



# Lead Scoring Case Study

By:

Sharath Kumar T S

Vijendra Kumar

# Problem Statement

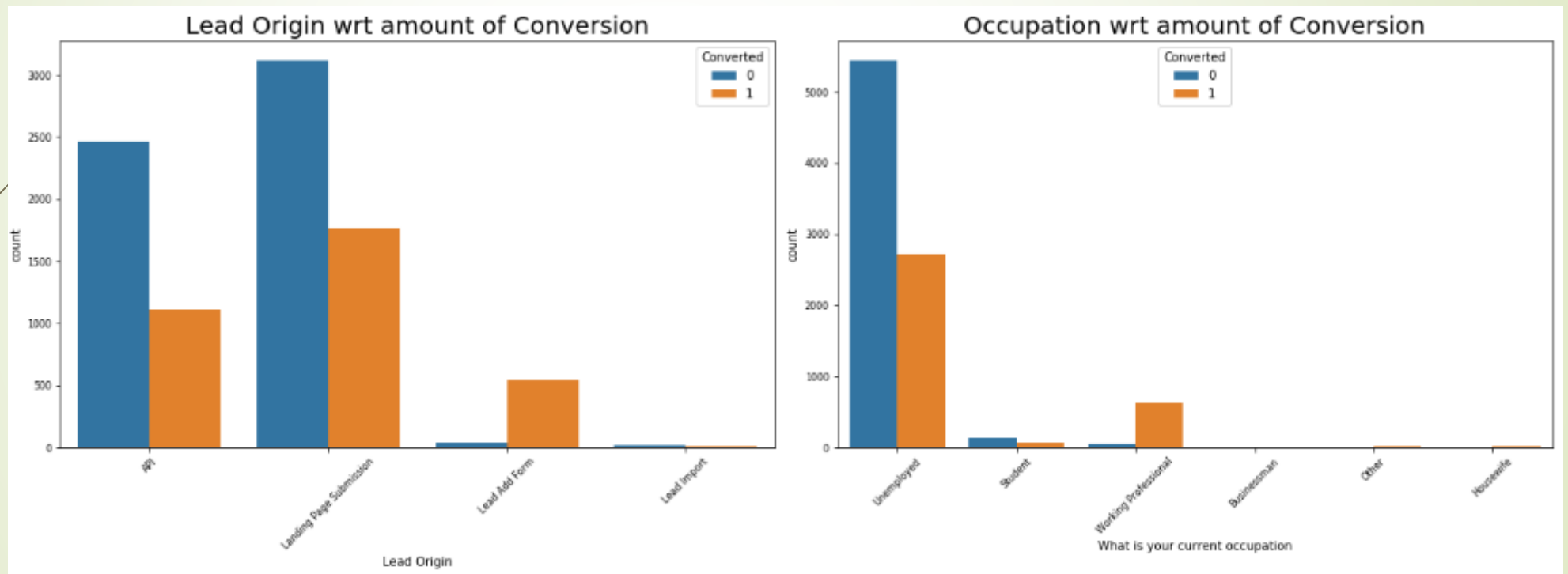
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these 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. A typical lead conversion process can be represented using the following funnel:
  - Lead Conversion Process - Demonstrated as a funnel
  - As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.
  - X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



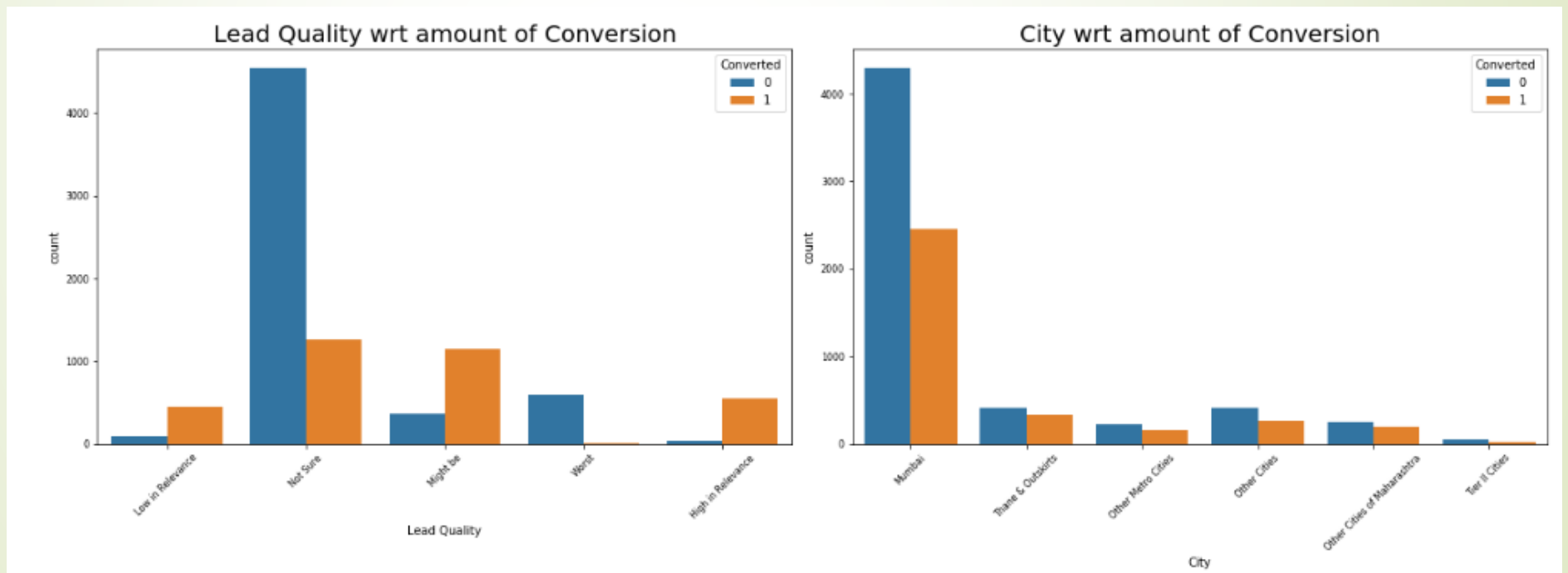
# Analysis Approach

- Here the data consists of 37 columns and 9240 rows, and this data includes both numerical and categorical data, it consists of many missing values.
- On this dataset first we check for missing values,
- This dataset consists select keyword in some of the columns which can be considered as null values, as the customer didn't select any options.
- We change all the select values to null values and then we check for percentage of null values in each column.
- We remove columns with more than 70% missing values, we perform the missing value treatment for the columns with less missing values, and we remove the data with less than 2% missing values in its columns.
- After that we check for outliers on the numeric columns as the outliers are there in some columns we perform outlier treatment on the dataset.
- Along with that remove some of the columns with varying values and whose data doesn't effect the data much.

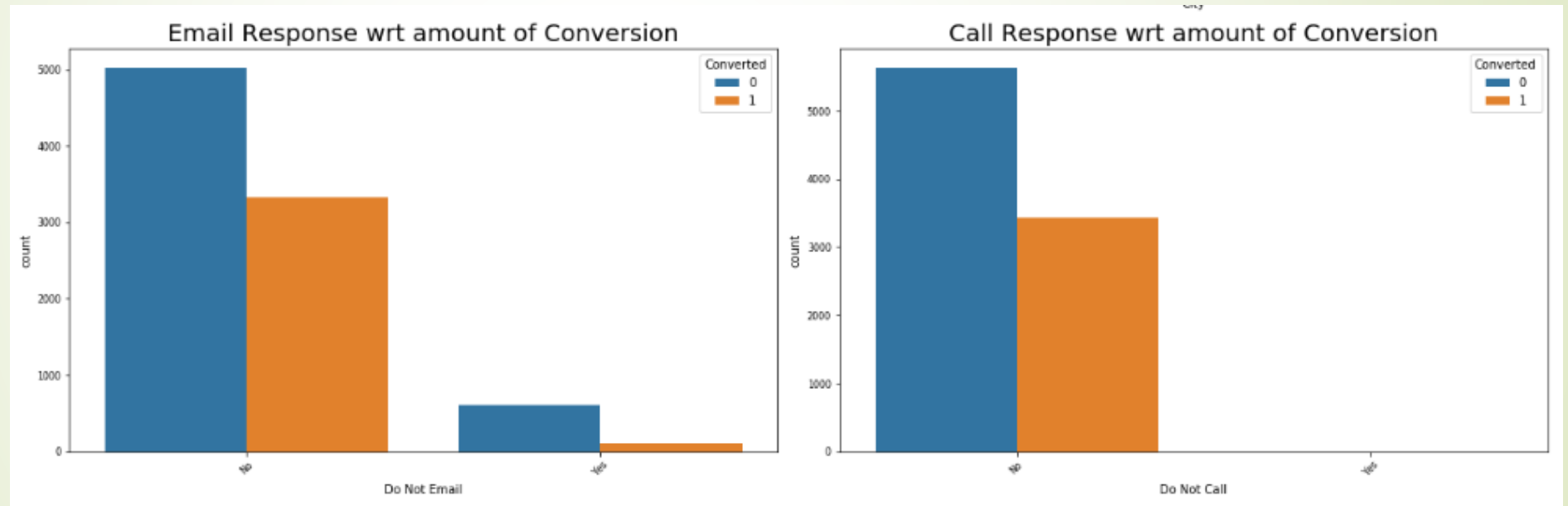
- Plotting the amount of conversion for the Lead Origin and the Occupation column.



