

**SOCIAL mEDIA TOURISM**

Data Science and Business Analytics | Capstone Project



July 7, 2024

NITYA SINGH

GREAT LEARNING

Table of Contents:

|  |  |
| --- | --- |
| **Topic** | **Page No** |
| **1) Introduction of the business problem** | **2** |
| Defining problem statement | 2 |
| Understanding business/social opportunity | 2 |
| Need of the study/project | 2 |
| **2)Data Report** | **3** |
| Understanding how data was collected in terms of time, frequency and methodology | 3 |
| Visual inspection of data (rows, columns, descriptive details) | 3 |
| Understanding of attributes (variable info, renaming if required) | 4 |
| **3) Exploratory data analysis** | **4** |
| Removal of unwanted variables (if applicable) | 4 |
| Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) | 7 |
| Outlier treatment (if required) | 11 |
| Missing Value treatment (if applicable) | 13 |
| Bivariate analysis (relationship between different variables, correlations) | 15 |
| **4) Business insights from EDA** | **18** |
| Is the data unbalanced? If so, what can be done? Please explain in the context of the business | 18 |
| Any business insights using clustering | 19 |
| Any other business insights | 19 |
| **5) Model building and interpretation.** | **20** |
| Build various models with model tuning | 20 |
| Interpretation of the most optimum model and its implication on the business | 47 |

**1. INTRODUCTION OF THE BUSINESS PROBLEM**

**Problem Statement**

An aviation company that provides domestic as well as international trips to the customers now wants to apply a targeted approach instead of reaching out to each of the customers. This time they want to do it digitally instead of tele calling. Hence they have collaborated with a social networking platform, so they can learn the digital and social behaviour of the customers and provide the digital advertisement on the user page of the targeted customers who have a high propensity to take up the product.

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

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

* **Understanding:**

An aviation company is looking to enhance its marketing strategy by shifting from traditional tele-calling methods to targeted digital advertising. To achieve this, they have partnered with a social networking platform to analyse the digital and social behaviour of their customers. The goal is to deliver digital advertisements on the user pages of customers who have a high propensity to purchase flight tickets.

* **Need of the study/project:**

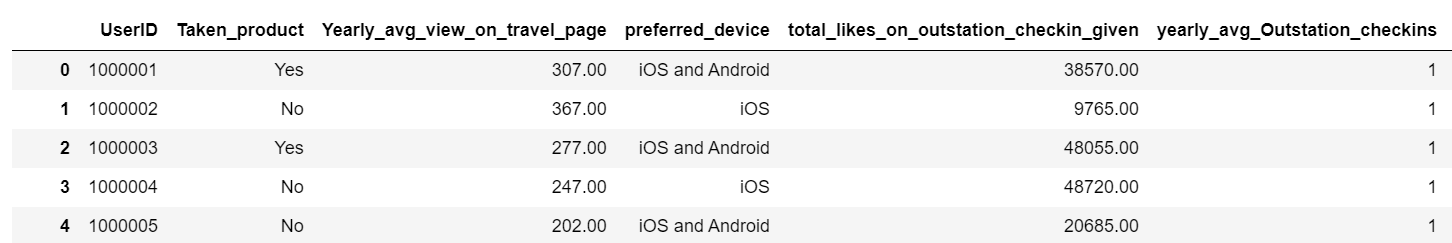
The need to study and undertake this project is driven by several factors. First, understanding customer behaviour through digital and social data allows for the creation of personalized and effective advertisements, thereby increasing engagement and conversion rates. Second, developing device-specific models ensures that the unique interaction patterns on laptops and mobile devices are appropriately addressed, enhancing the precision of the targeting strategy. Given the high cost of digital advertisements, it is imperative to accurately identify and target only those customers who are most likely to buy tickets, ensuring efficient use of marketing resources and maximizing return on investment.

Moreover, a data-driven approach to decision-making enables the company to rely on empirical evidence rather than intuition, leading to more informed and effective marketing strategies.

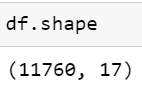
**2. Data Report:**

* **Understanding the data:**

Data was collected from the past records of the customers. Sample of the data available:

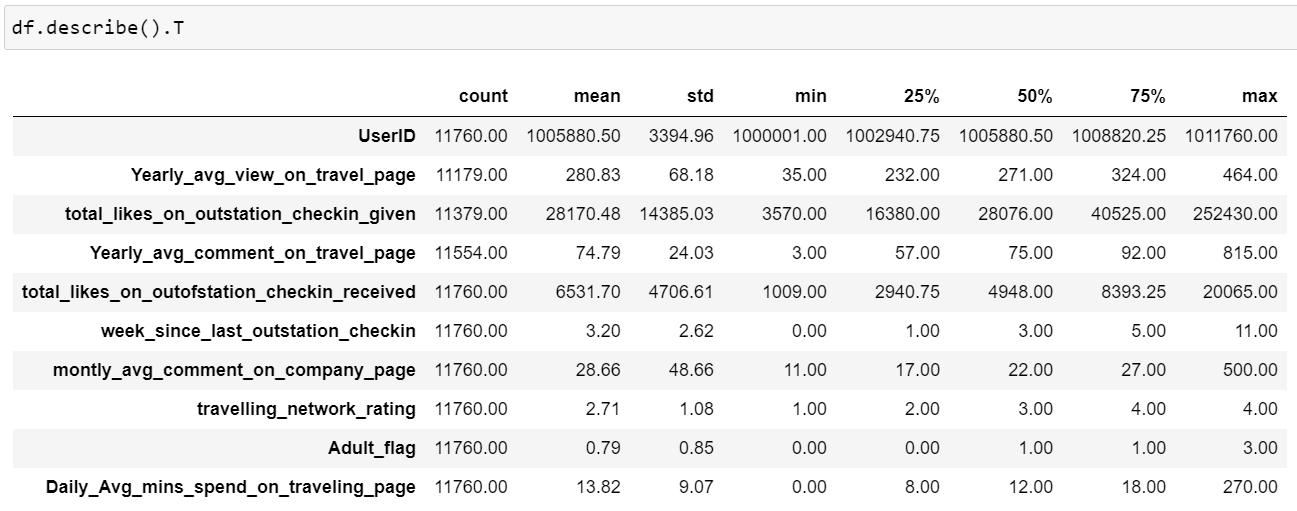


* **Visual inspection of data:**

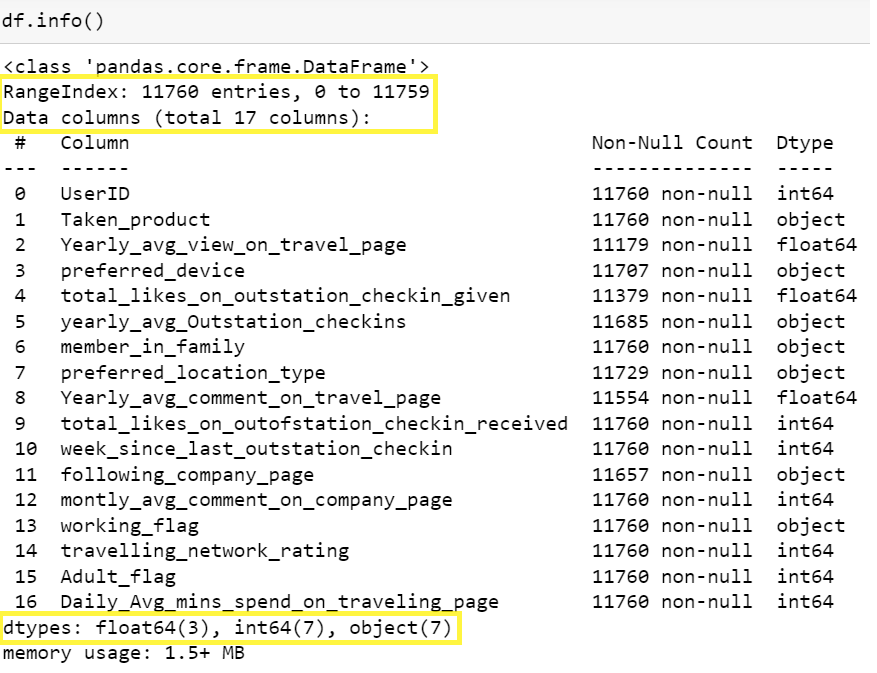


There are 11760 rows and 17 columns in the dataset

Using describe() to get the summary statistics of a dataset, including measures like mean, median, standard deviation, and percentiles:



* **Understanding of attributes:**



Total entries in dataset: 11760 entries

Total columns: 17 columns

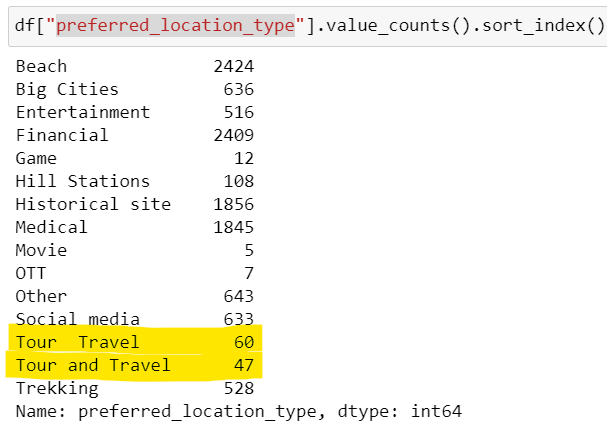
Data types: There are 3 float values, 7 integer values and 7 object type values

**3. Exploratory data analysis:**

Dropping column ‘UserID’ as its not required for the analysis.

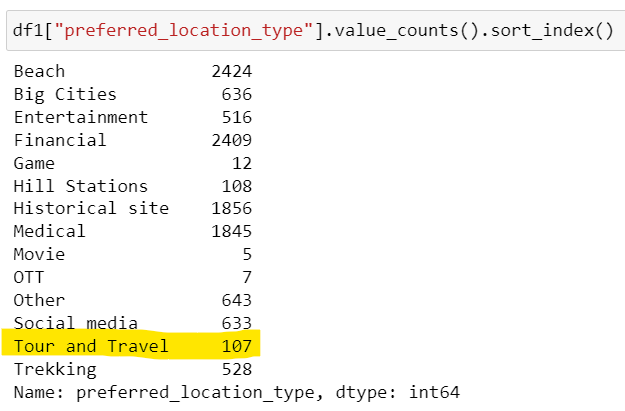
Adjusting incorrect values in dataset,

Column ‘**preferred\_location\_type’** contains two entries ‘Tour Travel’ and ‘Tour and Travel’ which should be the same entry.

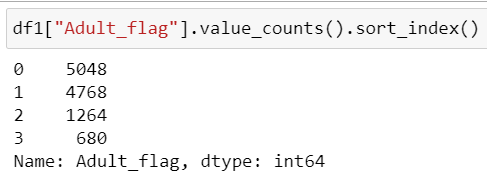


Fixing the duplicate values,



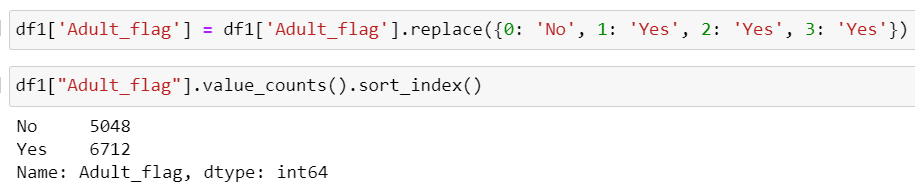


Column ‘**Adult\_flag’** should be a categorical column that should contain values ‘Yes’ or ‘No’ but the column is numerical type with multiple values:

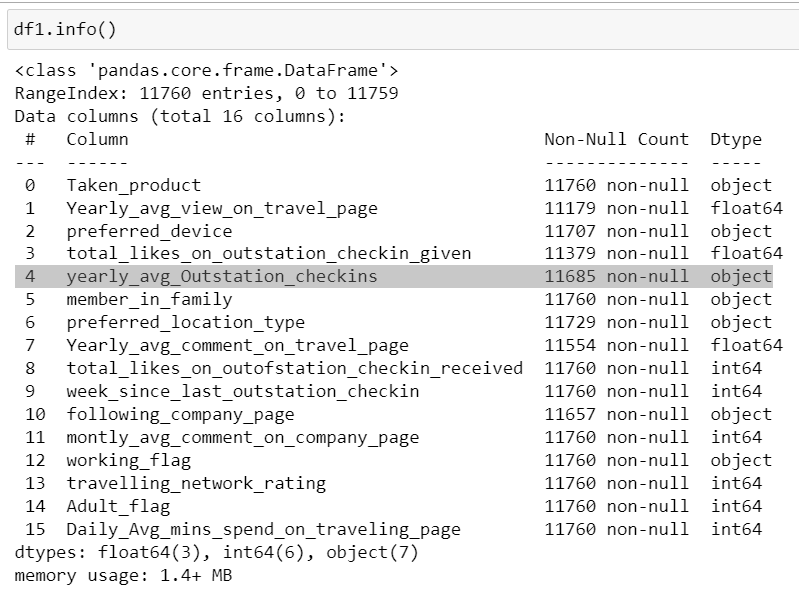


Column should be changed to categorical. Value 0 should be changed to ‘No’ and all the values greater than 0 should be changed to ‘Yes’





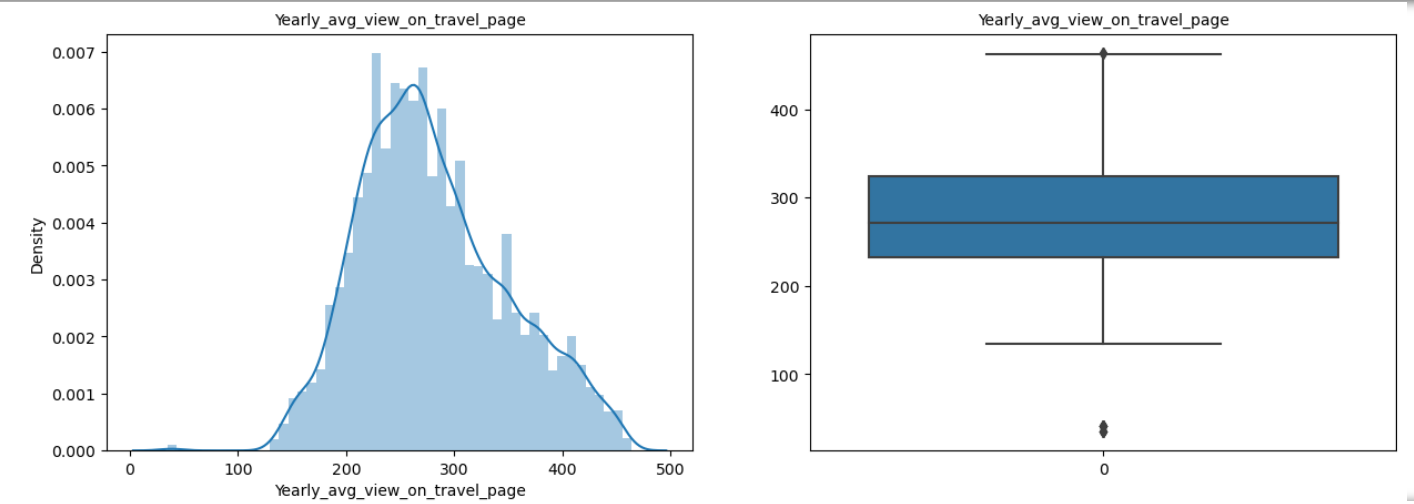
Column **‘yearly\_avg\_Outstation\_checkins’** is categorical type but should be changed to numeric type

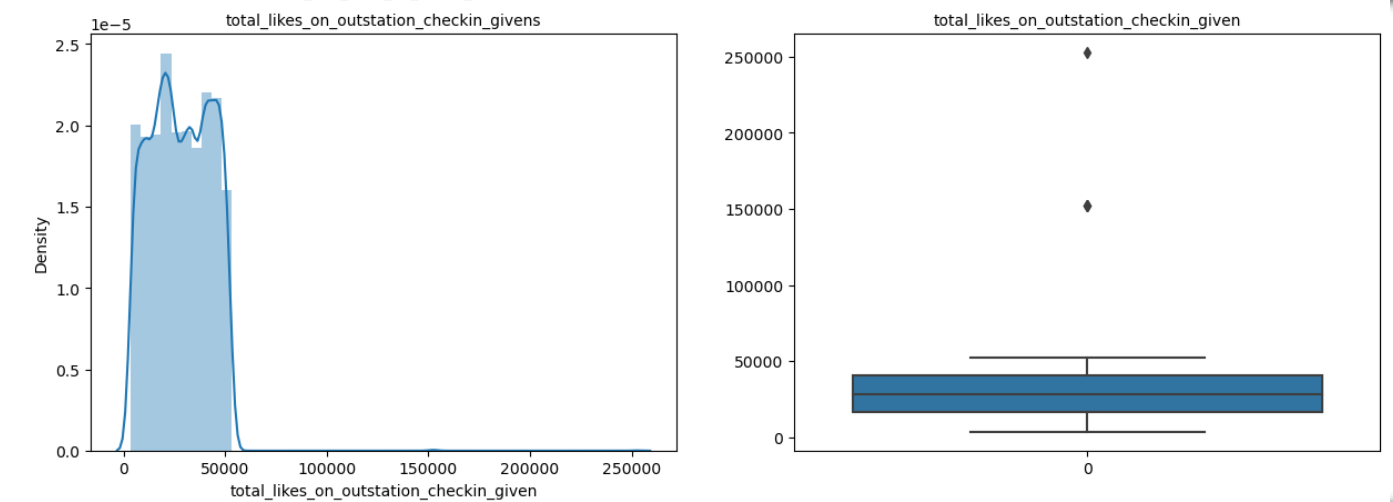


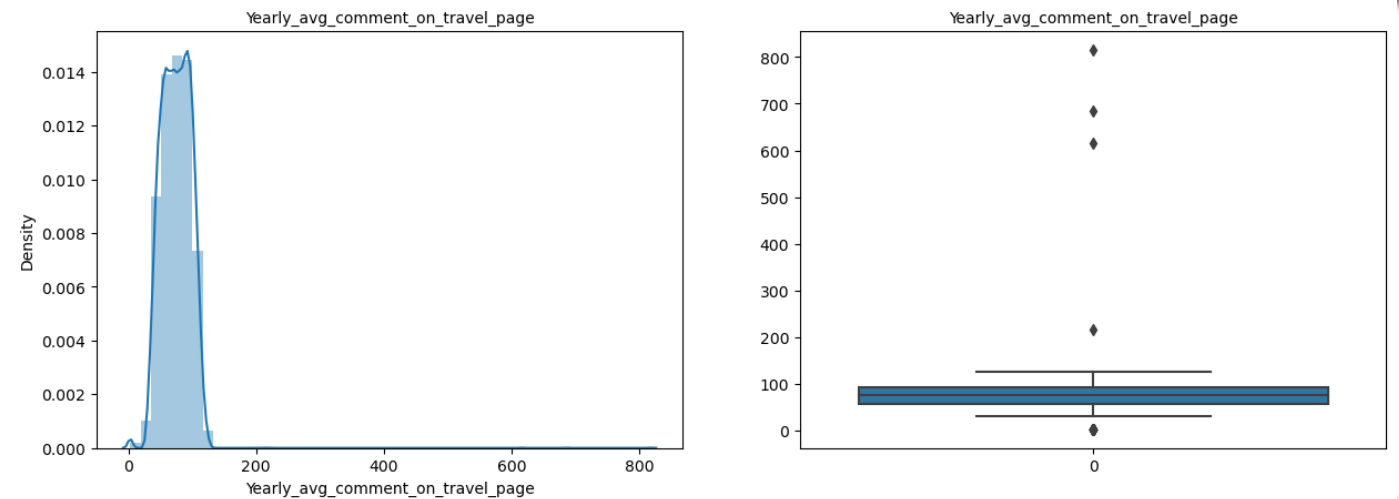


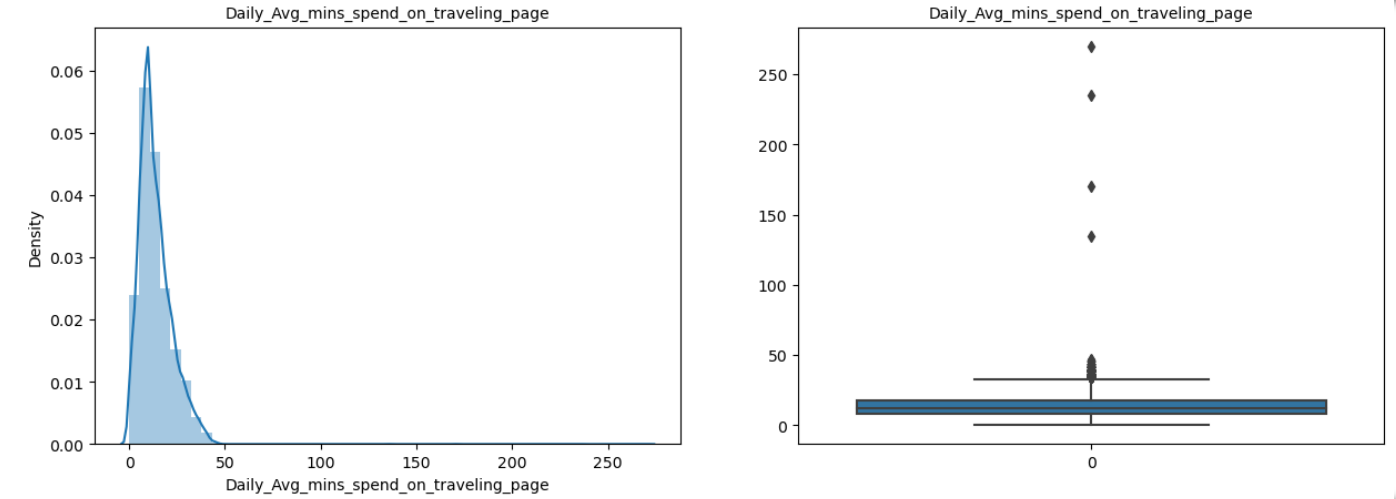
* **Data Visualisation- Univariant Analysis**

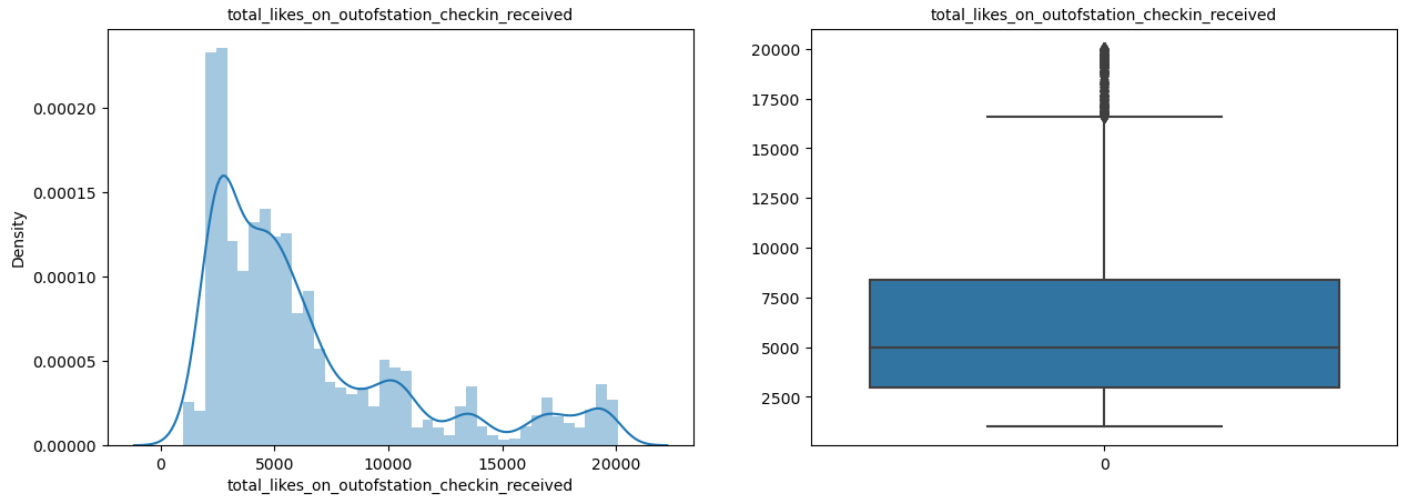
Numeric Data:

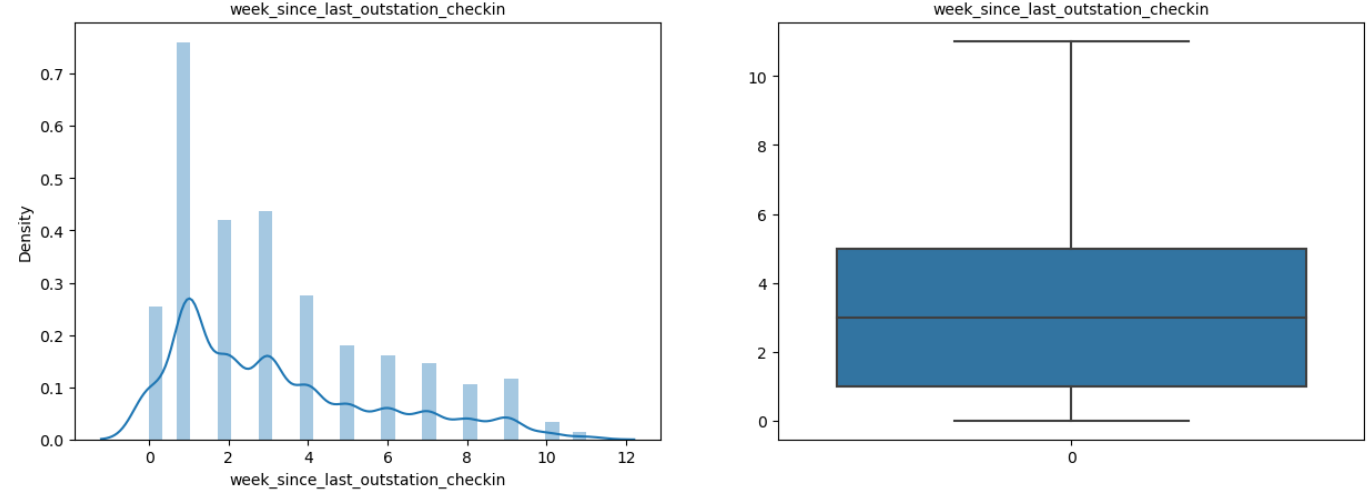
****

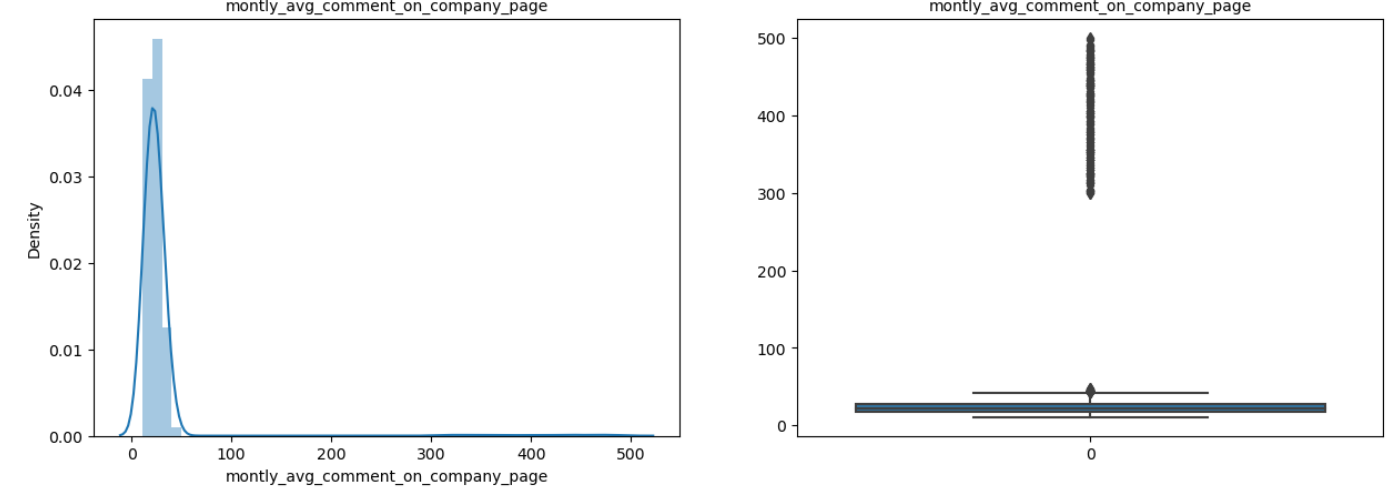
****

****

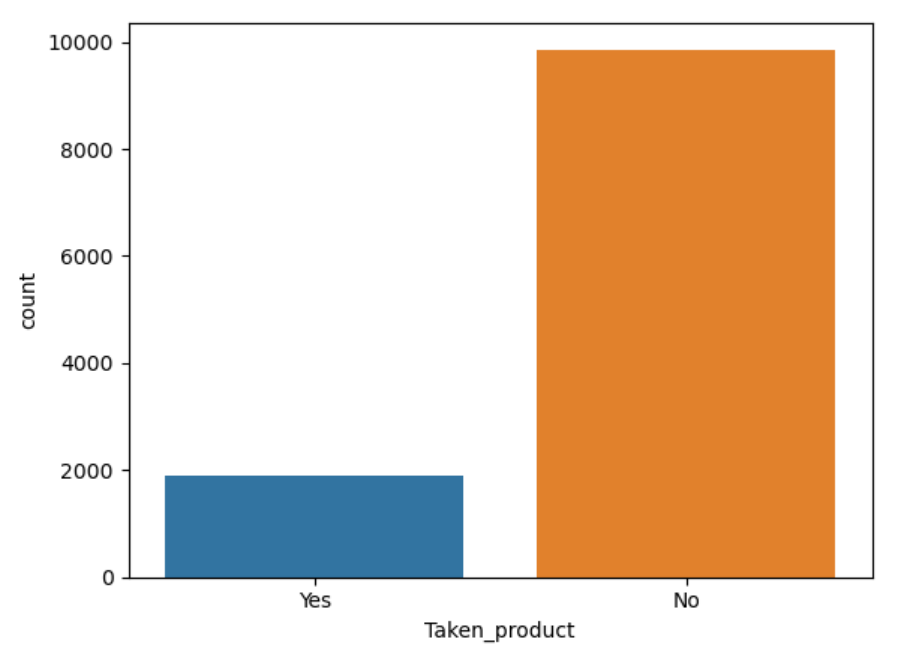
****

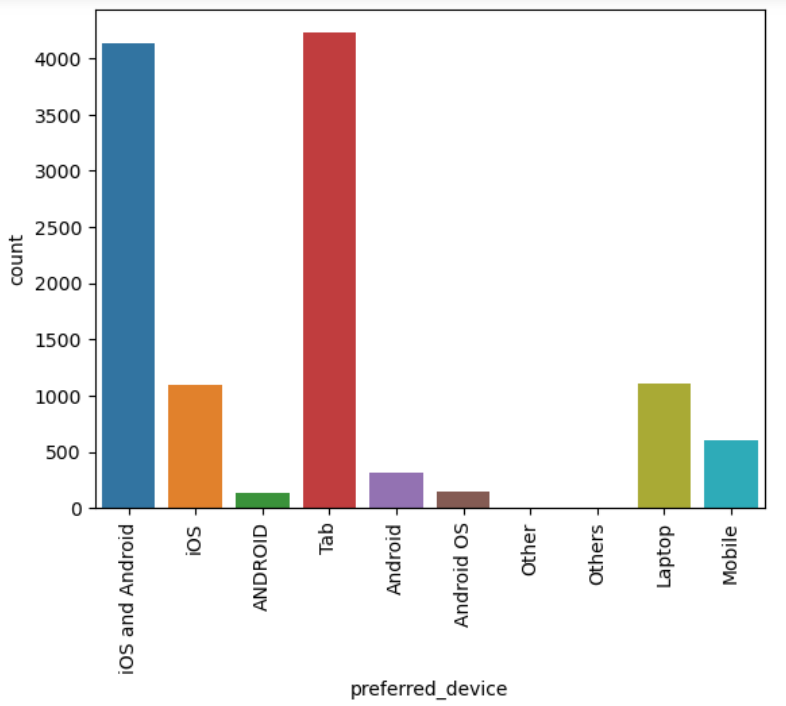
****

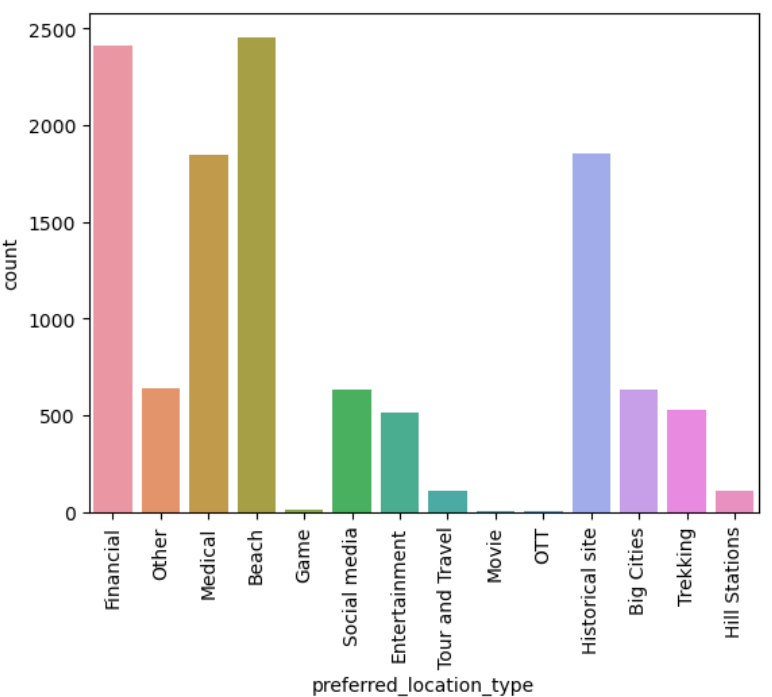
****

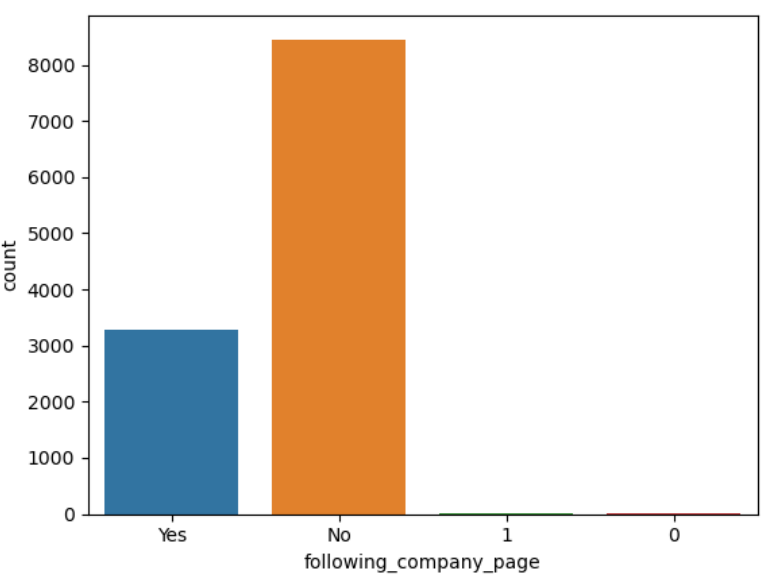
****

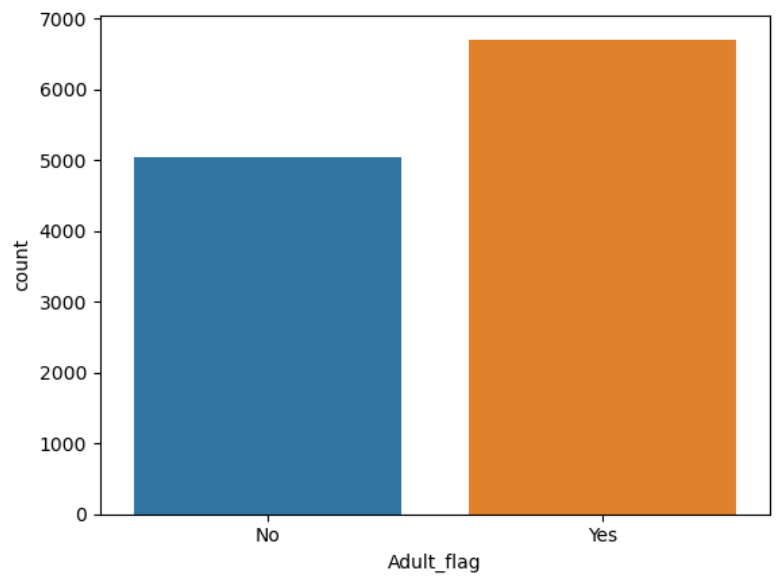
Categorical Data:





