class based on this observation.

Recall and F1 score are also higher for random forest models.

As a result, this should be taken as 100% of total consumers who use Laptops and were expected to purchase the product actually bought the product.

Similarly, among all customers who use mobile phones, 99 percent of all customers forecasted to purchase the product do so. Which is the highest level of precision that the model can achieve.

Business Implications

The prediction capabilities of a model are considerably improved by selecting the proper model. This increases the model's dependability for decision making. When we select the Random Forest model, we train the model using the entire dataset as the train set, and the resulting model is ready to generate predictions with independent variables.

In our case, given the social media components of a person, such as the time spent on travel websites, the number of likes received and given, and so on, we will now be able to predict the likelihood of that customer purchasing the travel packages offered by the aviation company with a 99 percent accuracy.

This, in turn, gave a tailored way for the aviation company to reach their consumer base, lowering costs and maximizing the return on investment in digital marketing initiatives.

Recommendations-

***The analysis on the dataset has given the company an understanding how the users behave on social media platform's and if the organization focuses on the 10 variables it can turn the revenue game in their favor.**

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

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

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

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

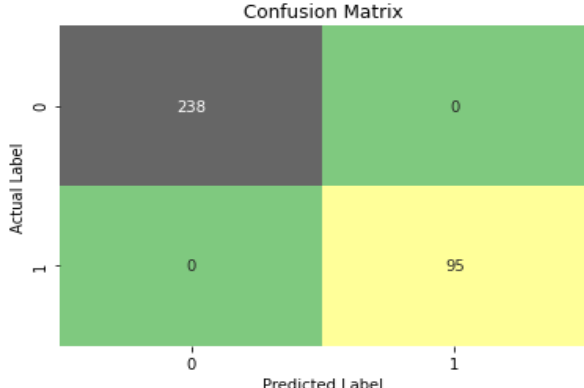
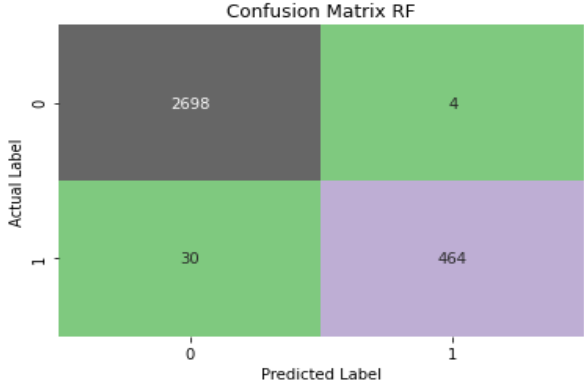
***For the customers to stay on their page and follow the page the company could use and introduce interactive games which offers the customer discounts or complimentary services or a free stay with an affiliated hotel and so on if they participate and win the games.**

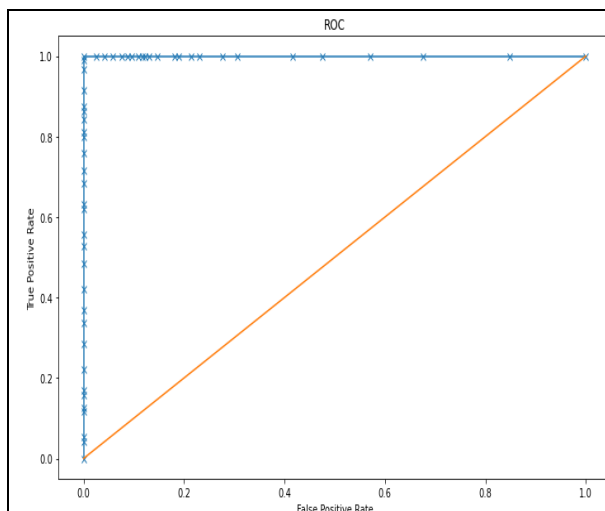
***They could ask the customers to write or comment on the company page and get points that can be redeemed in the future to get travel discounts.**

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

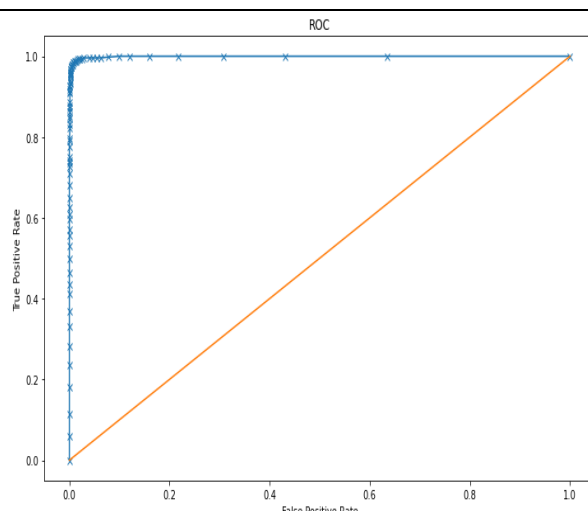
Appendix-

Confusion Metrics and ROC and AUC for various models-

Baseline Random Forest model									
Laptop users					Mobile users				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	238	0	0.99	1.00	0.99	2702
1	1.00	1.00	1.00	95	1	0.99	0.94	0.96	494
accuracy			1.00	333	accuracy			0.99	3196
macro avg	1.00	1.00	1.00	333	macro avg	0.99	0.97	0.98	3196
weighted avg	1.00	1.00	1.00	333	weighted avg	0.99	0.99	0.99	3196
Test Accuracy					Test Accuracy				
<pre>rfcl.score(X_test,y_test)</pre>					<pre>rfcl.score(X_test,y_test)</pre>				
1.0					0.9893617021276596				
									



Area under Curve is 1.0



Area under Curve is 0.999177022868051

Logistic Regression for Laptop dataset

0.7537537537537538

[[233 5]
[77 18]]

	precision	recall	f1-score	support
0	0.75	0.98	0.85	238
1	0.78	0.19	0.31	95
accuracy			0.75	333
macro avg	0.77	0.58	0.58	333
weighted avg	0.76	0.75	0.69	333

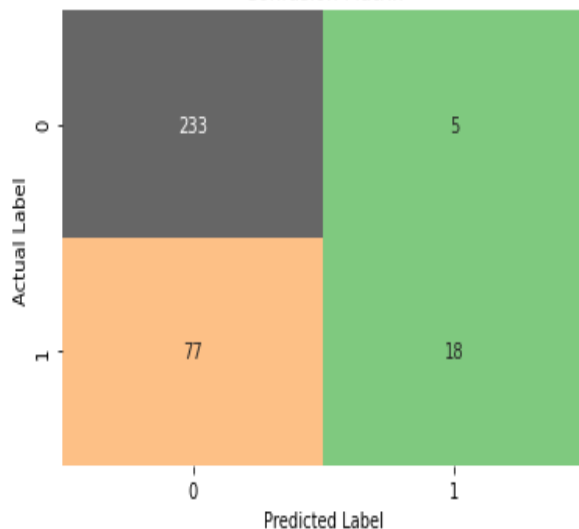
Logistic Regression for Mobile dataset

0.8444931163954944

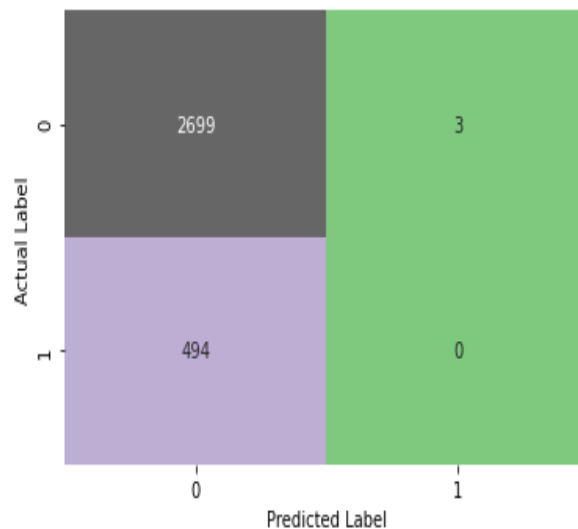
[[2699 3]
[494 0]]

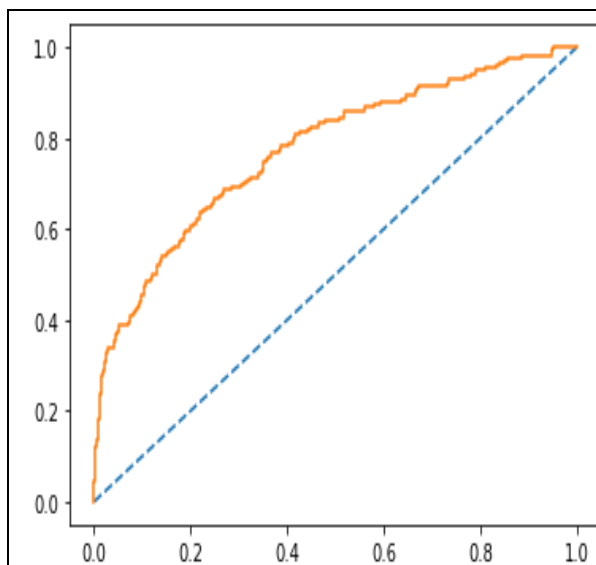
	precision	recall	f1-score	support
0	0.85	1.00	0.92	2702
1	0.00	0.00	0.00	494
accuracy			0.84	3196
macro avg	0.42	0.50	0.46	3196
weighted avg	0.71	0.84	0.77	3196

Confusion Matrix

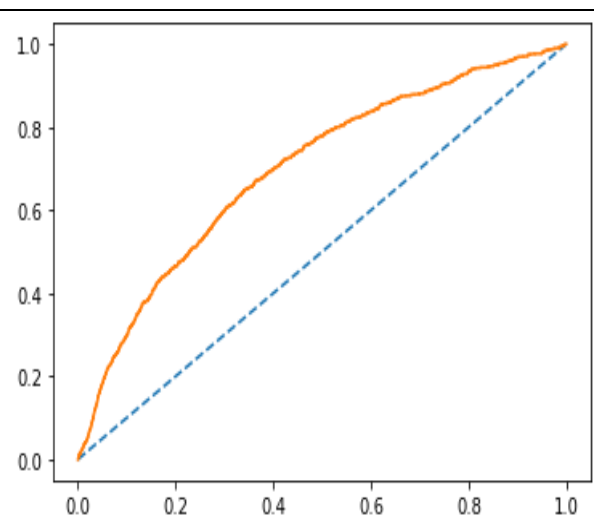


Confusion Matrix Mobile





AUC: 0.792



AUC: 0.692

Discriminant Analysis LDA for Laptop dataset

0.7957957957957958

[[227 11]

[57 38]]

	precision	recall	f1-score	support
0	0.80	0.95	0.87	238
1	0.78	0.40	0.53	95
accuracy			0.80	333
macro avg	0.79	0.68	0.70	333
weighted avg	0.79	0.80	0.77	333

Discriminant Analysis LDA for Mobile dataset

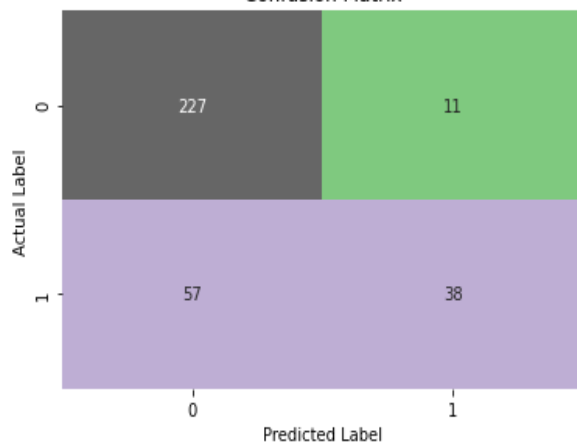
0.8620150187734669

[[2666 36]

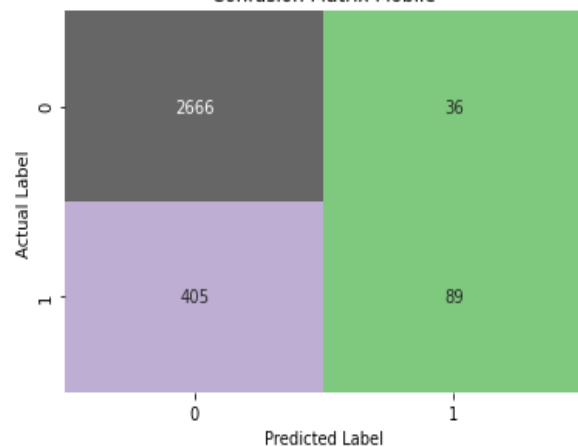
[405 89]]

	precision	recall	f1-score	support
0	0.87	0.99	0.92	2702
1	0.71	0.18	0.29	494
accuracy			0.86	3196
macro avg	0.79	0.58	0.61	3196
weighted avg	0.84	0.86	0.83	3196

Confusion Matrix



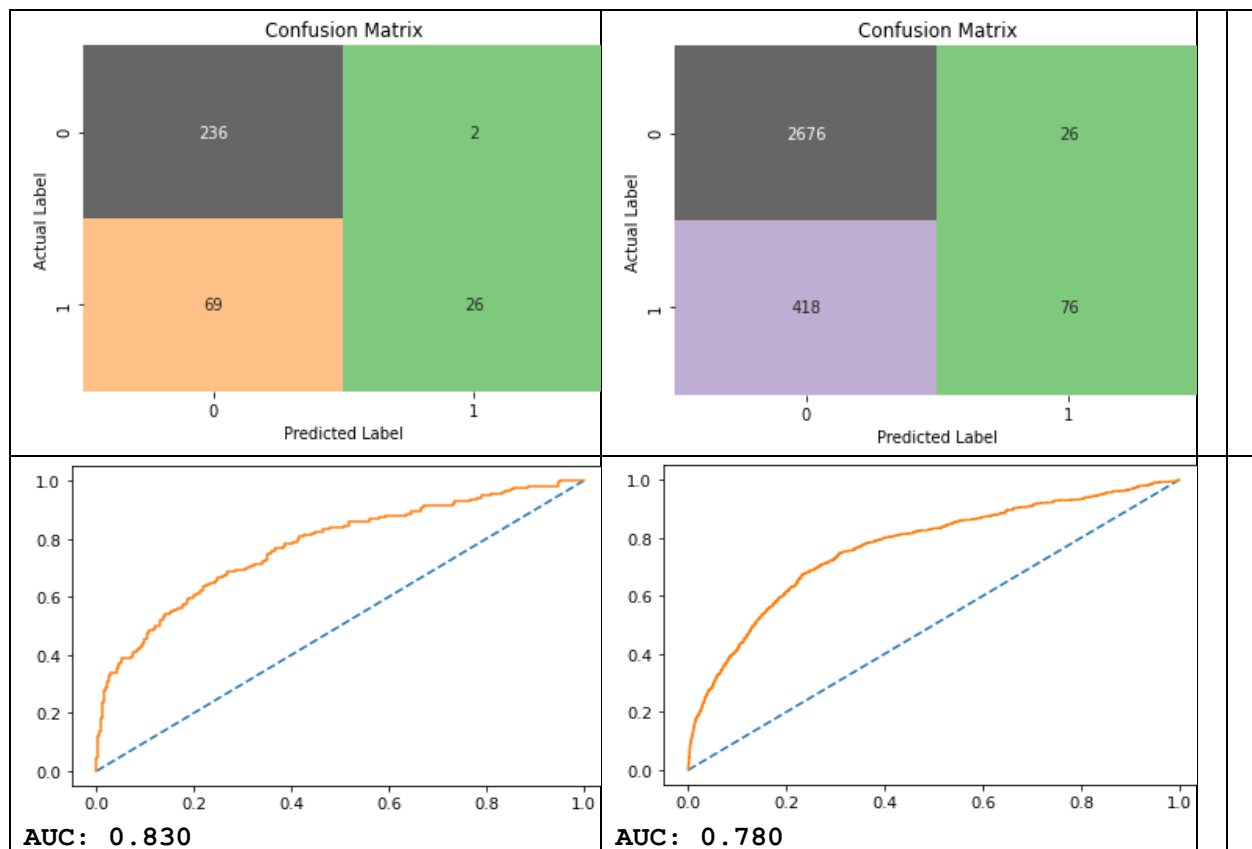
Confusion Matrix Mobile



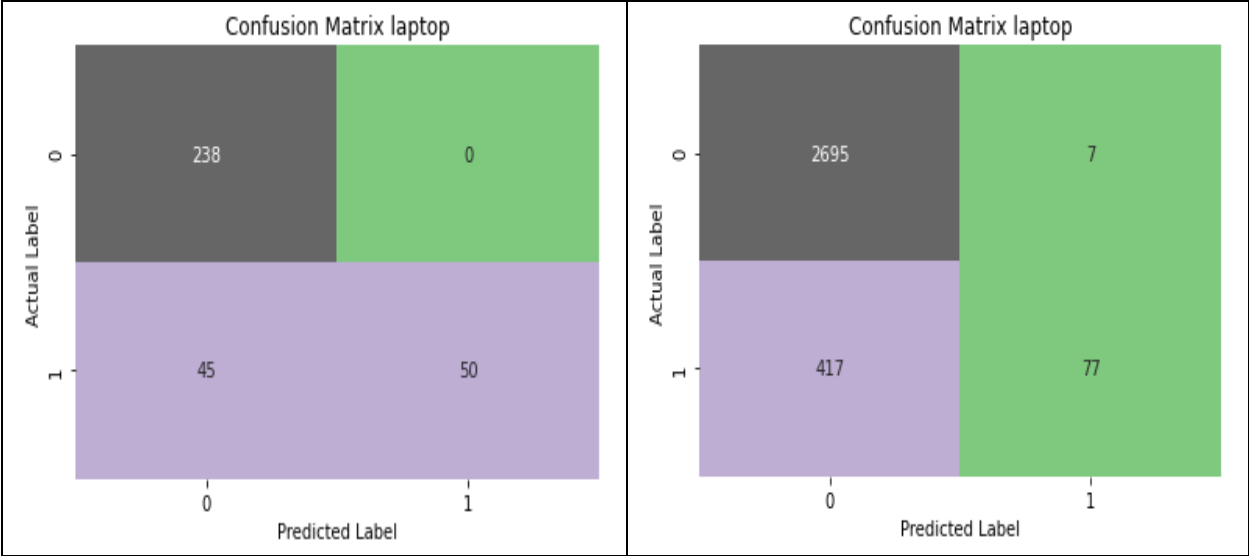
Naive Bayes for laptop dataset					Naive Bayes for Mobile dataset				
0.7897897897897898					0.8560700876095119				
[[201 37]					[[2690 12]				
[33 62]]					[448 46]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.84	0.85	238	0	0.86	1.00	0.92	2702
1	0.63	0.65	0.64	95	1	0.79	0.09	0.17	494
accuracy			0.79	333	accuracy			0.86	3196
macro avg	0.74	0.75	0.75	333	macro avg	0.83	0.54	0.54	3196
weighted avg	0.79	0.79	0.79	333	weighted avg	0.85	0.86	0.80	3196