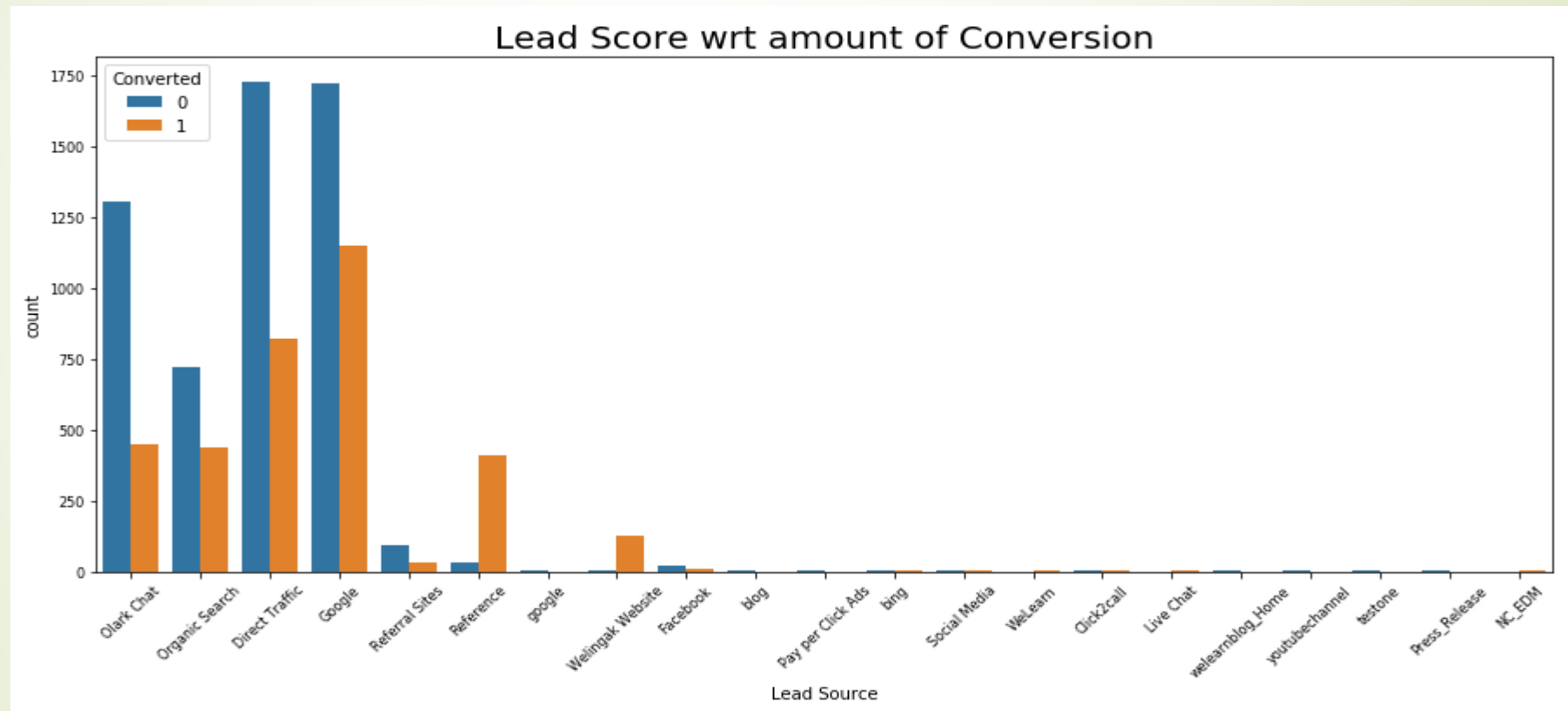
- Plotting the amount of conversion for the Lead Quality and the City column.



- Plotting the amount of conversion for the 'Do not Email' and the 'Do not call' columns.

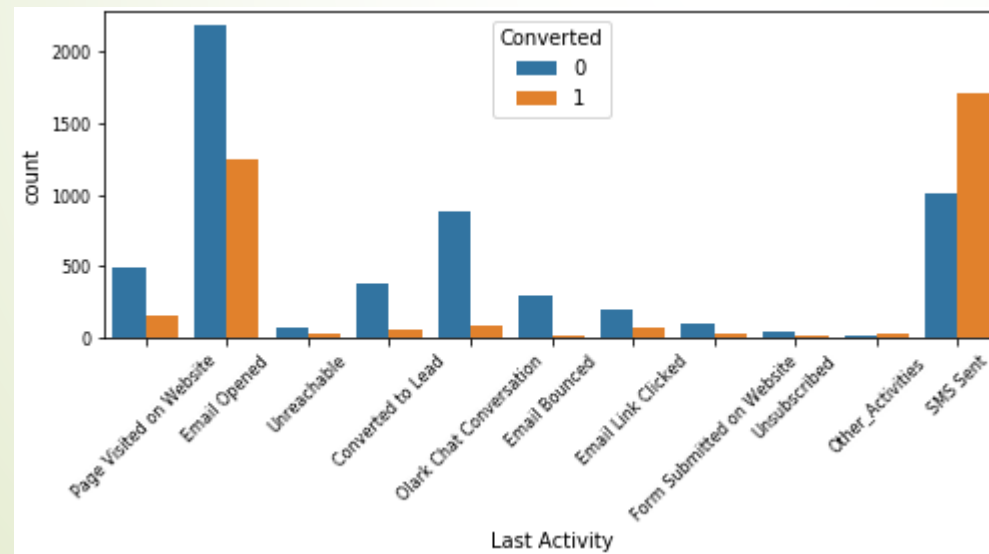
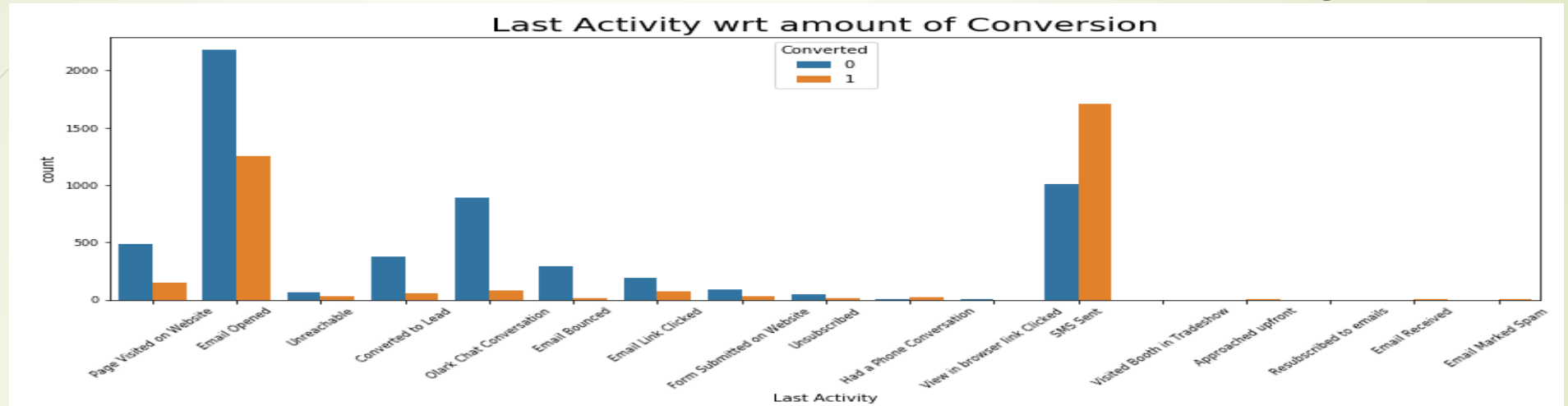


- Plotting the amount of conversion for the 'Lead Source' column.



