LEAD SCORING CASE STUDY

By : Pallavi Thakur

Anamika Nayak

Batch: DS C37

PROBLEM STATEMENT

- An Education Institute sells online courses to industry professionals. Lets call that Education institute as 'X'.
- 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.

Business Objective:

X education wants to know most promising leads, for which they want to build a Model which will identify the hot leads. The model to be built in such a way that it can be used for future purpose as well.

STEPS FOLLOWED:

- 1. Understanding and Inspecting the Dataset
- II. Cleaning and Preparation of the Data:
  - Converting the 'Select' values to NaN
  - Dropping all the columns that wont be used for further analysis and the columns that might contain duplicate values
  - Dropping the columns that would have large number of missing values.
  - Imputating the values as Mean/Median – Numeric Variables and Mode() – Categorical Variables
  - Checking for Outliners and handling them

III. Performing EDA both Univariative and Bi-Variative and drawing initial conclusion.

IV. Feature Scaling and creation of Dummy variables for the Categorical variables.

V. Regressions Classification Technique is used for building the Model.

VI. Validating the Model.

VII. Model Presentation

VIII. Drawing Conclusion.

DATA PREPARATION AND DATA CLEANING :

- As part of Data Preparation and Cleaning the below columns –

'Prospect ID', 'Lead Number', 'Asymmetrique Profile Index','Asymmetrique Activity Index',

'Asymmetrique Activity Score','Asymmetrique Profile Score','Lead Profile','Lead Quality',

'How did you hear about X Education','City','Country','Tags','What matters most to you in choosing a course',

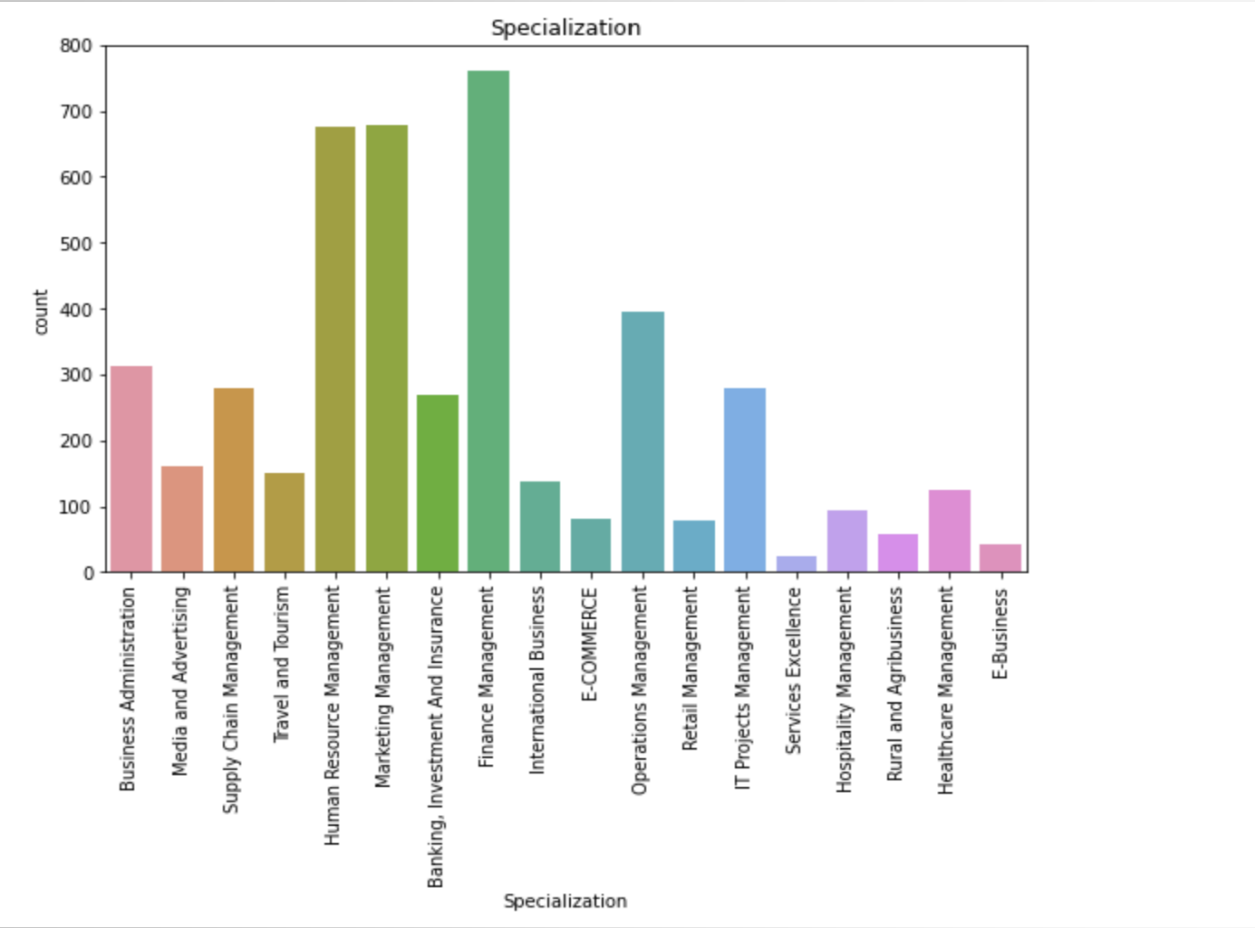
'What is your current occupation' are dropped.

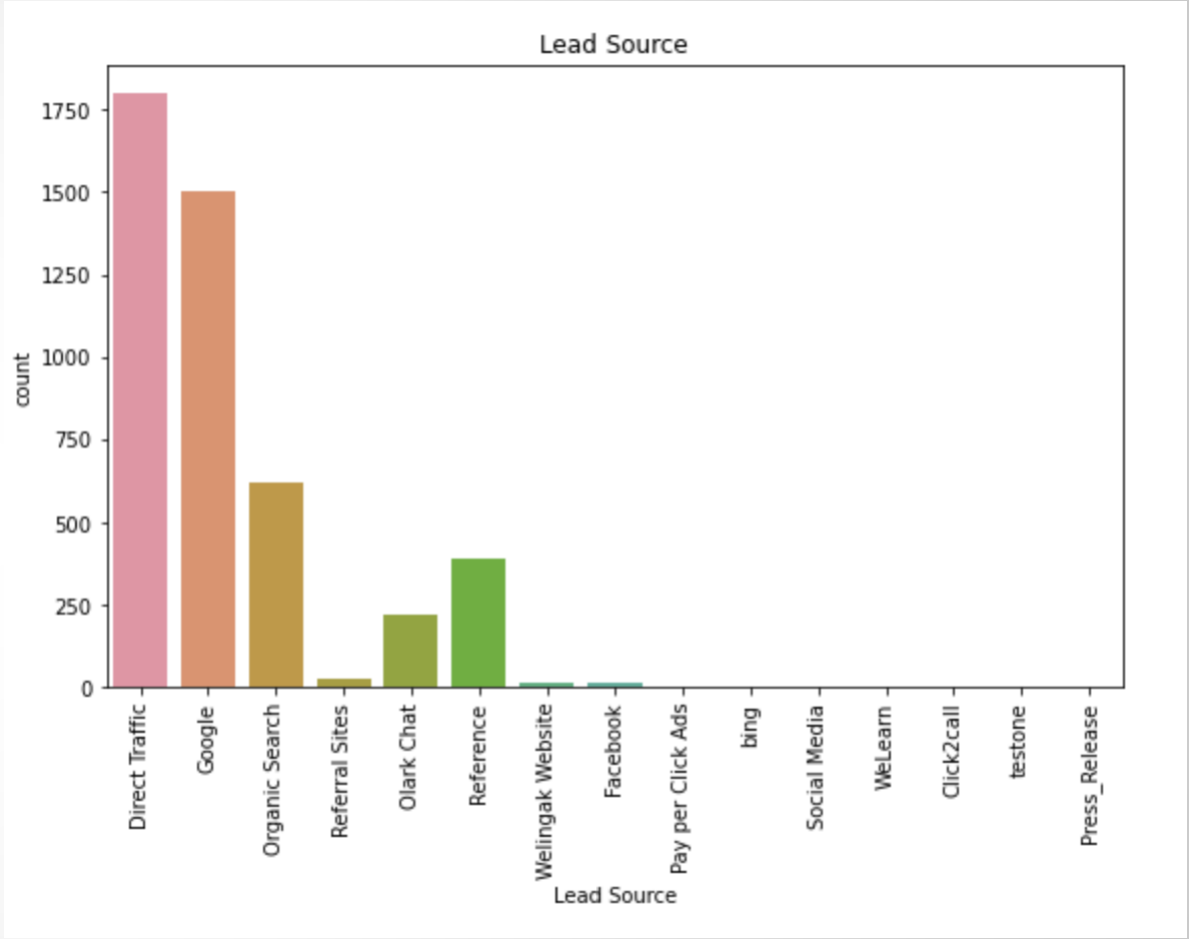
- All the Null values for the column 'Specialization' are replaced with 'not provided'; for columns 'Lead Source', 'TotalVisits', 'Page Views Per Visit', 'Last Activity' the Null values are replaced with Mean() and Mode().

where 'Last Activity', 'Lead Source' is replaced with Mode

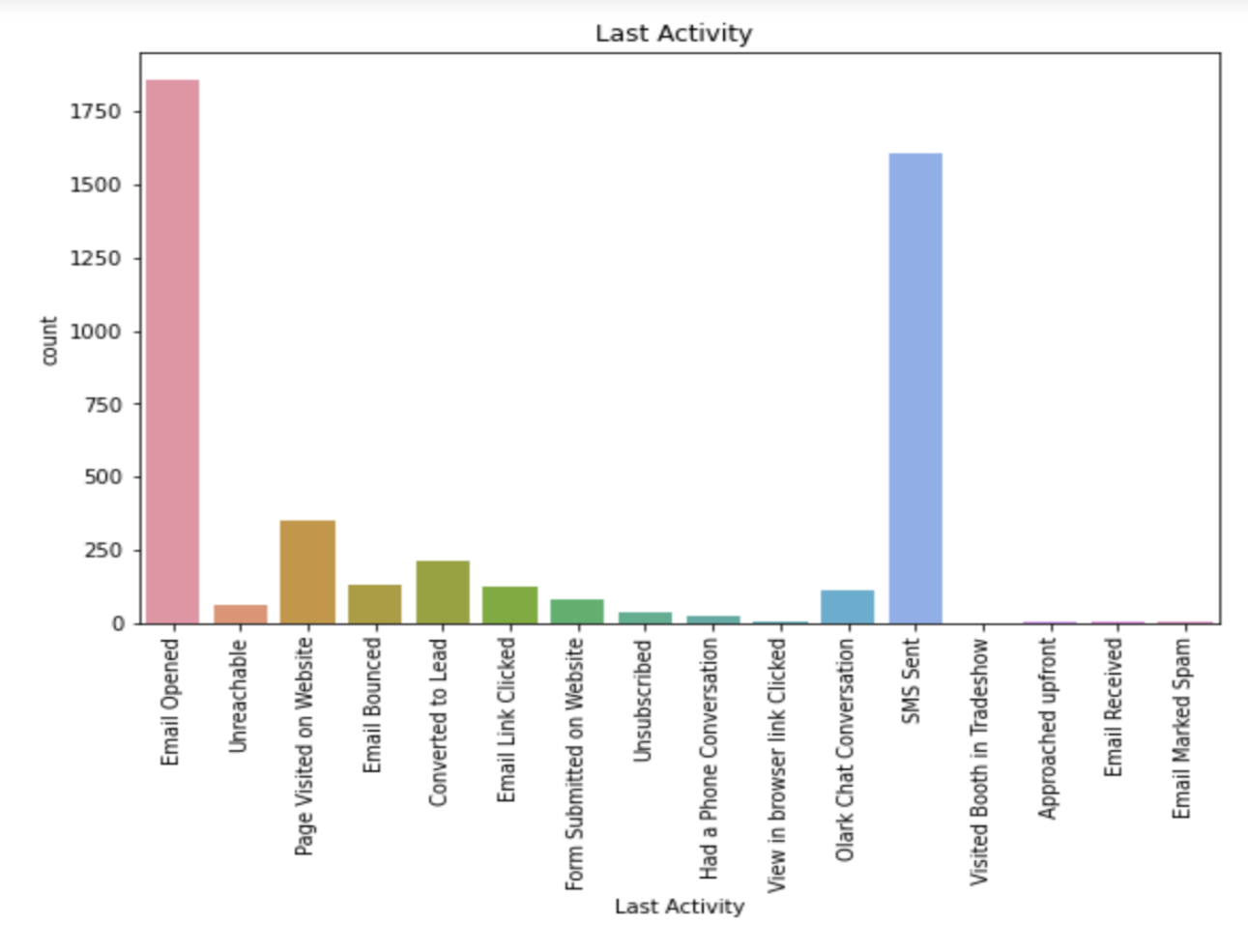
'Page Views Per Visit',' TotalVisits' is replaced with the Median

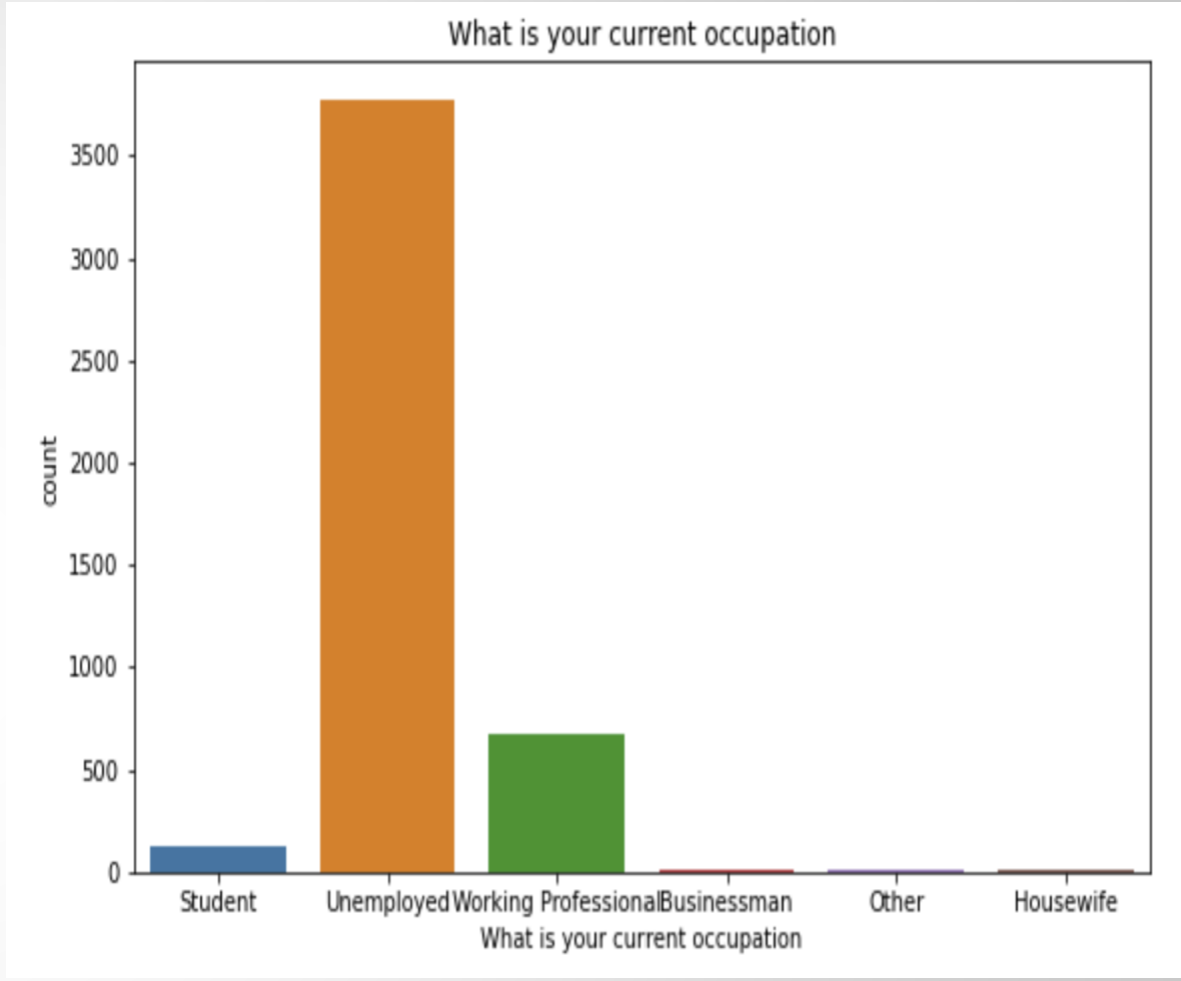
UNIVARIATE ANALYSIS:





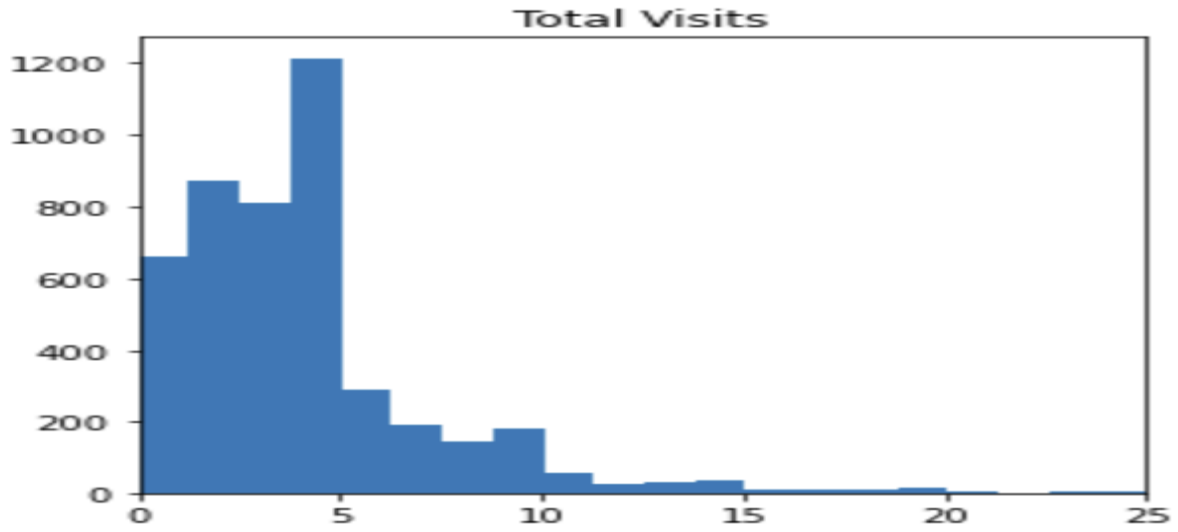
People who have specialization with Management background and people whose lead source was Direct Traffic or Google can be considered as Leads.

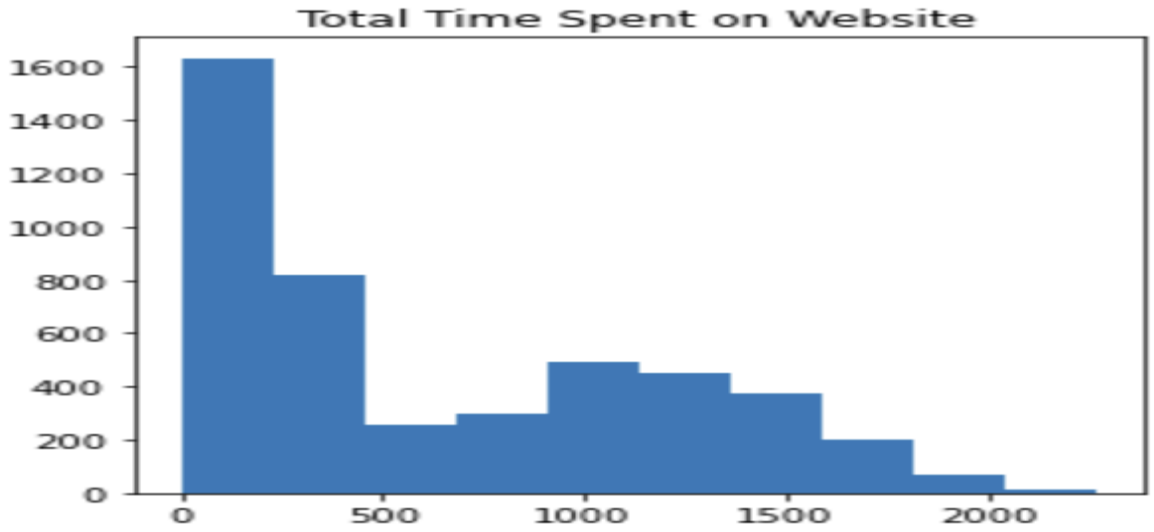


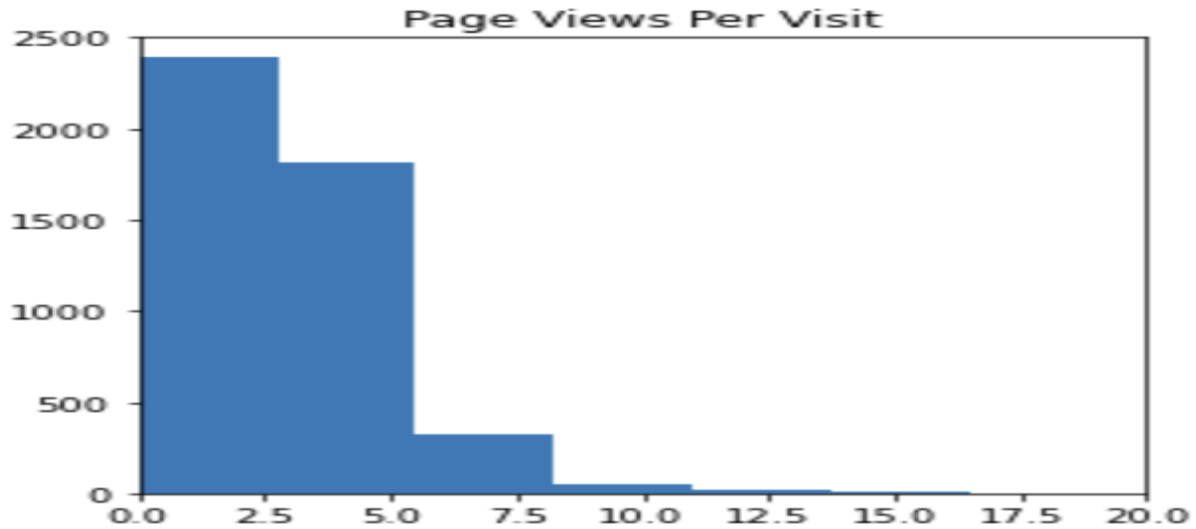


People who have opened the email or have sent SMS can be considered as a Potential Lead. Also, who are unemployed can be considered as Potential lead; Working professionals count is not as much as Unemployed individuals but attempt can be made.

