1/22/2023

Ajinkya Godbole

A Report on Social Media Tourism

Social Media Data

Table of Contents

[Q1. Problem Statement 4](#_Toc125311714)

[Problem Statement 4](#_Toc125311715)

[Project Need 4](#_Toc125311716)

[Understanding Business/Social Opportunity 4](#_Toc125311717)

[Understanding data 5](#_Toc125311718)

[Data Description 6](#_Toc125311719)

[Q2. Exploratory Data Analysis 7](#_Toc125311720)

[Members in Family 7](#_Toc125311721)

[Feature – Preferred Location 9](#_Toc125311722)

[Feature – Following Company Page 12](#_Toc125311723)

[Feature – Working Flag 13](#_Toc125311724)

[Feature – Travelling Network Rating 14](#_Toc125311725)

[Feature - Adult Flag 15](#_Toc125311726)

[Feature – Week since Last Check-in 16](#_Toc125311727)

[Feature – Continuous variables 17](#_Toc125311728)

[Feature – Taken Product 19](#_Toc125311729)

[Correlation Plot 21](#_Toc125311730)

[Q3. Data Cleaning and Pre-processing 21](#_Toc125311731)

[Treating Outliers 21](#_Toc125311732)

[Imputing Missing Values 22](#_Toc125311733)

[Categorical variables 22](#_Toc125311734)

[Continuous variable 22](#_Toc125311735)

[Label Encoding 22](#_Toc125311736)

[Creating Subset of Dataset 23](#_Toc125311737)

[Scaling 23](#_Toc125311738)

[Model Building 24](#_Toc125311739)

[Removing Unnecessary Columns 24](#_Toc125311740)

[Splitting into Train and Test 24](#_Toc125311741)

[Random Forest 24](#_Toc125311742)

[Mobile Subset 25](#_Toc125311743)

[Laptop Subset 25](#_Toc125311744)

[Naïve Bayes 26](#_Toc125311745)

[Mobile Subset 26](#_Toc125311746)

[Laptop 27](#_Toc125311747)

[KNN Classifier 28](#_Toc125311748)

[Mobile 28](#_Toc125311749)

[Laptop 29](#_Toc125311750)

[Logistic Regression 31](#_Toc125311751)

[Mobile 31](#_Toc125311752)

[Laptop 32](#_Toc125311753)

[Summary of Model Performances 34](#_Toc125311754)

[Business Insights 34](#_Toc125311755)

[Recommendations 34](#_Toc125311756)

List of Figures

[Figure 1 Features Information 6](#_Toc125311429)

[Figure 2 Missing Values Information 7](#_Toc125311430)

[Figure 3 Five-Point Summary 7](#_Toc125311431)

[Figure 4 Unique values in Members in Family feature 7](#_Toc125311432)

[Figure 5 Countplot of Members in Family 8](#_Toc125311433)

[Figure 6 Families with 10 Members 8](#_Toc125311434)

[Figure 7 Market Share According to Family Sizes 9](#_Toc125311435)

[Figure 8 Unique Values in Preferred Location Feature 9](#_Toc125311436)

[Figure 9 Value Count of Each value 10](#_Toc125311437)

[Figure 10 Final Preferred Preferences 11](#_Toc125311438)

[Figure 11 Market Share Preferred Location Wise. 12](#_Toc125311439)

[Figure 12 Unique Values in Following Company Page 12](#_Toc125311440)

[Figure 13 Countplot of Following Company Page 13](#_Toc125311441)

[Figure 14 Unique Values for Working Flag 13](#_Toc125311442)

[Figure 15 Countplot of Working Flag 14](#_Toc125311443)

[Figure 16 Travelling Network Rating Unique Values 14](#_Toc125311444)

[Figure 17 Countplot of Travelling network Rating 15](#_Toc125311445)

[Figure 18 Unique Values of Adult Flag 16](#_Toc125311446)

[Figure 19 Countplot of Adult\_flag 16](#_Toc125311447)

[Figure 20 Countplot of Week since Last Check-in 17](#_Toc125311448)

[Figure 21 Distribution of Continuous Variables 18](#_Toc125311449)

[Figure 22 Boxplot of Continuous Variables 18](#_Toc125311450)

[Figure 23 Description of Continuous variables 19](#_Toc125311451)

[Figure 24 Correlation Plot 21](#_Toc125311452)

[Figure 25 Boxplot after Treating Outliers 22](#_Toc125311453)

[Figure 26 Unique Values for Preferred Device 23](#_Toc125311454)

[Figure 27 Mobile Subset 23](#_Toc125311455)

[Figure 28 Laptop Subset 23](#_Toc125311456)

[Figure 29 Confusion Matrix Random Forest Mobile 25](#_Toc125311457)

[Figure 30 Classification Report Random Forest Mobile 25](#_Toc125311458)

[Figure 31 Confusion Matrix Random Forest Laptop 26](#_Toc125311459)

[Figure 32 Classification Report Random Forest Laptop 26](#_Toc125311460)

[Figure 33 Confusion Matrix Naive Bayes Mobile 27](#_Toc125311461)

[Figure 34 Classification Report naïve Bayes Mobile 27](#_Toc125311462)

[Figure 35 Confusion Matrix Naive Bayes Laptop 28](#_Toc125311463)

[Figure 36 Classification Report Naive Bayes Laptop 28](#_Toc125311464)

[Figure 37 Confusion matrix KNN Mobile 29](#_Toc125311465)

[Figure 38 Classification Report KNN Mobile 29](#_Toc125311466)

[Figure 39 Confusion Matrix KNN Laptop 30](#_Toc125311467)

[Figure 40 Classification Report KNN Laptop 30](#_Toc125311468)

[Figure 41 Accuracy Vs K-value Plot 31](#_Toc125311469)

[Figure 42 Confusion Matrix LR Mobile 32](#_Toc125311470)

[Figure 43 Confusion Matrix LR Laptop 33](#_Toc125311471)

[Figure 44 Classification Report LR Laptop 33](#_Toc125311472)

[Figure 45 Performance Summary 34](#_Toc125311473)

List of Tables

[Table 1 Data Overview 5](#_Toc125311518)

[Table 2 Merged Preferences 10](#_Toc125311519)

[Table 3 Train Test Split shape 24](#_Toc125311520)

Social Media Data

# Q1. Problem Statement

## Problem Statement

An aviation company wants to apply a targeted marketing strategy using a person’s social media record. The social media record is collected by collaborating with a social media platform. Based on the social and digital record, a potential customer needs to be predicted.

