Raghad saleh alkhathran
Project proposal
14 October

Online Shoppers Intention

reminder of our project goals:

In this project I will analyze ‘Online Shoppers Intention ‘data to try improve the customers experience with the website, to gain more revenues

And answer these few questions:

- Is the weekend a factor to raise the revenue?
- the closeness of the site visiting time to a specific special day (e.g., Mother’s Day, Valentine's Day) effect the revenue?
- What features has the strong correlation with revenue?

Prepossessing:

Started by looking for null values and drop the duplicate record, from the figures below we can clearly see that there is no missing data or a any duplicate value.

```
df.info() #no Null value
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

Figure 1 Missing value

```
df.drop_duplicates()
print('The raw data is ', df.shape[0], 'rows and', df.shape[1], 'columns.' ) #no duplicates Record
```

```
The raw data is  12330 rows and 18 columns.
```

Figure 2 Remove the duplicate

Raghad saleh alkhathran
Project proposal
14 October

Then converted the categorical value into a numeric, Started by converted the month name into 1,2,3..etc. Also the visitor type, in the ending changed the Boolean into 0s and 1s , the code shown in the figures below.

```
df['Revenue'] = df['Revenue'].astype(int)
df['Weekend'] = df['Weekend'].astype(int)
#true and false into 0s and 1s
```

```
# month into numbers
dic = {'Jan' : 1 , "Feb" : 2 , 'Mar' : 3 , 'Apr' : 4 , 'May' : 5 , "June" : 6 , 'Jul' : 7 , 'Aug' : 8
      , 'Oct' : 10 , 'Nov' : 11 , 'Dec' : 12}


df['Month'] = df.Month.map(dic)

df['Month'].unique()
```

```
array([ 2,  3,  5, 10,  6,  7,  8, 11,  9, 12])
```

```
# month into numbers
dic = {'Returning_Visitor' : 1 , 'New_Visitor' : 2 , 'Other' : 3 }


df['VisitorType'] = df.VisitorType.map(dic)

df['VisitorType'].unique()
```

```
array([1, 2, 3])
```

Figure 3 convert into numerical

Exploratory Data analysis (EDA) :

Started visualizing the data to learn about the data:

Figure 4 Revenue Count

From the figure above we now that the data is imbalance, since the “Zero”(False) value is large than the “1” (True).

Figure 5 Revenue VS Weekend

From the graph above, we can say that the weekend not a factor to raise the revenue and the visitor in the weekday is actually more than in the weekend , and from this graph is also answered the first question ‘Is the weekend a factor to raise the revenue?’

Raghad saleh alkhathran
Project proposal
14 October

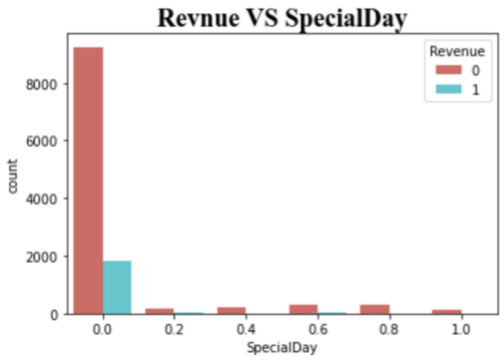

Figure 6 Revenue VS Special day

The special day represent the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day). It's a non-zero value if it close to a special day and 0 if it's not. And most of the revenue done in special day value equal to zero where it is considered as a day that not close to any special day. From the graph above answered the second question 'the closeness of the site visiting time to a specific special day (e.g., Mother's Day, Valentine's Day) effect the revenue?'.

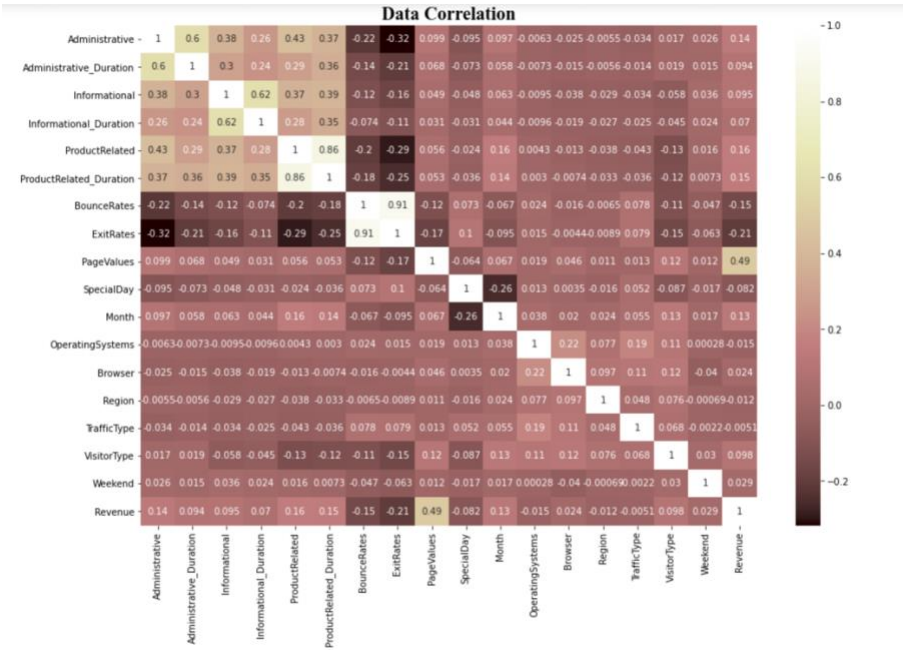

Figure 7 Heatmap

From the heatmap above we can see the data correlation and we can identify the feature with strong correlation to answer the third question 'What features has the strong correlation with revenue?' Actually revenue has no strong correlation since there is no correlation above 0.70 but the most strongest correlation was with 'page value' equal to 0.49 where is the consider as a weak correlation, In addition a lot of features have correlation smallest than 0.3 such as 'Browser'...etc.it consider that features without any correlation with revenue or very weak correlation and we can later try to fit the model with and without these features.

Raghad saleh alkhathran
Project proposal
14 October

References :

Strength of Relationship : https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf