LEAD SCORING

PROBLEM STATEMENT :-

An X education company need to help most promising leads. The leads that are most likely to convert into playing customers. The company requires us to build a model where in 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 customer 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%

Business Objectives:-

1) X education want a model to assign a lead score to know more promising or hot leads.
2) T here are some more problem presented by the company which the model should be able to adjust to if the company's requirements change.
3) The model should be that it can be used accurately when it is deployed in future.

- Mythology used to drive conclusion

Step :-1

Data import, understand, check null value, Cleaning, Model Bulding

a) important the data
b) Handling the duplicate data.
c) Handling the null values.
d) Drop the unnecessary column.
e) imputing the value where are required.
f) Handling the outlier.

Step :- 2

Start to Data Analysis- Exploration

- ❖ Univariate Analysis :-
  1) Categorical Variable.
  2) Numerical Variable.
- ❖ Bivariate Analysis :-
  1) co-relation coefficient and pattern between the viables etc.

Step:-3

- Model building preparation & Validation
- Dummy variable's
- Train – Test split.
- Scaling

Step :- 4

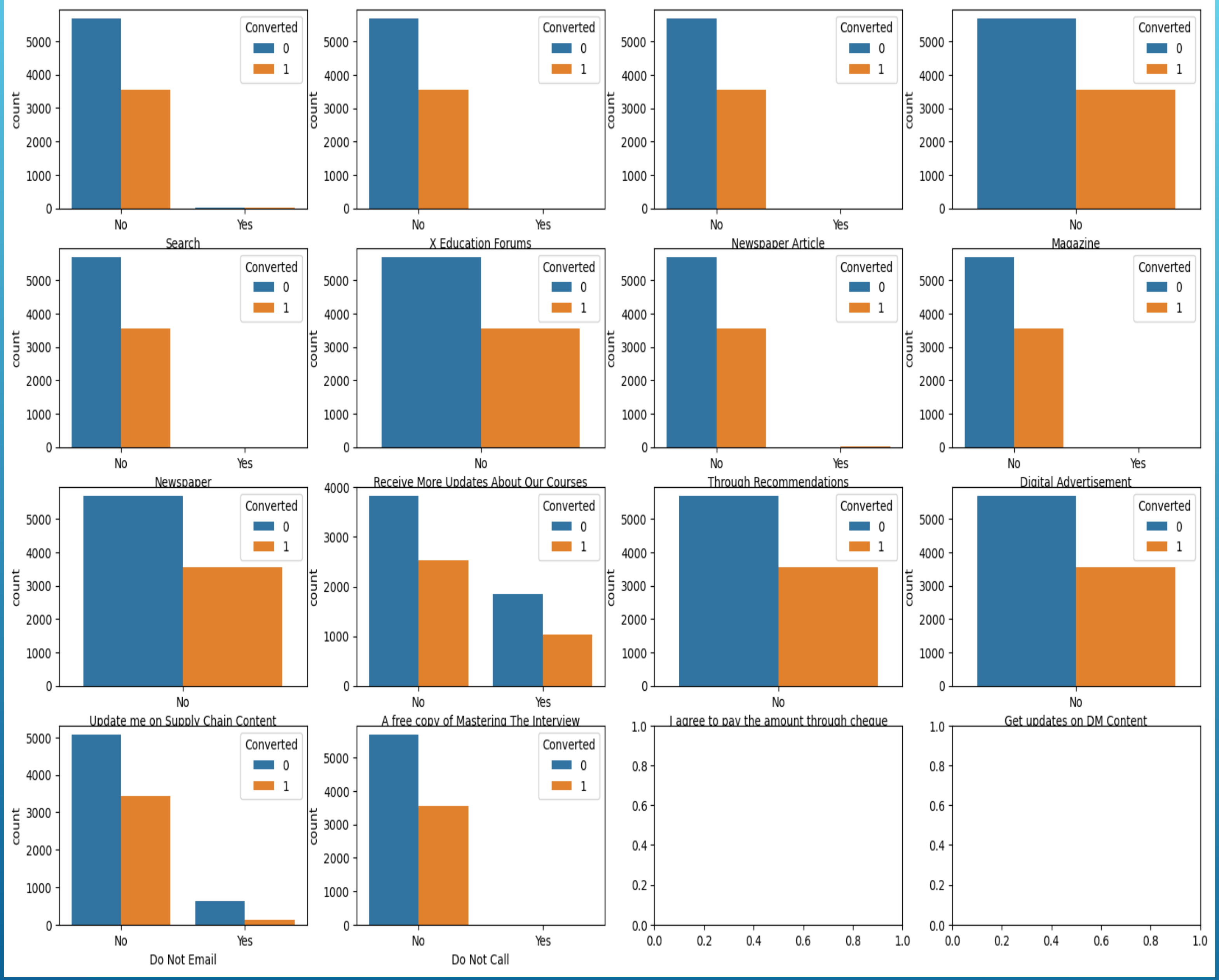
- Model Evaluation.
- Creating a data frame With the actual conversion flag and predicted probability.
- Creating a new column ‘predicted’.
- Finding the optimal cutoff
- Precision Recall view.

Step:- 5

Making Predictions based on the test set

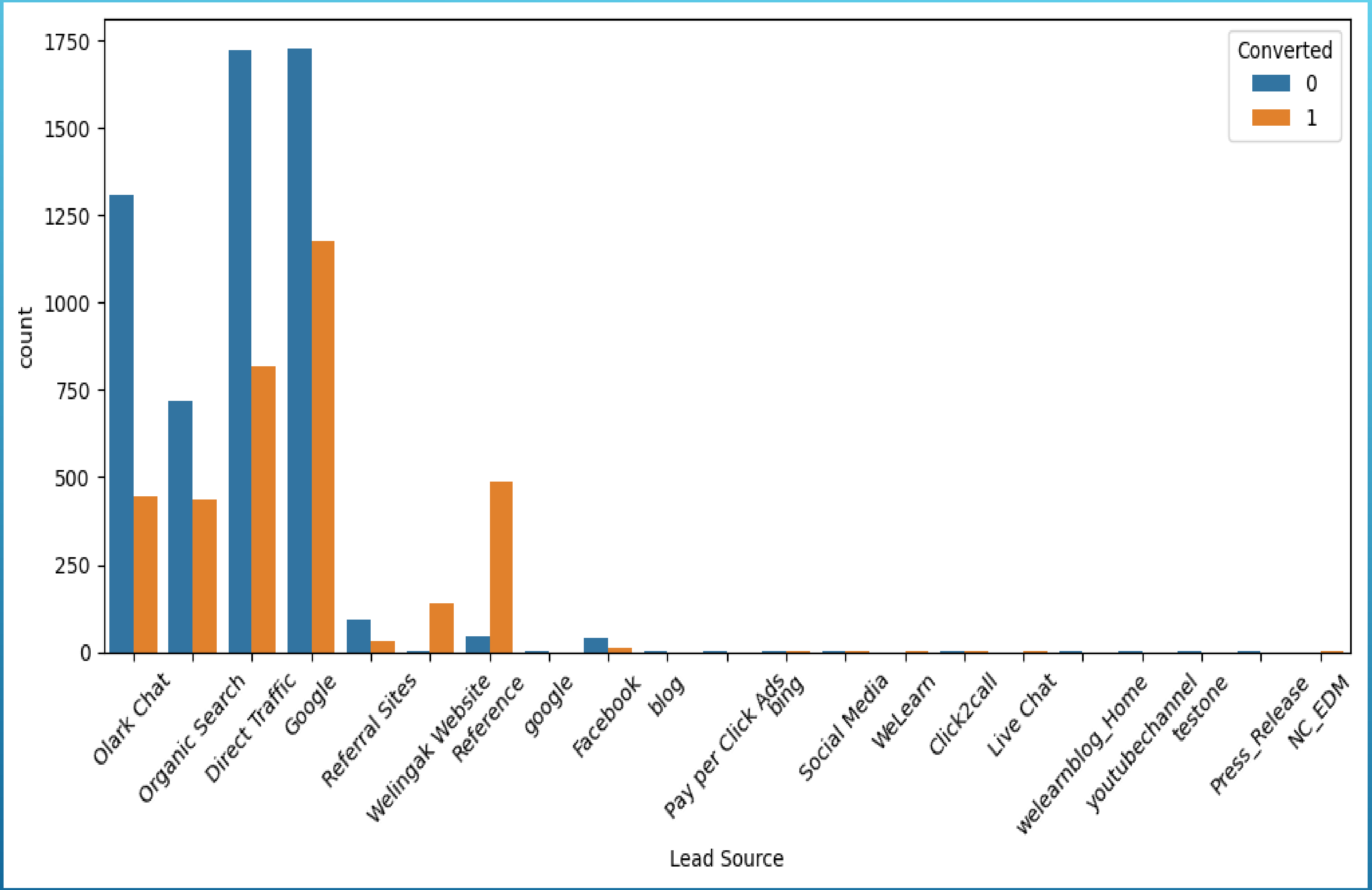
Step:- 6

Deriving conclusion & Recommendation based on model



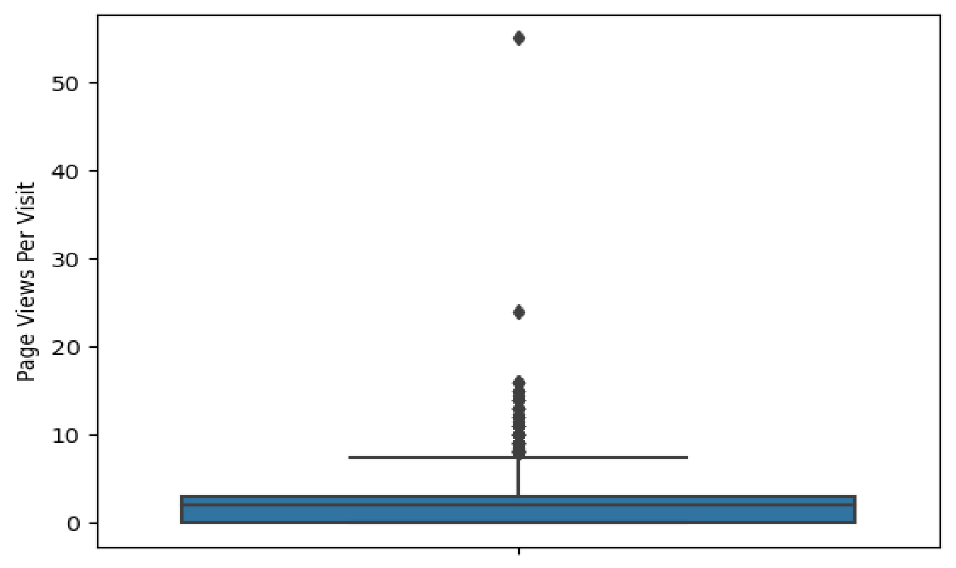
DATA CLEANING AND PREPARATION

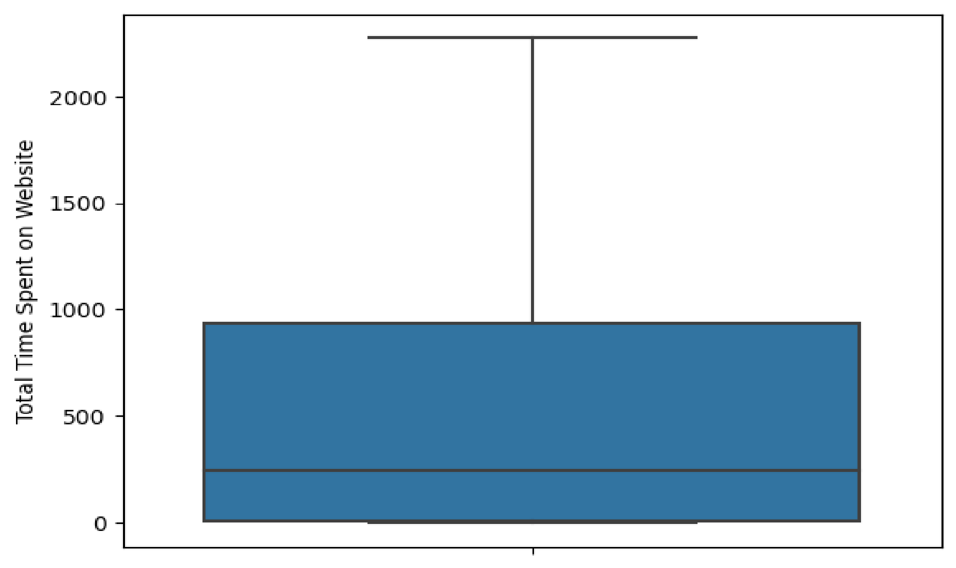
- ❖ Firstly we drop the columns which have less than 30% null values have been removed
- ❖ Let us check the remaining columns and accordingly remove columns which are not required for our analysis.
- ❖ We we look into the excel data we can see that there are many columns with no and yes as data. Let us visually represent these columns
- ❖ Next the column with null value , we imputed them with 0.0
- ❖ We checked other remaining column and drop column which are not required.

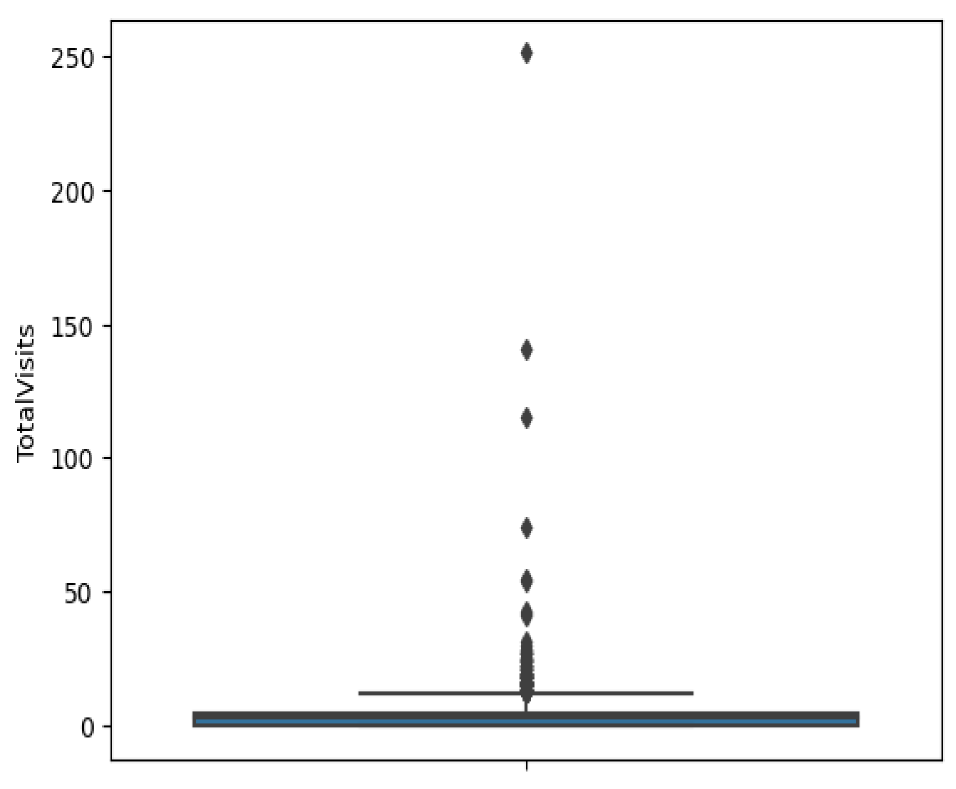


2) LEAD SOURCE

We can observe that most Number of conversions are for Google. Welingak websites and References have a high conversion rate. Direct Traffic and Olark chat have a lower conversion rate.





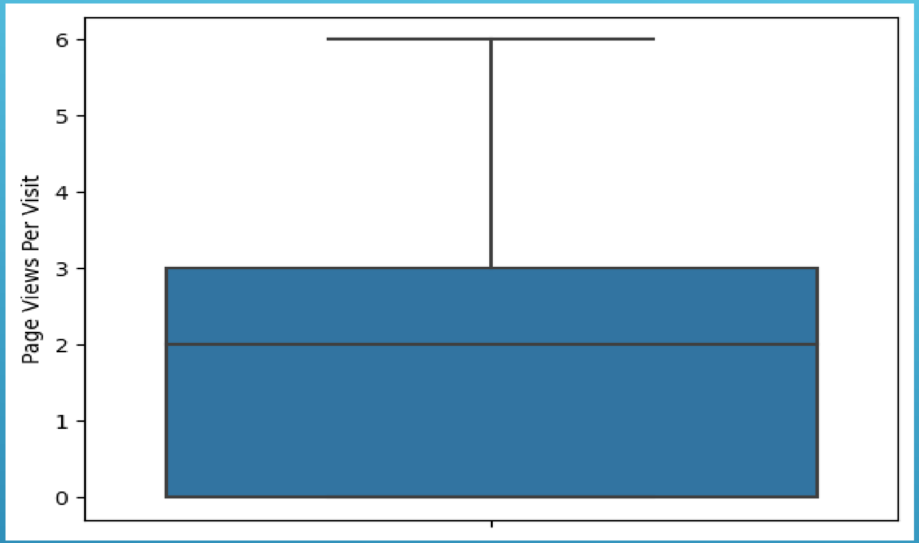


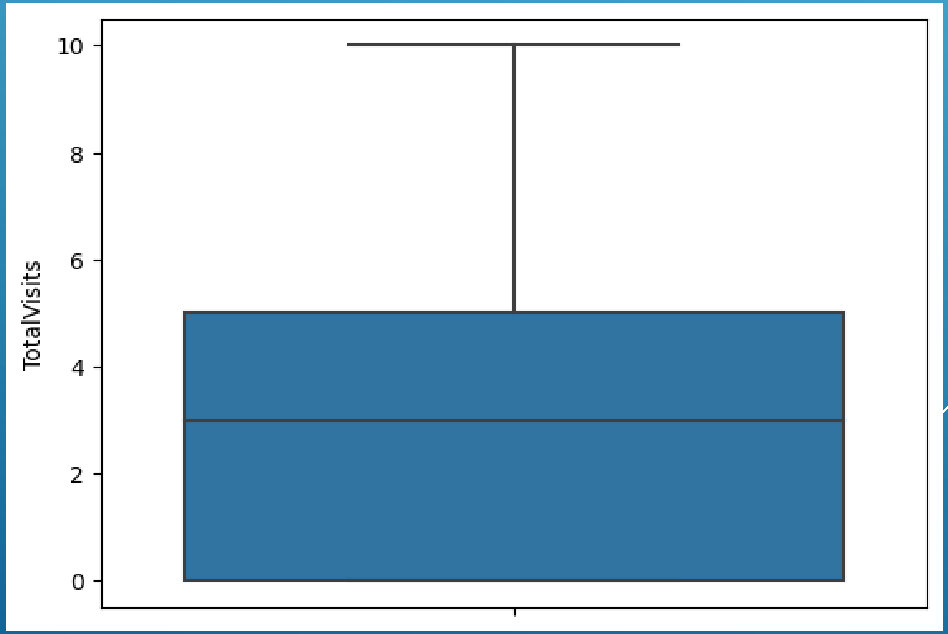
We can observe outliers in the above boxplot We can see that "Page Views Per Visit" and "Total Visits" have outliers and hence they need to be treated.

Data Conversion:-

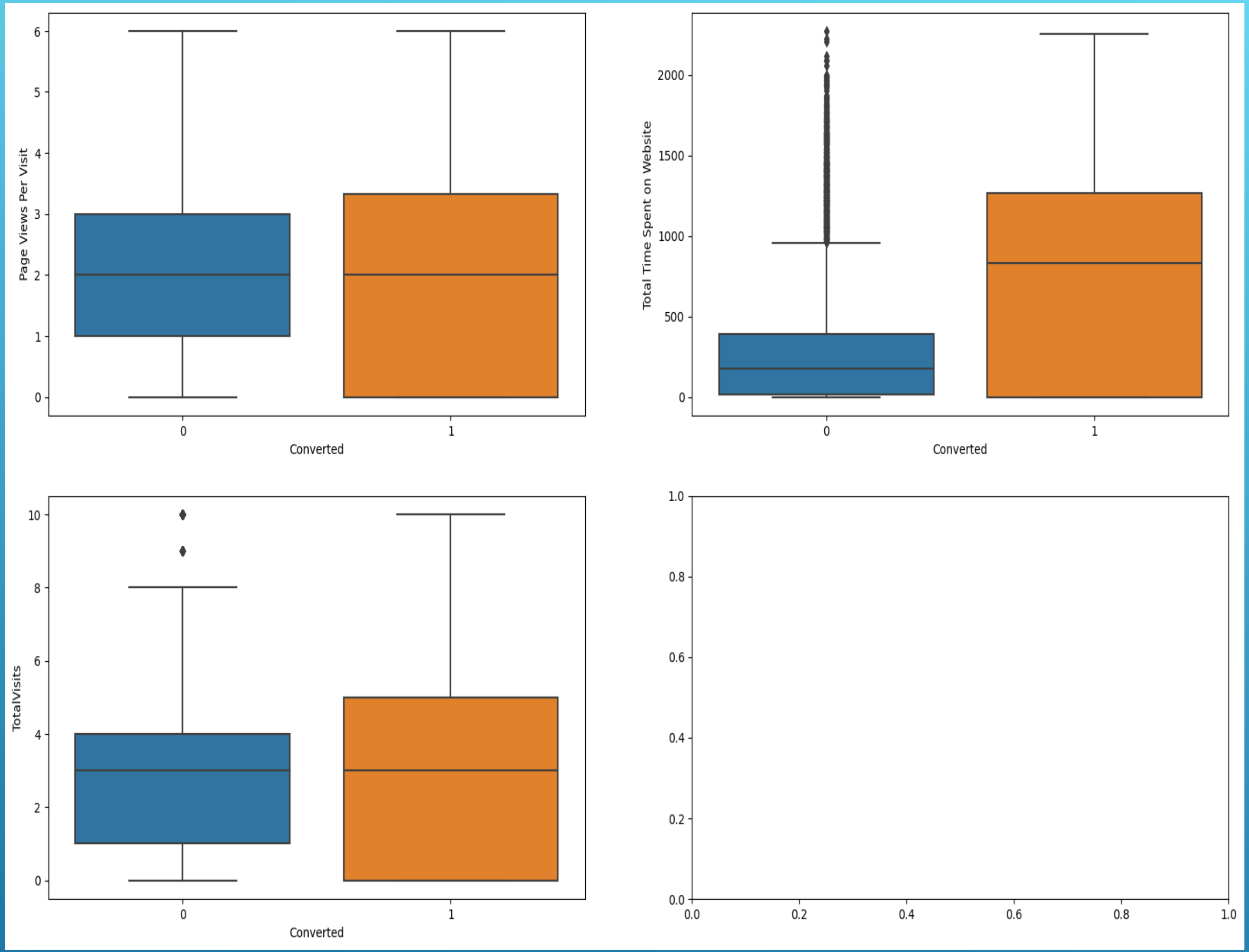
- Numerical Variable normalized.
- Outliers Handel.
- Dummy Variable Created.
- Feature Scaling.
- Correlation Searched and found.

We can observe now that there are no outliers.



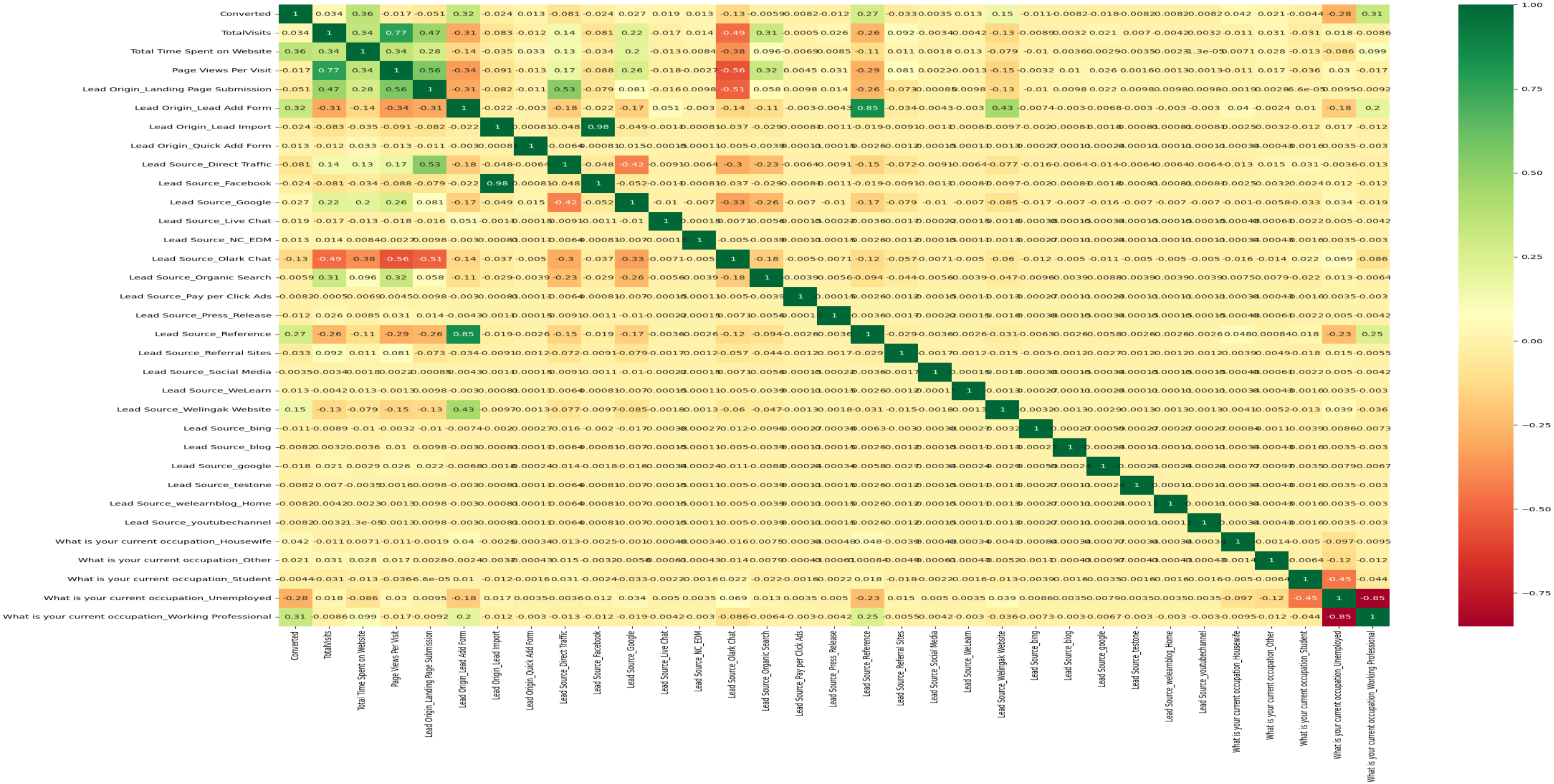


LET US NOW PLOT ALL THE NUMERICAL VARIABLES AND CHECK WHICH AMONG THEM HAVE THE HIGHEST CONVERSIONS.



From the above we can observe that the Total time spent on Website has the highest conversion of leads. The rest of the columns have a moderate conversion of leads.

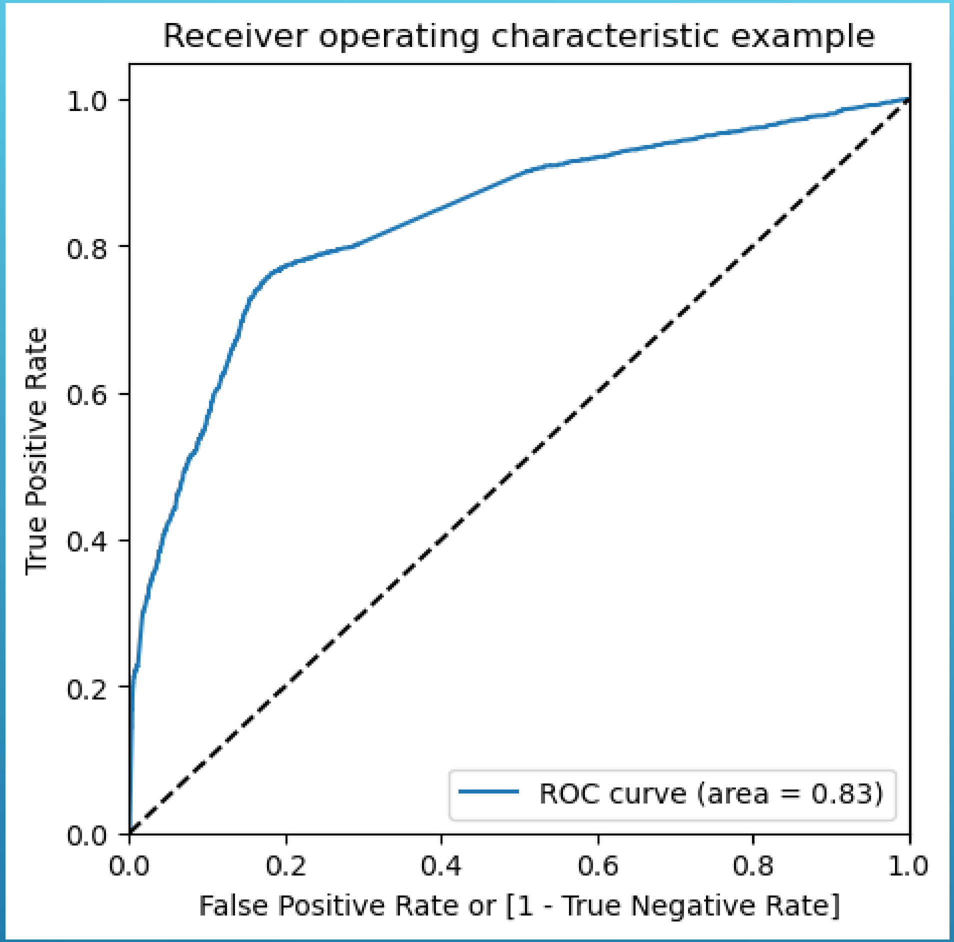
STEP 5: CORRELATIONS LET US PLOT A HEATMAP AND OBSERVE THE CORRELATIONS BETWEEN DIFFERENT COLUMNS



MODEL BUDLING

- Splitting the data into Training and Test sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio
- Generalize linier model regression results.
- Feature Selection Using REF
- Budling Model
- Assessing the model.
- Prediction made based on test data set.

ROC Curve



For an ideal model, the ROC curve has area of 1. We have got an ROC area of 0.83. Closer a model is to 1 the better the model is considered. The ROC of 0.83 is an indication of a good model.

Conclusions & Predictions

- The Accuracy, precision and recall score we got from the test date are in the acceptable region.
- Accuracy, Sensitivity and specificity value of test set are around 76% and 77% which are approximately closer to the respective value calculated using trained set.
- Also, the lead score calculated in the trained set of data show the conversion rate on the final predicted model is closed to 80%, ~78%
- Hance overall all the model seem to be good.
- A customer lead score by ‘welingak Website’ is a hot lead.
- A customer who is currently ‘Working professional’ or ‘unemployed’ is a hot lead.
- Total time spent on web site high conversion.

$$ln(p/1-p)=-0.4024 + 1.0960 \times \mathit{TotalTimeSpentonWebsite} + 3.0447 \times \mathit{LeadOriginLeadAddFormy} - 0.9683 \times \mathit{LeadSourceDirectTraffic} - 0.9582 \times \mathit{LeadSourceFacebook} - 0.5735 \times \mathit{LeadSourceGoogler} - 0.7163 \times \mathit{LeadSourceOrganicSearch} - 1.1980 \times \mathit{LeadSourceReferralSites} + 1.9739 \times \mathit{LeadSourceWelingakWebsite} + 1.9739 \times \mathit{WhatisyourcurrentoccupationWorkingProfessional}$$