LEAD SCORING CASE STUDY

SUBMITTED BY:

MAMLAKATOI HAIDAROVA

AKSHADA JOSHI

PROBLEM STATEMENT

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, has given a ballpark of the target lead conversion rate to be around 80%.

GOALS AND OBJECTIVE OF CASE STUDY

There are quite a few goals for this case study.

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

INSPECTING AND UNDERSTANDING DATASET

1. The dimension/shape of the dataset is (9240, 37), i.e. it contains 9249 rows and 37 columns.
2. There is no duplicate values present in the dataset.
3. We checked columns for unique value- Dropping redundant columns like 'Prospect ID', 'Lead Number', 'Country', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'City'.
4. Checked for the missing values in the data frame and dropped the values with >30% of missing value.

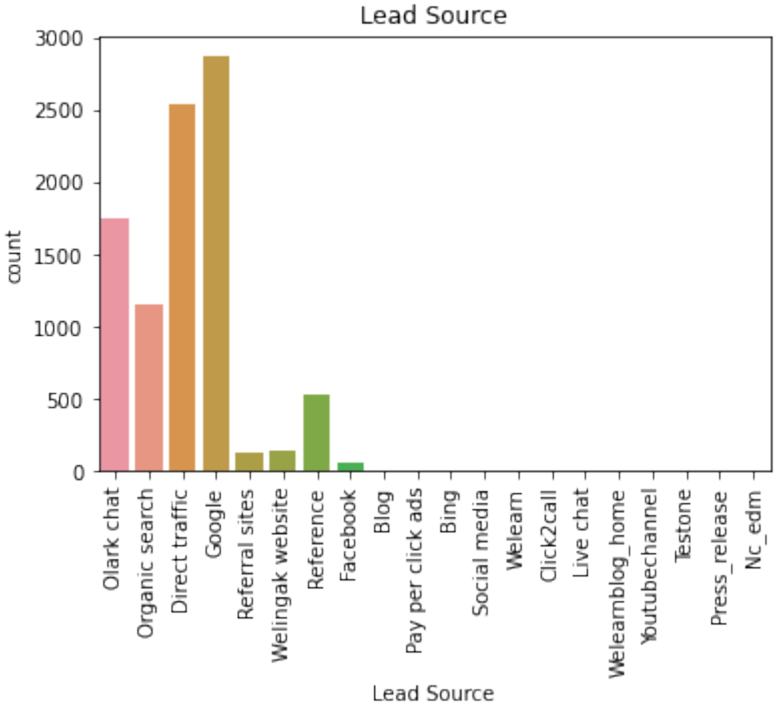
Following columns have null values :

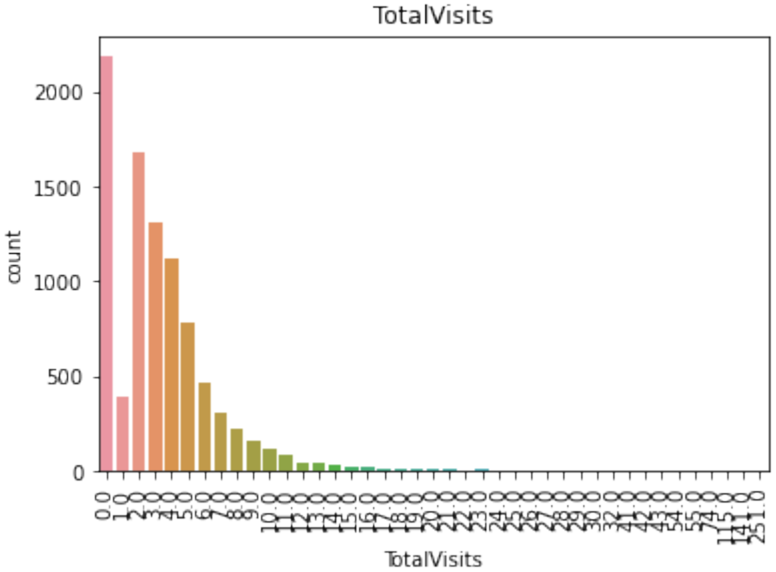
- Lead Source
- Total Visits
- Page Views Per Visit
- Last Activity
- Specialization
- How did you hear about X Education
- What is your current occupation
- What matters most to you in choosing a course
- Lead Profile

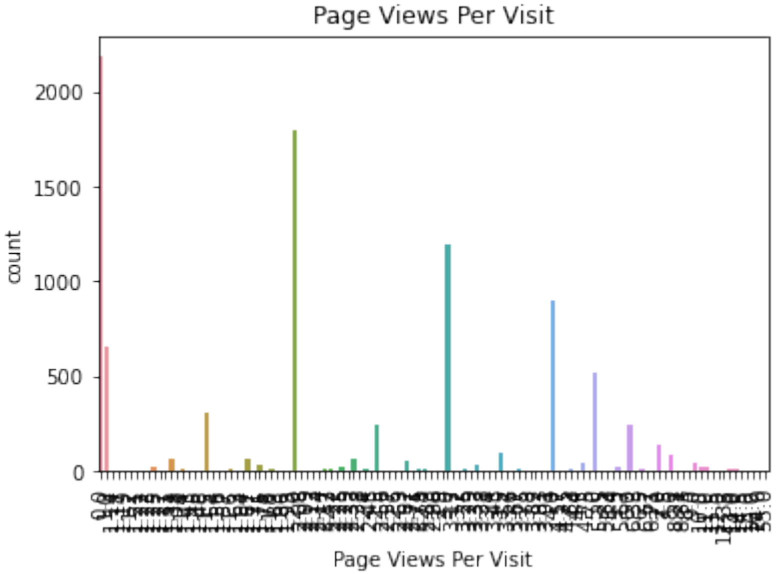
5.After cleaning data the data frame size reduced to

Univariate Analysis for categorical data

1. For Lead Source data we gathered the insight that -Google has the highest number of occurrences, hence need to impute the missing values with label 'Google’.
2. 0.0 has the highest number of occurrences, hence need to impute the missing values with label '0.0’.
3. 0.0 has the highest number of occurrences, hence need to impute the missing values with label '0.0’.



