Lead Scoring Case Study using logistic regression

SUBMITTED BY :

- Surendra Dhondale
- Surbhi Yadav
- Surej Prabhu

Contents

- Problem Statement
- Business Objective
- Problem Approach
- EDA
- Correlations
- Model Evaluation
- Conclusion

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. They have process of form filling on their website after which the company classifies that individual as a lead.
- 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, this means 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

Business Objective

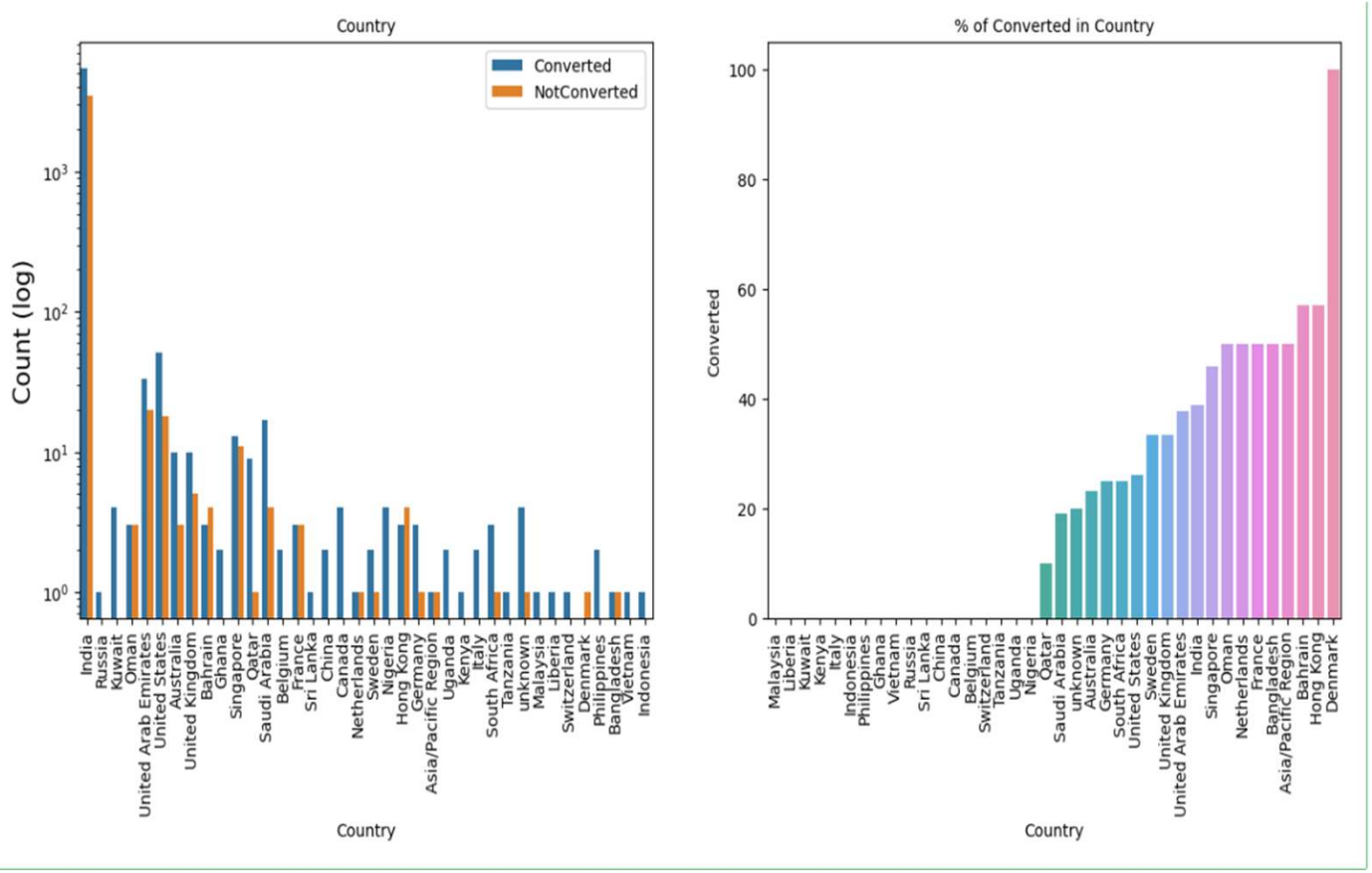
- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full manpower and after achieving target what should be the approaches.

Problem Approach

- **Importing the Data and inspecting the Data frame.**
- **Data preparation**
- **EDA**
- **Numerical Column Analysis**
- **Dummy Variable Creation**
- **Train-Test Split & Logistic Regression Model Building**
- **Model Building using stats model and RFE**
- **Model Evaluation**
- **Making Prediction on Test Set**

EDA – Data Cleaning

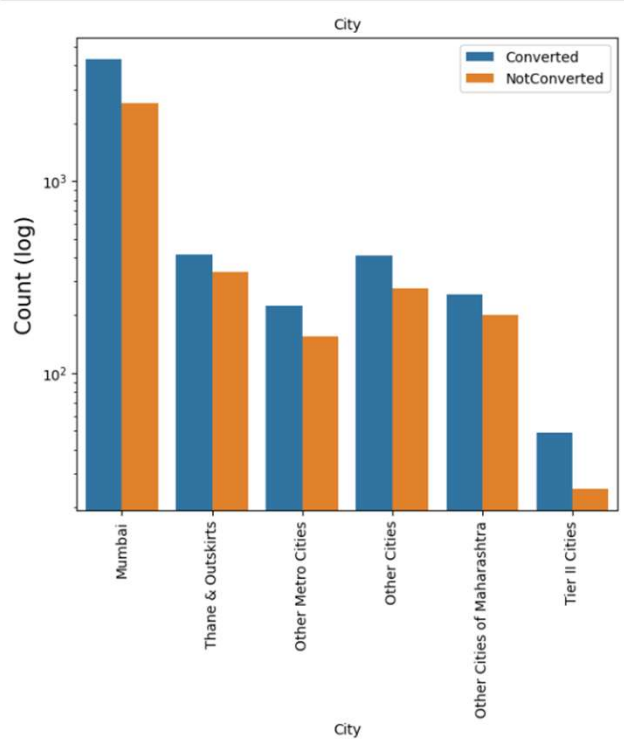
- As the data has lot of outliers - using logarithmic representation for better view of data in graph

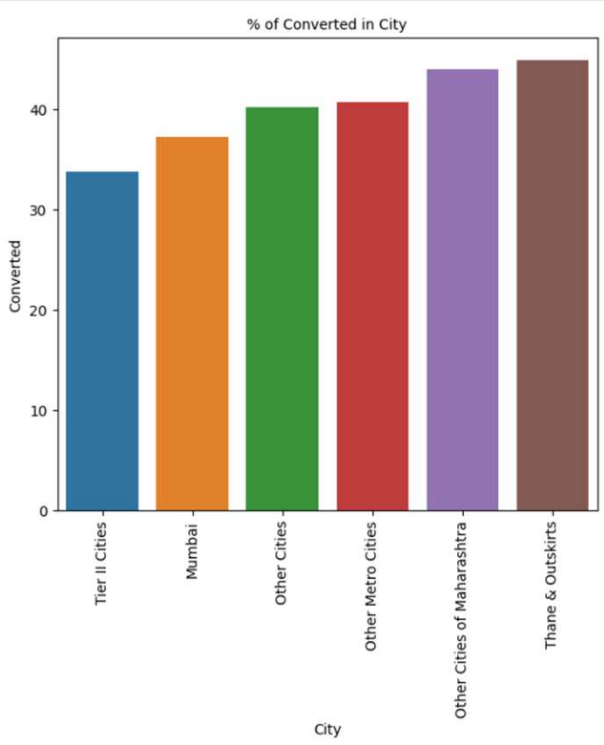


Analysis based on the Graph Figure:

•Conversion rate seem to be highest for Asia Pacific and Middle East countries
•European nations have moderate conversion percentage
•Considering the data distribution however, we are looking at dropping the country column

Analysing the column City and plotting the graph to check the distribution



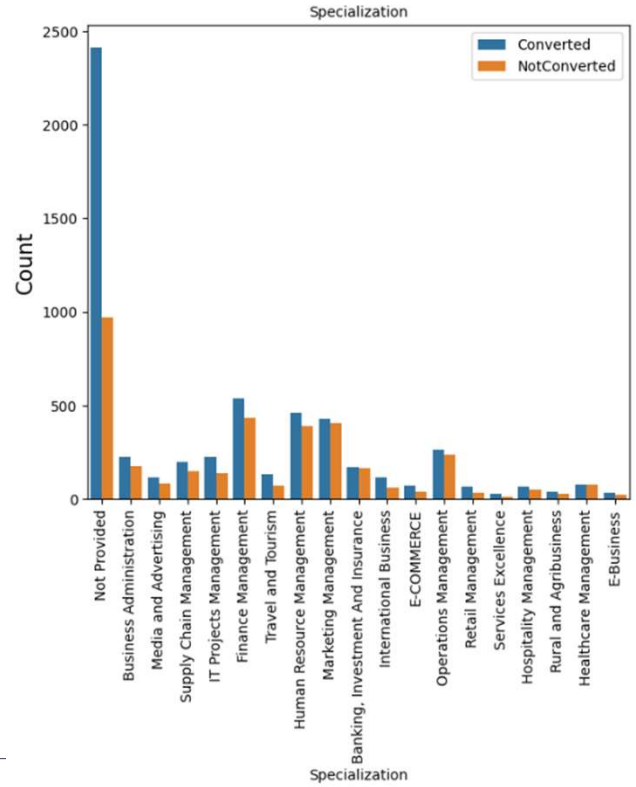


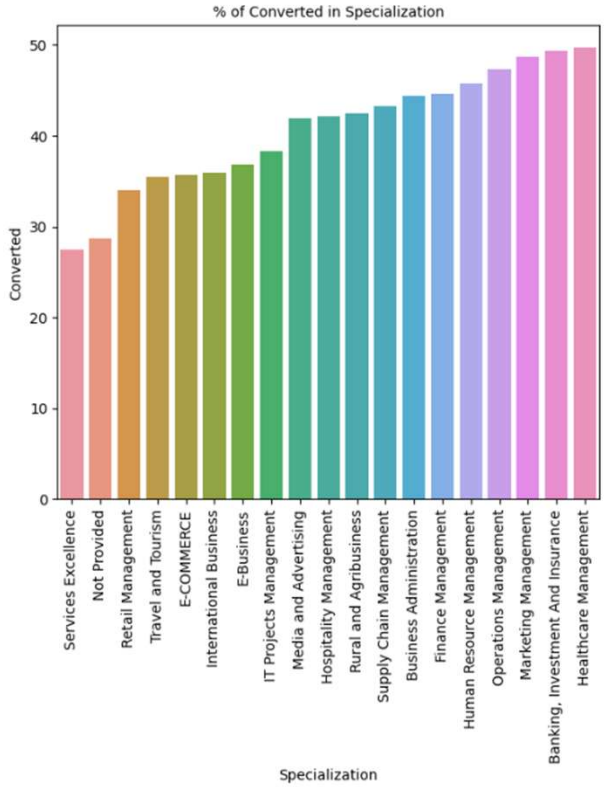
Analysis based on the Graph Figure:

- We see that the most number of people converted is from Mumbai, but the conversion percentage is high in Thane and Outskirts

Analyzing the column Specialization

Plotting the graph to check the distribution



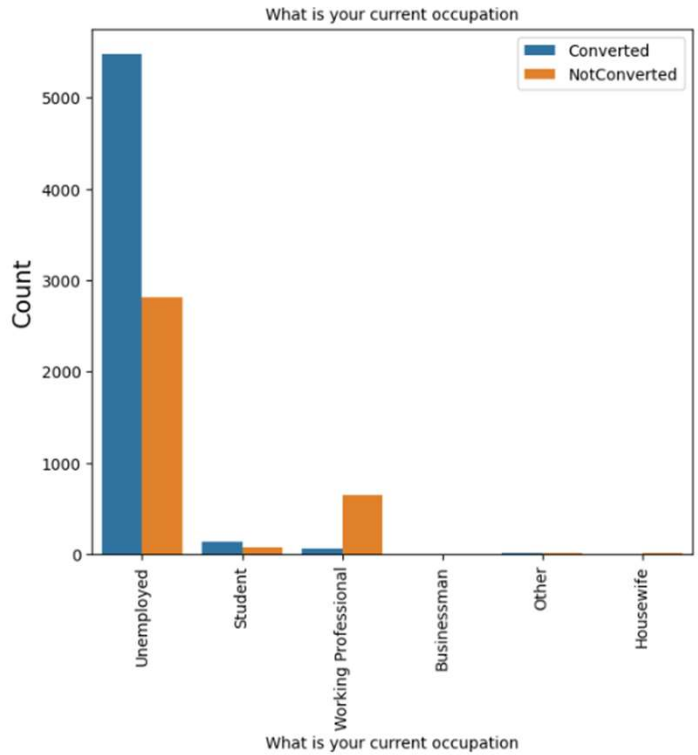


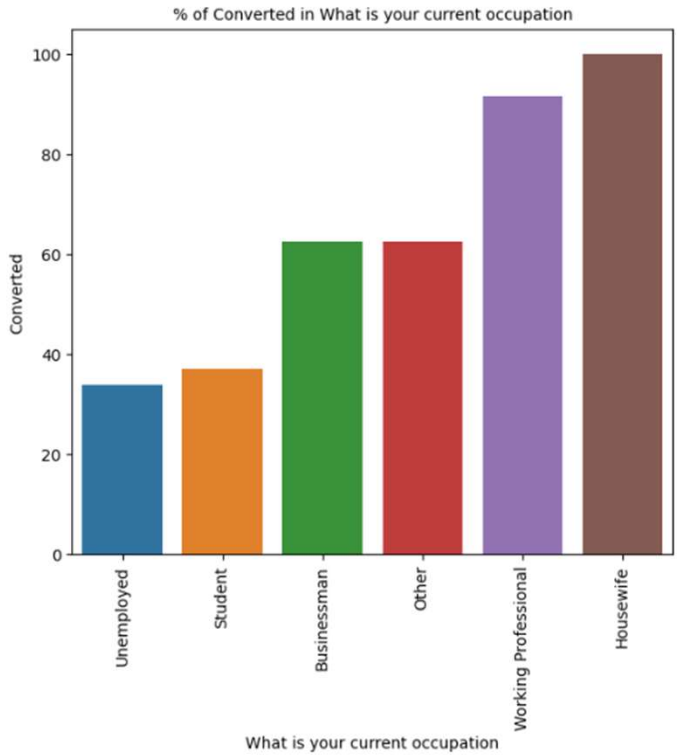
Analysis based on the Graph Figure:

- Conversion rate seem to be very good for potential candidates in Health Care management
- potential candidates who have not specified their specialization do not seem to look very promising

Analyzing the column 'What is your current occupation’

Plotting the graph to check the distribution



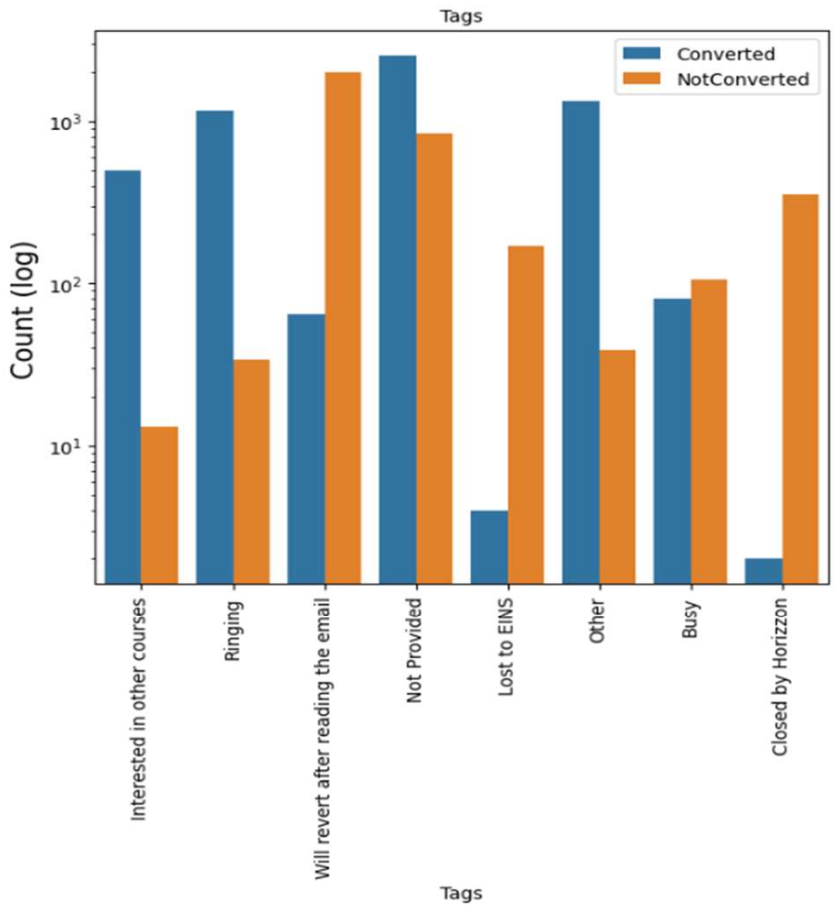


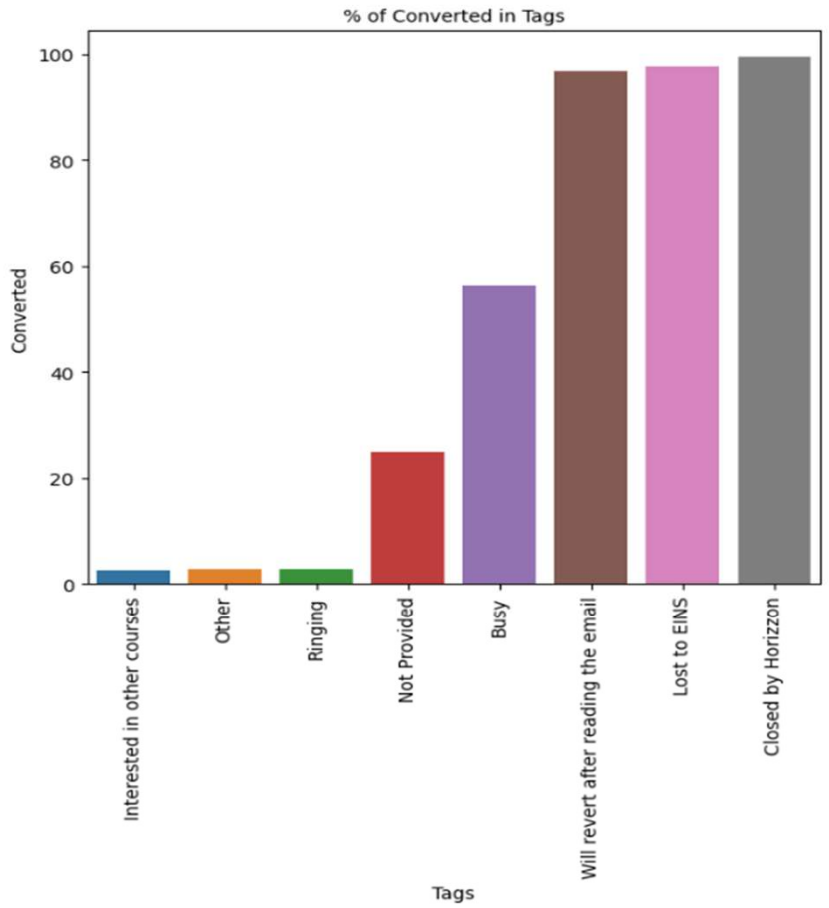
<u>Analysis based on the Graph Figure:</u>

•Conversion rate seem to be very good for Working professionals, Housewives, Businessman, Other

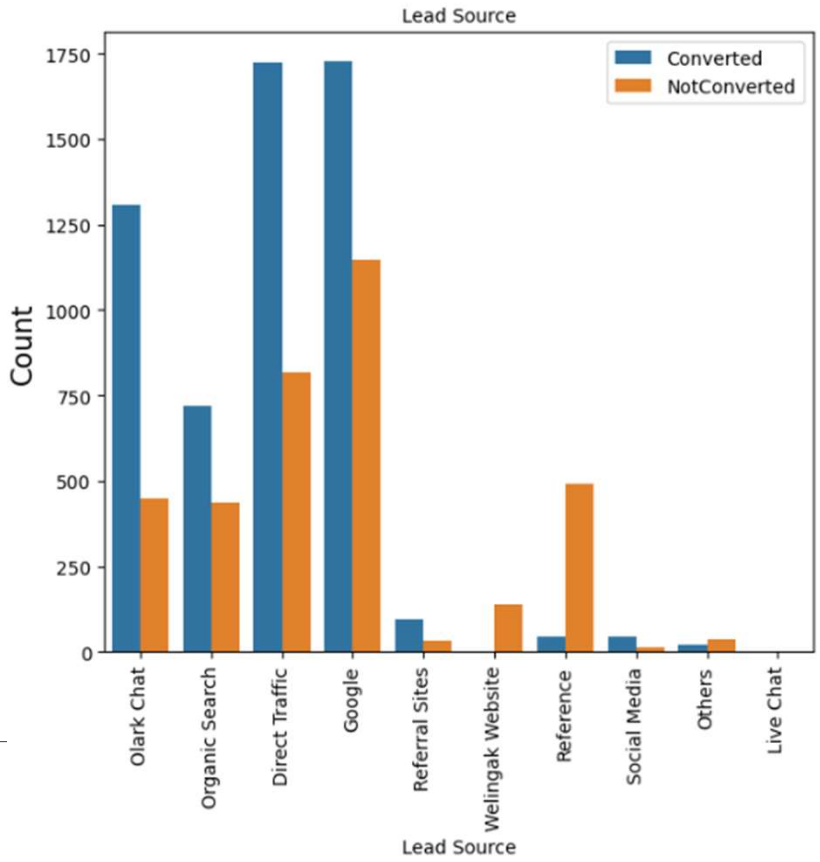
•potential candidates who have not specified their specialization do not seem to look very promising

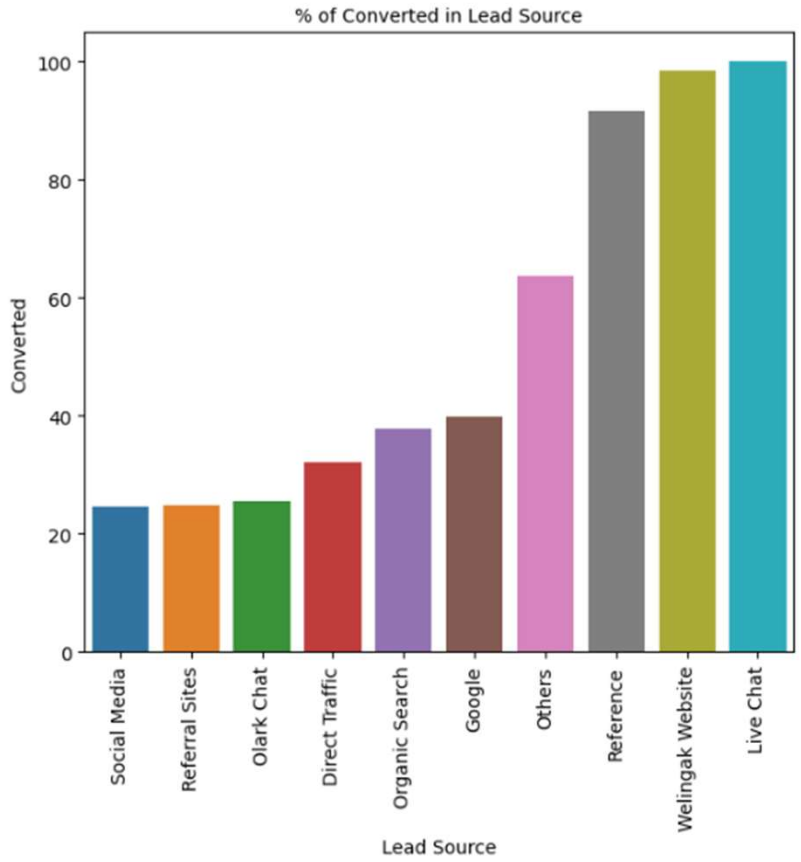
Analyzing the column - "What matters most to you in choosing a course"





Analyzing the column - Lead Source

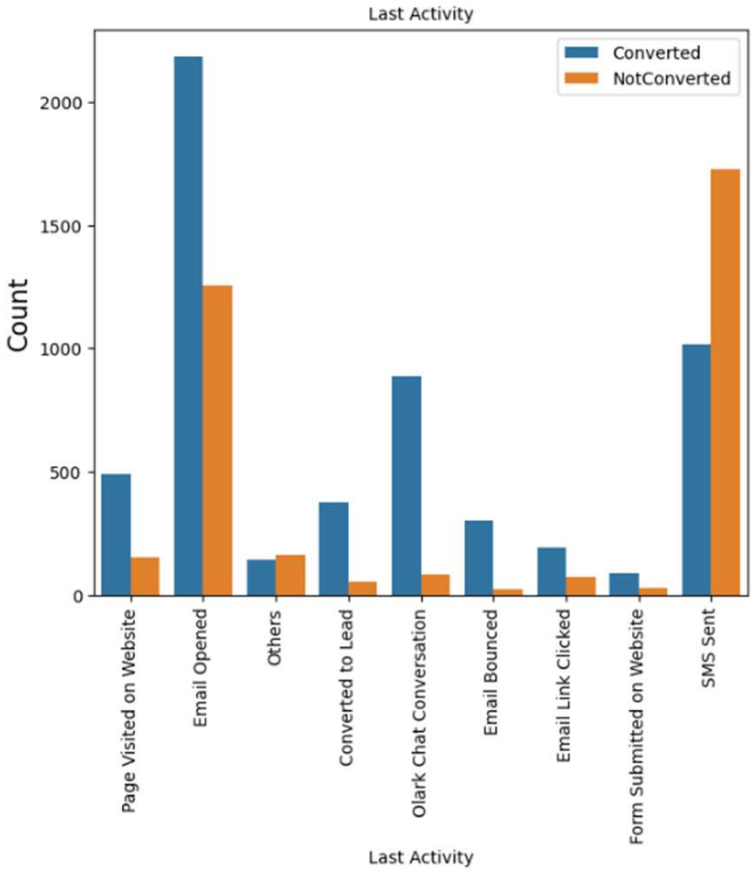


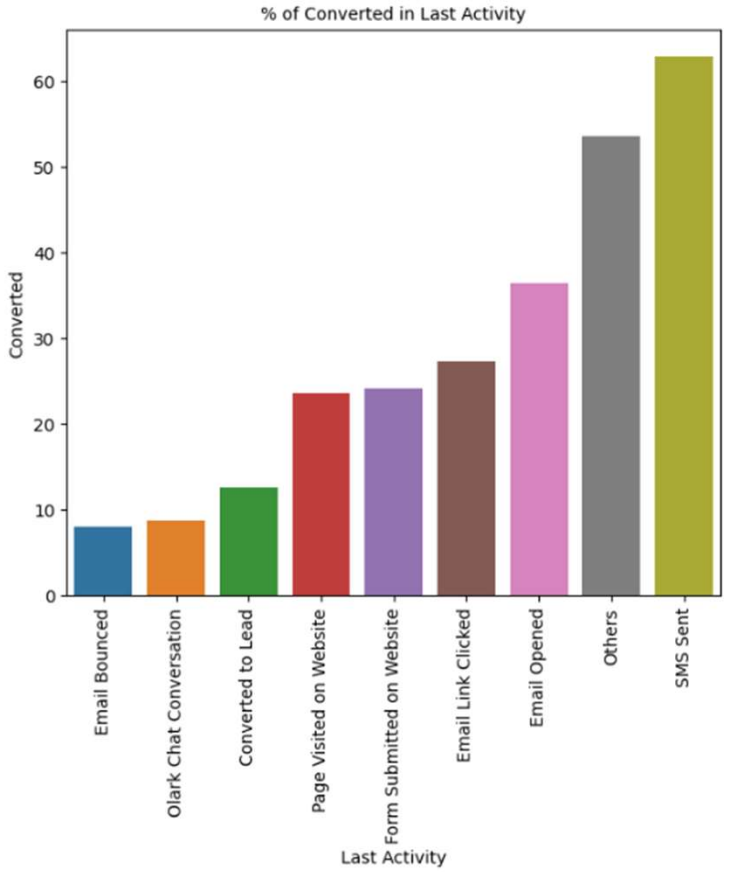


Analysis based on the Graph Figure:

- Conversion rate seem to be very good for folks that come through Reference, Google, Others, Welingak Website
- LiveChat although shows high conversion .. the number of candidates influx is very low

Analyzing the column - Last Activity



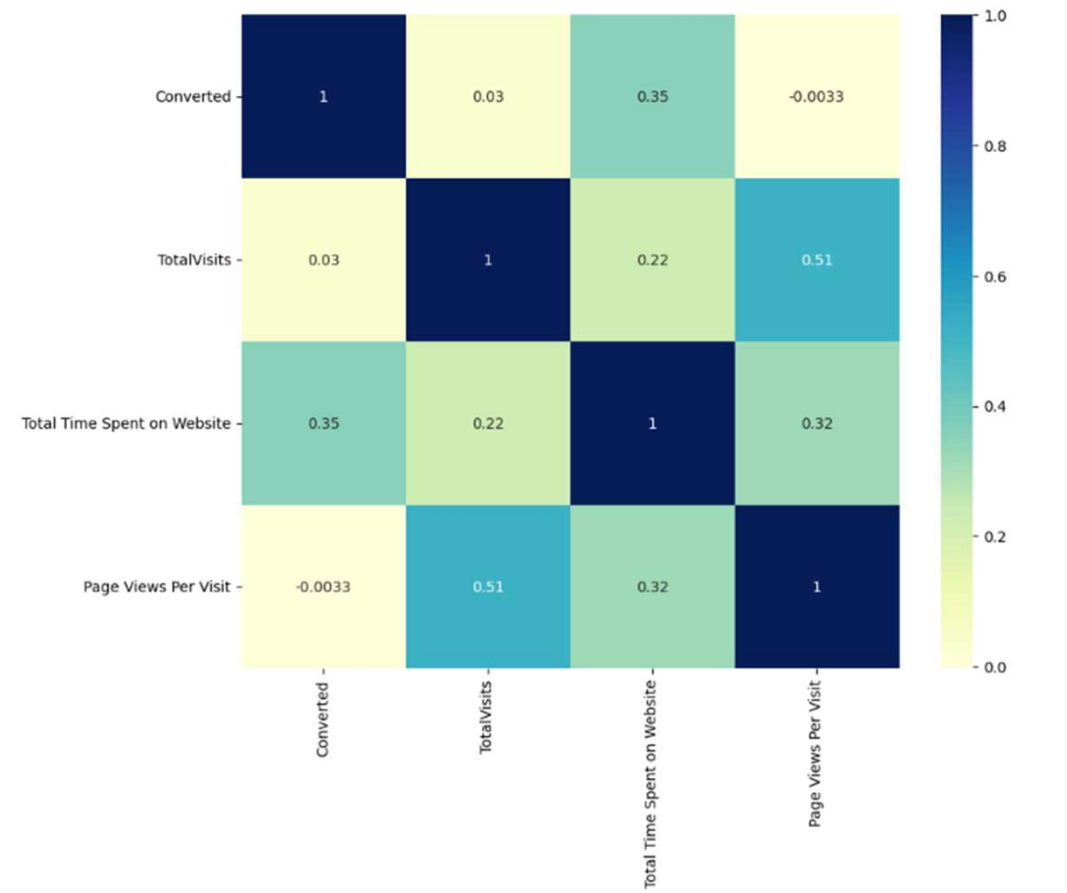


Analysis based on the Graph Figure:

•Conversion rate seem to be very good for folks that have last activity as: SMS Sent, Others, Email Opened, Email Link Clicked

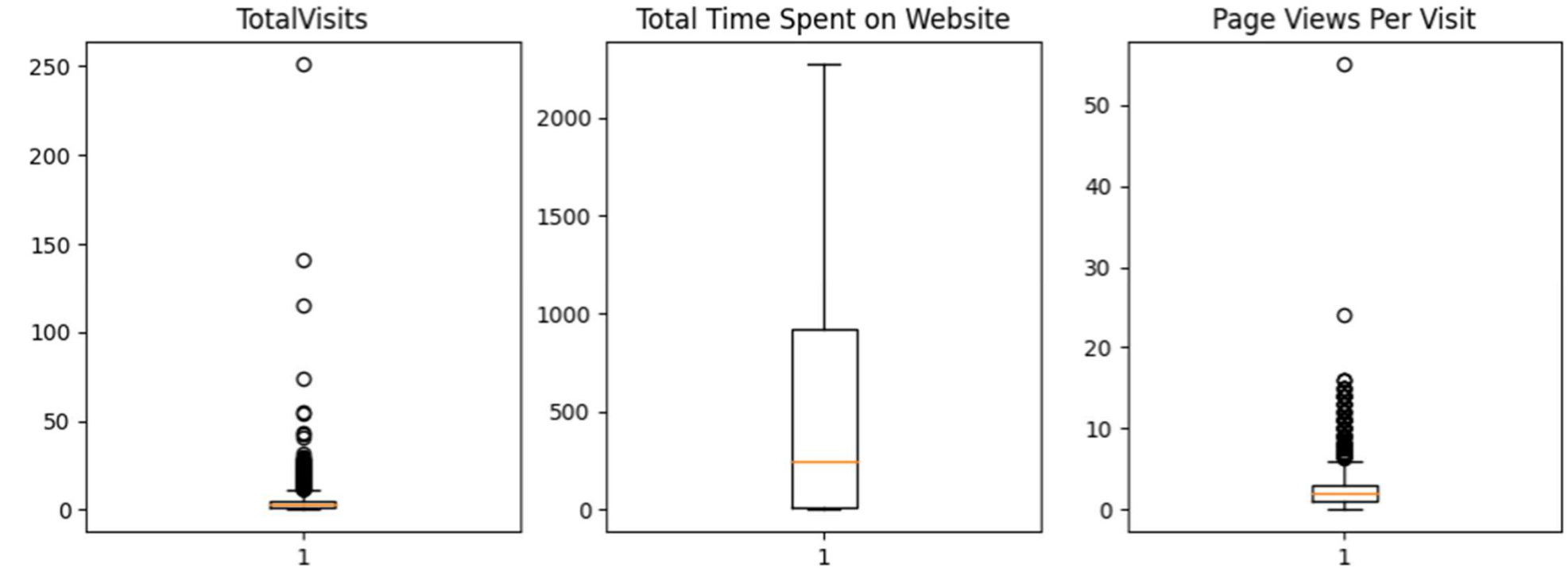
Correlation

Numerical columns analysis: correlation of the columns

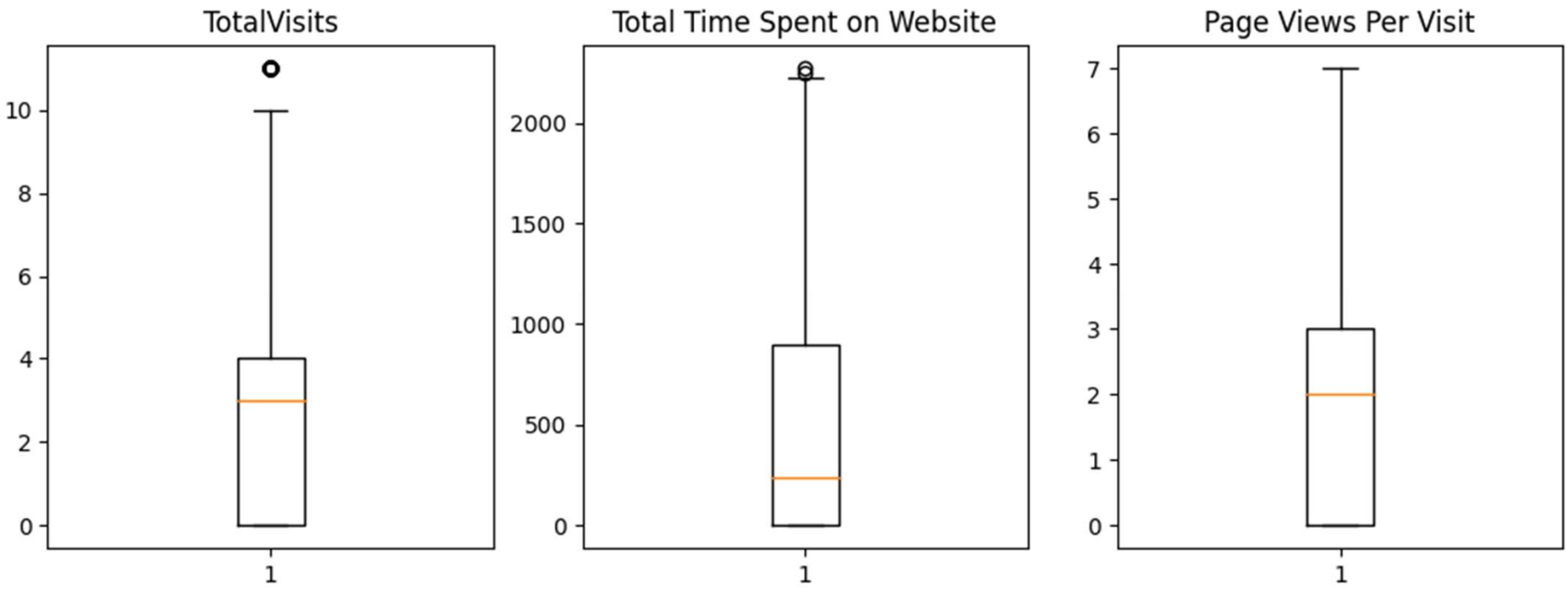


Outliers

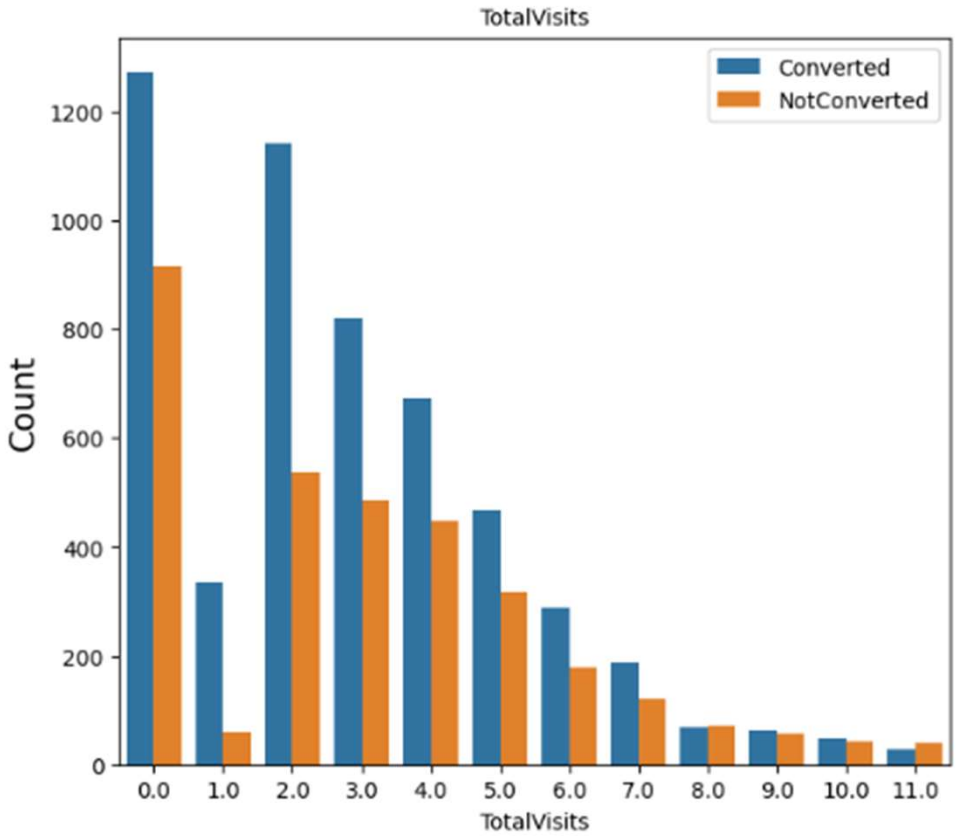
Boxplot for Outlier check

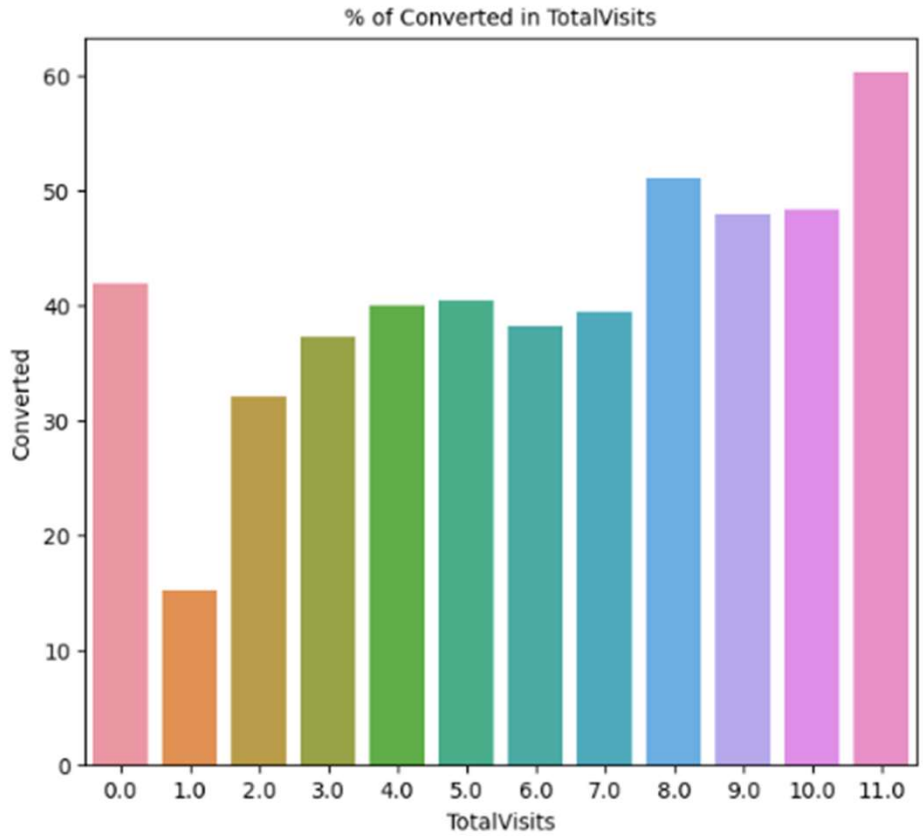


After handling the Outliers

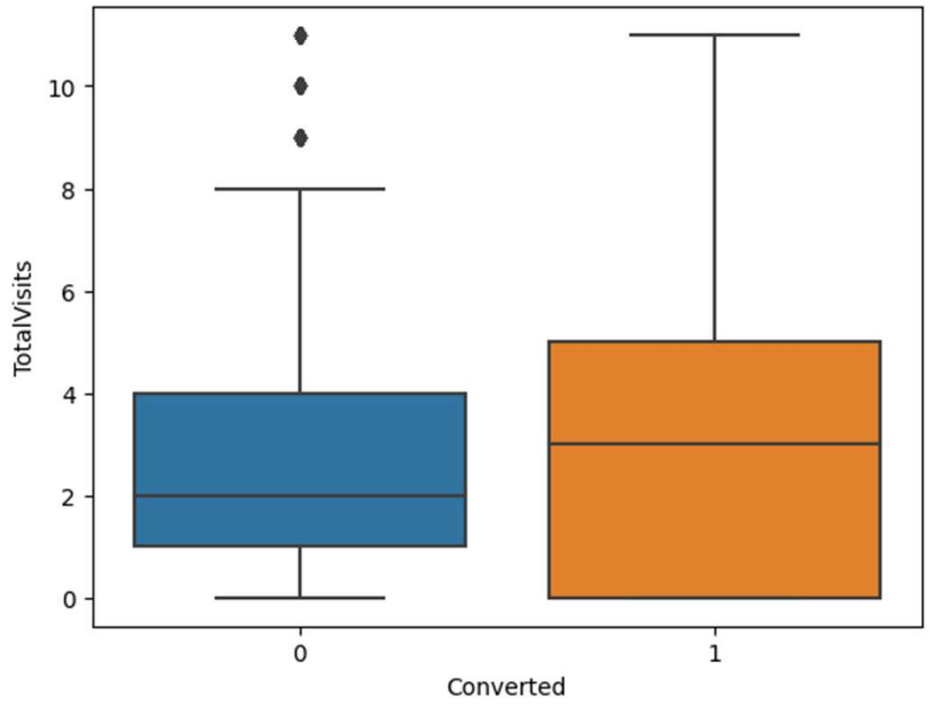


Bar Graph Analysis for Total Visits after handling the Outliers





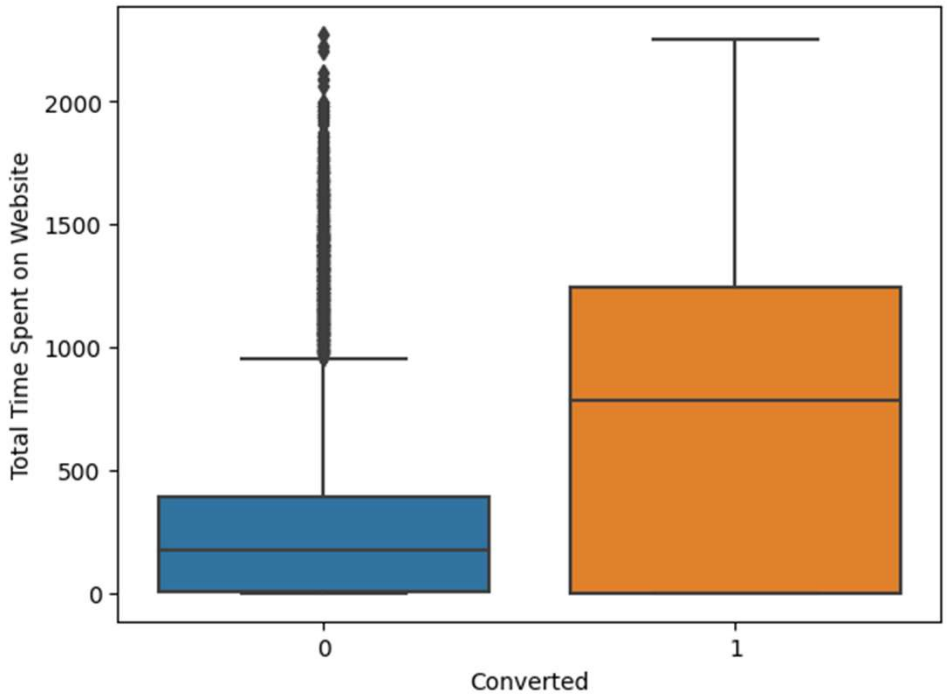
Box plot analysis for Total Visits after handling the Outliers



Analysis based on the Graph Figure:

•Total Visits do not seem to influence conversion as folks who have not visited the site at all are also getting converted and their number is high as well

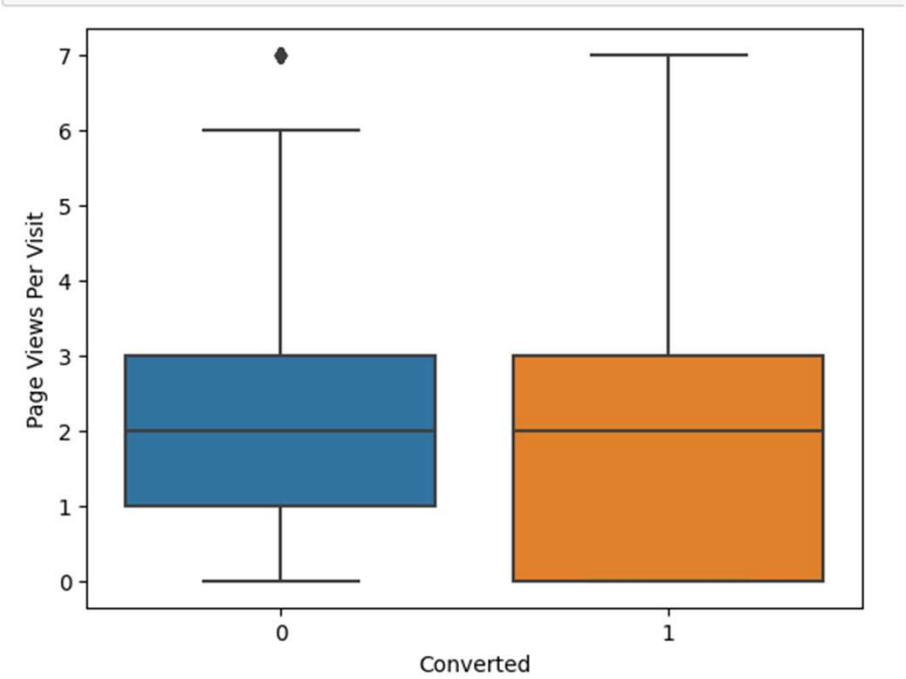
Box plot analysis for Total Time Spent on Website after handling the Outliers



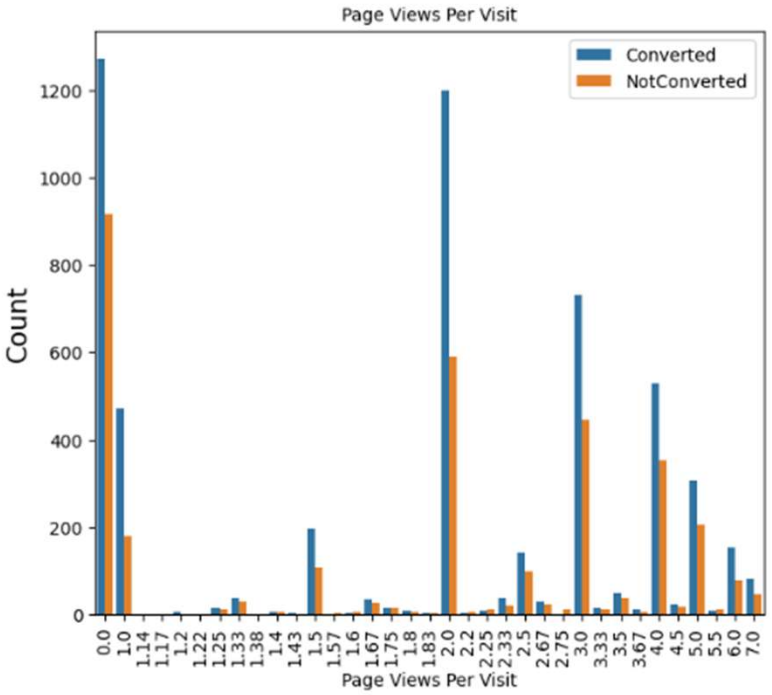
Analysis based on the Graph Figure:

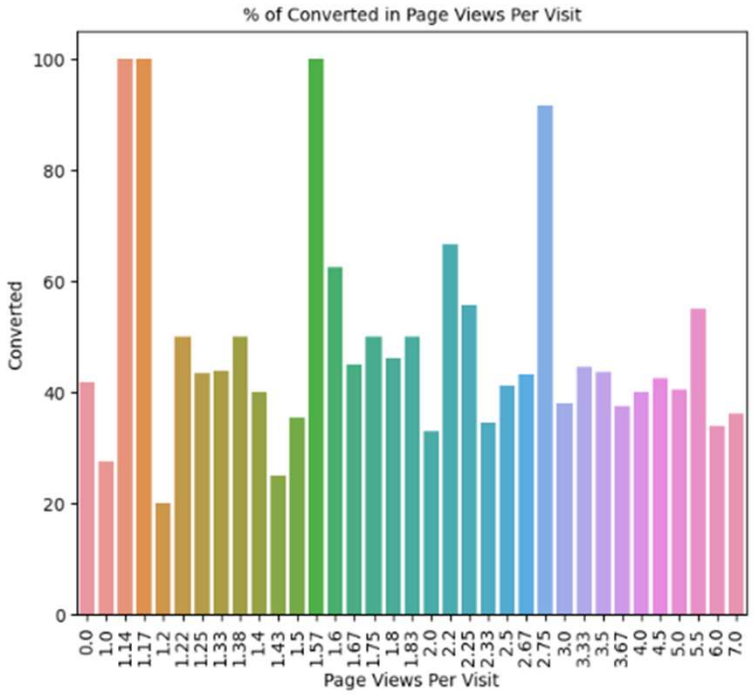
•The time spent on website has some impact on the converted ratio, as the median for converted ratio is higher

Box plot analysis for Page Views Per Visit after handling the Outliers



Bar Graph analysis for Page Views Per Visit after handling the Outliers

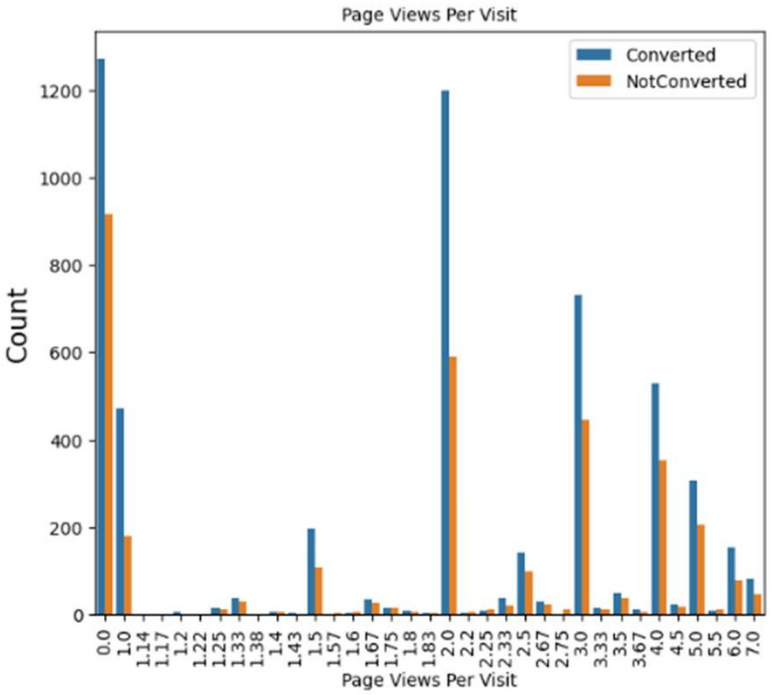


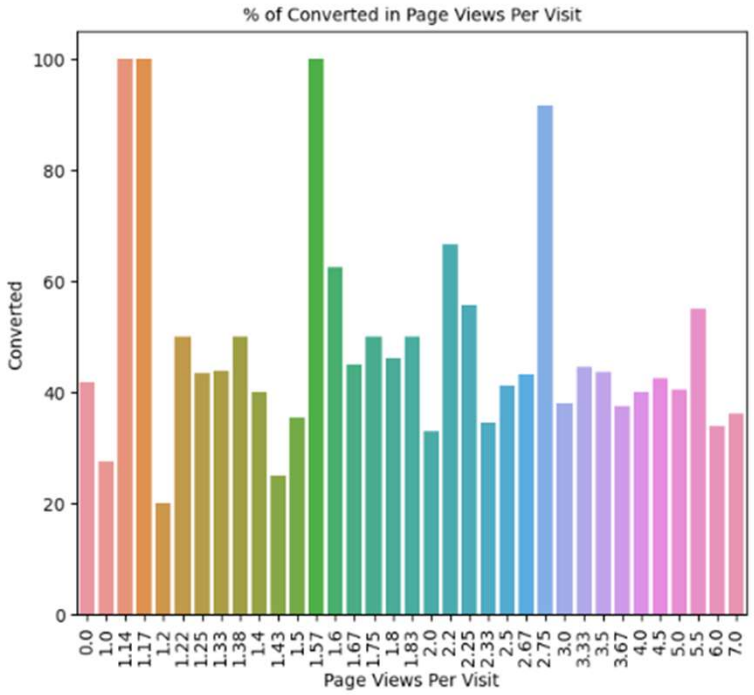


Analysis based on the Graph Figure:

- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

Box plot analysis for Page Views Per after handling the Outliers



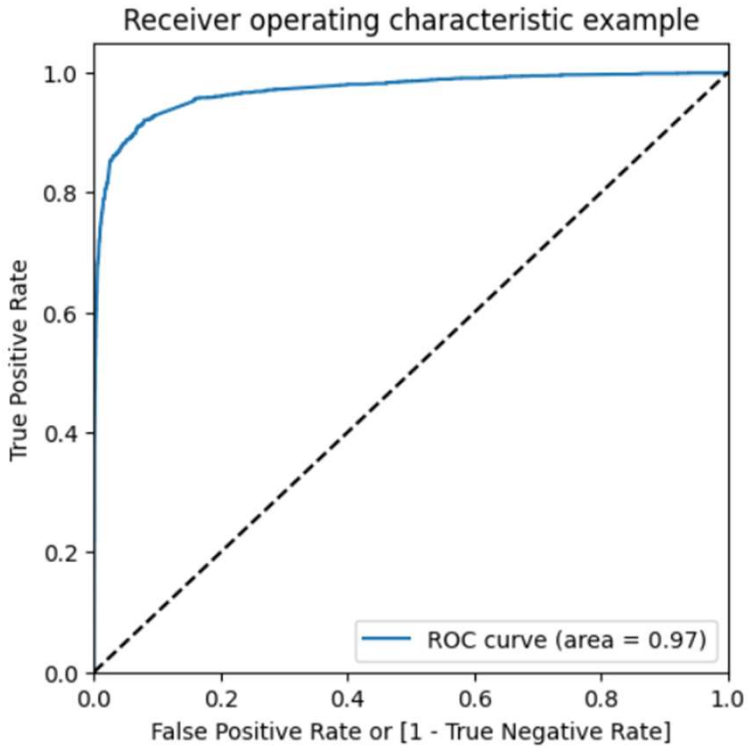


Analysis based on the Graph Figure:

- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

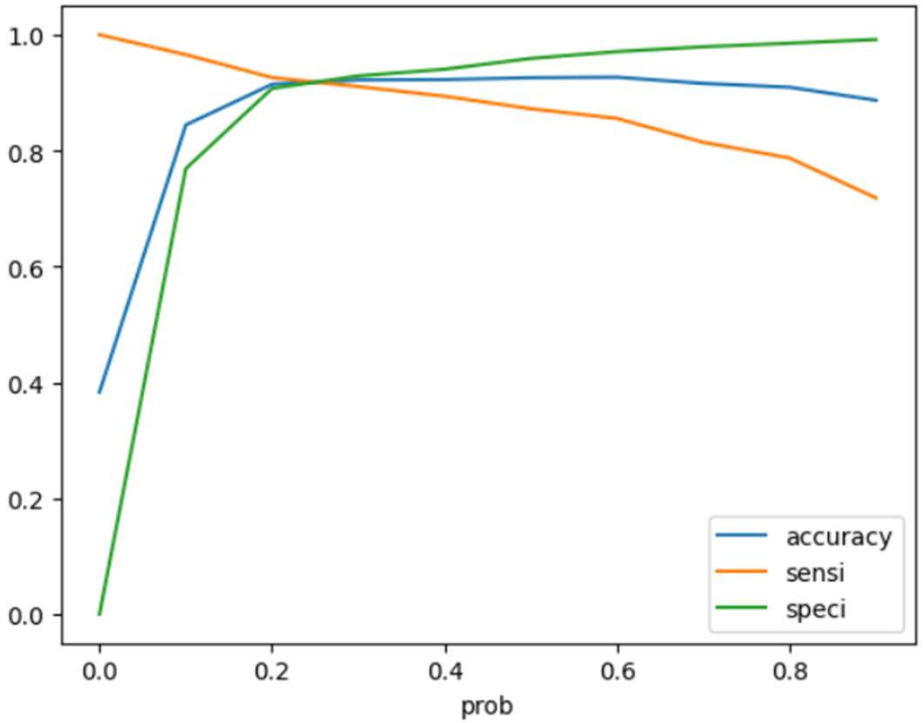
Model Building using Stats model and RFE

Plotting the ROC curve



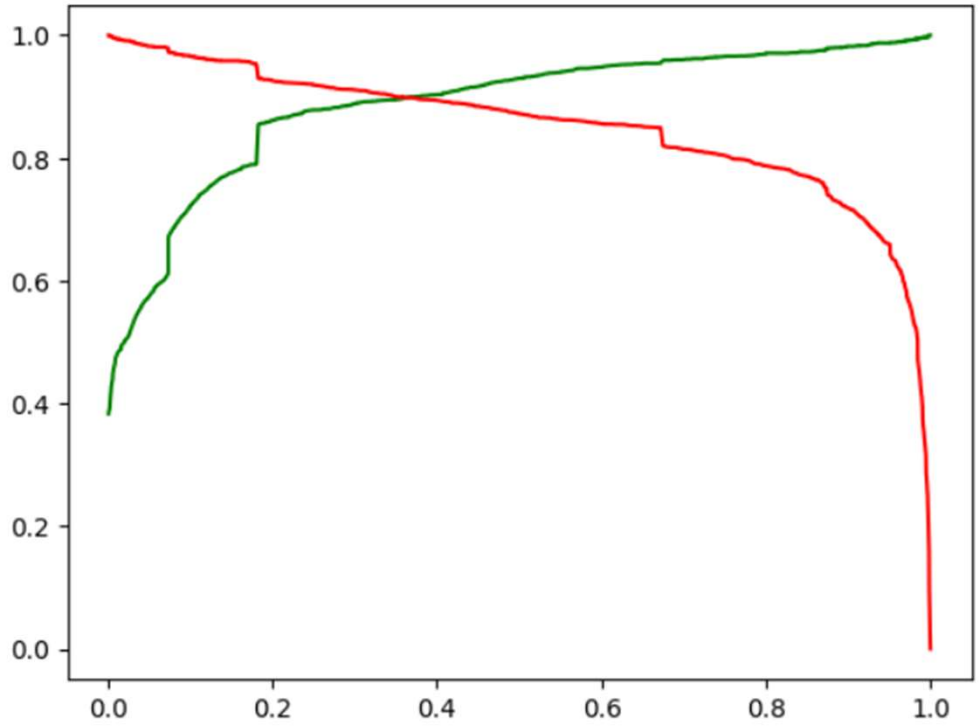
The ROC curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

Plotting accuracy, sensitivity and specificity for various probabilities



The best fit cut off for this model looks to be around 0.3 considering the best of all the metrices.

Precision Recall Curve



Conclusion

- We see that the most number of people converted is from Mumbai, but the conversion percentage is high in Thane and Outskirts.
- We see that conversion rate is high for people belonging to Asia Specific and Middle East Countries
- We see that Conversion rate seem to be very good for Working professionals, Housewives, Businessman, Other.
- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website more likely to convert.
- Model seems to well predictive as the ROC curve value is 0.96 which is very close to 1.