LEAD CASE STUDY

PROBLEM STATEMENT

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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%.

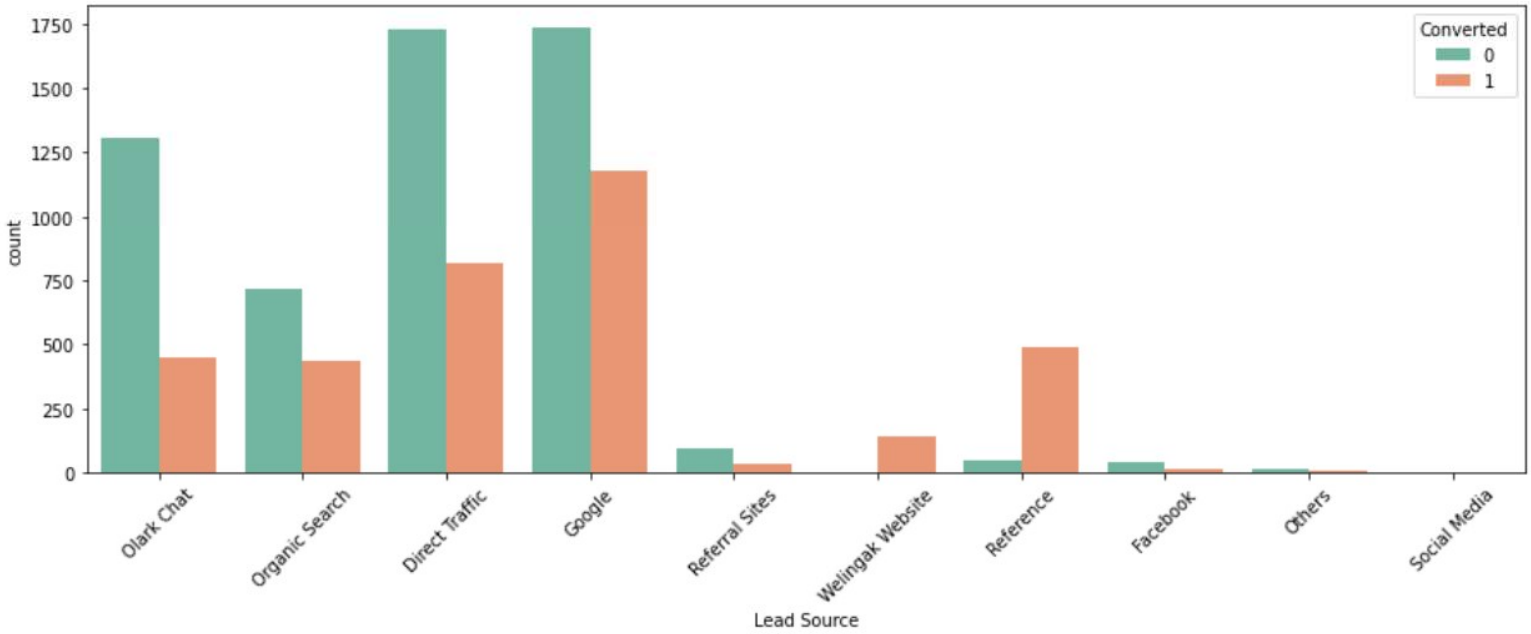
GOALS OF CASE STUDY

- 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.

How the analysis is done?

