Lead Score Case Study

SAKSHI

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the Case Study

There are quite a few goals for this case study:

1. 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.
2. There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

Approach

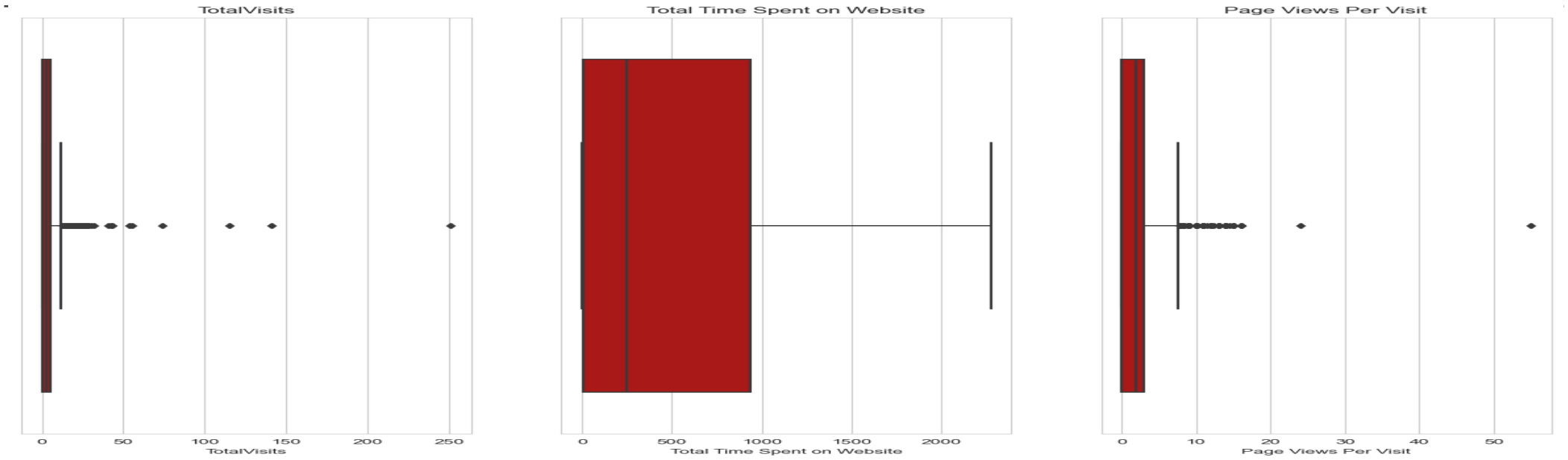
- Source the data for analysis
- Reading and understanding the data
- Data cleaning
- EDA
- Feature Scaling
- Splitting the data into train and test dataset
- Prepare the data for modeling
- Model building
- Model Evaluation – Accuracy, Sensitivity and specificity
- Precision Recall
- Making prediction on the test dataset

Data Sourcing, Cleaning & Preparation

- Read the data from CSV file
- Outlier Treatment
- Data Cleaning, Handling Null values and Removing higher null value data
- Removing Redundant columns in the data
- Imputing Null values
- Exploratory Data Analysis – Approx conversion rate is 38%
- Feature Standardization

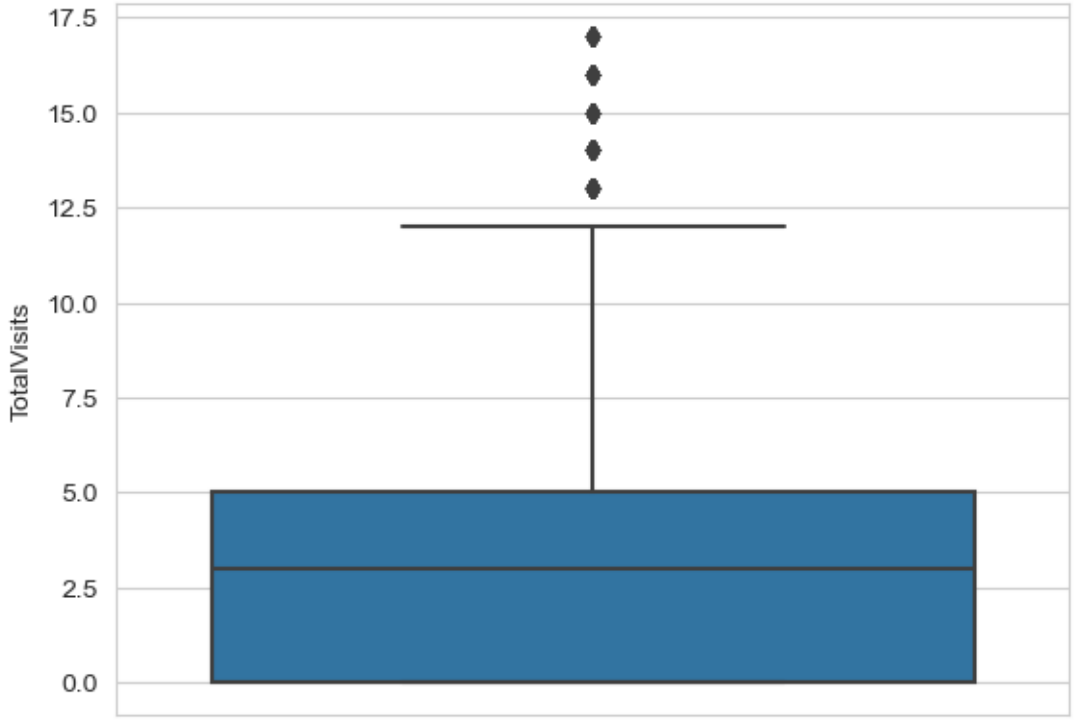
Outliers:

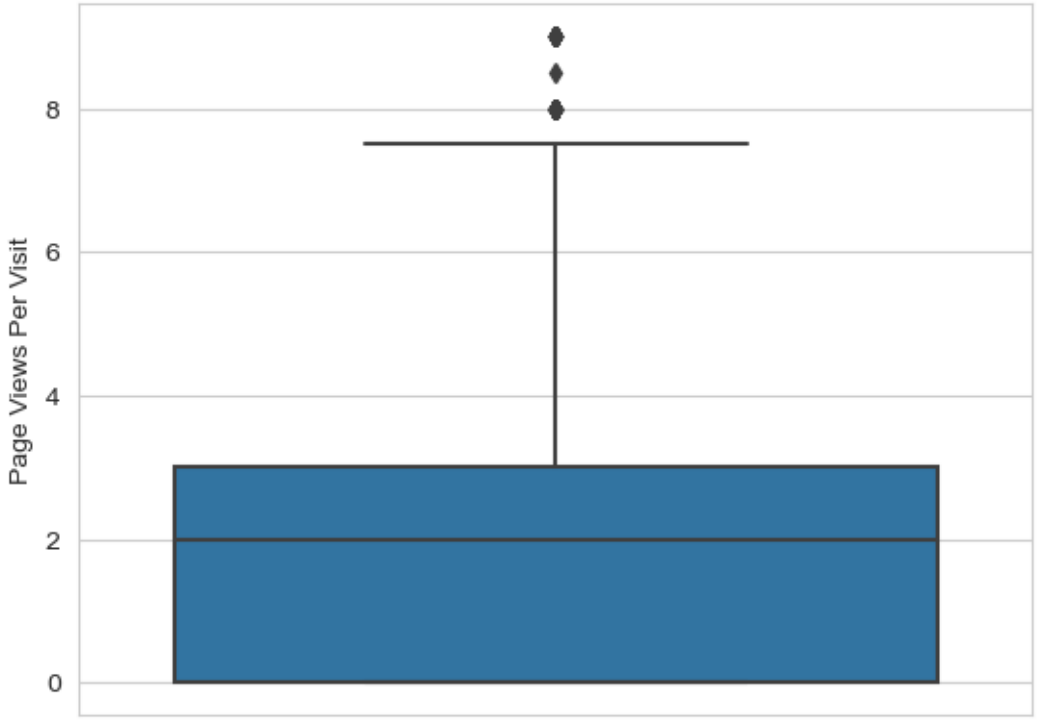
Total visits, Total time spent on website, Page views per visit have outliers.