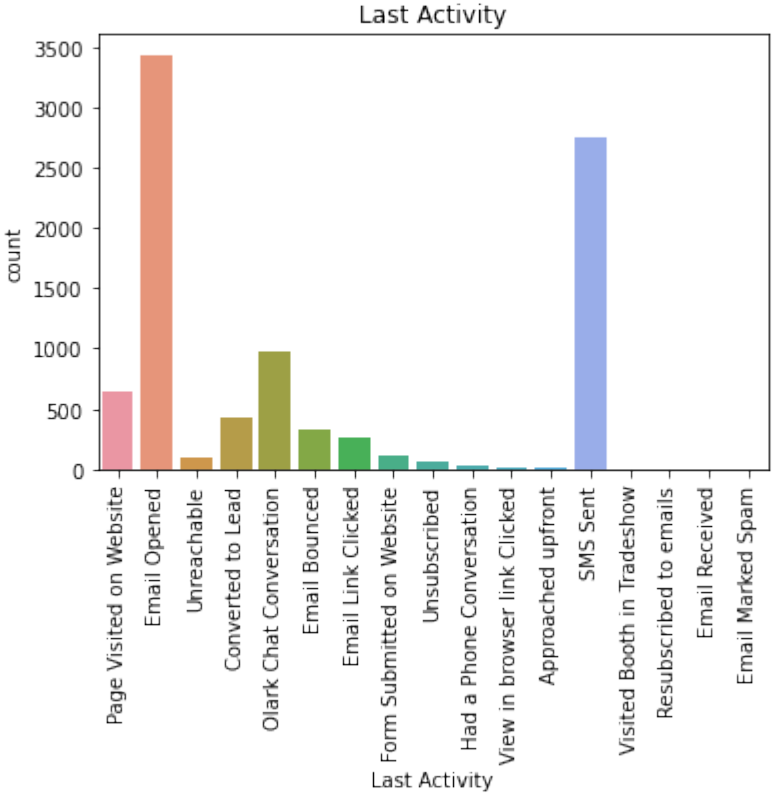


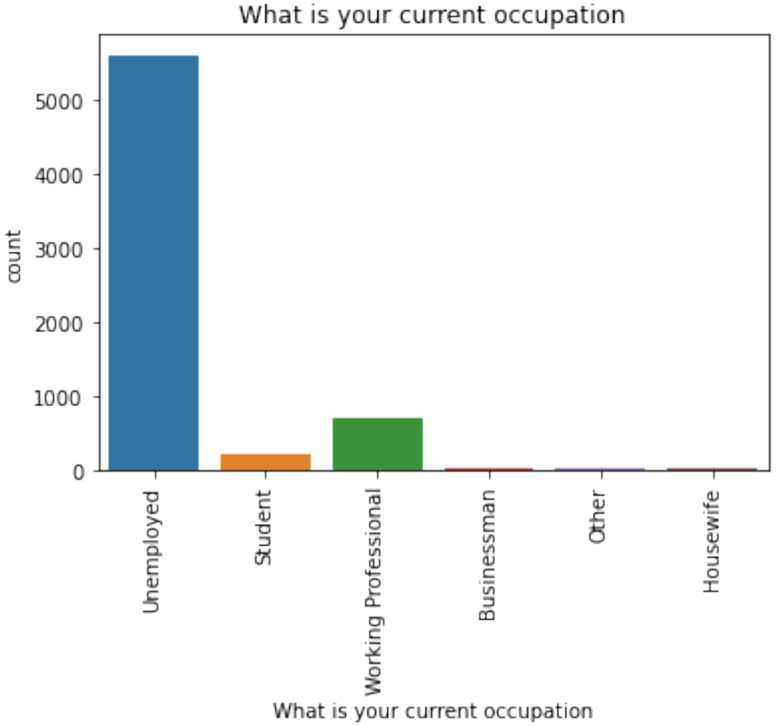


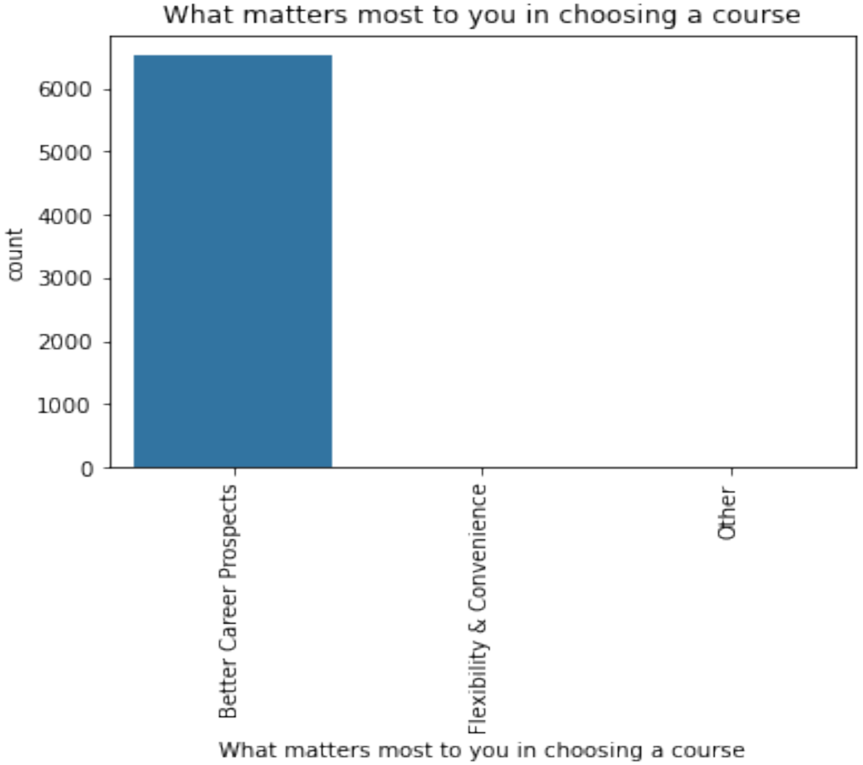
4. Insights for last activity Email Opened has the highest number of occurrences, hence need to impute the missing values with label 'Email Opened'.