## Project Need

Randomly contacting customers is a very expensive and time consuming marketing process. Besides, randomly and repetitively calling customers can even hurt the reputation of the brand.

So, in order to reduce the cost of marketing and targeting the potential customer, a machine learning model needs to be developed which will predict the potential customers for the company.

## Understanding Business/Social Opportunity

In any business, targeting the right customer at the right time is very important. By collecting the social media data, company can not only find the potential customer base but also see how frequently a customer travels, what is the preferred locations of most of the customers, what is the nature of their travel and so on. Based on these many packages can be designed to suit the needs of the customers. Also, by knowing all these information, we can contact the right customer just at the right time.

## Understanding data

The below table explains each feature.

Table Data Overview

|  |  |
| --- | --- |
| **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 check-ins 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\_pa ge | Average yearly comments on any travel related page by user |
| total\_likes\_on\_outofstation\_checki n\_received | Total number of likes received by a user on out of station check-ins in last year |
| week\_since\_last\_outstation\_checki n | 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\_compan y\_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 the highest and 4 is the lowest. |
| Adult\_flag | Weather the customer is adult or not |
| Daily\_Avg\_mins\_spend\_on\_travelin g\_page | Average time spent on the company page by user on daily basis. |

The data is collected on Daily, weekly, monthly, and yearly as per the requirements. Like the data is collected on yearly basis for these features:

* Yearly\_avg\_view\_on\_travel\_page
* total\_likes\_on\_outstation\_checkin\_ given
* yearly\_avg\_Outstation\_checkins
* Yearly\_avg\_comment\_on\_travel\_page
* total\_likes\_on\_outofstation\_checkin\_ received

Features having data collected on monthly basis:

* montly\_avg\_comment\_on\_company\_page

Features having data collected on Weekly basis:

* week\_since\_last\_outstation\_checkin

Features having data collected on daily basis:

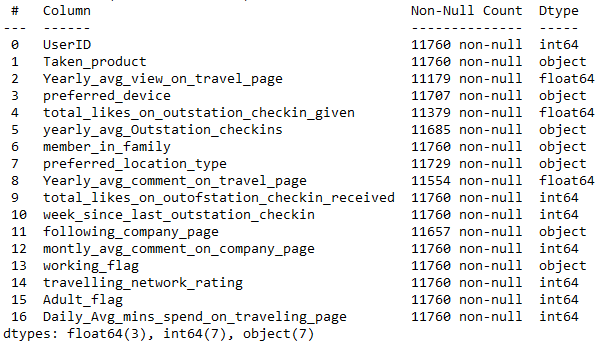
* Daily\_Avg\_mins\_spend\_on\_traveling\_page

As it can be seen that most of the data is collected on a **yearly** basis. There are very few features collected on different time scales.

## Data Description

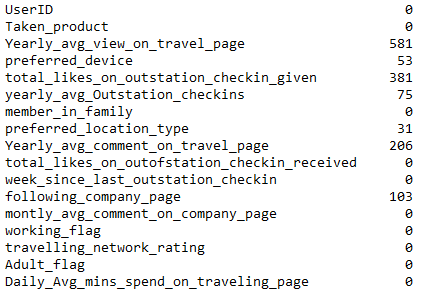
The data set has **17** features and **11760** observations. Most of the features are of float or integer datatype. The below image shows the feature information.

Figure Features Information



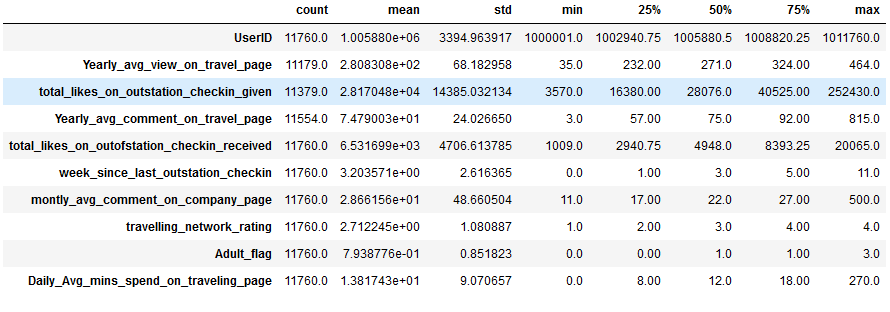
There are 7 features of object datatype and need to be converted into integer or float datatype. Also, it can be seen that there are missing values in the dataset. The below image shows feature wise missing values.

Figure Missing Values Information



Missing values are present in 6 features. The proportion of missing values is very less, so we will impute the missing values instead of dropping.

Figure Five-Point Summary



This image shows the five-point summary of the features. We check outliers and data distribution from this image.

# Q2. Exploratory Data Analysis

We will do the EDA feature-wise and also, based on the EDA we will clean and treat the variables.

## Members in Family

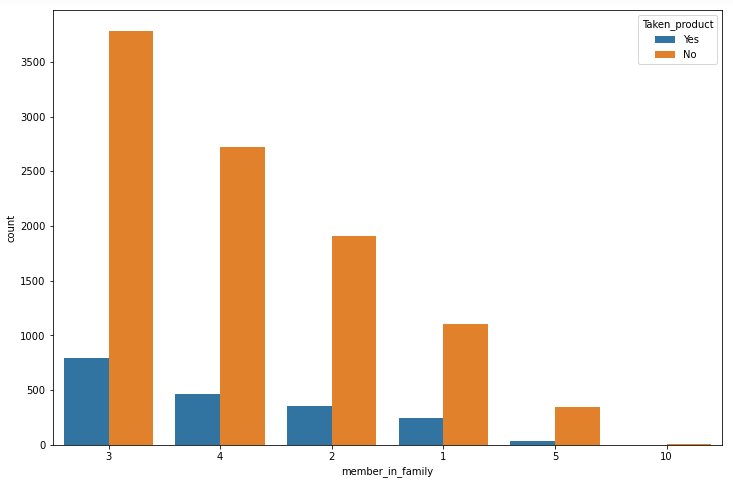
This feature gives count of members in a family. The unique values in this column is

Figure Unique values in Members in Family feature



We will first replace all the instances of “Three” to “3”. There are no missing values in this feature. As this a categorical variable, let’s check the distribution of the data in it.

Figure Countplot of Members in Family



Now, it can be seen that the values are 1,2,3,4,5,10. Most of the families consists of 3 or 4 members, and very few families have 10 members in it. The 10 entry might be done by mistake as it is very rare. Let’s check the count of families with 10 members in it.

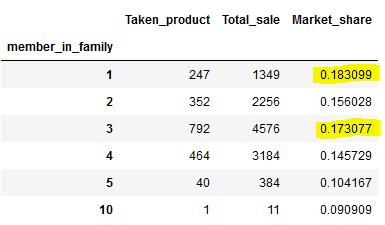
Figure Families with 10 Members



As we can see there are only 11 families with 10 members which is very small. So, for simplicity we will replace 10 members with 5 for simplicity.

The below image shows the percentage of bookings received from families of different sizes.

Figure Market Share According to Family Sizes



Most of the families have 3 members. Out of all those families only 17% preferred our travelling agency. Also, 18% of solo travelers preferred our travel company.

**Travel Company must offer attractive offers for families of size 2 and 3.**

## Feature – Preferred Location

This is a categorical variable. The unique values in this column are

Figure Unique Values in Preferred Location Feature



First, we will replace “Tour Travel” with “Tour and Travel”. We will also check the value counts of each unique value. If the count is very less we will merge into a similar category.

Figure Value Count of Each value



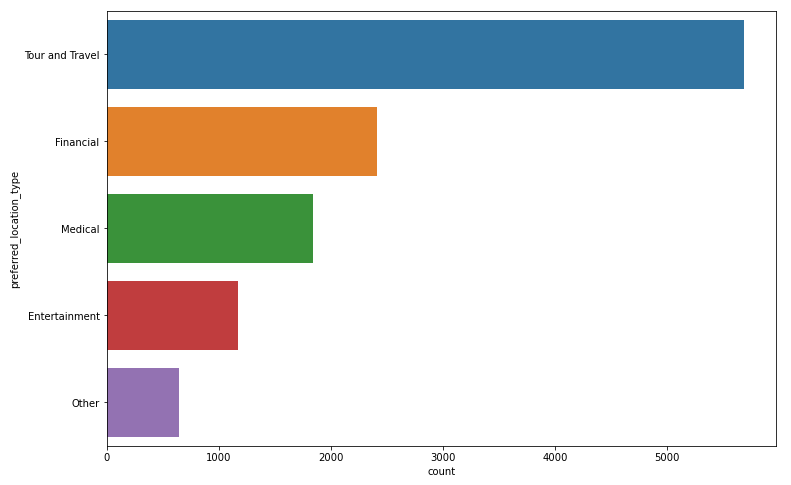
The below table shows the merged preferences.

Table Merged Preferences

|  |  |
| --- | --- |
| **Merged Into** | **Merged Preferred Location Type** |
| Tour and Travel | Tour Travel, Beach, Big Cities, Historical site, Hill Stations |
| Entertainment | Game, Movie, OTT, Social media, |

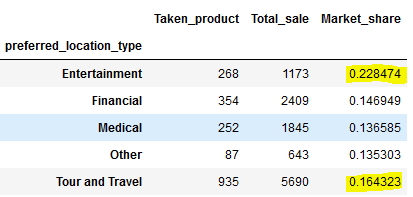
We have merged most of the vacation related travels into Tour and Travel category. Also, Game, Movie, OTT, and Social media are merged into Entertainment category. The merger is done in such a way that the original distribution is not disturbed. Below image shows the final categories and their value counts.

Figure Final Preferred Preferences



Let’s check how much market share is available for our company. The below image shows percentage of product taken category wise.

Figure Market Share Preferred Location Wise.



Our company has the largest market share in Entertainment category, 22%. The most popular travel type is “Tour and Travel” and we have only 16% market share in this category.

**Company must aggressively promote their holiday packages.**

## Feature – Following Company Page

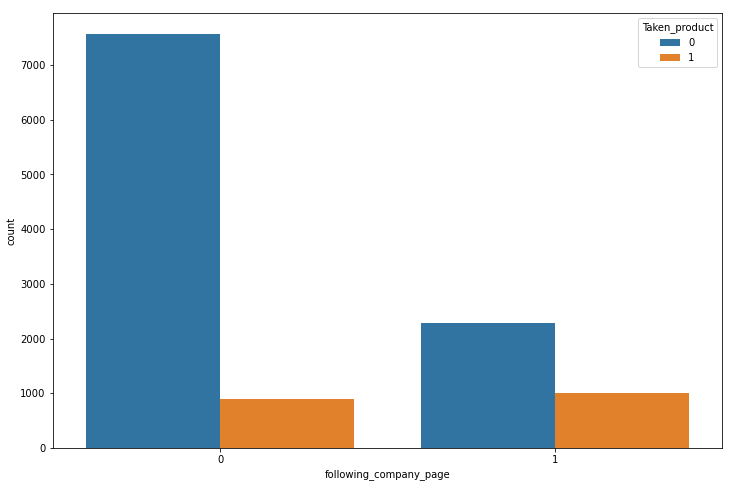
Following Company Page is also a categorical variables. The unique values are

Figure Unique Values in Following Company Page



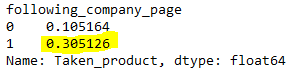
Firstly, we will impute the missing values with “0”. “0” is the most frequently occurring category. Also, we will replace “Yes” with “1” and “No” with “0”. The below image shows the distribution after cleaning the variable.

Figure Countplot of Following Company Page



From the above image, it is clear that most of the customers are not following the company page.

It is observed from the data that people who follow our company page are more likely to buy our products than people who do not follow our page. Below table shows the percentage of people buy our products given that they are following our company page.



**The company must try to increase followers as it can be seen in the image that people who follow our page are more likely to buy our products.**

## Feature – Working Flag

This is also a categorical variable. The unique values in this feature are

Figure Unique Values for Working Flag



We will replace all instances of “Yes” with “1” and “No” with “0”.

Figure Countplot of Working Flag

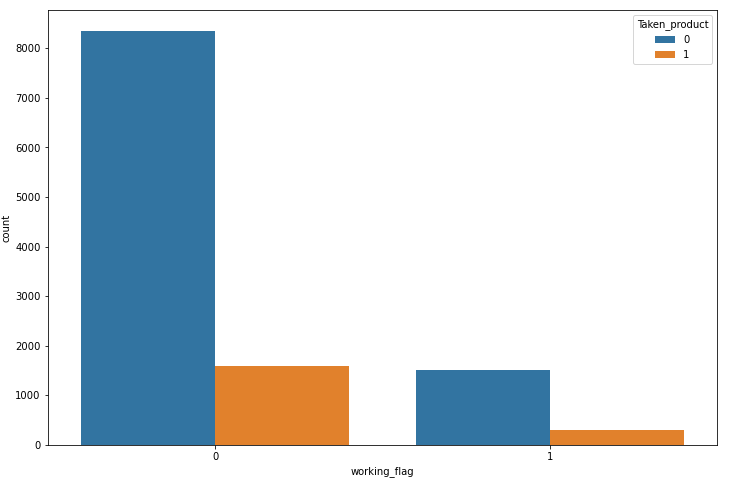


Image shows that most of the customers from the database are not working. This means most of the customer are youngsters.

## Feature – Travelling Network Rating

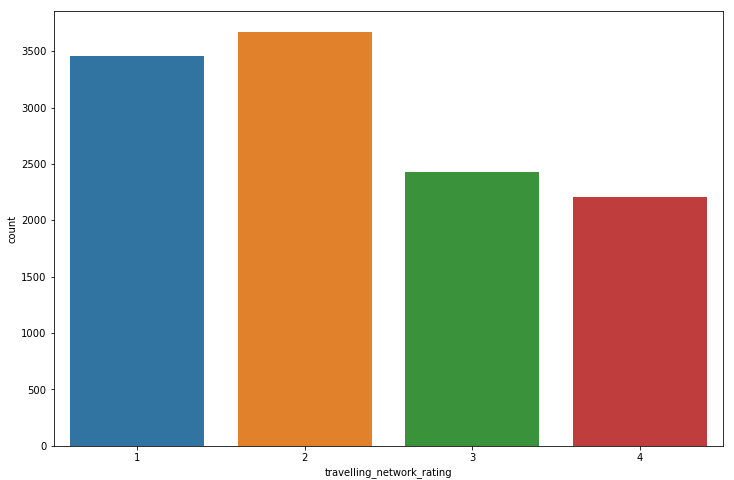
Travelling Network Rating is a categorical variable. It has four categories 1, 2, 3, and 4. 1 is the highest and 4 is the lowest.

Figure Travelling Network Rating Unique Values

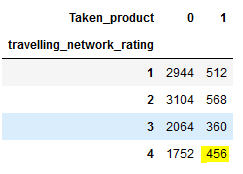


As we know, machine learning models put weightage based on the number in increasing order, so we will reverse and make **4** as the highest rating and **1** as the lowest rating. Below chart shows count of customers fall in each category.

Figure Countplot of Travelling network Rating



As we can see most of the customers have 2 rating which is average. Also, let’s check the conditional probability the product taken given that the customer is having 4 rating.



The probability can be calculated as (456/2208)\*100 = 20.65%. Here 2208 is the total number of customers having 4 rating.

**Only 20% of customers having 4, highest rating, have purchased our products. Company can offer discounts based on traveler’s rating.**

## Feature - Adult Flag

Adult flag is also a categorical variable. The unique values are

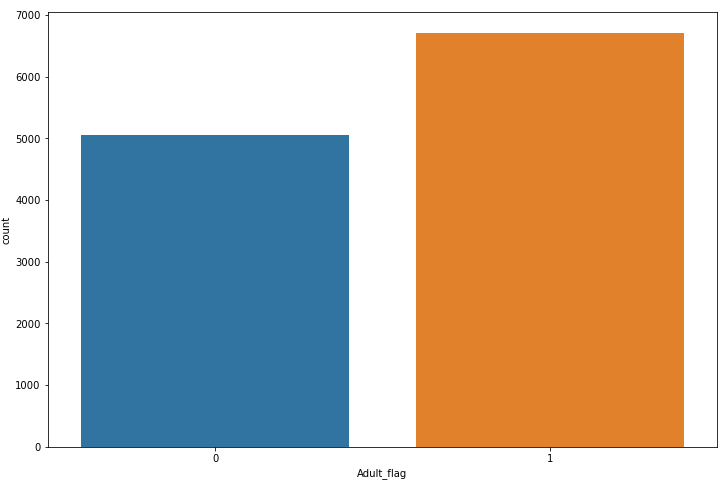
Figure Unique Values of Adult Flag



Since adult flag determines whether the customer is adult or not, 4 unique values don’t make any sense. Assuming “0”s are non-adults and “1”s are adults, we will replace “2”s with 1 and “3”s with 1.

Let’s check the distribution after merging 2’s and 3’s into 1.

Figure Countplot of Adult\_flag



**According to working flag plot, most of the customers are not working, and according to Adult flag plot, most of the customers are adults that means above the age of 18. From these two observations, it can be concluded that most of the customers are between the ages 18 and 25.**

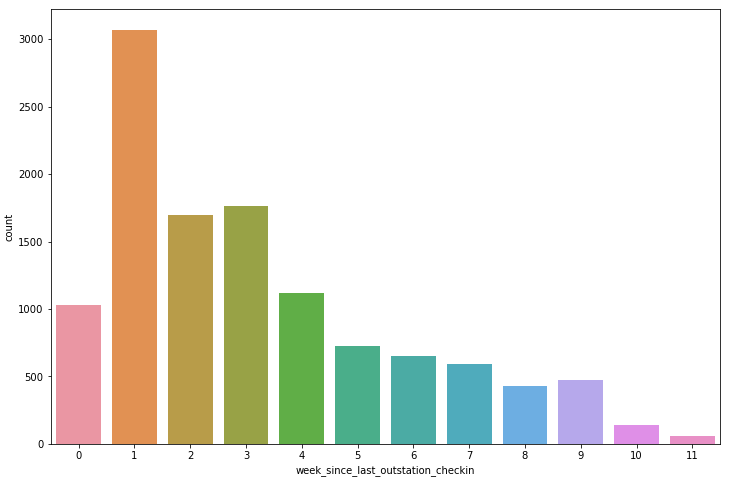
## Feature – Week since Last Check-in

This is also a categorical variable. The unique values in this variable are



Let’s also check the count plot of this variable.

Figure Countplot of Week since Last Check-in



Most of the people have checked-in very recently. Maybe many customers are frequent travelers.

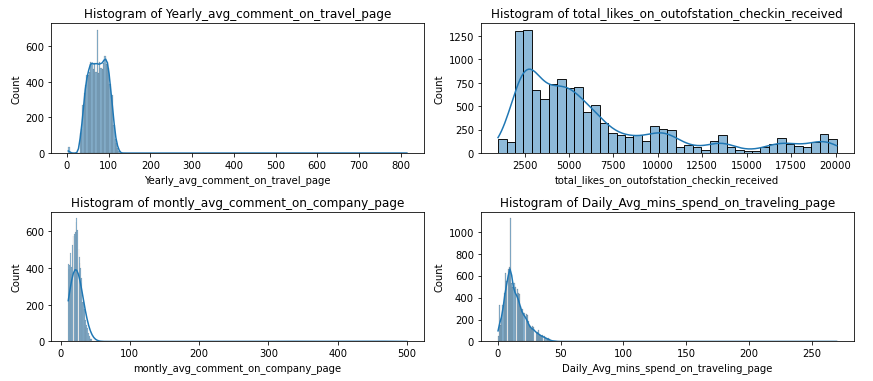
## Feature – Continuous variables

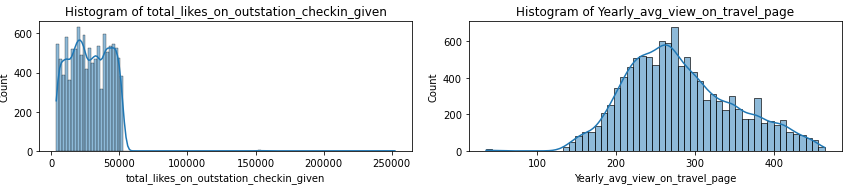
Here is the list of continuous Variables:

1. Yearly average comments on Travel page
2. Total likes on out-of-station check-in received
3. Monthly average comment on company page
4. Daily average minutes spent on travelling page

Below image shows the distribution of continuous variables.

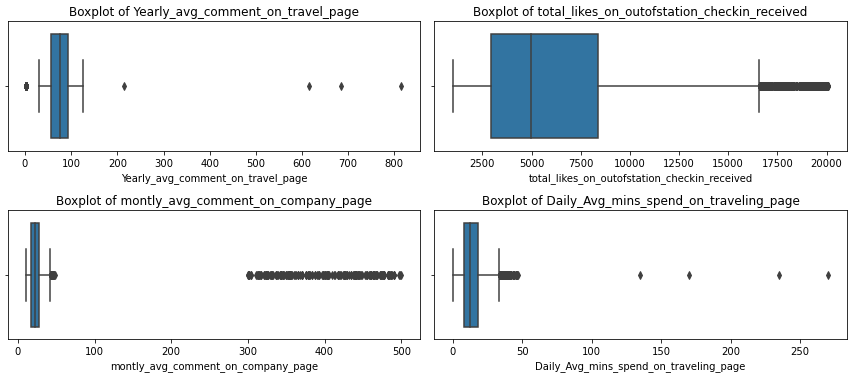
Figure Distribution of Continuous Variables





Almost all the continuous variables have outliers. Outliers can impact the performance of models based on linear relationships like Logistic regression. Let’s also check boxplots of all these variables.

Figure Boxplot of Continuous Variables



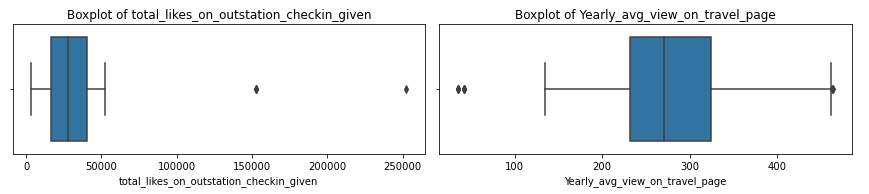
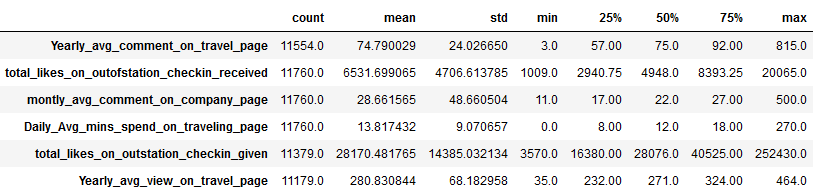


Figure Description of Continuous variables



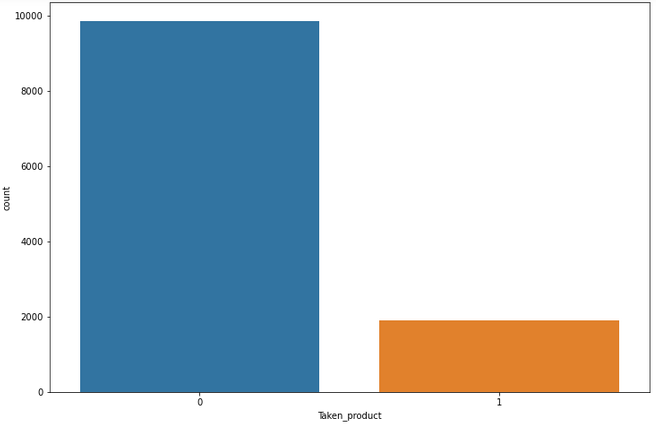
It is clear that all the features are slightly right skewed. That means mean is greater than median. The difference between the means and the medians are not much.

Also, the distances between minimum values and Q1 is less than the distances between Q3 and maximum values. This clearly indicates that the distribution is right skewed.

**Note:** Outlier treatment is mandatory while building classification based on Linear Regression like Logistic Regression, Linear Discriminant Analysis and so on.

## Feature – Taken Product

This a target column. There are only two categories “Yes” or “No”. Let’s see the distribution of the target variable. We will replace “Yes” with 1 and “No” with 0. **Since we want to predict the potential customer, we are replacing “Yes” with 1**.

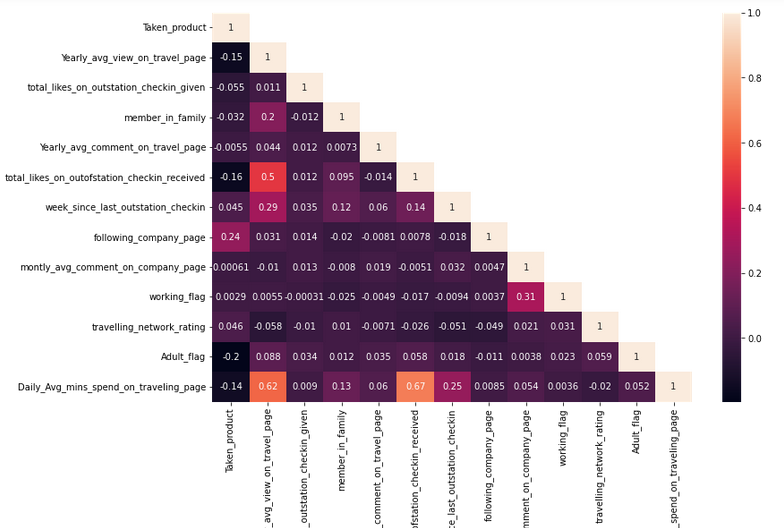


|  |  |
| --- | --- |
| **Product Taken** | **Percentage** |
| Yes | 16.12% |
| No | 83.87% |

The data is highly imbalanced. Only 16% of the total customers have purchased our products in the past.

## Correlation Plot

Figure Correlation Plot



There are very few features that have strong positive and negative relationships.

The correlation between dependent and independent variable is very less. Also, there is no much multi-collinearity in the dataset.

# Q3. Data Cleaning and Pre-processing

## Treating Outliers

We will use an empirical rule for detecting outliers. Whenever a data point is outside the range of Q1 – (1.5\*IQR) and Q3 + (1.5\*IQR), it will be considered as an outlier.

Where,

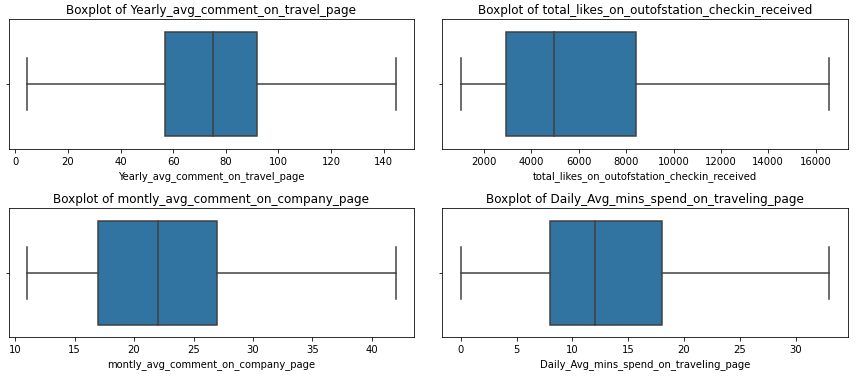
IQR = Inter Quartile Range (The range between 25th percentile and 75th percentile)

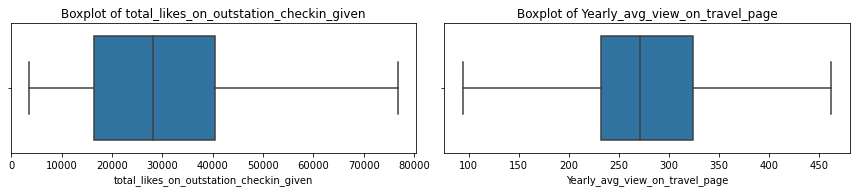
Q1 = 25th percentile value

Q3 = 75th percentile value

Below image shows the boxplots of variables after treating outliers.

Figure Boxplot after Treating Outliers





The outliers are treated, but the data is still skewed.

## Imputing Missing Values

### Categorical variables

For categorical variables, we have used mode value for imputing missing values.

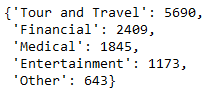
### Continuous variable

For continuous variables, we have median value for imputing missing values. We have also converted special characters found in the variables into Nan values and imputed the missing value. For example, we found “\*” in ‘yearly\_avg\_Outstation\_checkins’ variable.

## Label Encoding

Most of the variables are binary variables that is, Yes or No. In this case, we have encoded with 0s and 1s. 1 for Yes and 0 for No.

In preferred\_location\_type, we have encoded using the value counts. Below image shows the dictionary of values used to encode.



If the categorical variables have more than 2 distinct values, we have encoded with number in ascending order. For example, 1 for lowest and 4 for highest.

## Creating Subset of Dataset

Unique values for Preferred Device variable are:

Figure Unique Values for Preferred Device



For simplicity, everything that is not laptop is considered to be a mobile. Based on Mobile and Laptop, we will divide the dataset into two subsets. The main reason for this division is propensity for buying tickets is different for different devices.

* The shape of Laptop user subset is 1108,16. That is 1108 rows and 16 columns.
* The shape of Laptop user subset is 10652,16. That is 10652 rows and 16 columns.

Below images show snapshot of two subsets.

Figure Mobile Subset

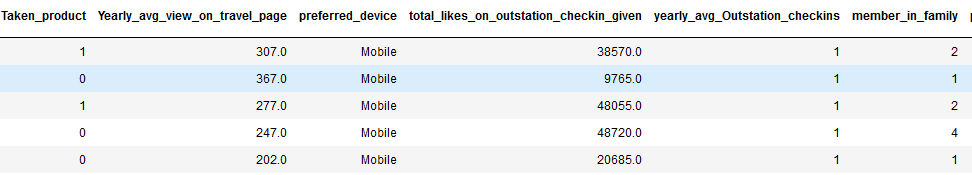
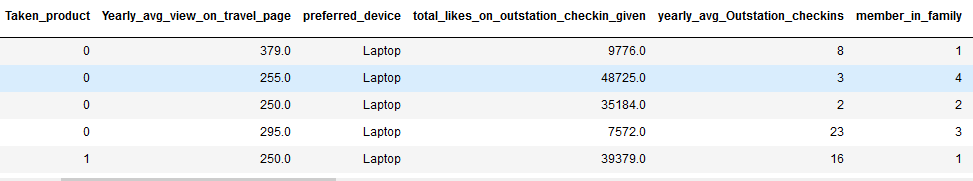


Figure Laptop Subset



## Scaling

Scaling is mandatory for building distance based algorithms or algorithm that uses gradient descent to find slope and intercept values.

Standard Scaler is used to scale the scale the variables.

# Model Building

## Removing Unnecessary Columns

We will remove Unnamed: 0","UserID", and “preferred\_device" columns from both the subsets. Ads these columns will not add any value while building models.

## Splitting into Train and Test

We have split the data into train and test using test size as 30%. Below table shows the shape and size of train and test data.

Table Train Test Split shape

|  |  |  |  |
| --- | --- | --- | --- |
| **Device** | **Type** | **Rows** | **Column** |
| Mobile | X\_train | 7456 | 14 |
| Mobile | X\_test | 3196 | 14 |
| Mobile | y\_train | 7456 | - |
| Mobile | y\_test | 3196 | - |
| Laptop | X\_train | 775 | 14 |
| Laptop | X\_test | 333 | 14 |
| Laptop | y\_train | 775 | - |
| Laptop | y\_test | 333 | - |

**NOTE**: The product taken values is proportionally divided between both mobile and Laptop subsets.

**Let’s start building the models. As the data is highly imbalance, Accuracy cannot be considered as a parameter for evaluation. We will consider Precision and Recall for model evaluation.**

Precision quantifies the number of positive class predictions that actually belong to the positive class.

Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

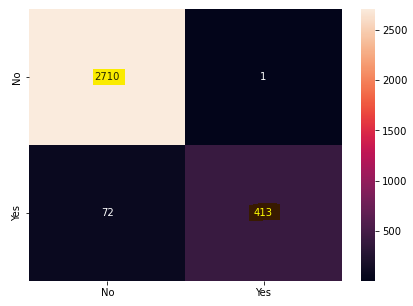
## Random Forest

Random Forest is a decision tree based classifier.

### Mobile Subset

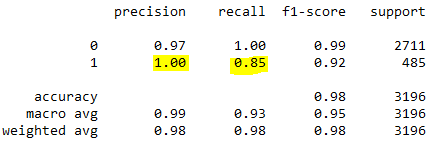
The below image shows the confusion matrix for Mobile Data set

Figure Confusion Matrix Random Forest Mobile



Here True positives are **413** and True Negatives are **2710**.

Figure Classification Report Random Forest Mobile

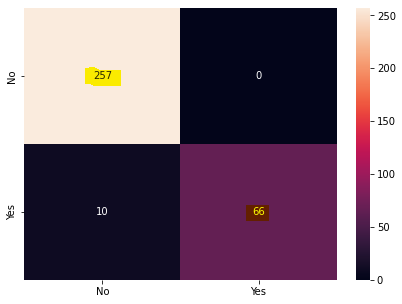


We are focusing only on 1 as 1 indicates potential customer. In this report both precision and recall is good.

### Laptop Subset

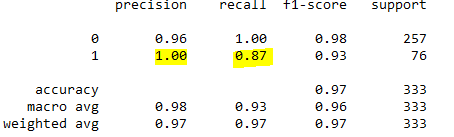
The below image shows the confusion matrix for Laptop Data set

Figure Confusion Matrix Random Forest Laptop



Here True positives are **66** and True Negatives are **257**.

Figure Classification Report Random Forest Laptop



In this report both precision and recall is good.

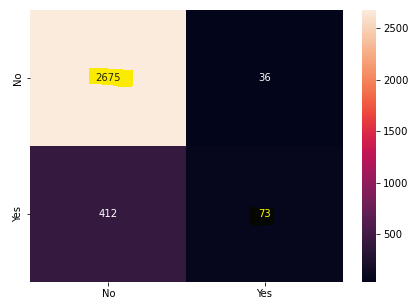
## Naïve Bayes

Naïve Bayes works on Bayes theorem.

### Mobile Subset

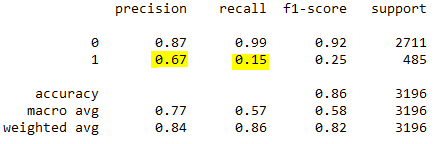
The below image shows the confusion matrix for Mobile Data set

Figure Confusion Matrix Naive Bayes Mobile



Here True positives are **73** and True Negatives are **2675**.

Figure Classification Report naïve Bayes Mobile



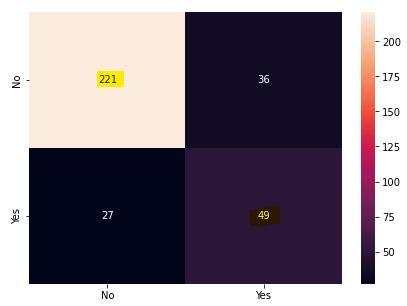
Precision and Recall is very bad. Potential reasons for bad precision and recall are:

1. Data is highly imbalanced.
2. All numerical and continuous variables are not properly normally distributed.

### Laptop

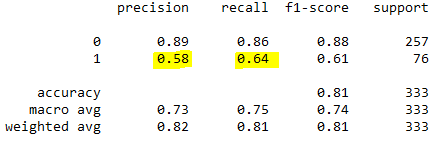
The below image shows the confusion matrix for Laptop Data set

Figure Confusion Matrix Naive Bayes Laptop



Here True positives are **49** and True Negatives are **221**.

Figure Classification Report Naive Bayes Laptop



Precision and Recall is slightly better than mobile data set but overall model doesn’t perform well in predicting potential customers.

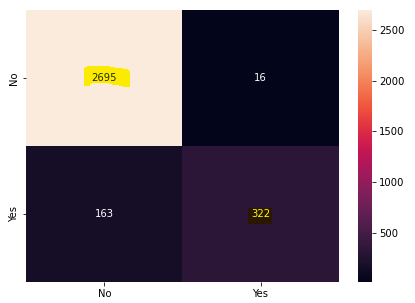
## KNN Classifier

KNN is a distance based algorithm. As this is a distance based algorithm, we will scale the train and test sets before applying the model.

### Mobile

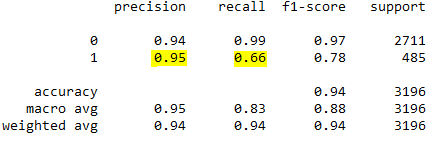
The below image shows the confusion matrix for Mobile Data set

Figure Confusion matrix KNN Mobile



Here True positives are **322** and True Negatives are **2695**.

Figure Classification Report KNN Mobile

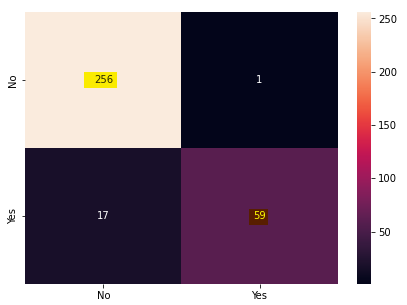


Precision and recall is better than Naïve Bayes classifier but not as good as Random Forest.

### Laptop

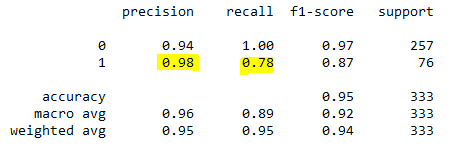
The below image shows the confusion matrix for Laptop Data set

Figure Confusion Matrix KNN Laptop



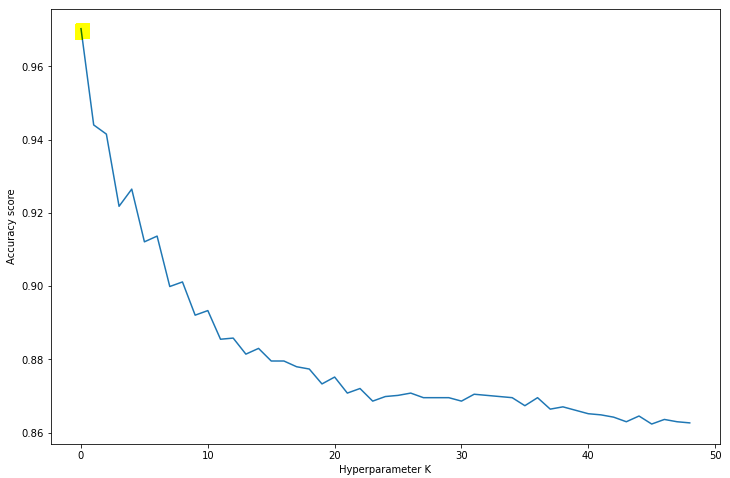
Here True positives are **59** and True Negatives are **256**.

Figure Classification Report KNN Laptop



The overall performance of KNN is satisfactory. In KNN n\_neighbor is a hyper parameter and can be set to maximize model performance. The below figure shows the best n\_neighbor value.

Figure Accuracy Vs K-value Plot



From the graph it is clear that the accuracy is maximum when n\_neighbor is between 1 and 3. We have selected k=2 while building the models.

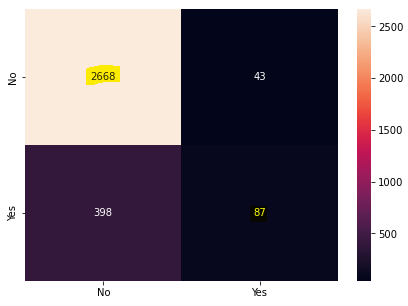
## Logistic Regression

Logistic Regression is based on Linear Regression.

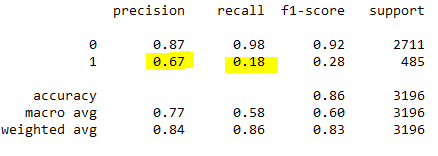
### Mobile

The below image shows the confusion matrix for Laptop Data set

Figure Confusion Matrix LR Mobile



Here True positives are **87** and True Negatives are **2668**.

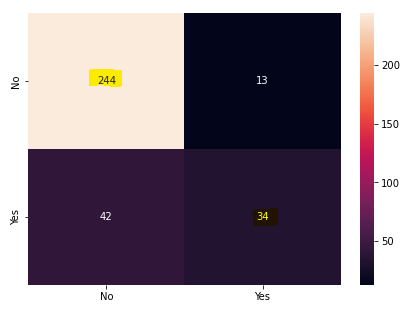


Precision and Recall are not good. Model did not predict the potential customers properly.

### Laptop

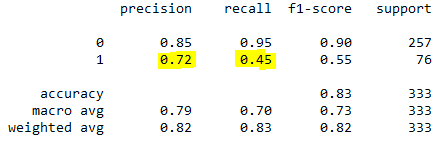
The below image shows the confusion matrix for laptop dataset

Figure Confusion Matrix LR Laptop



Here True positives are **34** and True Negatives are **244**.

Figure Classification Report LR Laptop



The model didn’t perform well on laptop dataset as well. The main reasons are:

1. Correlation is very week between dependent and independent variables.
2. The continuous variables are not properly normally distributed.

# Summary of Model Performances

Figure Performance Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Device | Precision | Recall |
| Random Forest | Mobile | 1 | 0.85 |
| Random Forest | Laptop | 1 | 0.87 |
| Naïve Bayes | Mobile | 0.67 | 0.15 |
| Naïve Bayes | Laptop | 0.58 | 0.64 |
| KNN | Mobile | 0.95 | 0.66 |
| KNN | Laptop | 0.98 | 0.78 |
| Logistic Regression | Mobile | 0.67 | 0.18 |
| Logistic Regression | Laptop | 0.72 | 0.45 |

Random Forest has performed better than all the models. The main reasons are:

1. Most of the variables are categorical in nature.
2. Random Forest doesn’t have any assumption of normality for continuous variables.
3. Random Forest usually performs better for binary classification problems.

# Business Insights

* Most of the families have 3 members. Out of all those families only 17% preferred our travelling agency. Also, 18% of solo travelers preferred our travel company.
* The most popular travel type is “Tour and Travel” and we have only 16% market share in this category.
* The company must try to increase followers as it can be seen in the image that people who follow our page are more likely to buy our products.
* Only 20% of customers having 4, highest rating, have purchased our products. Company can offer discounts based on traveler’s rating.
* Only 20% of customers having 4, highest rating, have purchased our products. Company can offer discounts based on traveler’s rating.
* According to working flag plot, most of the customers are not working, and according to Adult flag plot, most of the customers are adults that means above the age of 18. From these two observations, it can be concluded that most of the customers are between the ages 18 and 25.

# Recommendations

* Social media marketing campaign must be launched as most of the people are not following company page on social media.
* 90% of the public login through mobile devices. So, mobile application must be developed. With the mobile app, notifications can be pushed.
* We must focus on destinations like beaches, historical sights. Business is second most important travel type, but socializing websites doesn’t have much working class people on it.
* Company can offer discounts on places like Goa, Andaman and Nicobar and so on that have both beaches and historical monuments.
* Travel videos must be produced to attract more customers on our company page.