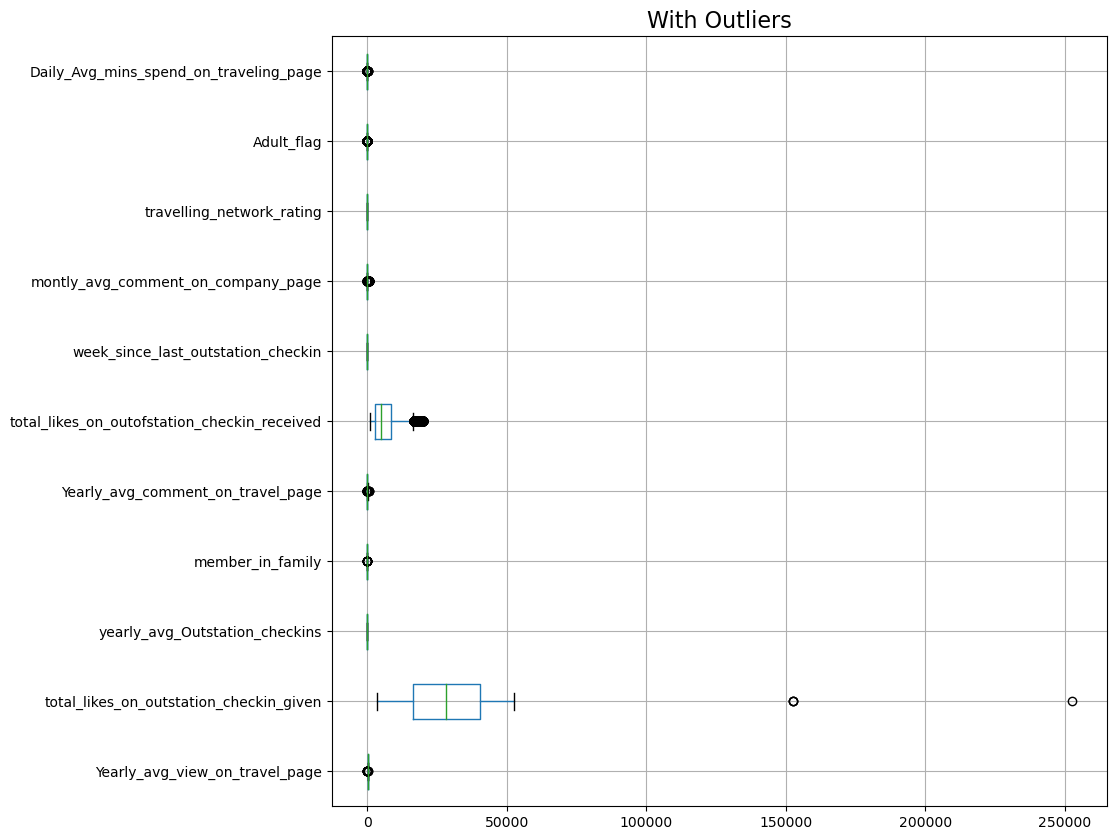




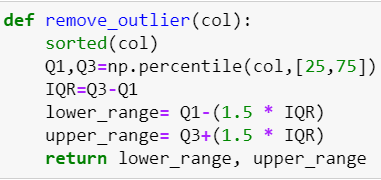


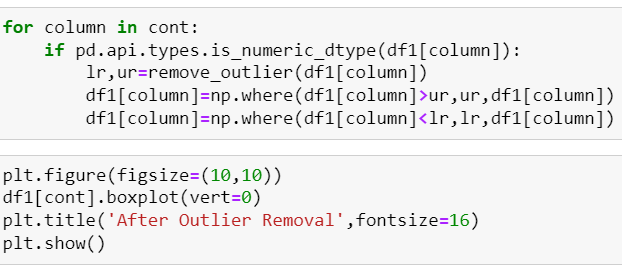
Outliners:

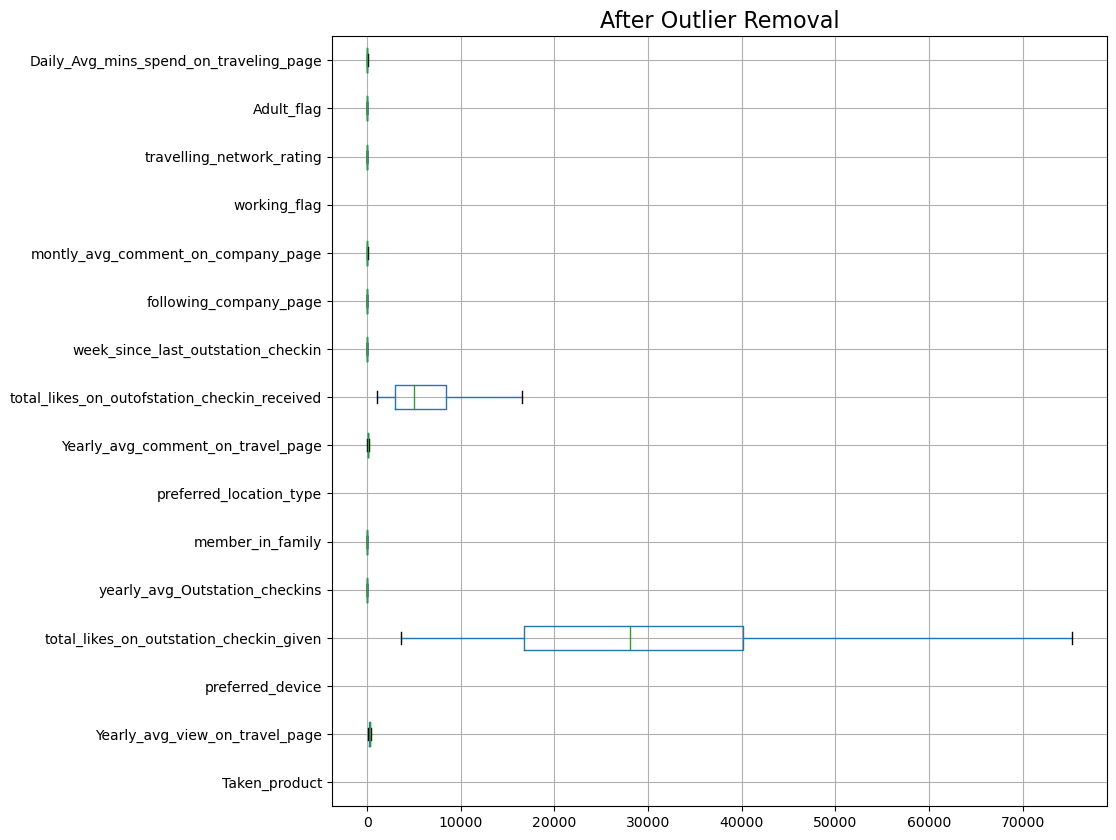
Checking outliners by constructing boxplot for continuous variables:



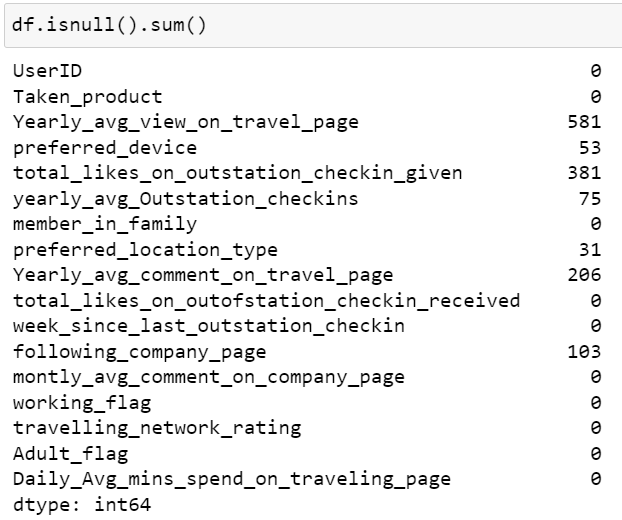
Treating outliners:



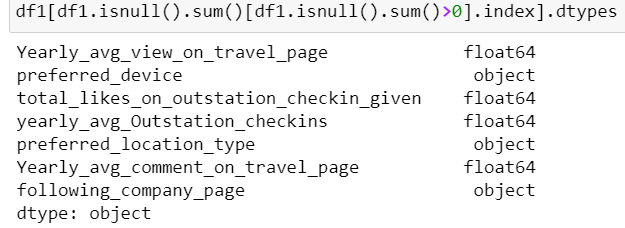




Missing Values in the dataset:

****

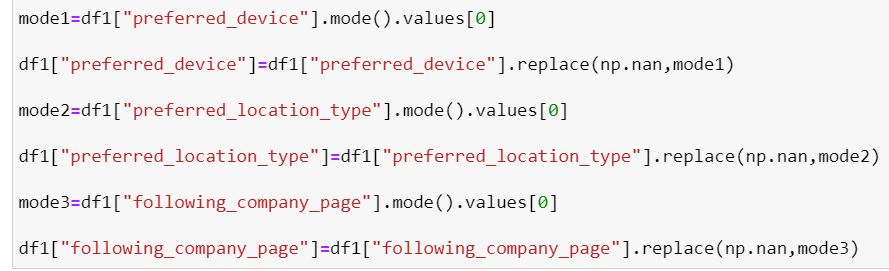
checking the data type of the missing columns:

****

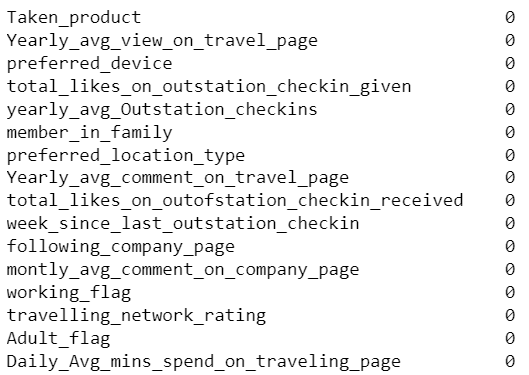
Replacing NULL values in Numerical Columns using Median:



Replacing NULL values in Categorical Columns using Mode:

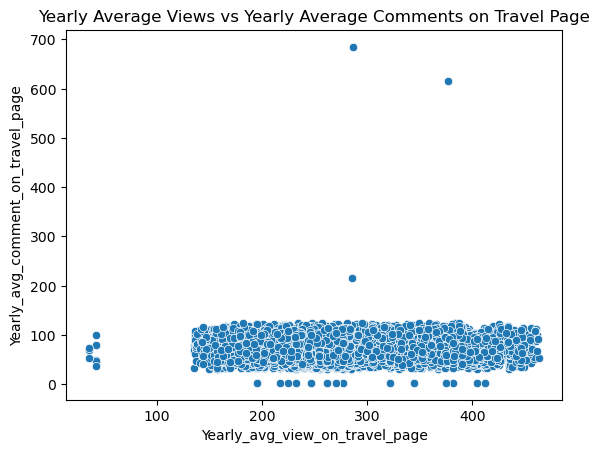


Missing values has been treated in the dataset:

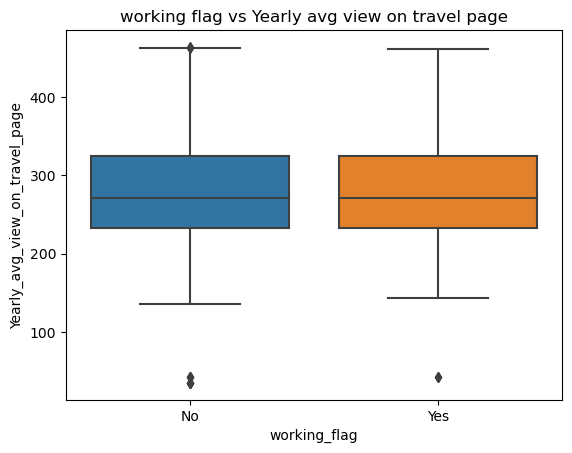


* **Bivariant Analysis:**

To see if there is a relationship between the average yearly views and average yearly comments on travel-related pages.