5. The count of Unemployed is the highest for the Occupation column.

6. The ‘Better career Prospects’ matters the most while choosing a career.







Data Preparation

1. Step 1- Converting some binary variables (Yes/No) to 0/1.

Following are the values to be converted:

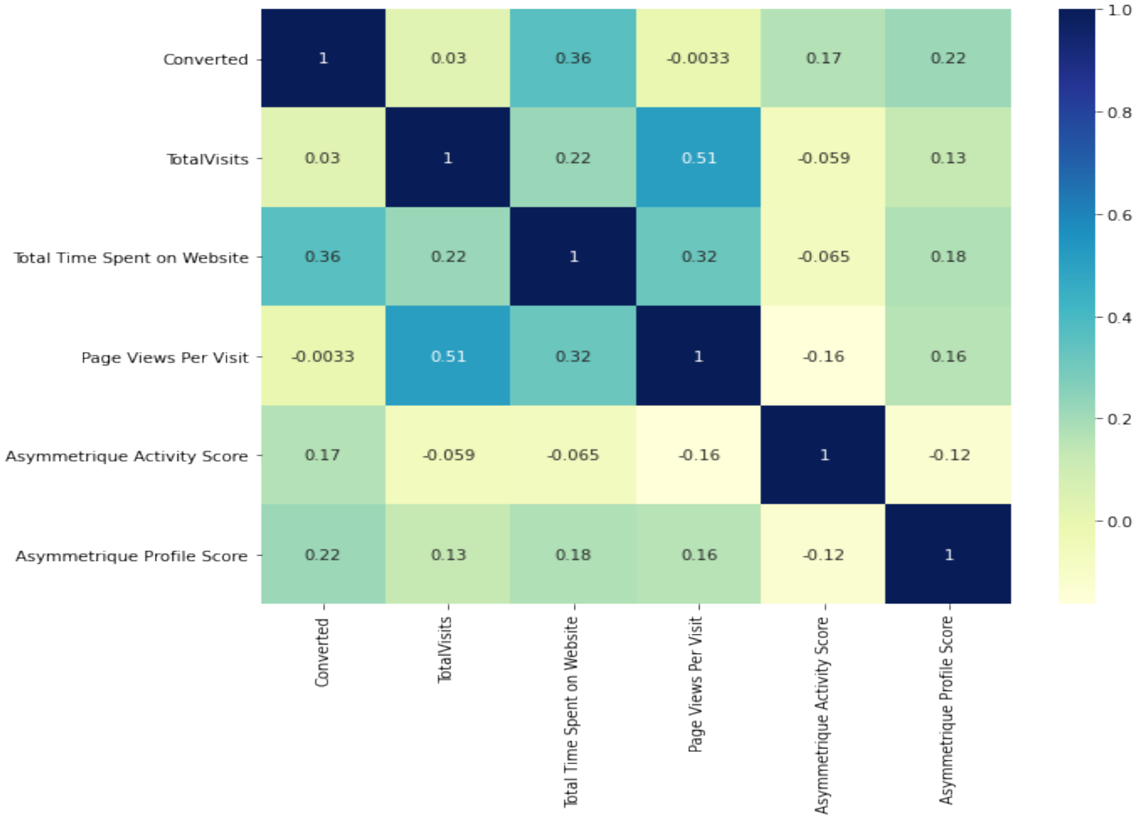
- 'Do Not Email’,
- 'Do Not Call’,
- 'Newspaper Article’,
- 'X Education Forums’,
- 'Newspaper’,
- 'Digital Advertisement’,
- 'Through Recommendations’,
- 'Receive More Updates About Our Courses’,
- 'Update me on Supply Chain Content’,
- 'Get updates on DM Content’

3. The insights gathered after converting the data- After converting the binary categories from 'Yes' to 1 and 'No' to 0, the dummy variables for multiple levels of categories will be used.