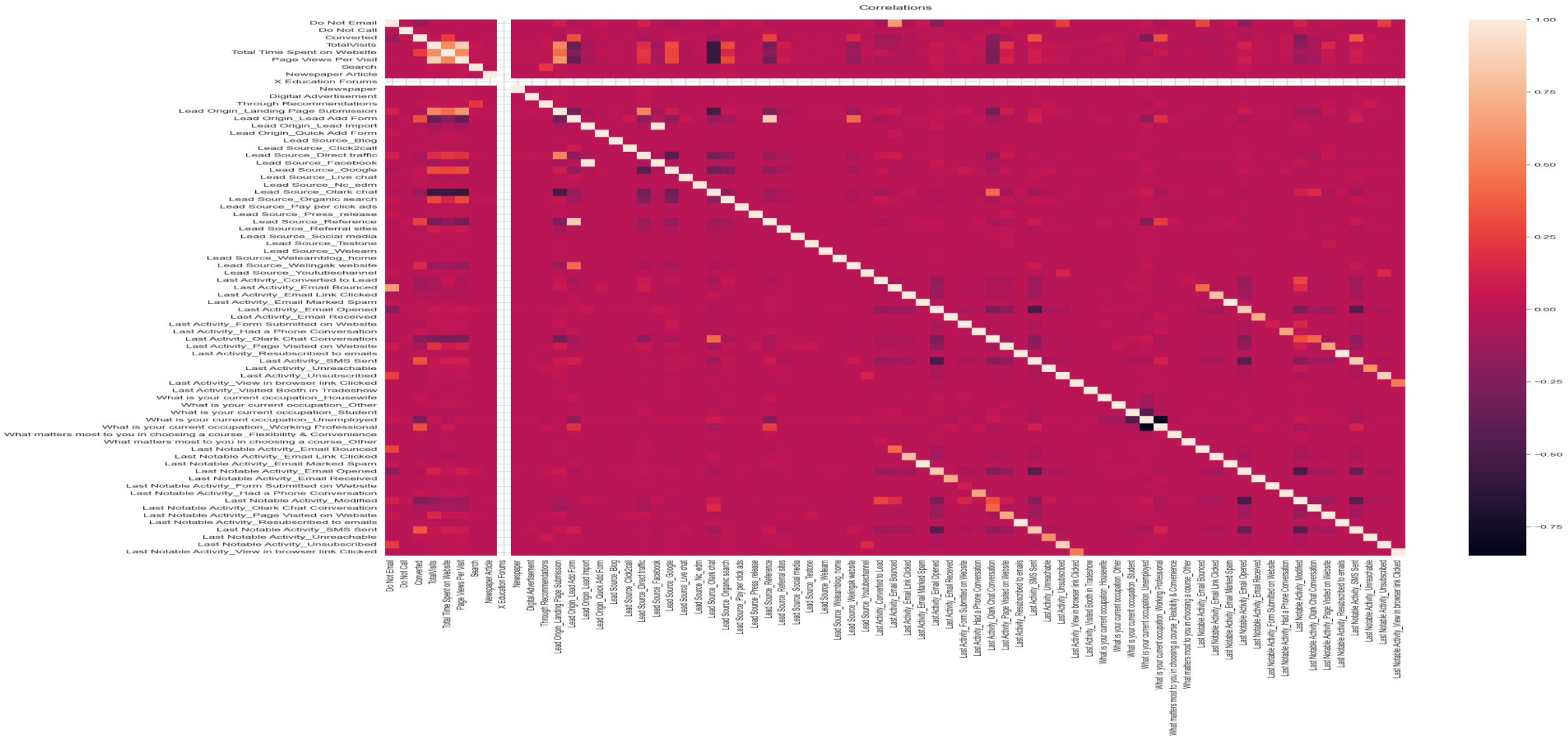
- From the above boxplots we can observe two outlier variables in our dataset ('TotalVisits' and 'Page Views Per Visit').
- We need to do a 0.99-0.1 analysis in order to correct the outliers.

Data Analysis

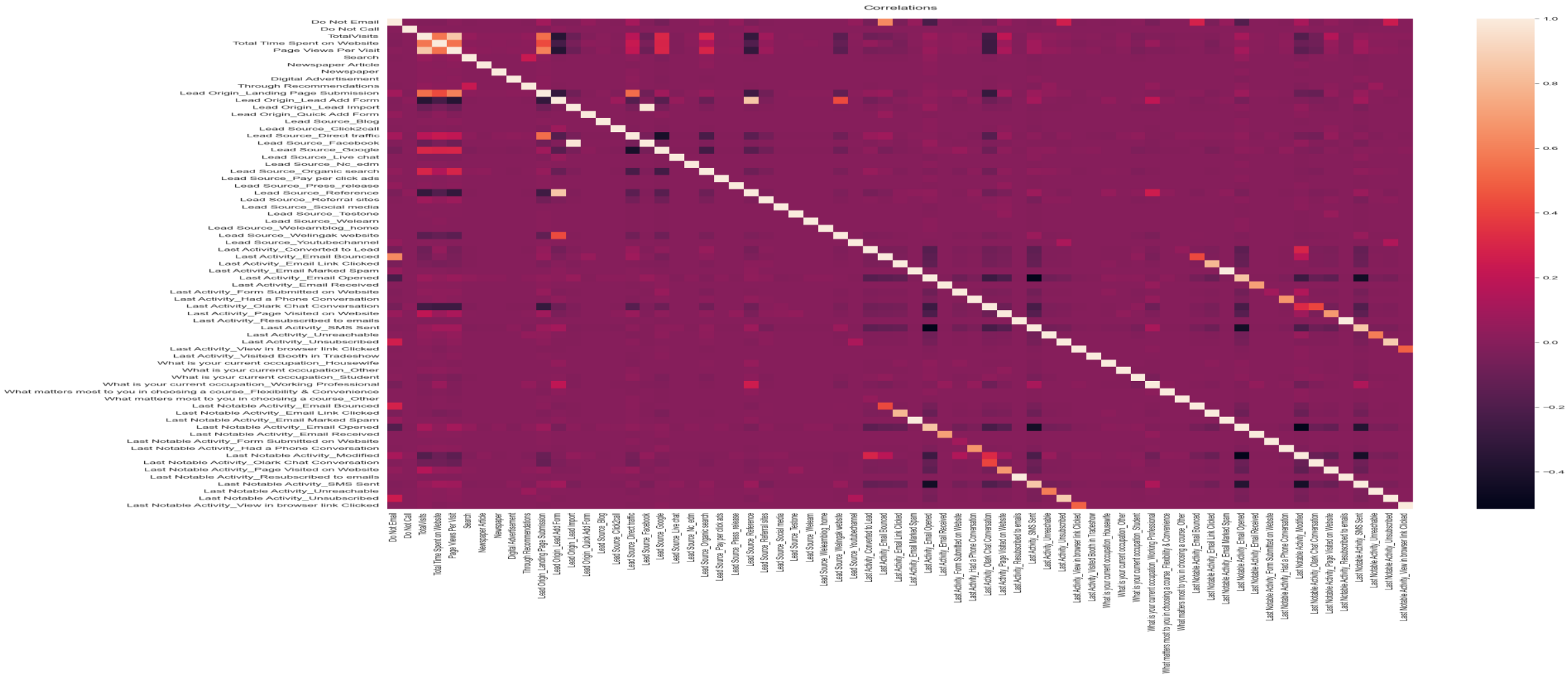




Correlation of the dataset¶



Here, X Education Forums has no data so, it is better to remove from the dataset.
Also, we need to remove the highly correlated value.



Data Preparation

- Converted binary variable into 0 and 1
- Created dummy variables for categorical variables

Feature Scaling and Splitting Train and Test datasets

- Feature Scaling of Numerical data
- Splitting Data into Train and Test data

Model Building

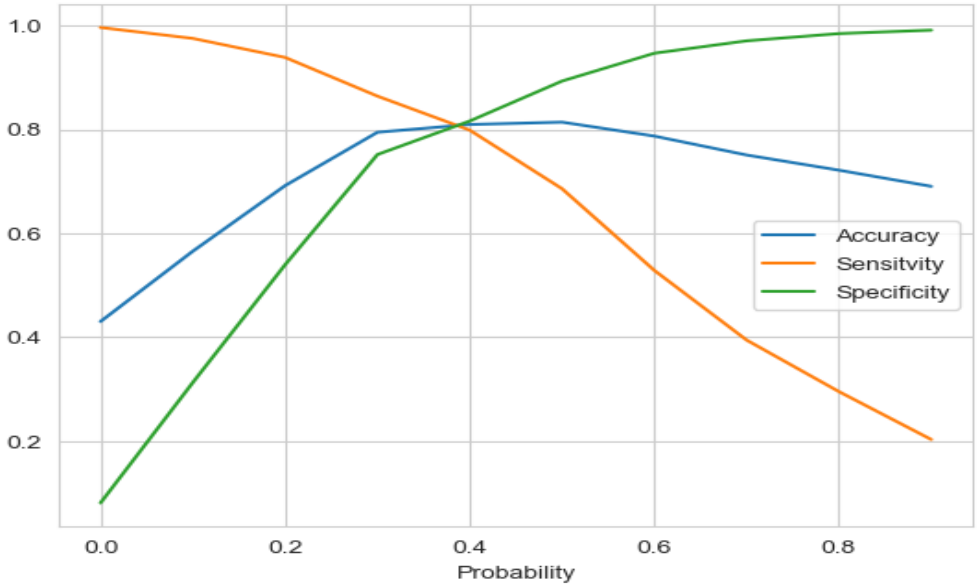
- Feature Selection using RFE
- Determined optimal Model using Logical Regression
- Calculated Accuracy, Sensitivity and Specificity
- Precision and Recall to evaluate Model