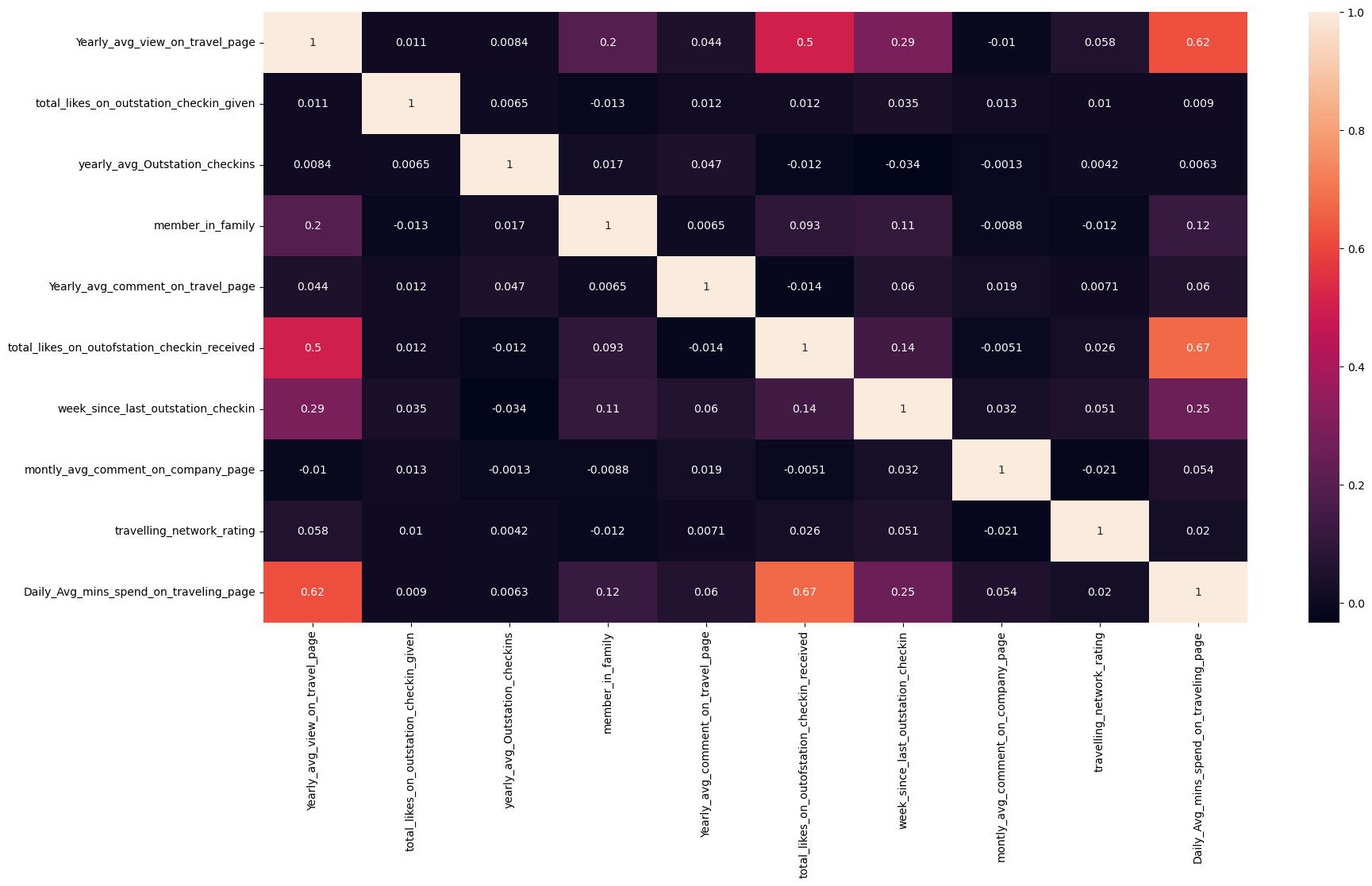


To see how the average yearly views on travel pages differ between working and non-working customers:

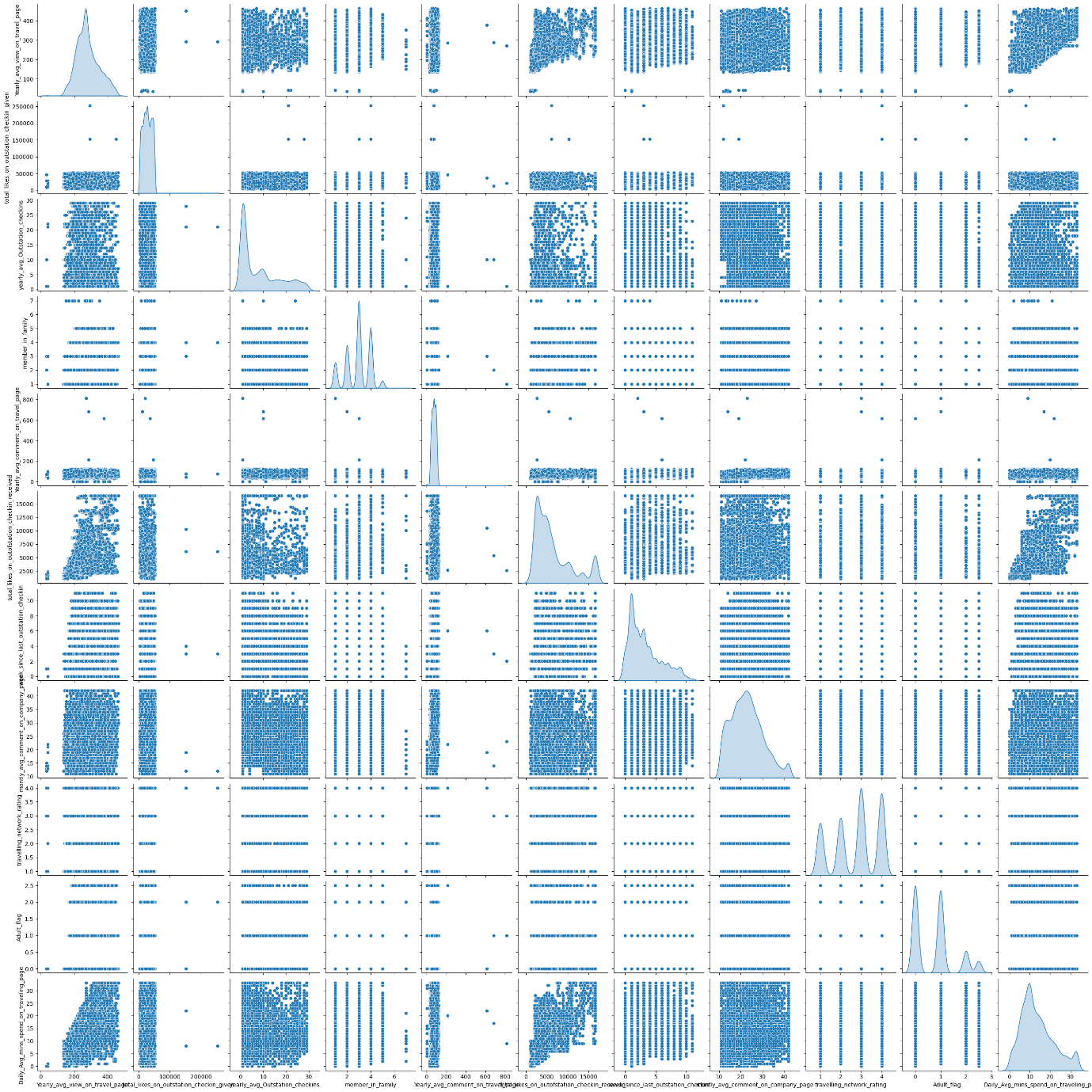


* **Multivariant Analysis:**

Heatmap:



Pairplot:



**4. Business insights from EDA**

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

Yes, data is unbalanced as the no to yes ratio is not equal. Smote should be done to treat it.

* **Any business insights using clustering**

Clustering is an unsupervised machine learning technique used to identify and group similar data points within larger datasets without focusing on a specific outcome. It is typically employed to classify data into more understandable and manageable structures. However, clustering is not needed for our purposes as our business objective is to make predictions.

* **Business insights:**
* We can observe that the most visited locations are beach and financial whereas the least visited place is hill station. This can be considered while targeting audience and providing discounts.
* Number of laptop users are higher than number of people using phone across all the different OS
* The count of non-adult consumers is high, though less than adults. Offers targeting non-adults can help to influence their family members.

**5. Model building and interpretation.**

We have trained the model using different techniques and algorithms on the prepared data. The process involves training a machine learning algorithm to predict labels from features, tuning it to meet business needs, and validating it on holdout data. The result is a trained model that can make predictions on new data points.

Following the exploratory data analysis (EDA), the data is now ready for model building. We have applied machine learning algorithms to the dataset.

Models used:

* Logistic Regression
* Random Forest
* K Nearest Neighbours (KNN)

Classification model is used to categorize data into specific classes or groups. The dependent variable is categorical binary.

The objective is to predict whether a product will be taken or not. The data is split between Laptop devices and Mobile devices. For model building I have divided the data into a 70:30 ratio. Where, Training data size is 70% and testing data size is 30% of data

**Data Split in Training and Testing**

The dependent variable is the column ‘Taken\_product’

The data is divided into two parts: a training set and a testing set. The training set contains 70% of the observations, while the testing set comprises the remaining 30%. The training set is used to fit our model and then evaluate its performance using the testing set.

* **Logistic Regression Model**

**LAPTOP USERS:**

Model Score (Train): 0.8206451612903226

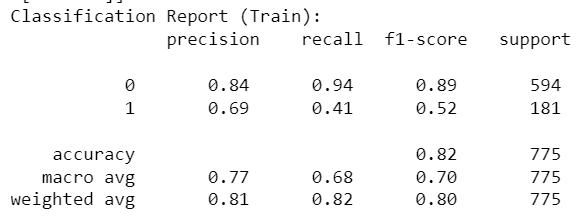
Model Score (Test): 0.8108108108108109

The model shows a slight drop in performance from the training set to the testing set, which is expected and indicates that the model generalizes reasonably well to unseen data.

Confusion Matrix (Train):

[[ **561 33** ]

[ **106 75** ]]



**Inference from Training Data:**

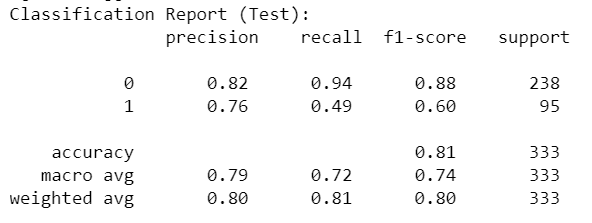
* Class 0 (Product not taken) has high precision (0.84) and recall (0.94), indicating the model is very good at identifying non-purchasers and making few mistakes.
* Class 1 (Product taken) has lower precision (0.69) and recall (0.41), suggesting that while the model correctly identifies a portion of actual purchasers, it misses many cases and makes a fair number of false positives.
* The overall accuracy of 82% is good, but the lower recall for the purchasing class indicates that improvements can be made, especially in capturing more positive cases.

Confusion Matrix (Test):

[[ **223 15** ]

[ **48 47** ]]

Classification Report (Test):

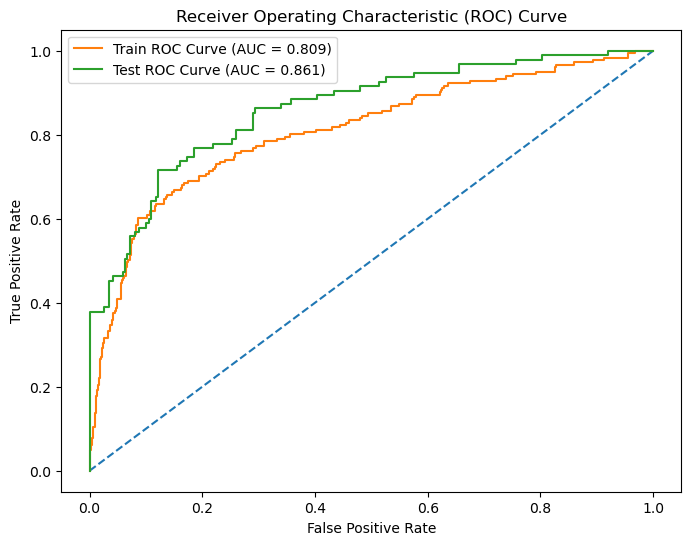


**Inference from Testing Data:**

* Class 0 (Product not taken) maintains high precision (0.82) and recall (0.94), similar to the training set, indicating the model performs consistently well in predicting non-purchasers.
* Class 1 (Product taken) has reduced precision (0.76) and recall (0.49) compared to training data, suggesting a decrease in performance on unseen data. The model is less effective at identifying purchasers, with a higher number of false negatives compared to the training set.

Train AUC: 0.809

Test AUC: 0.861



The **AUC (Area Under the Curve)** is higher on the test set (0.861) than on the train set (0.809). This indicates that the model has a better discriminatory power on the testing data compared to the training data. It shows the model's ability to distinguish between the classes is slightly better on new data

**MOBILE USERS:**

Model Score (Train): 0.8489806866952789

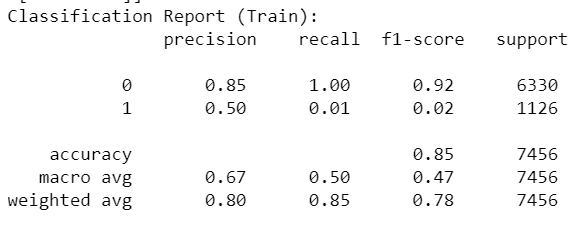
Model Score (Test): 0.8438673341677096

Confusion Matrix (Train):

[[**6320 10**]

[**1116 10**]]

Classification Report (Train):



**Inference from Training Data:**

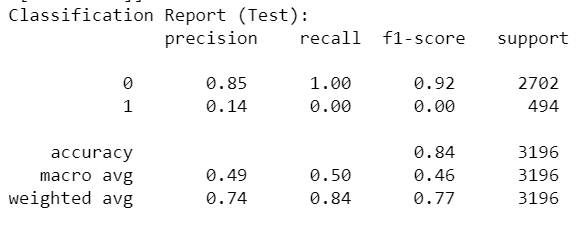
* Class 0 (Not purchasing) has high precision (0.85) and perfect recall (1.00), showing that the model is very effective at predicting non-purchasers.
* Class 1 (Purchasing) shows poor performance with a very low recall (0.01) and precision (0.50), indicating that the model struggles significantly to identify actual purchasers.
* The overall accuracy of 85% is high, but the very low recall for the purchasing class means the model fails to capture most of the positive cases, resulting in a large number of false negatives.

Confusion Matrix (Test):

[[**2696 6**]

[ **493 1**]]

Classification Report (Test):



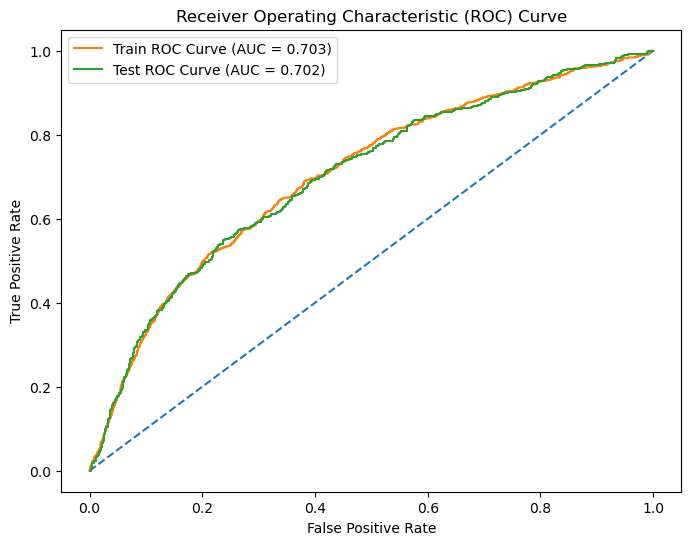
**Inference from Testing Data:**

* Class 0 (Not purchasing) maintains high precision (0.85) and perfect recall (1.00), similar to the training set, indicating consistent performance in predicting non-purchasers.
* Class 1 (Purchasing) has very low precision (0.14) and recall (0.00) on the test set, showing that the model is nearly ineffective at predicting purchasers, with nearly all actual purchasers being missed.
* The accuracy of 84% is good, but this figure is misleading due to the model's failure to detect the positive class. Most of the positive cases are classified as negative, resulting in a significant number of false negatives.

Train AUC: 0.703

Test AUC: 0.702

The **AUC (Area Under the Curve)** is very similar for both the train and test sets (around 0.70). This suggests that the model’s ability to distinguish between the classes is relatively stable across the training and testing datasets.



The model performs well in terms of accuracy for both training and testing data but is heavily biased towards the negative class.

**USING MODEL TUNING**

Model Tuning is basically enhancing the model using hyper parameters. Newer features are used and feeded to the model and the computation is done on all the features and the best parameters or values are used to build a model and then train on it. GridSearch and best\_estimators are used for the same

**LAPTOP USERS:**

Model Score (Train): 0.7858064516129032

Model Score (Test): 0.7357357357357357

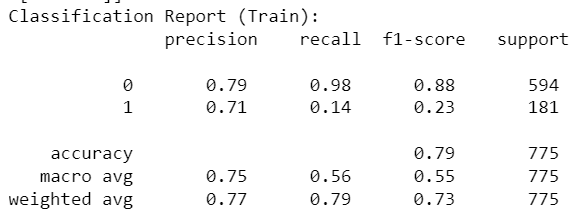
There is a noticeable drop in the model score from the training set to the testing set, indicating some degree of overfitting or challenges in generalizing to unseen data.

Confusion Matrix (Train):

[[**584 10**]

[**156 25**]]

Classification Report (Train):



**Inference from Training Data:**

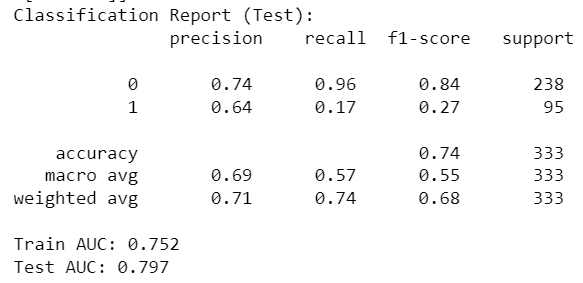
* Class 0 (Not purchasing) has high precision (0.79) and very high recall (0.98), showing that the model is effective at predicting non-purchasers.
* Class 1 (Purchasing) shows poor performance with low recall (0.14) and moderate precision (0.71), indicating that the model still struggles to identify actual purchasers.
* The overall accuracy of 79% is decent, but the very low recall for the purchasing class means the model still misses many positive cases, resulting in a significant number of false negatives.

Confusion Matrix (Test):

[[**229 9**]

[ **79 16**]]

Classification Report (Test):

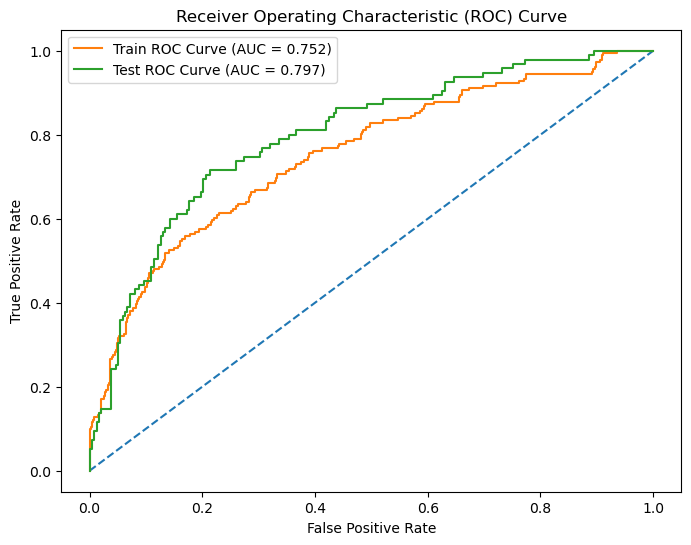


**Inference from Testing Data:**

* Class 0 (Not purchasing) maintains high precision (0.74) and very high recall (0.96), similar to the training set, indicating consistent performance in predicting non-purchasers.
* Class 1 (Purchasing) still has low precision (0.64) and very low recall (0.17) on the test set, showing that the model is still ineffective at predicting purchasers, with a majority of actual purchasers being missed.
* The accuracy of 74% is good, but similar to the training data, this figure is misleading due to the model's failure to detect the positive class. Most positive cases are classified as negative, resulting in a significant number of false negatives.

Train AUC: 0.752

Test AUC: 0.797



**Inference from AUC (Area Under the Curve):**

The AUC (Area Under the Curve) is somewhat higher for the test set (0.797) than the train set (0.752), indicating that the model has a reasonable ability to distinguish between classes. However, the performance discrepancy between precision and recall for the purchasing class suggests room for improvement.

**MODEL TUNING FOR MOBILE USERS:**

Model Score (Train): 0.8492489270386266

Model Score (Test): 0.8454317897371715

Confusion Matrix (Train):

[[6330 0]

[1124 2]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
|  |
| 0 0.85 1.00 0.92 6330 |
| 1 1.00 0.00 0.00 1126 |
|  |
| accuracy 0.85 7456 |
| macro avg 0.92 0.50 0.46 7456 |
| weighted avg 0.87 0.85 0.78 7456 |

Confusion Matrix (Test):

[[2702 0]

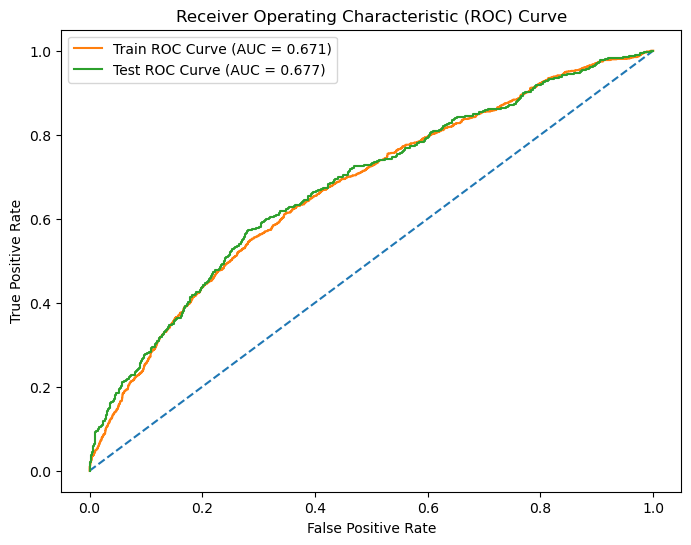
[ 494 0]]

Classification Report (Test):

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

Train AUC: 0.671

Test AUC: 0.677



**Inference:**

* Class 0 is perfectly predicted with 100% recall.
* Class 1 is entirely missed, indicating the model is biased towards predicting non-purchasers.
* The AUC scores are low, suggesting poor overall model performance in distinguishing between the two classes.
* **Random Forest Model**

**LAPTOP USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.975975975975976

Confusion Matrix (Train):

[[594 0]

[ 0 181]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
|  |
| 0 1.00 1.00 1.00 594 |
| 1 1.00 1.00 1.00 181 |
|  |
| accuracy 1.00 775 |
| macro avg 1.00 1.00 1.00 775 |
| weighted avg 1.00 1.00 1.00 775 |

**Training Performance:**

* The model perfectly classifies both classes in the training data, achieving 100% precision, recall, and F1-scores for both classes. This indicates the model has fully memorized the training data, showing an ideal training accuracy of 1.00.
* The AUC score of 1.000 confirms perfect discrimination ability in the training set.

Confusion Matrix (Test):

[[238 0]

[ 8 87]]

Classification Report (Test):

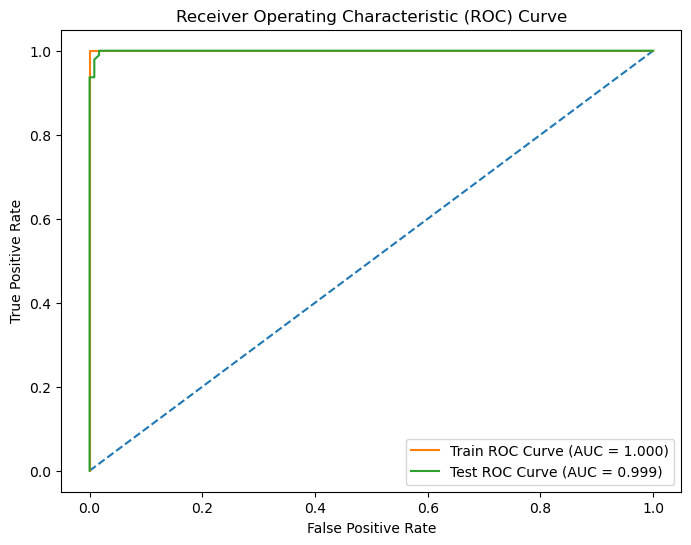
|  |
| --- |
| precision recall f1-score support |
|  |
| 0 0.97 1.00 0.98 238 |
| 1 1.00 0.92 0.96 95 |
|  |
| accuracy 0.98 333 |
| macro avg 0.98 0.96 0.97 333 |
| weighted avg 0.98 0.98 0.98 333 |

**Testing Performance:**

* The overall test accuracy is extremely high at 0.976, indicating the model generalizes well to unseen data.
* The test AUC score of 0.999 indicates the model has an excellent ability to distinguish between the two classes in the test set.

Train AUC: 1.000

Test AUC: 0.999



**Observations:**

* The model exhibits strong performance on both the training and testing datasets, suggesting effective learning from the data.
* The high precision and recall for both classes in the test set imply the model is well-calibrated and effective at correctly identifying both purchasers and non-purchasers.
* The near-perfect AUC scores suggest that the model is highly reliable in distinguishing between the classes, both in training and testing scenarios.

**MOBILE USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.980287859824781

Confusion Matrix (Train):

[[6330 0]

[ 0 1126]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 6330 |
| 1 1.00 1.00 1.00 1126 |
|  |
| accuracy 1.00 7456 |
| macro avg 1.00 1.00 1.00 7456 |
| weighted avg 1.00 1.00 1.00 7456 |

**Training Performance:**

* The model perfectly classifies both classes in the training data, achieving 100% precision, recall, and F1-scores for both classes, indicating the model has fully memorized the training data, showing an ideal training accuracy of 1.00.
* The AUC score of 1.000 confirms perfect discrimination ability in the training set.

Confusion Matrix (Test):

[[2702 0]

[ 63 431]]

Classification Report (Test):

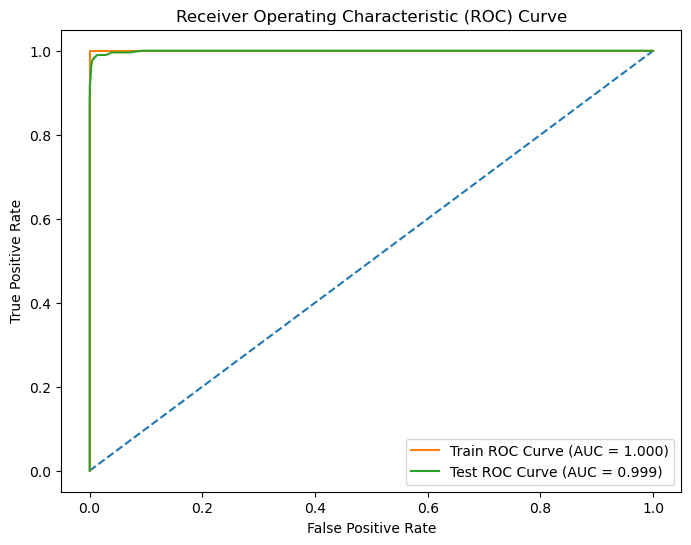
|  |
| --- |
| precision recall f1-score support |
| 0 0.98 1.00 0.99 2702 |
| 1 1.00 0.87 0.93 494 |
|  |
| accuracy 0.98 3196 |
| macro avg 0.99 0.94 0.96 3196 |
| weighted avg 0.98 0.98 0.98 3196 |

**Testing Performance:**

* The overall test accuracy is extremely high at 0.980, indicating the model generalizes well to unseen data.
* The test AUC score of 0.999 indicates the model has an excellent ability to distinguish between the two classes in the test set.

Train AUC: 1.000

Test AUC: 0.999



**Observations:**

* The model exhibits strong performance on both the training and testing datasets, suggesting effective learning from the data.
* The high precision and recall for both classes in the test set imply the model is well-calibrated and effective at correctly identifying both purchasers and non-purchasers.
* The near-perfect AUC scores suggest that the model is highly reliable in distinguishing between the classes, both in training and testing scenarios.

**MODEL TUNING:**

**LAPTOP USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.987987987987988

Confusion Matrix (Train):

[[594 0]

[ 0 181]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 594 |
| 1 1.00 1.00 1.00 181 |
|  |
| accuracy 1.00 775 |
| macro avg 1.00 1.00 1.00 775 |
| weighted avg 1.00 1.00 1.00 775 |

**Training Performance:**

* The model achieves perfect precision, recall, and F1-scores for both classes in the training data, indicating it has memorized the training data completely. This results in an ideal training accuracy of 1.00.
* The AUC score of 1.000 confirms the model's perfect discrimination ability in the training set.

Confusion Matrix (Test):

[[238 0]

[ 4 91]]

Classification Report (Test):

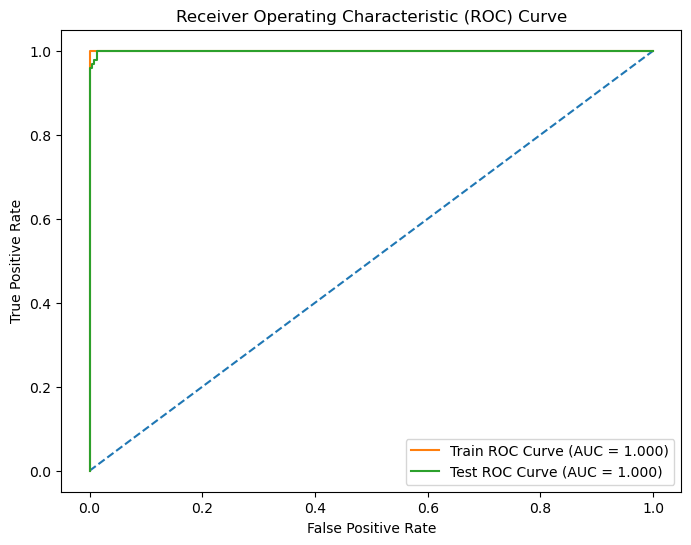
|  |
| --- |
| precision recall f1-score support |
| 0 0.98 1.00 0.99 238 |
| 1 1.00 0.96 0.98 95 |
|  |
| accuracy 0.99 333 |
| macro avg 0.99 0.98 0.99 333 |
| weighted avg 0.99 0.99 0.99 333 |

**Testing Performance:**

* The overall test accuracy is very high at 0.988, demonstrating the model generalizes extremely well to unseen data.
* The test AUC score of 1.000 indicates the model has a perfect ability to distinguish between the two classes in the test set.

Train AUC: 1.000

Test AUC: 1.000



**Observations:**

* The model shows excellent performance on both the training and testing datasets, indicating effective learning from the data.
* The high precision and recall for both classes in the test set imply the model is well-calibrated and effective at correctly identifying both purchasers and non-purchasers.
* The perfect AUC scores suggest the model is highly reliable in distinguishing between the classes, both in training and testing scenarios.

**MODEL TUNING FOR MOBILE USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.9793491864831039

Confusion Matrix (Train):

[[6330 0]

[ 0 1126]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 6330 |
| 1 1.00 1.00 1.00 1126 |
|  |
| accuracy 1.00 7456 |
| macro avg 1.00 1.00 1.00 7456 |
| weighted avg 1.00 1.00 1.00 7456 |

**Training Performance:**

* The model achieves perfect precision, recall, and F1-scores for both classes in the training data, indicating it has memorized the training data completely. This results in an ideal training accuracy of 1.00.
* The AUC score of 1.000 confirms the model's perfect discrimination ability in the training set.

Confusion Matrix (Test):

[[2702 0]

[ 66 428]]

Classification Report (Test):

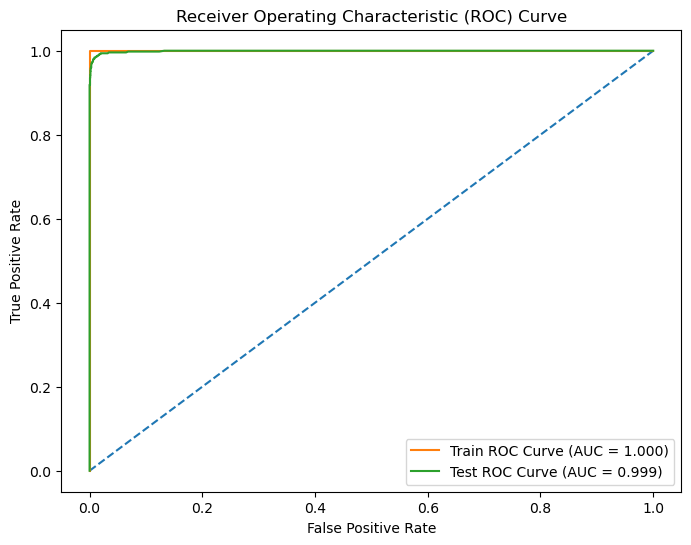
|  |
| --- |
| precision recall f1-score support |
| 0 0.98 1.00 0.99 2702 |
| 1 1.00 0.87 0.93 494 |
|  |
| accuracy 0.98 3196 |
| macro avg 0.99 0.93 0.96 3196 |
| weighted avg 0.98 0.98 0.98 3196 |

**Testing Performance:**

* The overall test accuracy is very high at 0.979, demonstrating the model generalizes extremely well to unseen data.
* The test AUC score of 0.999 indicates the model has an almost perfect ability to distinguish between the two classes in the test set.

Train AUC: 1.000

Test AUC: 0.999



**General Observations:**

* The model shows excellent performance on both the training and testing datasets, indicating effective learning from the data.
* The high precision and recall for both classes in the test set imply the model is well-calibrated and effective at correctly identifying both purchasers and non-purchasers.
* The near-perfect AUC scores suggest the model is highly reliable in distinguishing between the classes, both in training and testing scenarios.
* However, the slight drop in recall for class 1 in the test set compared to the training set indicates some room for improvement in identifying all purchasing instances.
* **K Nearest Neighbours (KNN)**

**LAPTOP USERS:**

Model Score (Train): 0.9483870967741935

Model Score (Test): 0.8558558558558559

Confusion Matrix (Train):

[[579 15]

[ 25 156]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 0.96 0.97 0.97 594 |
| 1 0.91 0.86 0.89 181 |
|  |
| accuracy 0.95 775 |
| macro avg 0.94 0.92 0.93 775 |
| weighted avg 0.95 0.95 0.95 775 |

**Training Performance:**

* The model achieves high precision, recall, and F1-scores for both classes in the training data, resulting in an accuracy of 0.95.
* The AUC score of 0.982 indicates excellent discrimination ability on the training set.

Confusion Matrix (Test):

[[221 17]

[ 31 64]]

Classification Report (Test):

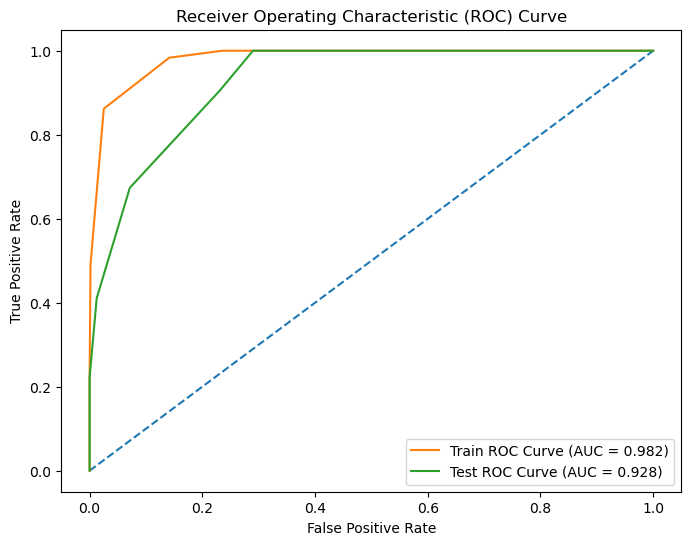
|  |
| --- |
| precision recall f1-score support |
| 0 0.88 0.93 0.90 238 |
| 1 0.79 0.67 0.73 95 |
|  |
| accuracy 0.86 333 |
| macro avg 0.83 0.80 0.81 333 |
| weighted avg 0.85 0.86 0.85 333 |

**Testing Performance:**

* The overall test accuracy is 0.856, demonstrating the model generalizes well to unseen data.
* The test AUC score of 0.928 confirms the model's strong ability to distinguish between the two classes in the test set.

Train AUC: 0.982

Test AUC: 0.928



**Observations:**

* The model shows high performance on both the training and testing datasets, indicating effective learning from the data.
* The high precision and recall for class 0 in the test set imply the model is well-calibrated and effective at correctly identifying non-purchasers.
* The good but lower recall for class 1 in the test set suggests the model could benefit from further tuning or additional features to better identify all purchasing instances.
* The high AUC scores in both training and testing scenarios indicate the model's robustness in distinguishing between the classes.

**MOBILE USERS:**

Model Score (Train): 0.9837714592274678

Model Score (Test): 0.9652690863579474

Confusion Matrix (Train):

[[6307 23]

[ 98 1028]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 0.98 1.00 0.99 6330 |
| 1 0.98 0.91 0.94 1126 |
|  |
| accuracy 0.98 7456 |
| macro avg 0.98 0.95 0.97 7456 |
| weighted avg 0.98 0.98 0.98 7456 |

**Training Performance:**

* The model achieves high precision, recall, and F1-scores for both classes in the training data, resulting in an accuracy of 0.984.
* The AUC score of 0.998 indicates near-perfect discrimination ability on the training set.

Confusion Matrix (Test):

[[2676 26]

[ 85 409]]

Classification Report (Test):

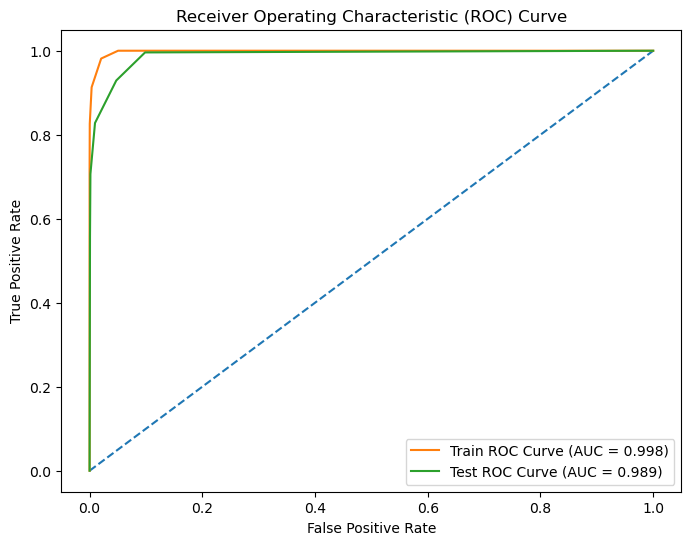
|  |
| --- |
| precision recall f1-score support |
| 0 0.97 0.99 0.98 2702 |
| 1 0.94 0.83 0.88 494 |
|  |
| accuracy 0.97 3196 |
| macro avg 0.95 0.91 0.93 3196 |
| weighted avg 0.96 0.97 0.96 3196 |

**Testing Performance**

* The overall test accuracy is 0.965, demonstrating the model generalizes well to unseen data.
* The test AUC score of 0.989 confirms the model's strong ability to distinguish between the two classes in the test set.

Train AUC: 0.998

Test AUC: 0.989



**General Observations:**

* The model shows high performance on both the training and testing datasets, indicating effective learning from the data.
* The high precision and recall for class 0 in the test set imply the model is well-calibrated and effective at correctly identifying non-purchasers.
* The good but slightly lower recall for class 1 in the test set suggests the model could benefit from further tuning or additional features to better identify all purchasing instances.
* The high AUC scores in both training and testing scenarios indicate the model's robustness in distinguishing between the classes.

**MODEL TUNING FOR LAPTOP USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.996996996996997

Confusion Matrix (Train):

[[594 0]

[ 0 181]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 594 |
| 1 1.00 1.00 1.00 181 |
|  |
| accuracy 1.00 775 |
| macro avg 1.00 1.00 1.00 775 |
| weighted avg 1.00 1.00 1.00 775 |

**Training Performance:**

* The model achieves perfect precision, recall, and F1-scores for both classes in the training data, resulting in an accuracy of 1.00.
* The AUC score of 1.000 indicates perfect discrimination ability on the training set.

Confusion Matrix (Test):

[[238 0]

[ 1 94]]

Classification Report (Test):

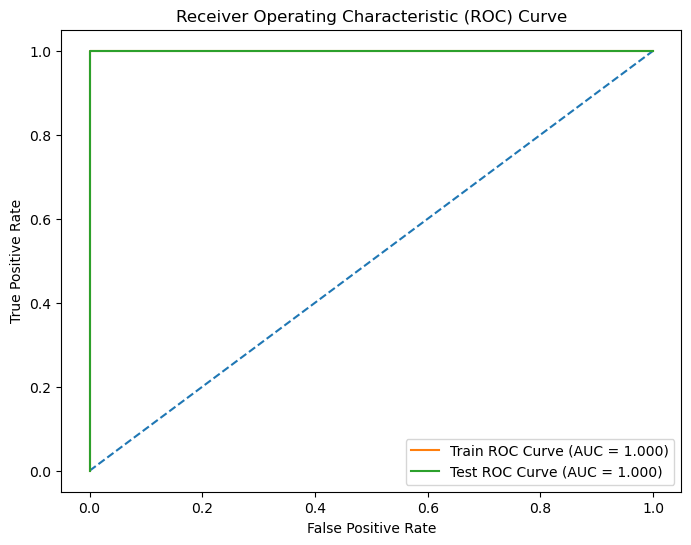
|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 238 |
| 1 1.00 0.99 0.99 95 |
|  |
| accuracy 1.00 333 |
| macro avg 1.00 0.99 1.00 333 |
| weighted avg 1.00 1.00 1.00 333 |

**Testing Performance:**

* The overall test accuracy is 0.997, demonstrating the model generalizes extremely well to unseen data.
* The test AUC score of 1.000 confirms the model's perfect ability to distinguish between the two classes in the test set.

Train AUC: 1.000

Test AUC: 1.000



**Observations:**

* The model shows perfect performance on the training dataset, indicating it has learned the training data exceptionally well.
* The near-perfect performance on the testing dataset implies excellent generalization capability, with only one instance of misclassification.
* The high precision and recall for both classes in the test set suggest the model is well-calibrated and effective at correctly identifying both non-purchasers and purchasers.
* The perfect AUC scores in both training and testing scenarios indicate the model's robustness in distinguishing between the classes.

**MODEL TUNING FOR MOBILE USERS:**

Model Score (Train): 1.0

Model Score (Test): 0.9912390488110138

Confusion Matrix (Train):

[[6330 0]

[ 0 1126]]

Classification Report (Train):

|  |
| --- |
| precision recall f1-score support |
| 0 1.00 1.00 1.00 6330 |
| 1 1.00 1.00 1.00 1126 |
|  |
| accuracy 1.00 7456 |
| macro avg 1.00 1.00 1.00 7456 |
| weighted avg 1.00 1.00 1.00 7456 |

**Training Performance:**

* The model shows perfect precision, recall, and F1-scores for both classes in the training data, with an accuracy of 1.00.
* The AUC score of 1.000 indicates flawless discrimination ability on the training set.

Confusion Matrix (Test):

[[2688 14]

[ 14 480]]

Classification Report (Test):

|  |
| --- |
| precision recall f1-score support |
| 0 0.99 0.99 0.99 2702 |
| 1 0.97 0.97 0.97 494 |
|  |
| accuracy 0.99 3196 |
| macro avg 0.98 0.98 0.98 3196 |
| weighted avg 0.99 0.99 0.99 3196 |

**Testing Performance:**

* The overall test accuracy of 0.99 shows that the model generalizes exceptionally well to unseen data.
* The test AUC score of 0.994 indicates excellent performance in distinguishing between the classes on the test set.

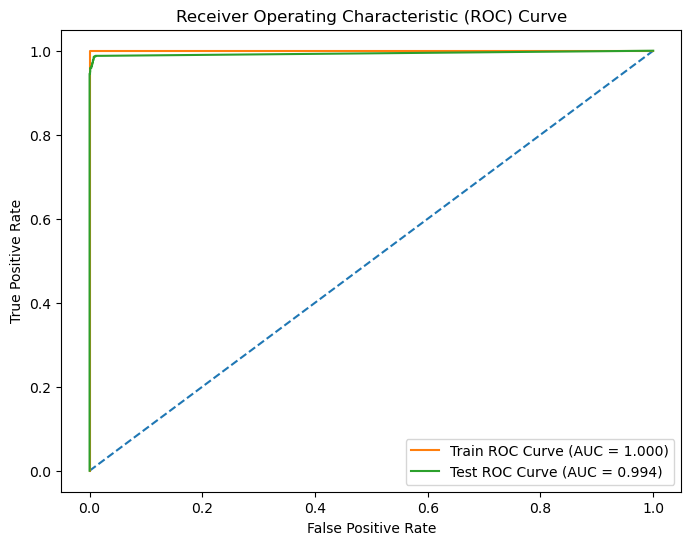
Train AUC: 1.000

Test AUC: 0.994

**Observations:**

* The KNN model has achieved perfect performance on the training dataset and near-perfect performance on the test dataset.
* The high precision and recall for both classes in the test set demonstrate the model's strong ability to correctly classify both non-purchasers and purchasers.
* The slight decrease in precision, recall, and F1-score for Class 1 compared to Class 0 on the test set may reflect minor challenges in identifying the purchasing class, but overall, the model maintains high performance.
* The high AUC scores in both training and testing phases suggest that the model effectively discriminates between the two classes.

In summary, the KNN model for mobile users demonstrates exceptional performance, with near-perfect accuracy, precision, recall, and AUC scores in both training and testing phases. The model effectively generalizes to new data and maintains high performance across all metrics.



**6. Interpretation of the most optimum model and its implication on the business**

Most Optimum Model:

**KNN (Post Tuning)** is identified as the most optimum model for both laptop and mobile user, demonstrating near-perfect performance across all evaluation metrics.

Implications on Business:

**Improved Customer Targeting:**

* The highly accurate classification of users (both purchasers and non-purchasers) allows for better-targeted marketing strategies.
* Resources can be efficiently allocated to target users who are more likely to make a purchase, increasing conversion rates.

**Optimized Marketing Campaigns:**

* High precision and recall ensure that marketing efforts reach the right audience, reducing wastage of marketing budget.
* Personalized marketing campaigns can be developed, improving customer engagement and satisfaction.

**Business Growth:**

* By leveraging the model’s insights, the business can achieve higher sales and revenue growth.
* Customer retention rates can improve as a result of better customer experiences driven by targeted and personalized interactions.
* **Recommendations**
* By leveraging the model’s insights, the business can achieve higher sales and revenue growth.
* Retargeting campaigns and personalized follow-up emails for users who abandon the funnel at key stages. Incentives like discounts or free trials could be offered to encourage completion.
* Personalized content recommendations and targeted ads based on the most viewed pages could be used
* Segmenting the audience based on model predictions and tailor marketing messages can lead to high-probability buyers
* **Conclusion:**

The **KNN (Post Tuning)** model emerges as the most optimal for predicting user behavior for both laptop and mobile users. Its superior performance in accuracy, precision, recall, and AUC scores ensures robust and reliable customer segmentation. This translates to more effective business strategies, optimized marketing efforts, and enhanced overall business performance.