4. For categorical variables with multiple levels, create dummy features, remove repeated columns.

5. Plotted heatmap for categorical data

Plotted Heatmap for Categorical data

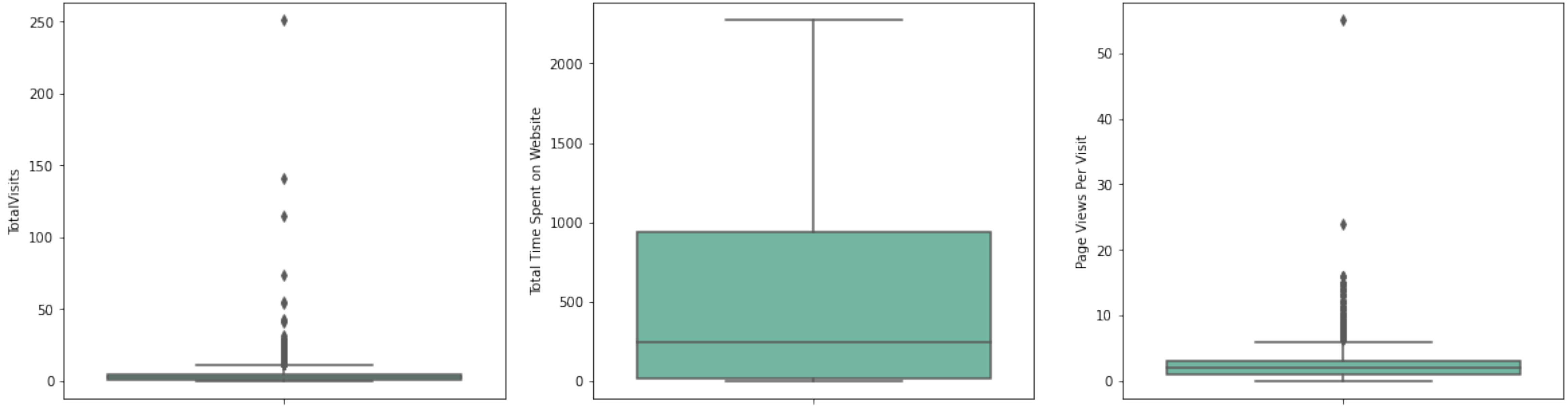


	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
Converted	1	0.03	0.36	-0.0033	0.17	0.22
TotalVisits	0.03	1	0.22	0.51	-0.059	0.13
Total Time Spent on Website	0.36	0.22	1	0.32	-0.065	0.18
Page Views Per Visit	-0.0033	0.51	0.32	1	-0.16	0.16
Asymmetrique Activity Score	0.17	-0.059	-0.065	-0.16	1	-0.12
Asymmetrique Profile Score	0.22	0.13	0.18	0.16	-0.12	1

Checking for outliers

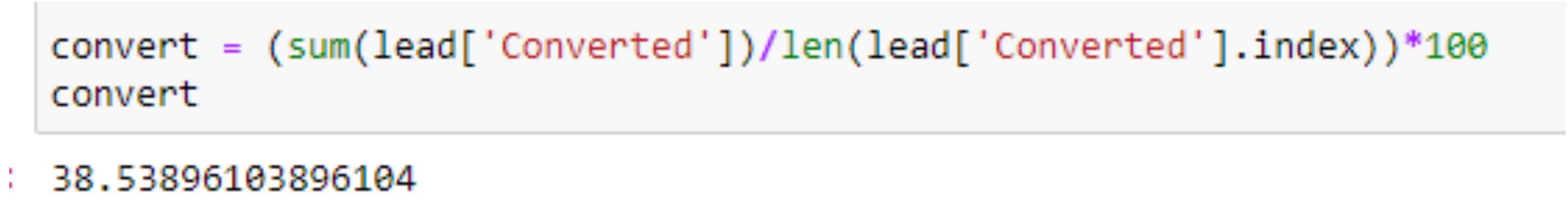
Insights:

The distribution is showing the outliers in this data.



TEST TRAIN SPLIT DATA

1. Separating target variable from dependent variable.
2. Putting target variable 'Converted' to a new series ‘y’.
3. Putting dependent variable in a new dataset called ‘X’.
4. Splitting the data into train and test set.
5. Rescaling the features with min max scaling.
6. We checked the converted rate

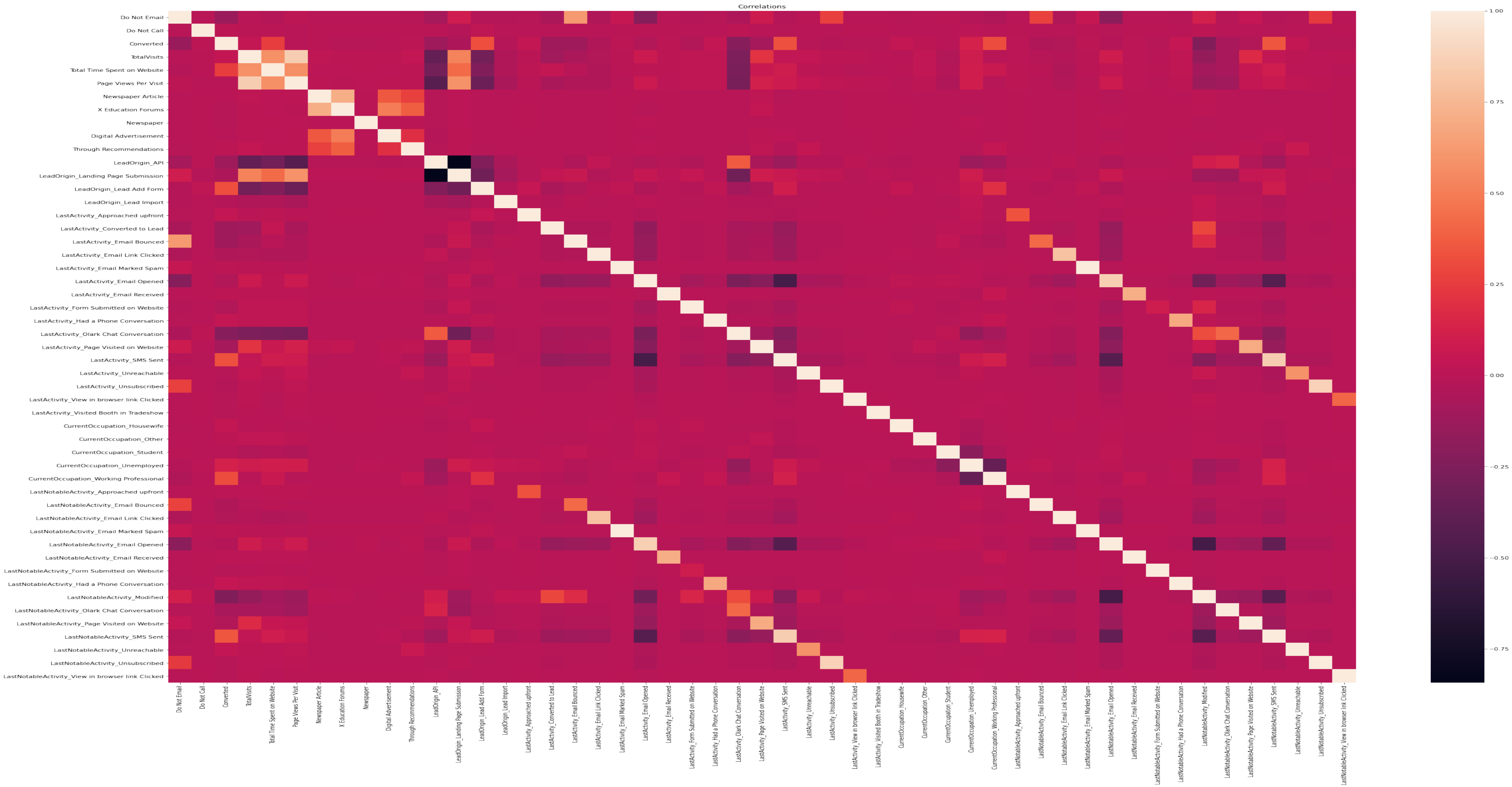


```
convert = (sum(lead['Converted'])/len(lead['Converted'].index))*100
convert
```

```
: 38.53896103896104
```

7. Insights- We have almost 39% converted rate.

CORRELATIONS -   This is the heatmap. So, proceeding with building a model based on the p-values and VIFs. Checking for correlation in the as heatmap, difficult to spot the highly correlated variables.



Model Building

1. Running the Initial Training Model.
2. Feature Selection Using RFE.
3. Rebuilding the models- Model 1, Model 2, Model 3, Model 4, Model 5, Model 6, Model 7.
4. Creating a data frame with predicted probabilities.
5. Creating a data frame with the actual churn flag and the predicted probabilities.
6. Insights gathered:

   - All features have VIF values less than 5, there is no multicollinearity issue in the dataset.
   - Dropping the highest in-significant features i.e 'What is your current occupation_Housewife' has 0.999 p-value.

7. Plotting the ROC Curve.

An ROC curve demonstrates several things:

- - It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Finding Optimal Cutoff Point:

Insights- Hence we can see that the final prediction of conversions have a target of 80% (79.8%) conversion as per the X Educations CEO's requirement . Hence this is a good model.

Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value, Negative Predicative Value on final prediction on train set

Making predictions on the test set:

Insights- Hence we can see that the final prediction of conversions have a target rate of 79% (78.5%) (Around 1 % short of the predictions made on training data set).

Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity on test set.

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the
  optimal.
  cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- The lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

THANK YOU