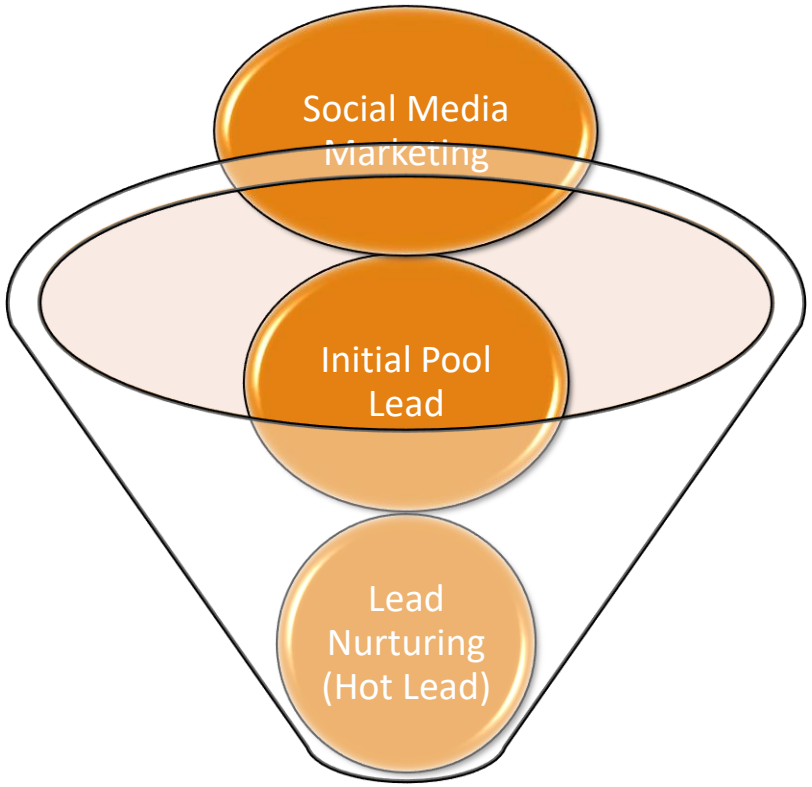
X EDUCATION CASE STUDY

IDENTIFICATION OF HOT LEADS – USING LOGISTIC REGRESSION

BY –DEEPIKA BHATT & SHREYAS R

PROBLEM STATEMENT



Converted Lead

Current conversion Rate is around 30%

Assign lead score to each lead

Find most potential leads (using assigned lead score)

APPROACH & METHODOLOGY

DESCRIPTION OF DATA

- Shape
- Info
- Data Type
- Statistical Summary of Data

DATA CLEANING

- Checking Null values
- Null value Treatment
- Dropping columns which are not useful for further analysis

EDA

- Univariate Analysis of all the variables (categorical variable- Bar plot, continuous variable- boxplot)
- Bivariate Analysis to find the correlation between variables by plotting Heat Map

DATA PREPARATION

- Outliers Treatment
- Formatting data for the model building process by converting binary variable into 0 and 1
- Creating dummy variable

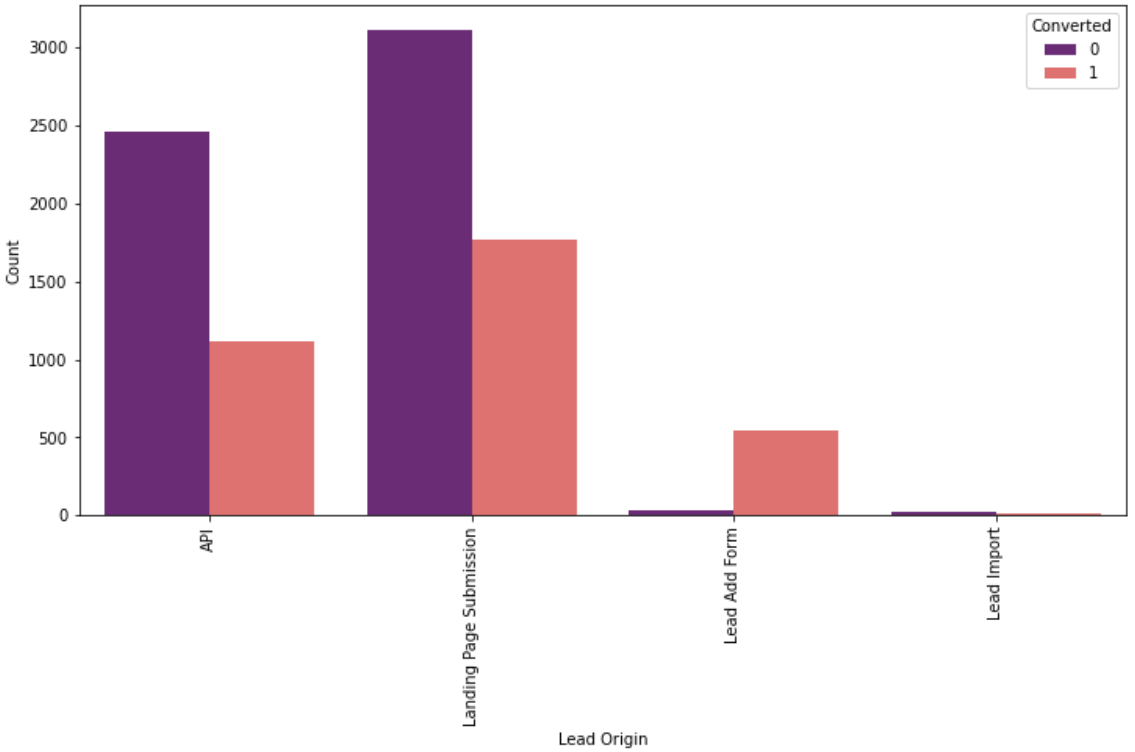
MODEL BUILDING

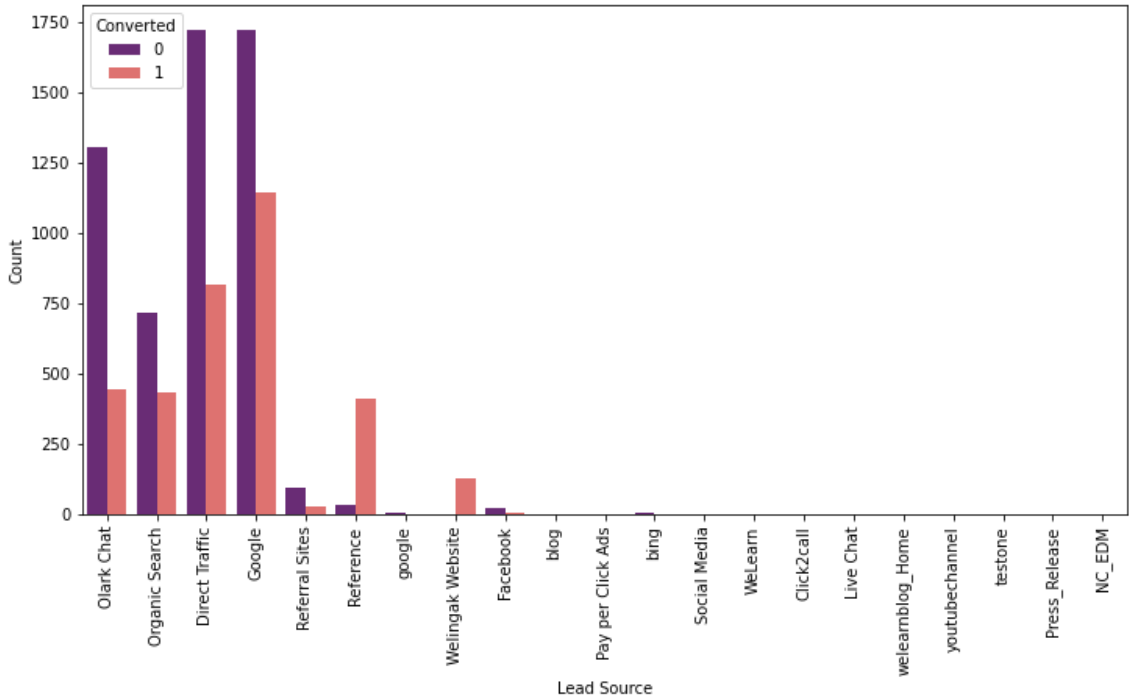
- Splitting data into train dataset and test dataset
- Feature scaling
- Coarse tunning – Automated approach of feature elimination (RFE)
- Fine tunning – Manual approach of feature selection (by observing p value and VIF)

MODEL EVALUATION & VALIDATION

- Accuracy
- Sensitivity
- Specificity
- Precision & Recall
- Validating the model using Test data set

UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE



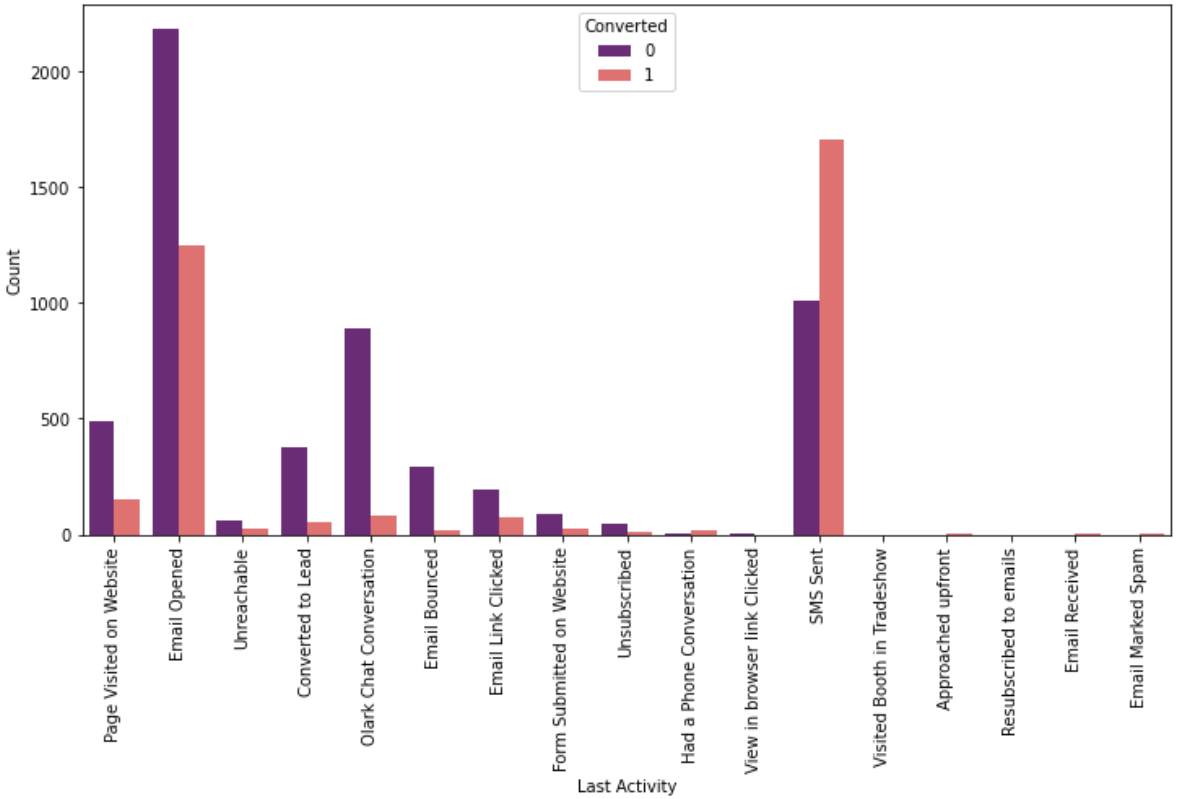


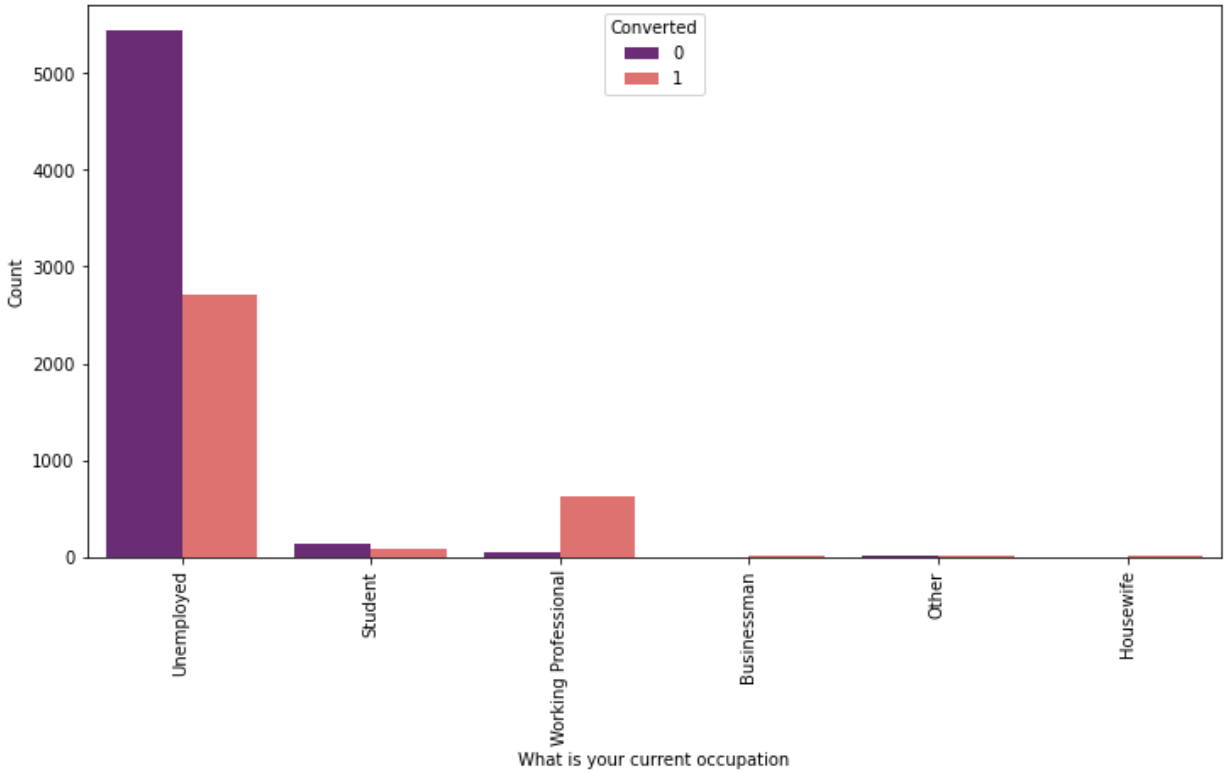
•The count of lead originating from 'API' and 'Landing Page Submission' are very significant and have around 30% conversion rate.

•'Lead Add Form' has a very high conversion rate(around 90%) but the count of lead is very low.

•The lead source is mostly from 'Google' followed by 'Direct traffic' but their conversion rate is pretty low. -'Reference' and 'welingak website' have a very high conversion rate but the number of leads generated is very low.

UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE

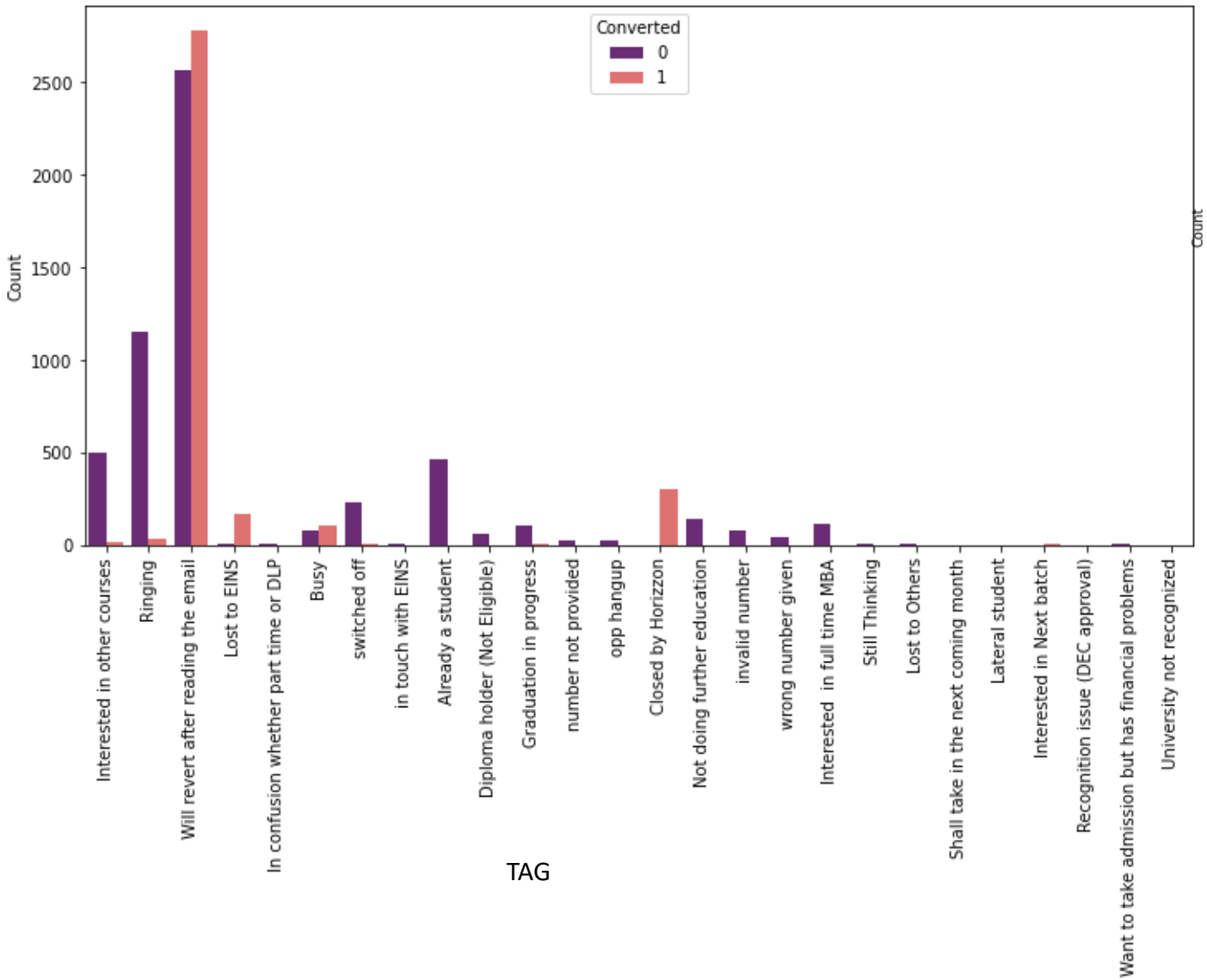


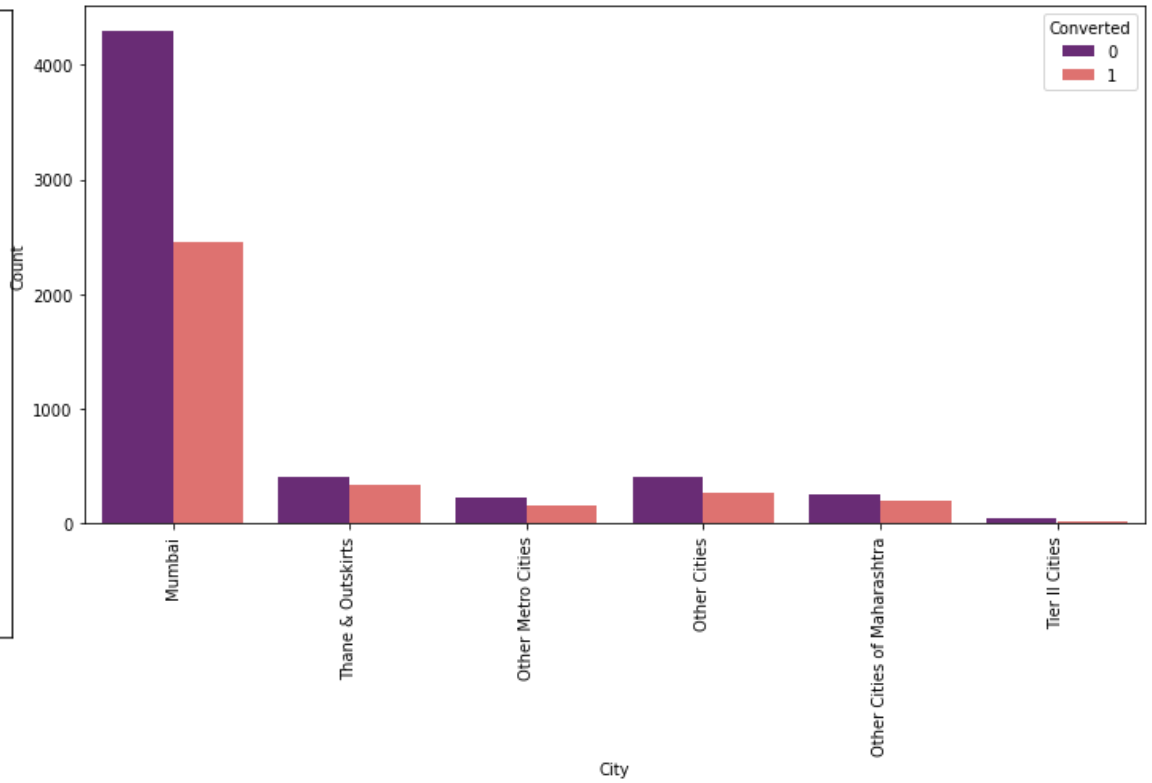


•Most of the leads last activity is 'Email Opened' followed by 'SMS Sent'. The conversion rate is pretty good for SMS sent category.

•Most of the leads are unemployed and they do not have a good conversion rate. Working professionals have a very good conversion rate but the number of leads is pretty low.

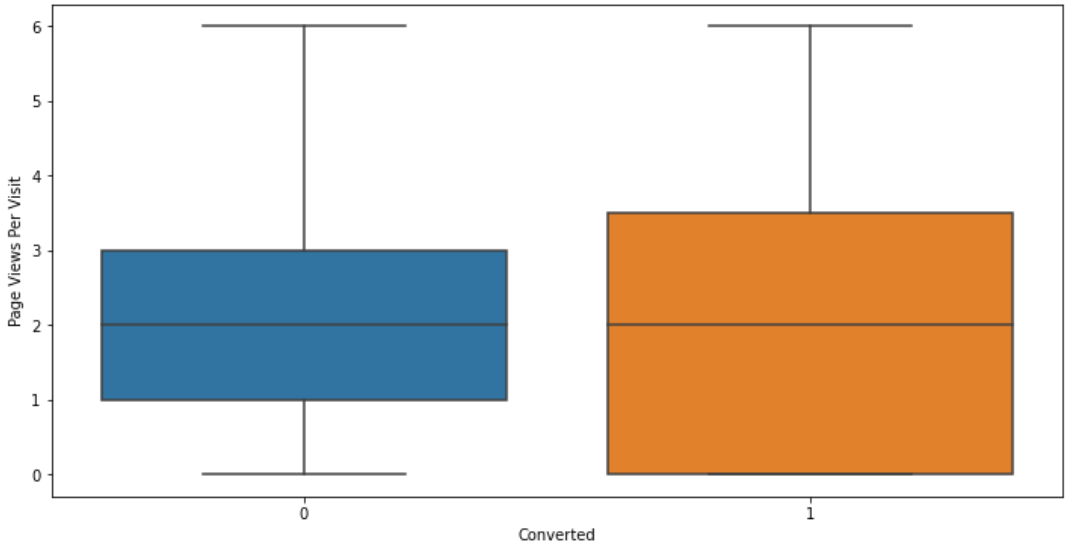
UNIVARIATE ANALYSIS – CATEGORICAL VARIABLE

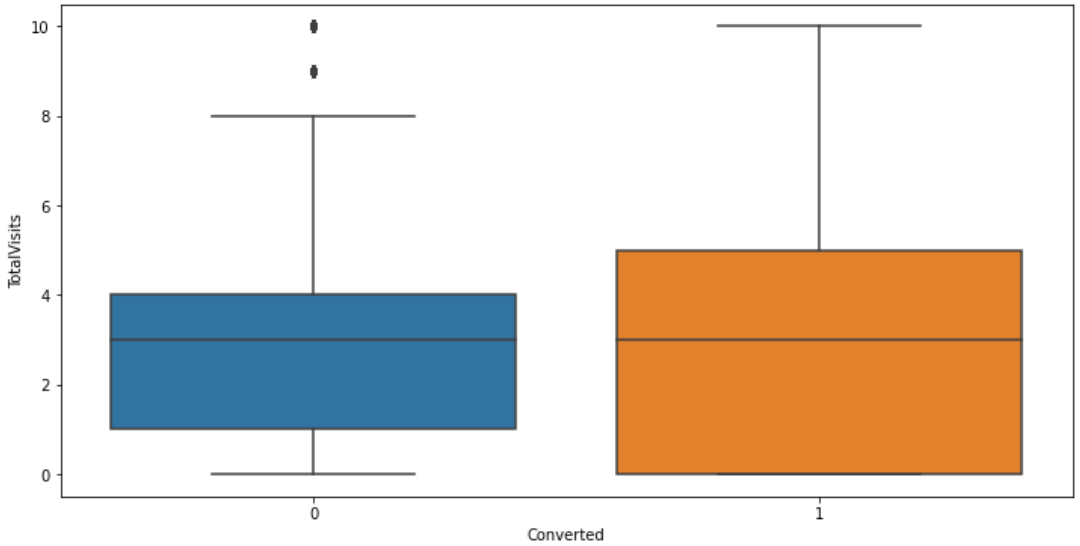


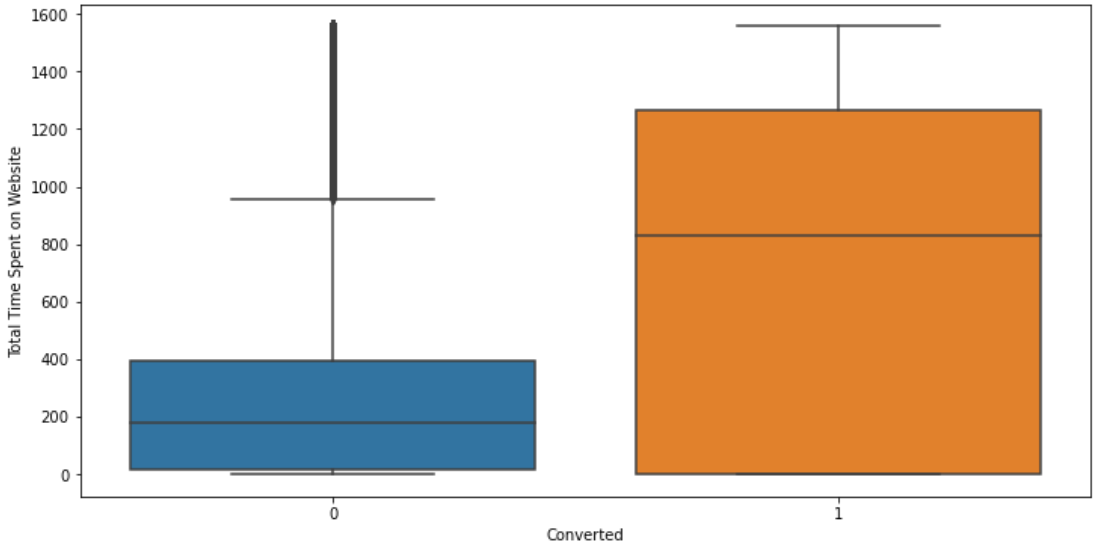


- Leads who are tagged as 'Will revert after reading email' have a good conversion rate and the number of leads are high as well.
- It is clearly evident that most of the leads are from Mumbai and they have around 30% conversion rate.

UNIVARIATE ANALYSIS – CONTINIOUS VARIABLE

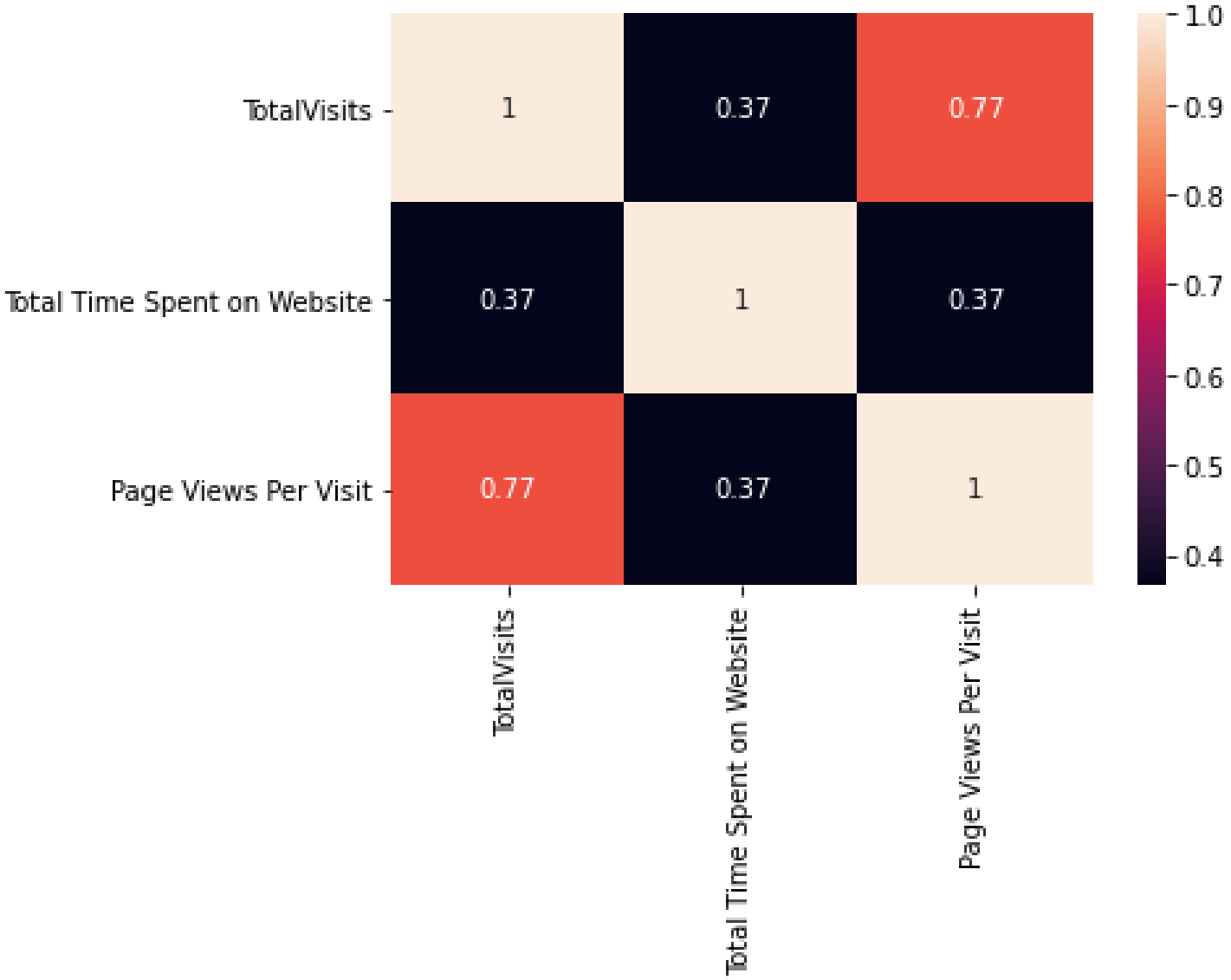






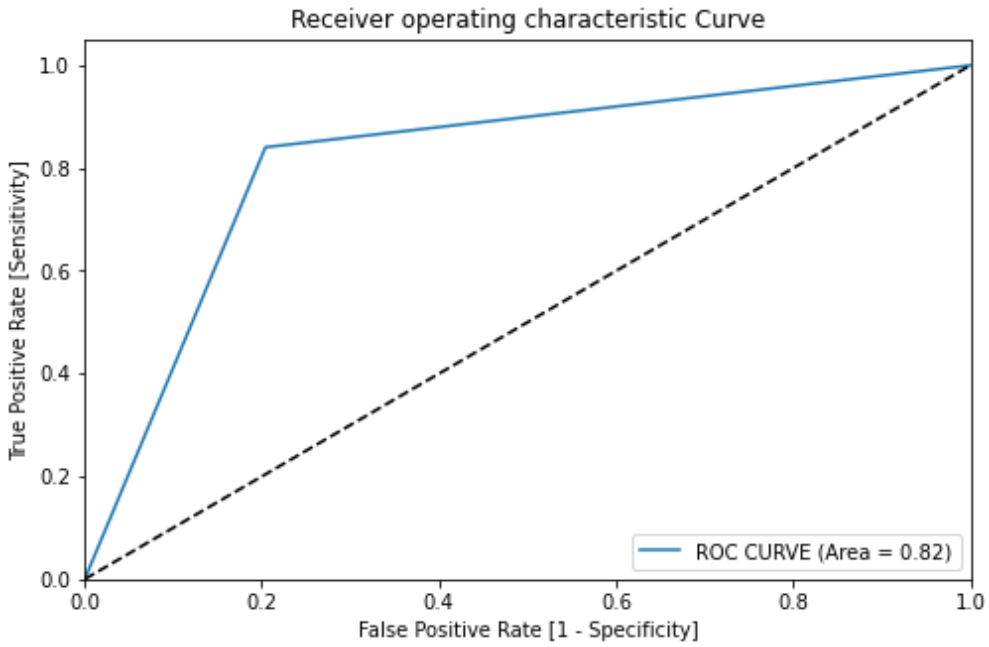
- The median value of total visits is similar for both converted and non converted leads
- The leads who are converted spent more time on the website as we can see a big increase in the median and IQR for the converted leads

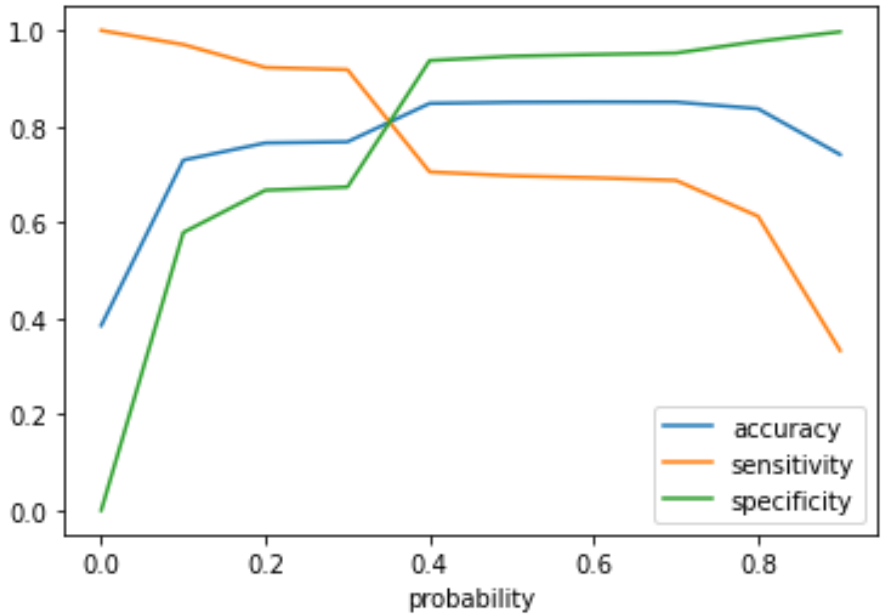
BIVARIATE ANALYSIS – CONTINIOUS VARIABLE

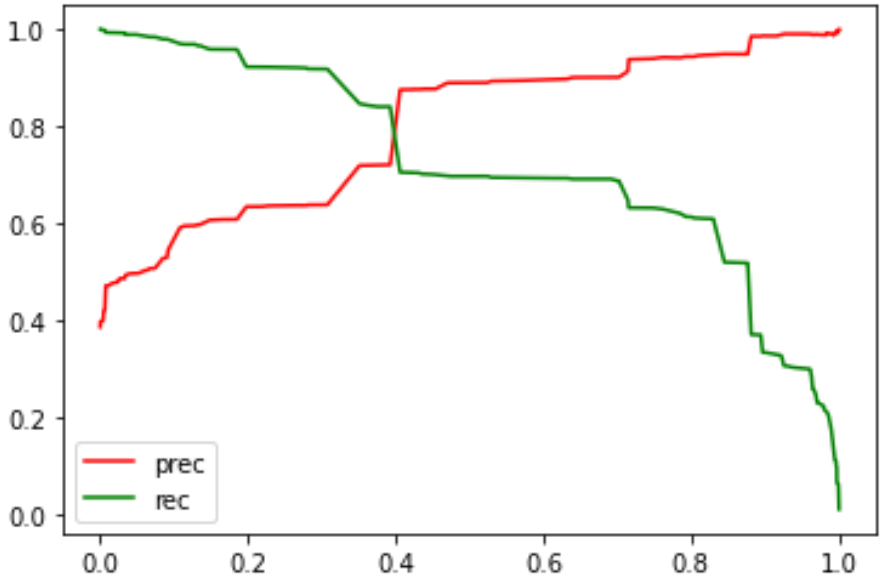


From Heat map we can observe that variable "Total visits" is highly positively correlated with "page Views Per Visit" with correlation coefficient of 0.77.

FINAL MODEL PARAMETERS







Logistic Regression Final Model Parameters

- ROC Curve area = 0.82
- With the help of accuracy , sensitivity and specificity we have finalized the cut off point as 0.385
- We can observe from precision and recall curve also that we are getting cut off point around 0.385

MODEL EVALUATION

Model Evaluation	Train Data	Test Data
Accuracy	0.81	0.81
Sensitivity	0.84	0.81
Specificity	0.79	0.80
Precision	0.72	0.70
Recall	0.84	0.81

- Our Logistic Regression Model is accurate and decent enough with high sensitivity as required. Since CEO in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- The values in the table shows that model is performing well on the test data set and is not over-trained.

CONCLUSION

The Education company needs to focus on the following factors in order to improve the conversion rate of leads.

- The leads who are tagged as 'Closed By Horizon' have a very high conversion rate.
- The leads who are tagged as 'Will revert after reading email' have a good conversion rate and the number of customers tagged to this are very high and therefore these leads need to be focused more.
- The leads with lead source 'Welingak website' have a high conversion rate.
- The leads with Last notable activity 'SMS sent' have a high conversion rate.
- The total time spent on the Website has an impact on the conversion rate.
- Leads who are working professionals have a high conversion rate.

THANK YOU