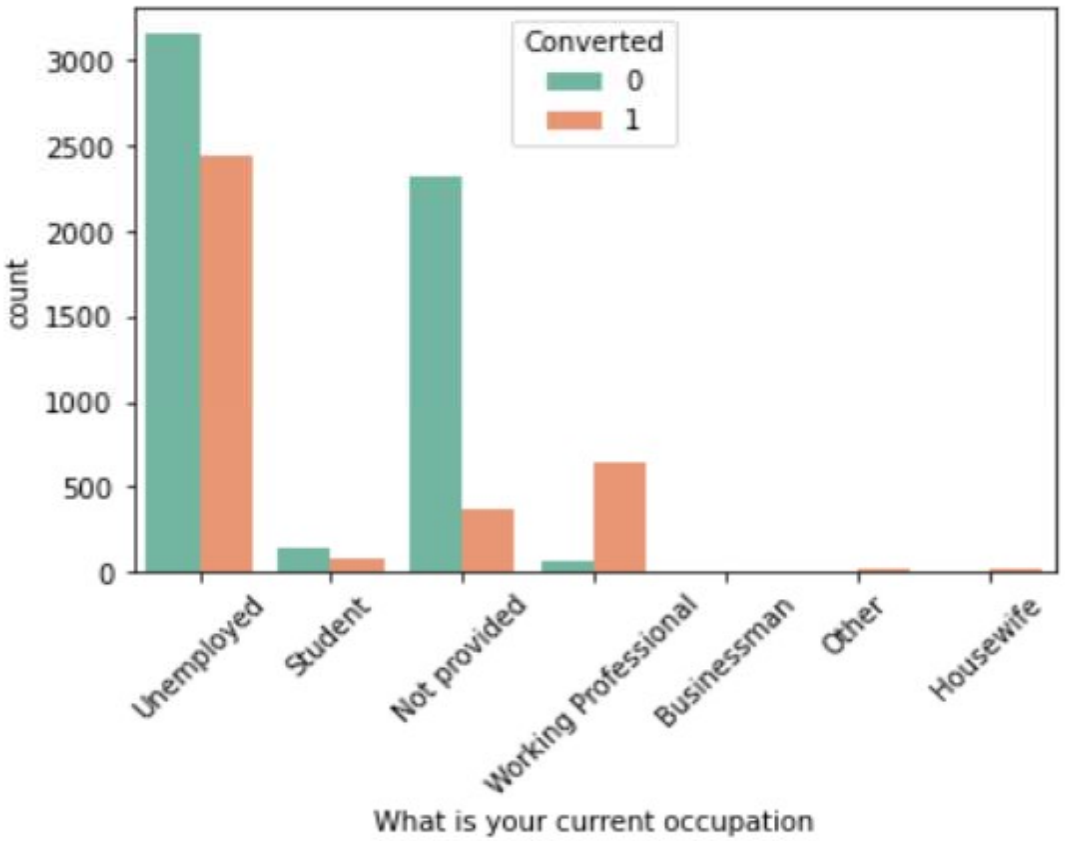
- Have Clear picture of the problem.
- Inspecting the dataframe.
- Data Cleansing.
- Data preparation.
- Test train split.
- Feature scaling.
- Model Building using Stats Model & RFE.
- Conclusion.

LEAD SOURCE ANALYSIS



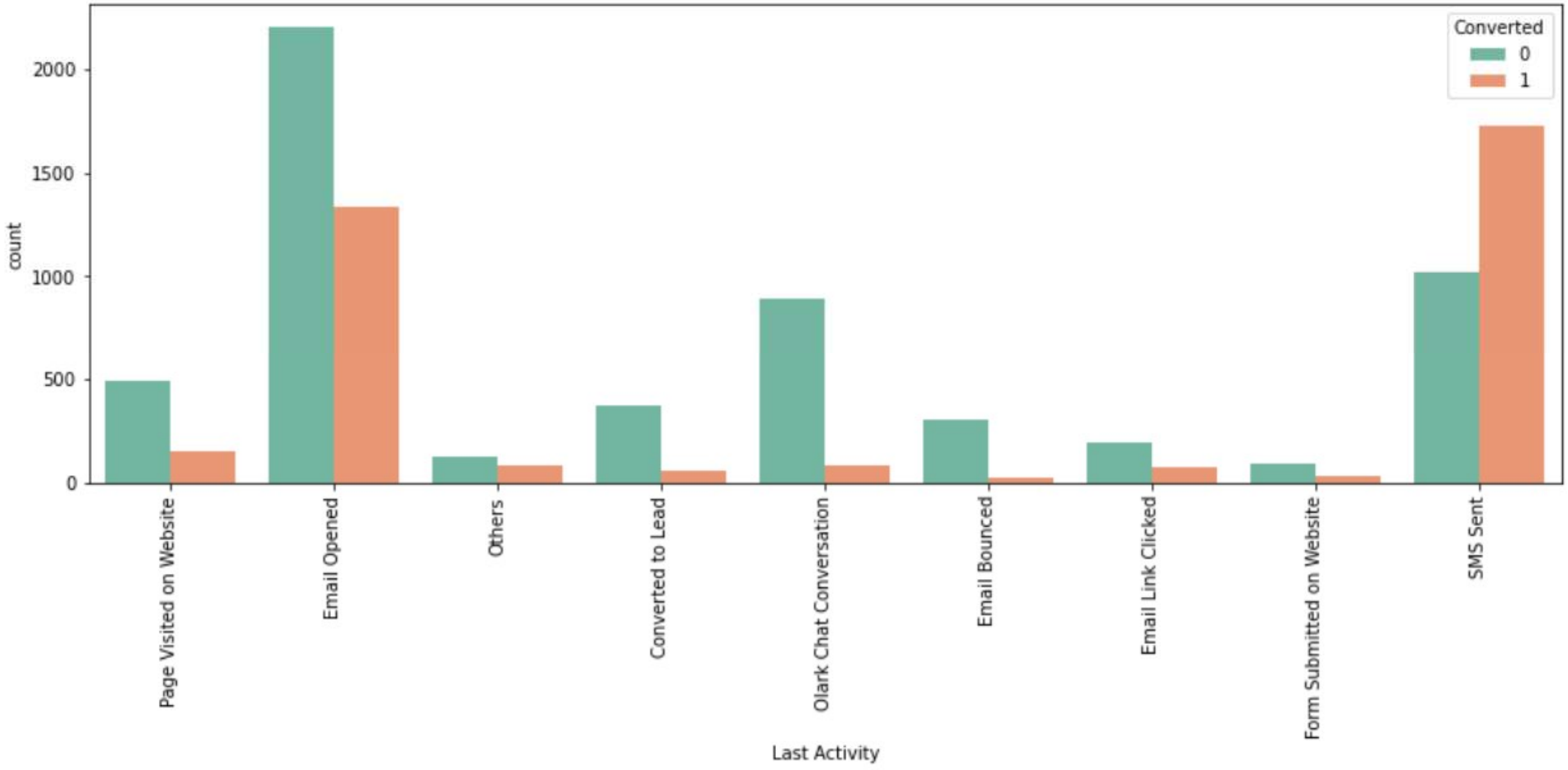
Maximum Leads are generated by Google and Direct Traffic.
Conversion rate of Reference leads and Welinkgak Website leads is very high.

LEADS BASED ON OCCUPATION



Maximum leads generated are unemployed and their conversion rate is more than 50%.
Conversion rate of working professionals is very high.

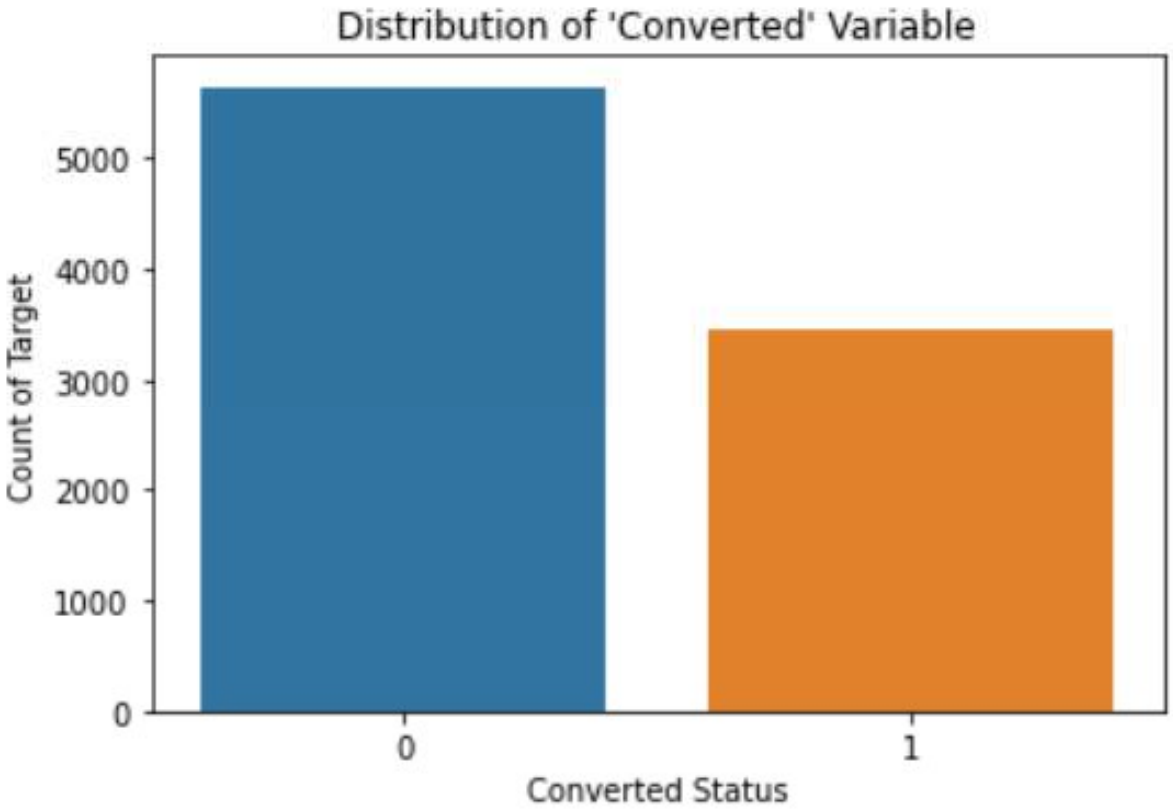
LEAD BASED ON LAST ACTIVITY



Maximum leads are generated having last activity as Email opened but conversion rate is not too good.

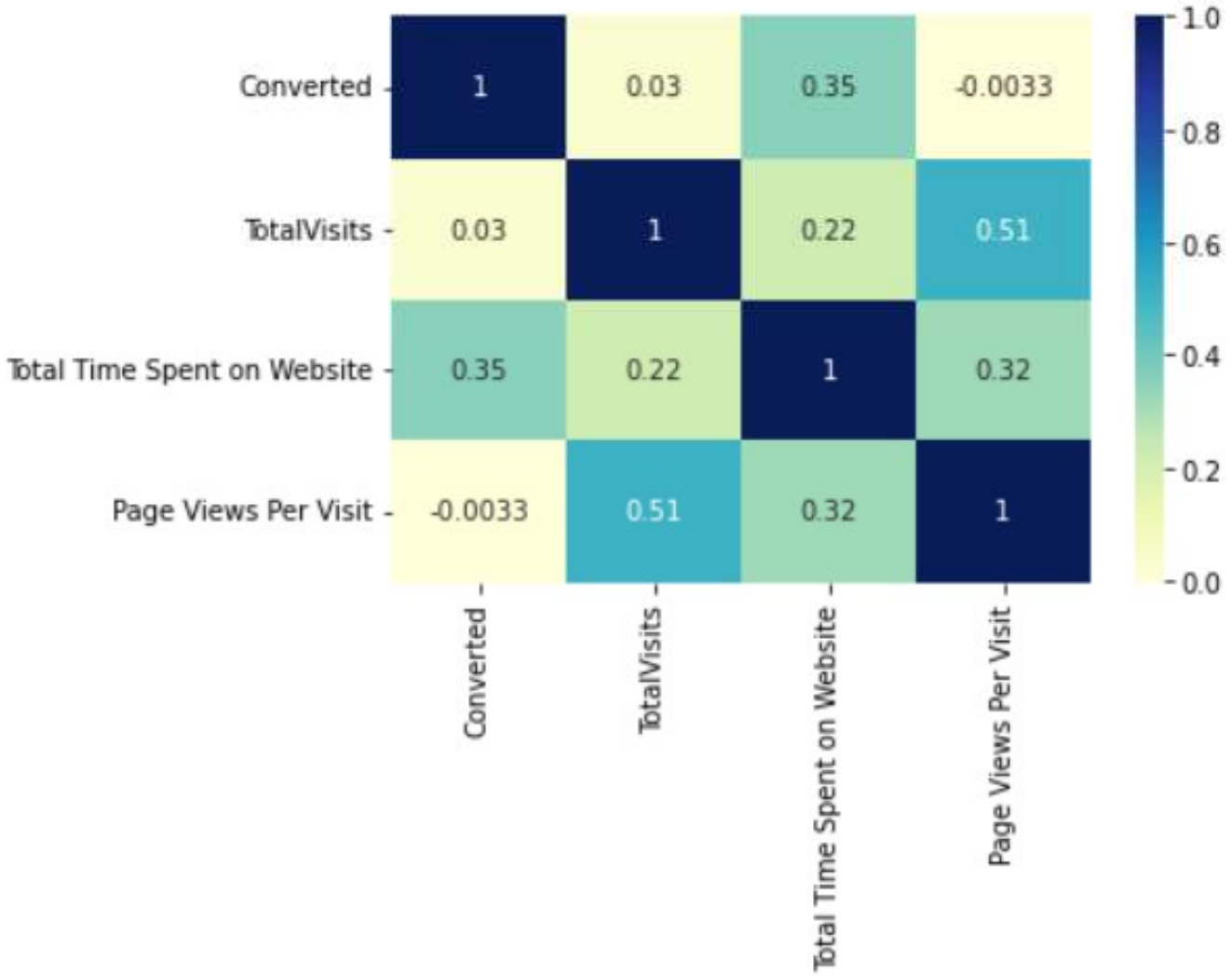
SMS sent as last acitivity has high conversion rate.

NUMERICAL ATTRIBUTE ANALYSIS

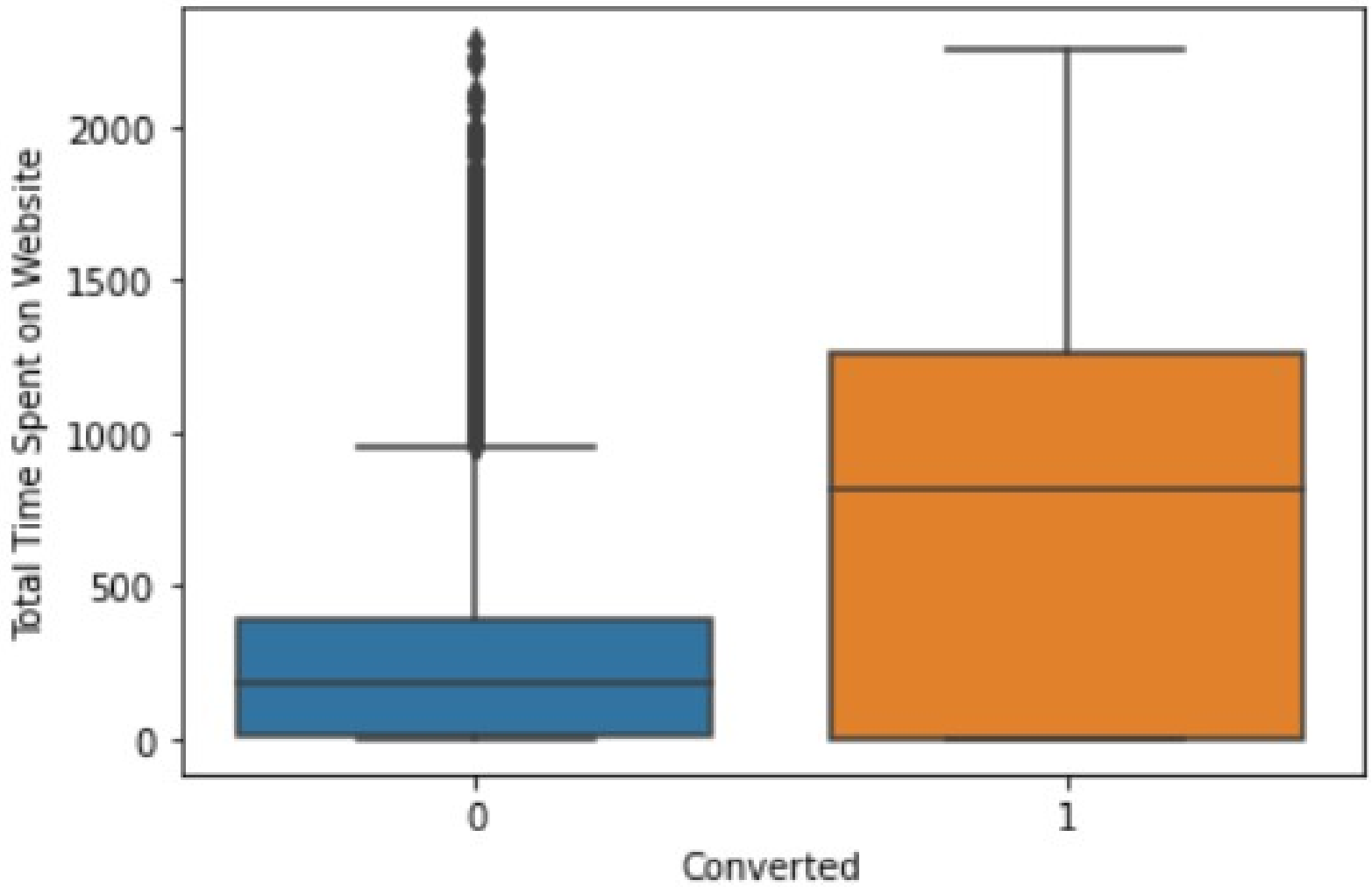


Currently, lead Conversion rate is 38% only

CORRELATIONS OF NUMERIC VALUES USING HEAT MAP

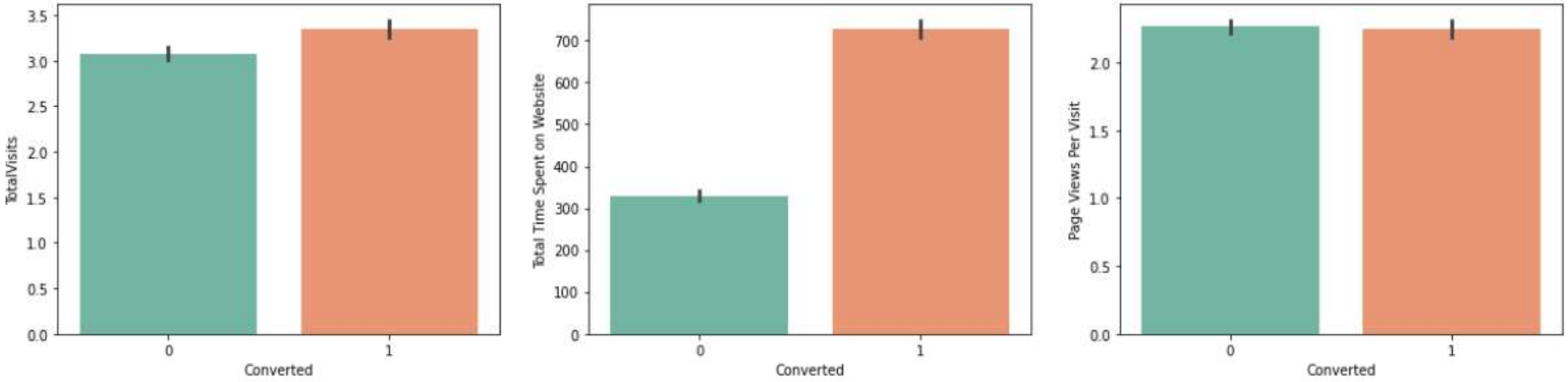


TOTAL TIME SPENT ON WEBSITE



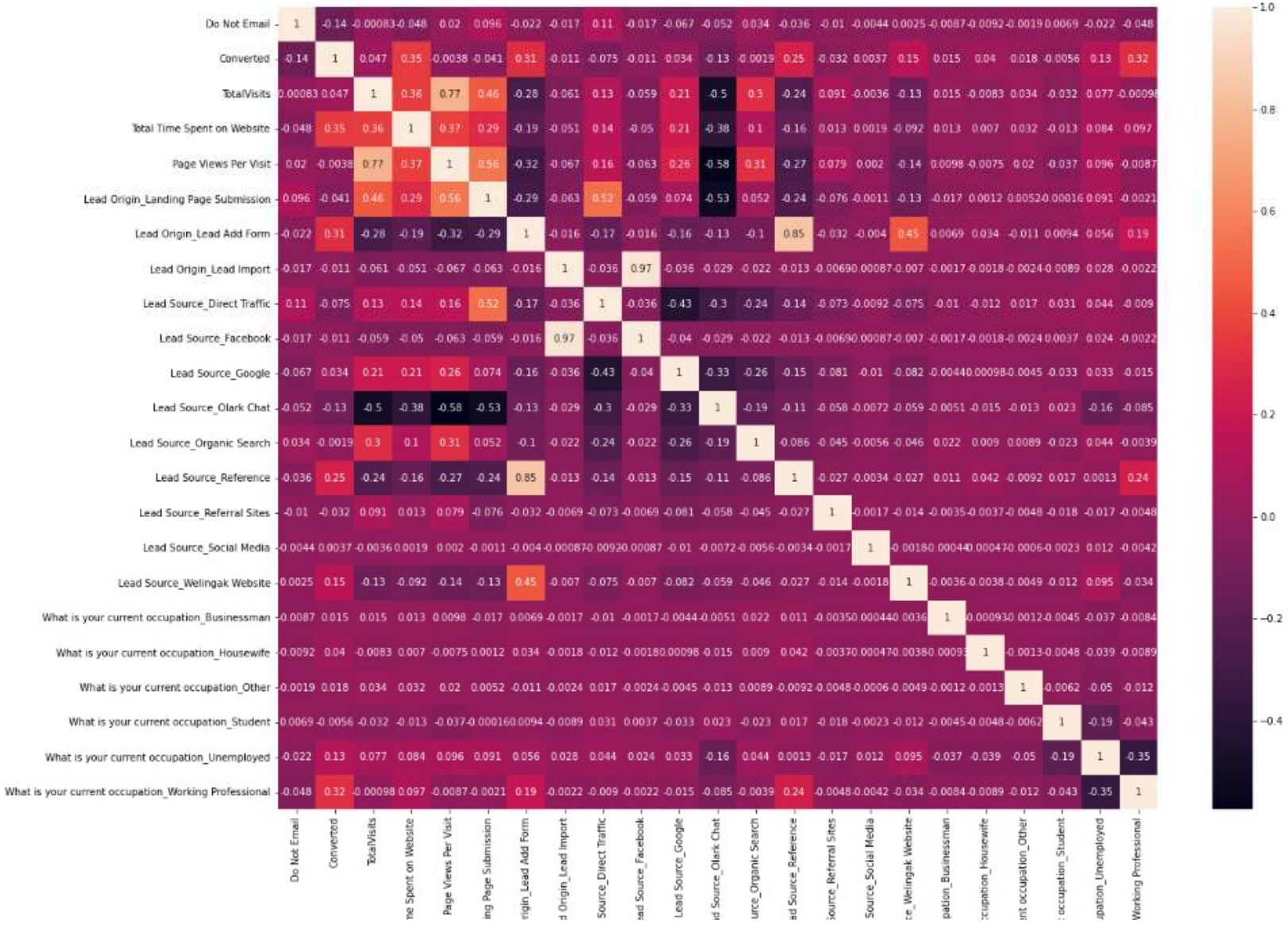
As can be seen, leads spending more time on website are more likely to convert , thus website should be made more enagaging to increase conversion rate

CHECKING CONVERSION OF NUMERICAL VALUES



- The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit

CORRELATION MATRIX

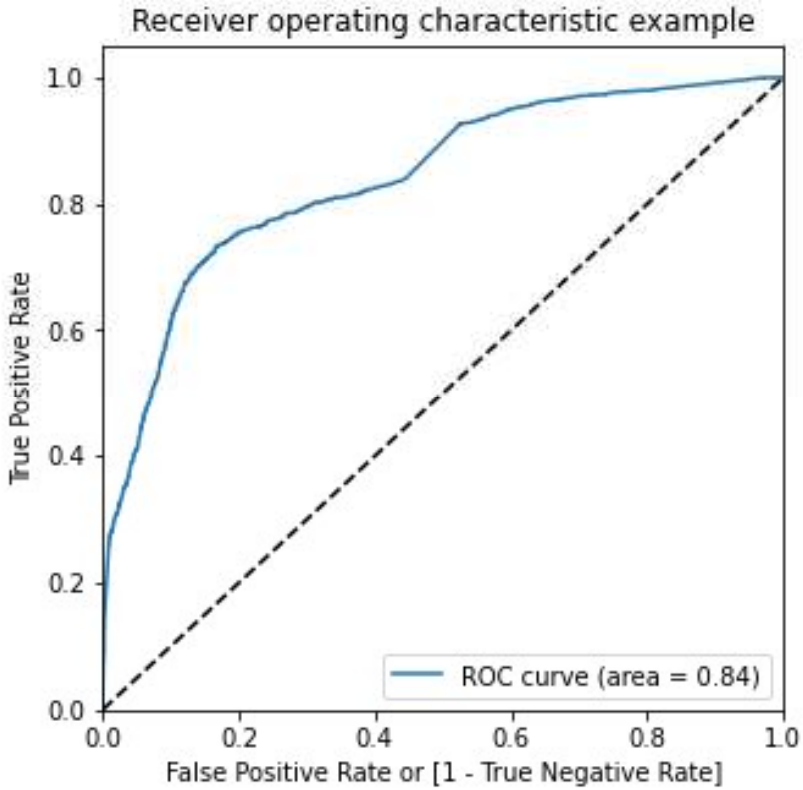


Model Building using Stats Model & RFE

	Features	VIF
2	Lead Origin_Lead Add Form	3.81
6	Lead Source_Reference	3.63
9	What is your current occupation_Unemployed	2.58
4	Lead Source_Google	1.70
3	Lead Source_Direct Traffic	1.67
5	Lead Source_Organic Search	1.31
10	What is your current occupation_Working Profes...	1.29
1	Total Time Spent on Website	1.12
8	What is your current occupation_Student	1.05
0	Do Not Email	1.03
7	Lead Source_Referral Sites	1.02

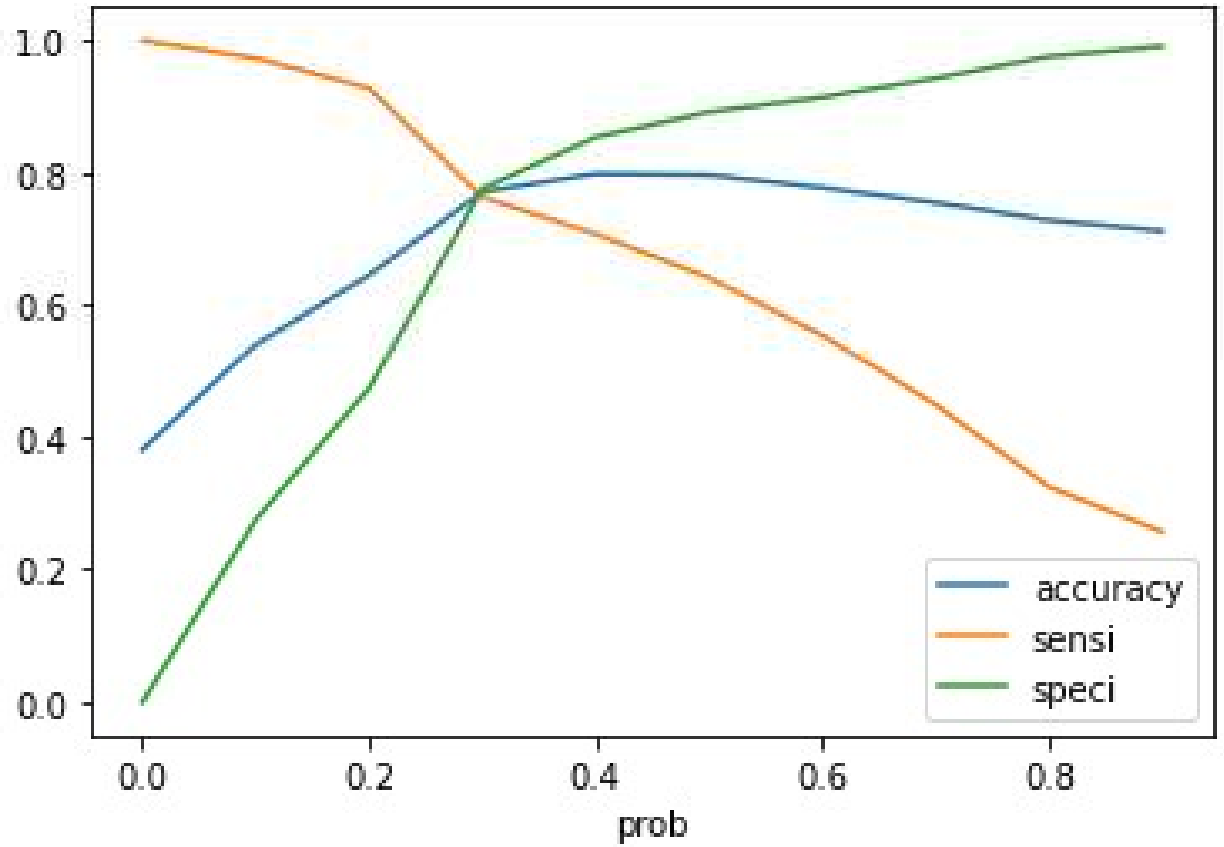
All variables have a good value of VIF. So we need not drop any more variables and we can proceed with making predictions using this model only

PLOTTING ROC CURVE



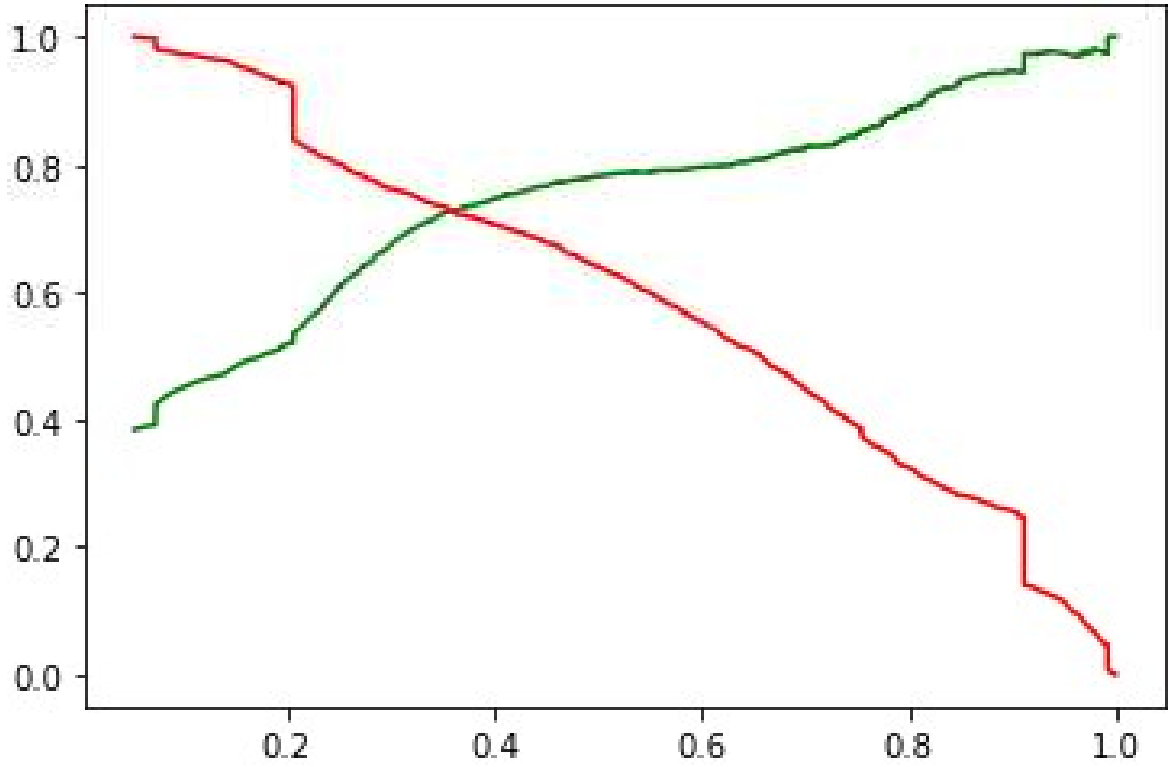
The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model.

Finding Optimal Cutoff Point



From the curve above, 0.3 is the optimum point to take it as a cutoff probability. Hence, we can see that the final prediction of conversions have a target of 83% conversion as per the X Educations CEO's requirement . Hence, we can say that this is a good model.

Precision and Recall Trade-off



the final prediction of conversions have a target rate of 83% (same as predictions made on training data set)

CONCLUSIONS

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.