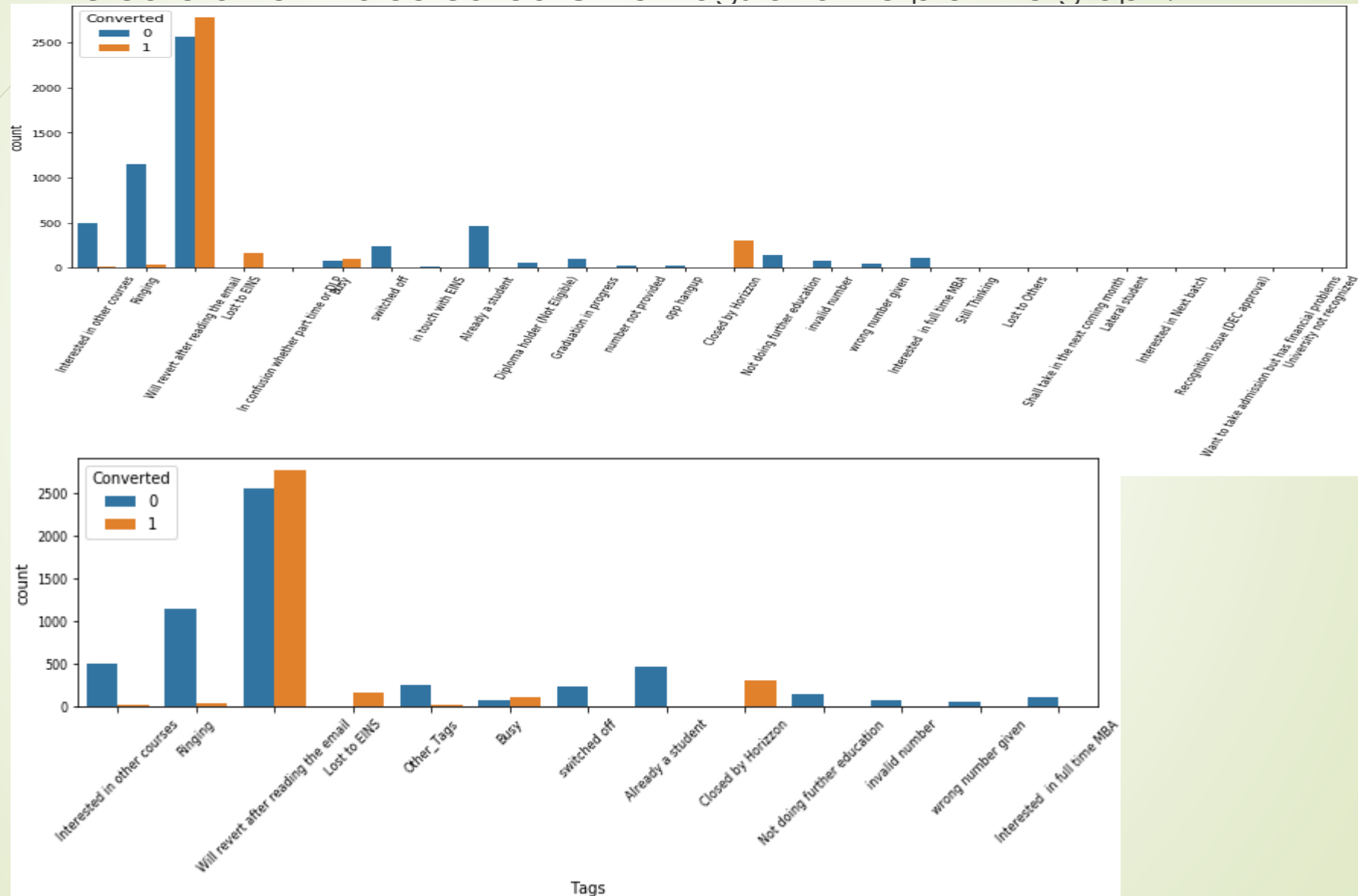


- Plotting the amount of conversion for the 'Last activity' column, as we can see many values in the column have very less count we will merge these and create a new value called Other Activities and we plot the graph.

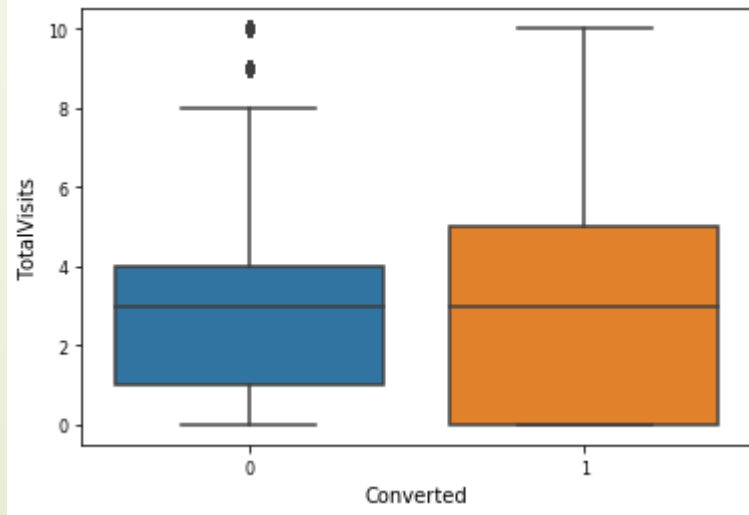
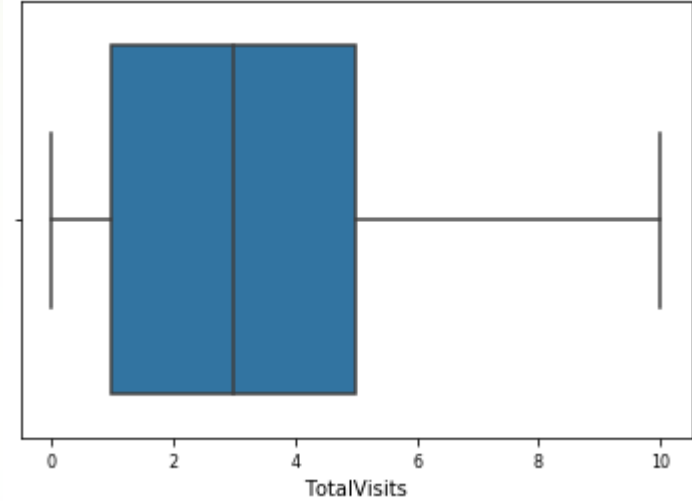
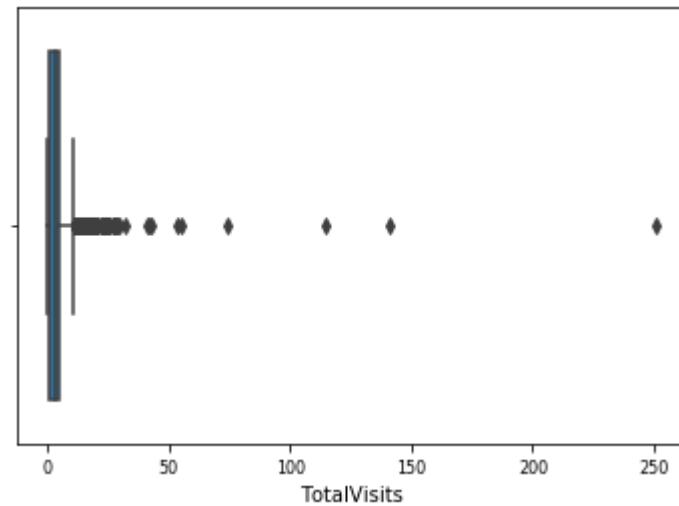




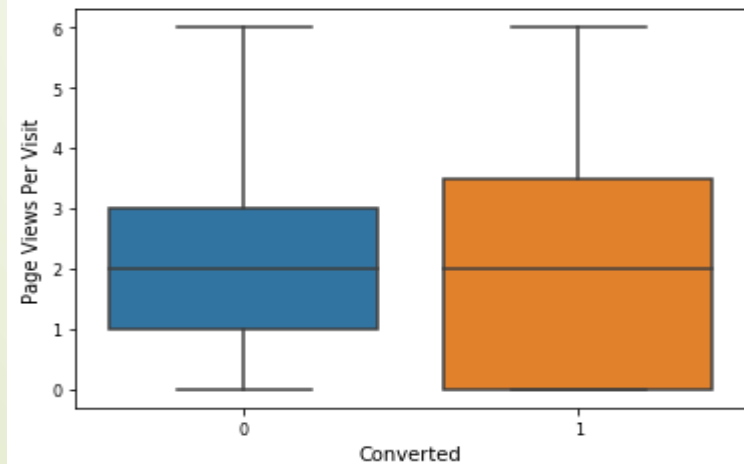
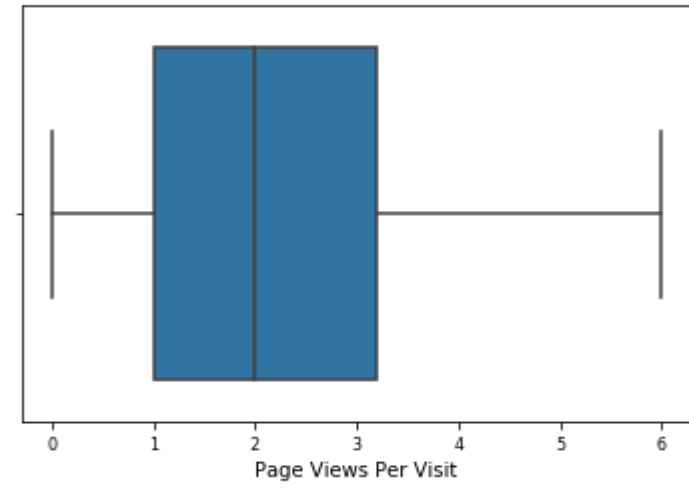
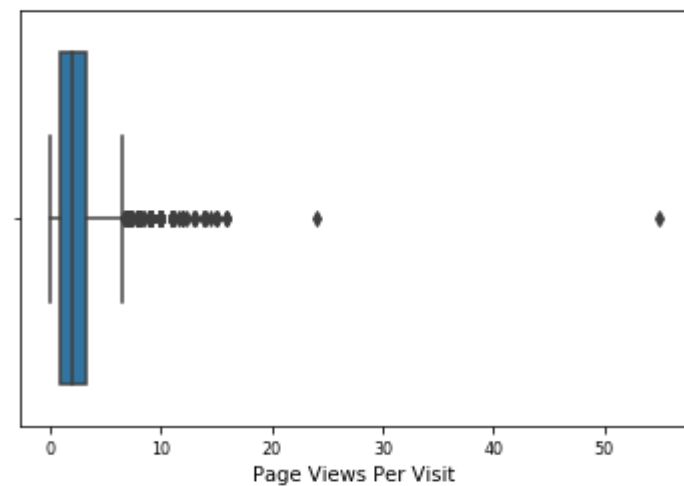
- Plotting the amount of conversion for the 'Tags' column, as we can see many values in the column have very less count we will merge these and create a new value called Other Tags and we plot the graph.


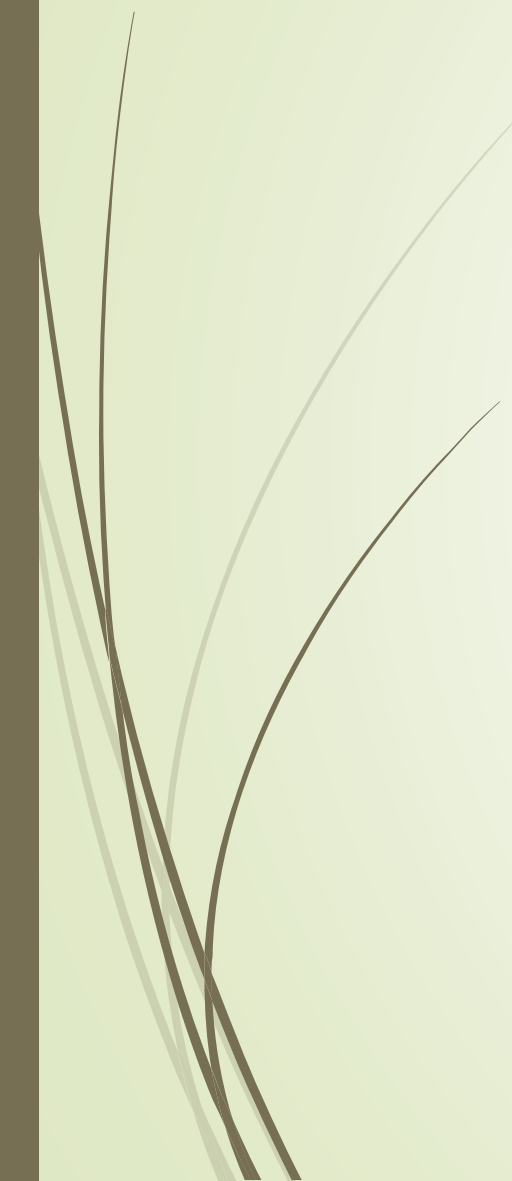


- Checking the outlier for the total visits column by plotting the box plot,
- After performing the outlier treatment we plot the boxplot again, and we also plot total visits wrt amount of conversion on a count plot.



- Checking the 'page views per visit' column for the outlier by plotting the box plot, and we can see outliers are there in this column.
- After performing the outlier treatment we again plot the box plot, along with that we also plot a count plot for the 'page views per visit' wrt conversion.



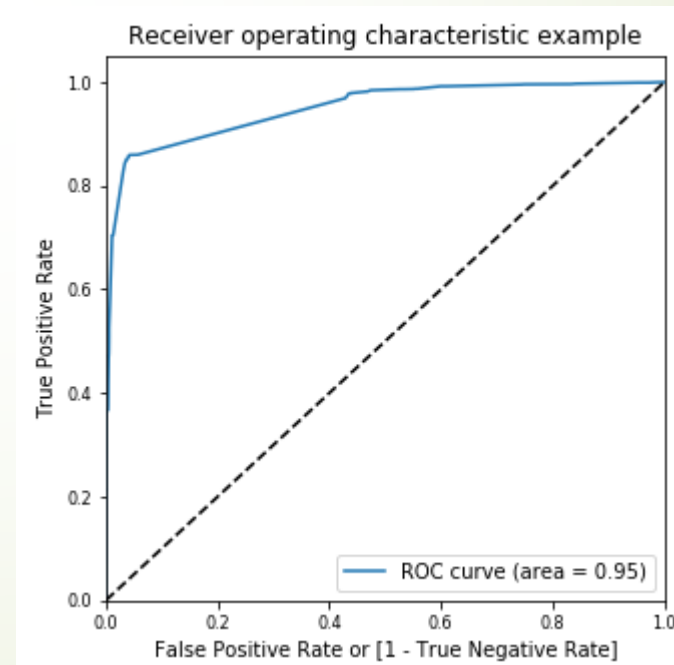
- 
- 
- After performing all the above mentioned tasks for the columns with only yes or no type values we'll do binary mapping with Yes as 1 and No as 0.
  - For the rest of the categorical variables we'll perform dummy variable creation, and then we'll drop the 1<sup>st</sup> column among them.
  - Then we merge all the dummy variables with the original dataset and we remove the columns for which we had already created the dummy variables.
  - We split the data into train and test datasets and for the X\_train dataset we scale the data by using standard scaler method
  - Then we first train the data and check the summary.
  - Then we perform the RFE with 15 variables as output on the X\_train dataset.
  - We consider only the variable which comes under top list, and on that dataset we perform the training and check the summary.
  - After removing the values with more p value, we perform the VFE on the data and check the colinearity between the variables.

- After training the data we check the predicted score, after that we'll put a threshold on the score as the one which is greater than 0.5 is considered as 1 and the rest as 0.

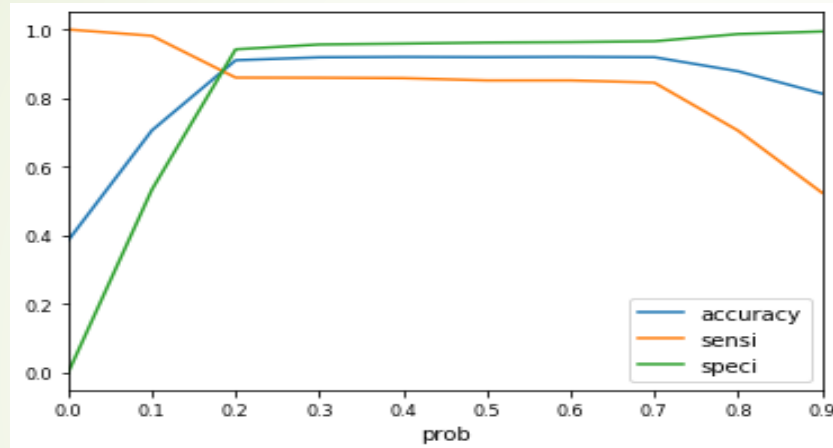
- We check for the confusion matrix.

```
[[3756 149]  
 [ 363 2083]]
```

- WRT the values in the confusion matrix we calculate the accuracy, sensitivity, specificity, false positive rate(FPR), Negative predicted value and Positive predicted value.
- We plot the ROC graph.



- Now we change the threshold from 0.1 to 0.9 and depending on the value we calculate the accuracy, sensitivity and specificity for various probabilities.
- Then we plot them on a graph to find the optimum threshold.



- From the above plot we can say that 0.2 is the optimum threshold.
- We change the threshold to 0.2, hence we get a different prediction, on the obtained data we again check the sensitivity, Specificity, FPR, Precision and Recall.
- After performing the training we test the data on the test dataset and predict the value for by considering 0.2 as the threshold, after obtaining the final data we again create the confusion matrix and check the accuracy, sensitivity and specificity of the data.





# Conclusion

- Here we used the RFE to select the top 15 columns.
- We had taken 0.2 as the threshold, if we want to increase the number of customers as 1 we need to decrease the threshold, similarly if we want to increase the true positive rate we need to increase the threshold.
- Here for 0.2 as threshold we are getting 90.6% as the accuracy,
- 84.3% as the Sensitivity and 94.2% as the Specificity.
- 89.3% as the precision and 84.3 % as the Recall.
- As everything is more than 80% we can say that the objective has been achieved.