Variables impacting the Conversion Rate

- Total visits
- Total Time spent on website
- Lead Source_Olark Chat
- Lead Origin_Lead Add Form
- Lead Source_Welingac Website
- Do not Email
- Lead Source_Referral sites etc.

Model Evaluation- Sensitivity & Specificity on Train data Set

Graph depicts an optimal Cut off of 0.37
Based on Accuracy, Sensitivity & Specificity



Accuracy = 78%

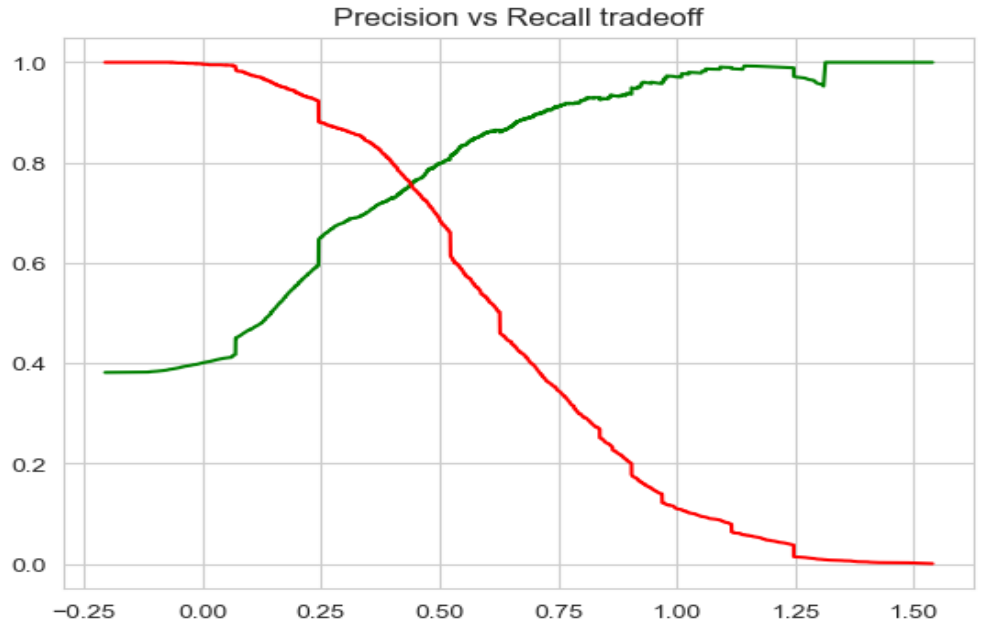
Sensitivity = 82%

Specificity = 76%

Model Evaluation

Precision & Recall on Train Dataset

The graph depicts optimal cut off of 0.42 based on Precision and Recall



Precision = 79%

Recall = 65%

Model Evaluation

Sensitivity & Specificity on Test Data Set

- Accuracy = 80%
- Sensitivity = 80.8%
- Specificity = 76.5%

Model Evaluation

Precision & Recall based on Test Data Set

```
Precision score in predicting test dataset: 0.725625539257981
Recall score in predicting test dataset: 0.788191190253046
```

Conclusion

- The Accuracy, Precision and Recall score we got from the test data are in the acceptable region.
- In business terms, this model has an ability to adjust with the company’s requirements in coming future.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:
    - Last Notable Activity_Had a Phone Conversation
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional.