Capstone Project

Project Notes -1

Contents:

Problem Statement.: Page 3

Introduction of the business problem:

- Defining problem statement. Page 4
- Need of the study/project. Page 4
- Understanding business/social opportunity. Page 4

Data Report:

- Understanding how data was collected in terms of time, frequency and methodology. Page 5
- Visual inspection of data. Page 5
  (rows, columns, descriptive details)
- Understanding of attributes. Page 5
  (variable info, renaming if required).

Exploratory data analysis:

- Removal of unwanted variables. Page 6
- Missing Value treatment. Page 6
- Outlier treatment. Page 6
- Variable transformation. Page 6
- Addition of new variables. Page 6
- Univariate analysis. Page 6
  (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones).
- Bivariate analysis. Page 7
  (relationship between different variables , correlations).

Business insights from EDA:

- Is the data unbalanced? If so, what can be done? Please explain in the context of the business. Page 8
- Any business insights using clustering. Page 8
- Any other business insights. Page 8

Problem Statement.

Business Objective:

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

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

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

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

Introduction of the business problem:

Introduction of the business problem

Defining problem statement:

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

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

Need of the study/project:

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

Understanding business/social opportunity:

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

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

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

Data Report:

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

Data includes:

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

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

Understanding of attributes (variable info, renaming if required)

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

Exploratory data analysis:

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

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

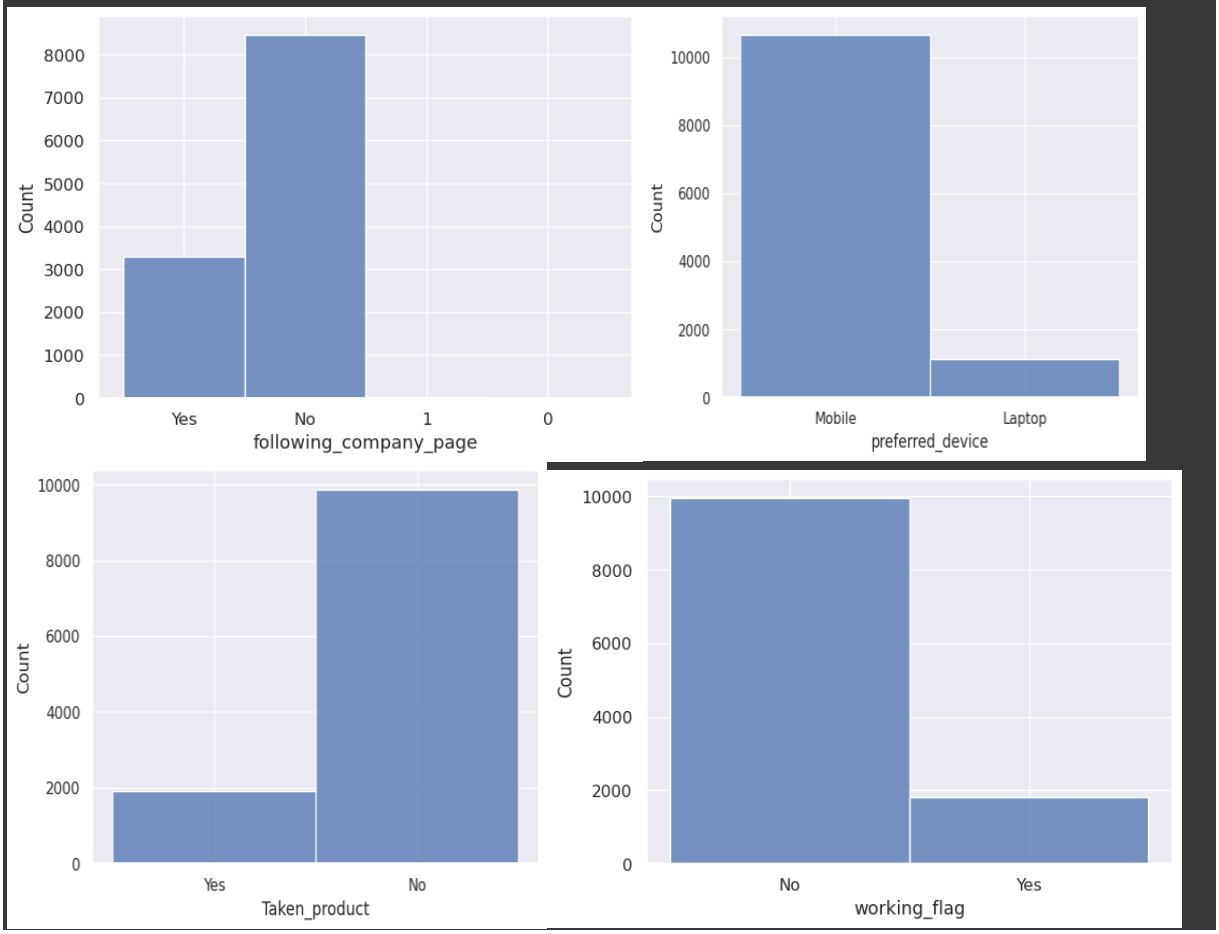
3. Missing values are treated with median and mode for both numerical and categorical columns respectively.

4. Outliers will be treated as per IQR.

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

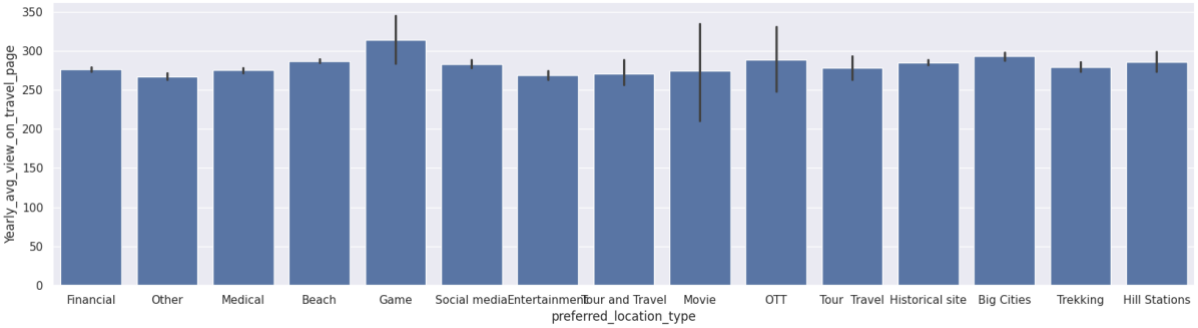
Univariate analysis.

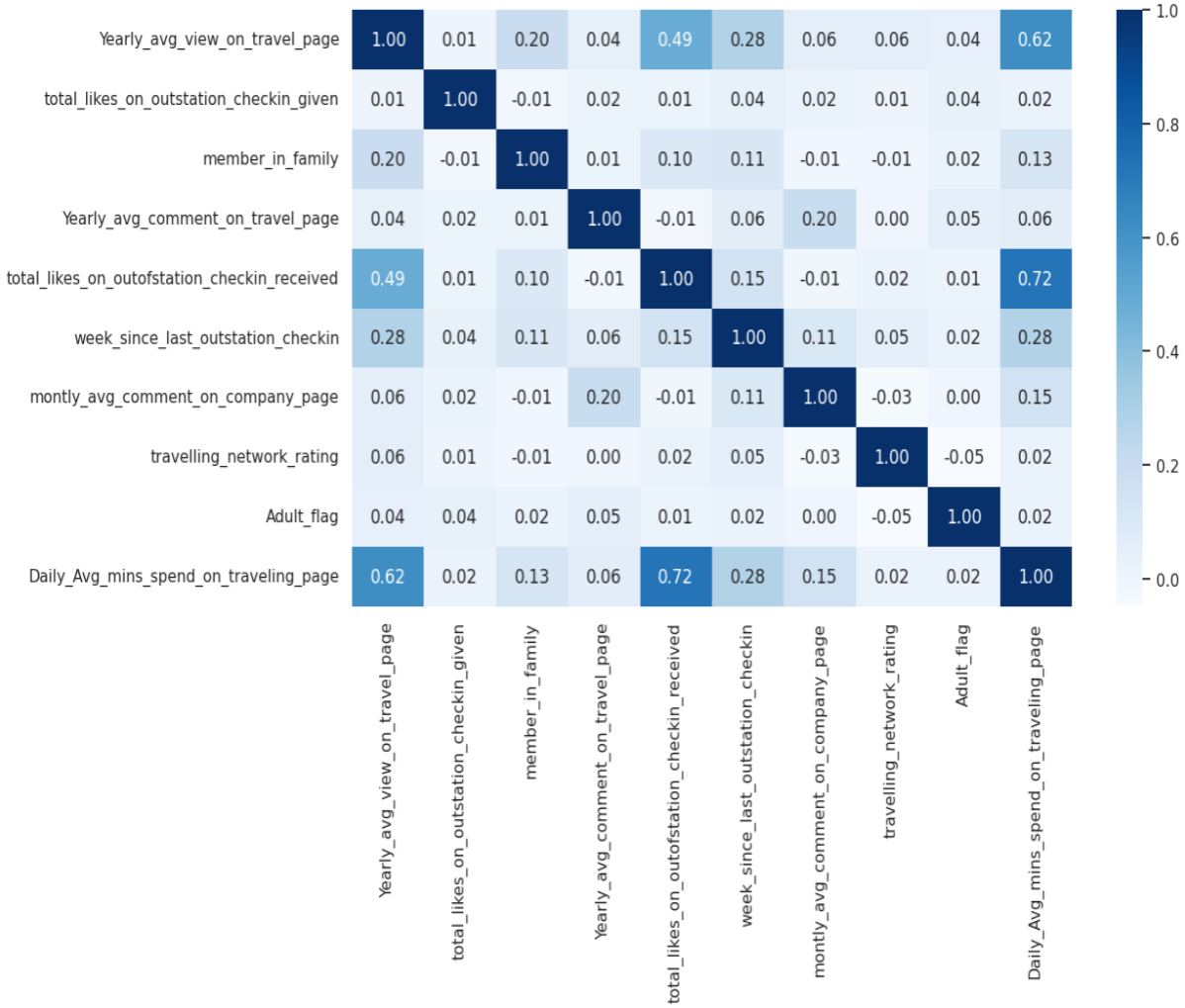
Some of the plots for reference:

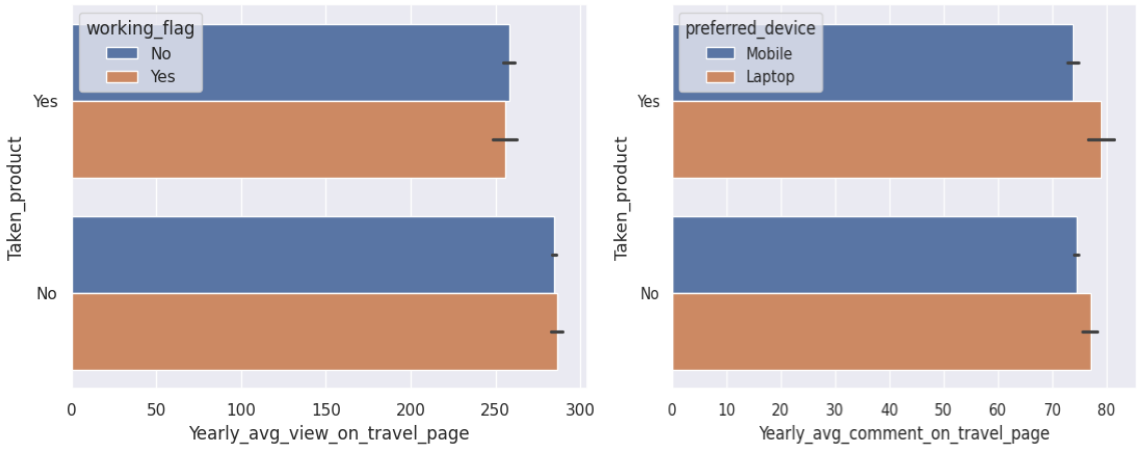


- With the above analysis it shows that more people uses mobile and less using laptops
- Product has been taken by less people.
- Most of the people visiting sites seems to be not working.
- Also, less people following the company.

Bivariate Analysis:







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

Business insights from EDA:

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