Confusion Matrix Naive Base			Confusion Matrix Mobile		
Actual Label	0	1	Actual Label	0	1
0	201	37	0	2690	12
1	33	62	1	448	46
	0	1		0	1
	Predicted Label			Predicted Label	

Grid search for Logistic model for laptop dataset					Grid search for Logistic model for Mobile dataset						
0.7867867867867868 [[236 2] [69 26]]					0.8610763454317898 [[2676 26] [418 76]]						
	precision	recall	f1-score	support		precision	recall	f1-score	support		
0	0.77	0.99	0.87	238	0	0.86	0.99	0.92	2702		
1	0.93	0.27	0.42	95	1	0.75	0.15	0.26	494		
accuracy			0.79	333	accuracy			0.86	3196		
macro avg	0.85	0.63	0.65	333	macro avg	0.80	0.57	0.59	3196		
weighted avg	0.82	0.79	0.74	333	weighted avg	0.85	0.86	0.82	3196		



Bagging using hyperparameters for laptop dataset					Bagging using hyperparameters for mobile dataset				
0.8648648648648649					0.867334167709637				
[[238 0]					[[2695 7]				
[45 50]]					[417 77]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.84	1.00	0.91	238	0	0.87	1.00	0.93	2702
1	1.00	0.53	0.69	95	1	0.92	0.16	0.27	494
accuracy			0.86	333	accuracy			0.87	3196
macro avg	0.92	0.76	0.80	333	macro avg	0.89	0.58	0.60	3196
weighted avg	0.89	0.86	0.85	333	weighted avg	0.87	0.87	0.82	3196



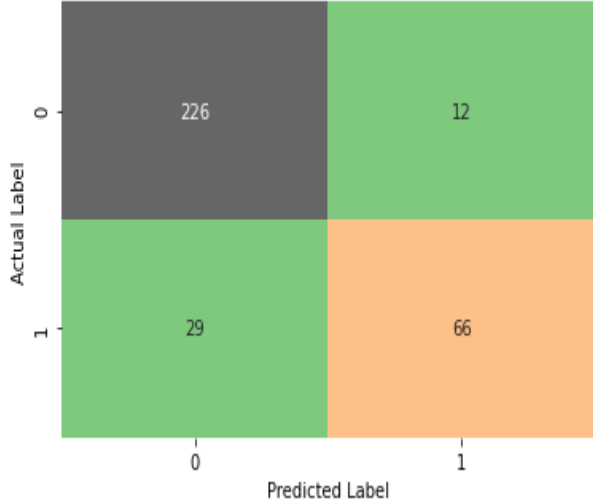
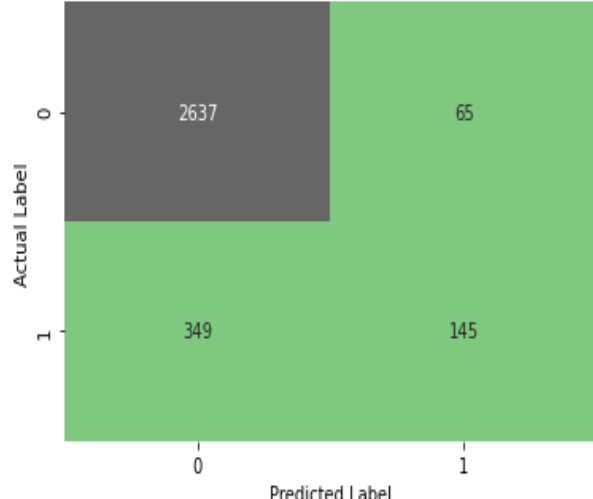
Bagging without grid search laptop dataset						Bagging without grid search Mobile dataset					
0.9819819819819819						0.9831038798498123					
[[238 0]						[[2692 10]					
[6 89]]						[44 450]]					
		precision	recall	f1-score	support			precision	recall	f1-score	support
	0	0.98	1.00	0.99	238		0	0.98	1.00	0.99	2702
	1	1.00	0.94	0.97	95		1	0.98	0.91	0.94	494
	accuracy			0.98	333		accuracy			0.98	3196
	macro avg	0.99	0.97	0.98	333		macro avg	0.98	0.95	0.97	3196
	weighted avg	0.98	0.98	0.98	333		weighted avg	0.98	0.98	0.98	3196

Confusion Matrix Baggingtest laptop

Actual \ Predicted	0	1
0	238	0
1	6	89

Confusion Matrix Baggingtest mobile

Actual \ Predicted	0	1
0	2692	10
1	44	450

ADA Boosting Model for laptop	ADA Boosting Model for mobile																																																												
0.8768768768768769 [[226 12] [29 66]]	0.8704630788485607 [[2637 65] [349 145]]																																																												
<table><tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr><tr><td>0</td><td>0.89</td><td>0.95</td><td>0.92</td><td>238</td></tr><tr><td>1</td><td>0.85</td><td>0.69</td><td>0.76</td><td>95</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.88</td><td>333</td></tr><tr><td>macro avg</td><td>0.87</td><td>0.82</td><td>0.84</td><td>333</td></tr><tr><td>weighted avg</td><td>0.87</td><td>0.88</td><td>0.87</td><td>333</td></tr></table>		precision	recall	f1-score	support	0	0.89	0.95	0.92	238	1	0.85	0.69	0.76	95	accuracy			0.88	333	macro avg	0.87	0.82	0.84	333	weighted avg	0.87	0.88	0.87	333	<table><tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr><tr><td>0</td><td>0.88</td><td>0.98</td><td>0.93</td><td>2702</td></tr><tr><td>1</td><td>0.69</td><td>0.29</td><td>0.41</td><td>494</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.87</td><td>3196</td></tr><tr><td>macro avg</td><td>0.79</td><td>0.63</td><td>0.67</td><td>3196</td></tr><tr><td>weighted avg</td><td>0.85</td><td>0.87</td><td>0.85</td><td>3196</td></tr></table>		precision	recall	f1-score	support	0	0.88	0.98	0.93	2702	1	0.69	0.29	0.41	494	accuracy			0.87	3196	macro avg	0.79	0.63	0.67	3196	weighted avg	0.85	0.87	0.85	3196
	precision	recall	f1-score	support																																																									
0	0.89	0.95	0.92	238																																																									
1	0.85	0.69	0.76	95																																																									
accuracy			0.88	333																																																									
macro avg	0.87	0.82	0.84	333																																																									
weighted avg	0.87	0.88	0.87	333																																																									
	precision	recall	f1-score	support																																																									
0	0.88	0.98	0.93	2702																																																									
1	0.69	0.29	0.41	494																																																									
accuracy			0.87	3196																																																									
macro avg	0.79	0.63	0.67	3196																																																									
weighted avg	0.85	0.87	0.85	3196																																																									
<p>Confusion Matrix ADABOOST laptop</p> 	<p>Confusion Matrix ADABOOST Mobile</p> 																																																												

Gradient Boosting for laptop					Gradient Boosting for Mobile				
0.954954954954955					0.9020650813516896				
[[235 3]					[[2669 33]				
[12 83]]					[280 214]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.95	0.99	0.97	238	0	0.91	0.99	0.94	2702
1	0.97	0.87	0.92	95	1	0.87	0.43	0.58	494
accuracy			0.95	333	accuracy			0.90	3196
macro avg	0.96	0.93	0.94	333	macro avg	0.89	0.71	0.76	3196
weighted avg	0.96	0.95	0.95	333	weighted avg	0.90	0.90	0.89	3196