NUMERICAL VARIABLES

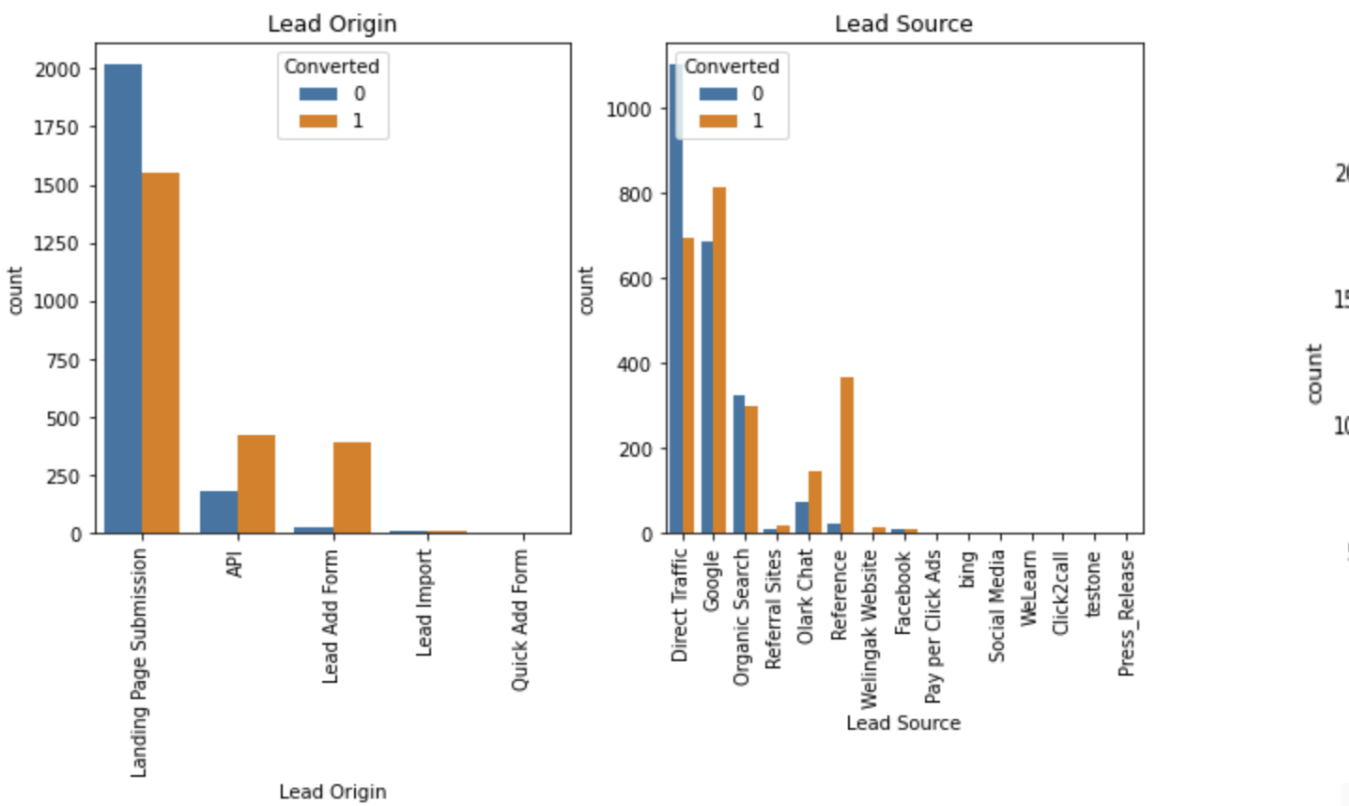


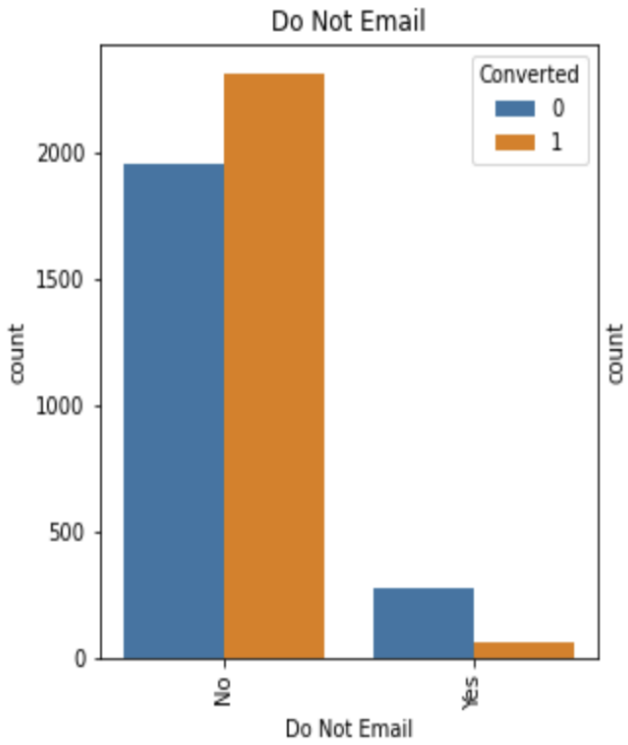


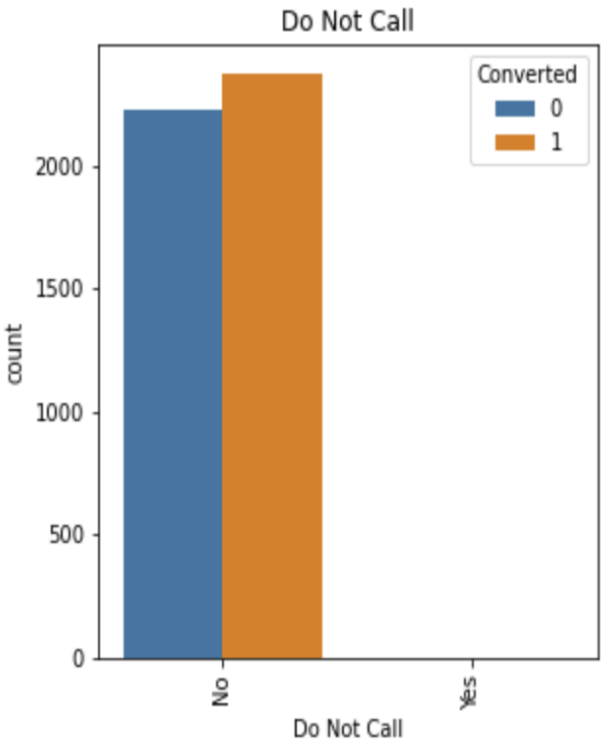


- As per the given graph for the Numerical variables we draw the conclusion :
- People who visit the website or have spent the maximum time on the website can be considered as the Potential leads as the time spent on the website can shown their potential interest in the course.

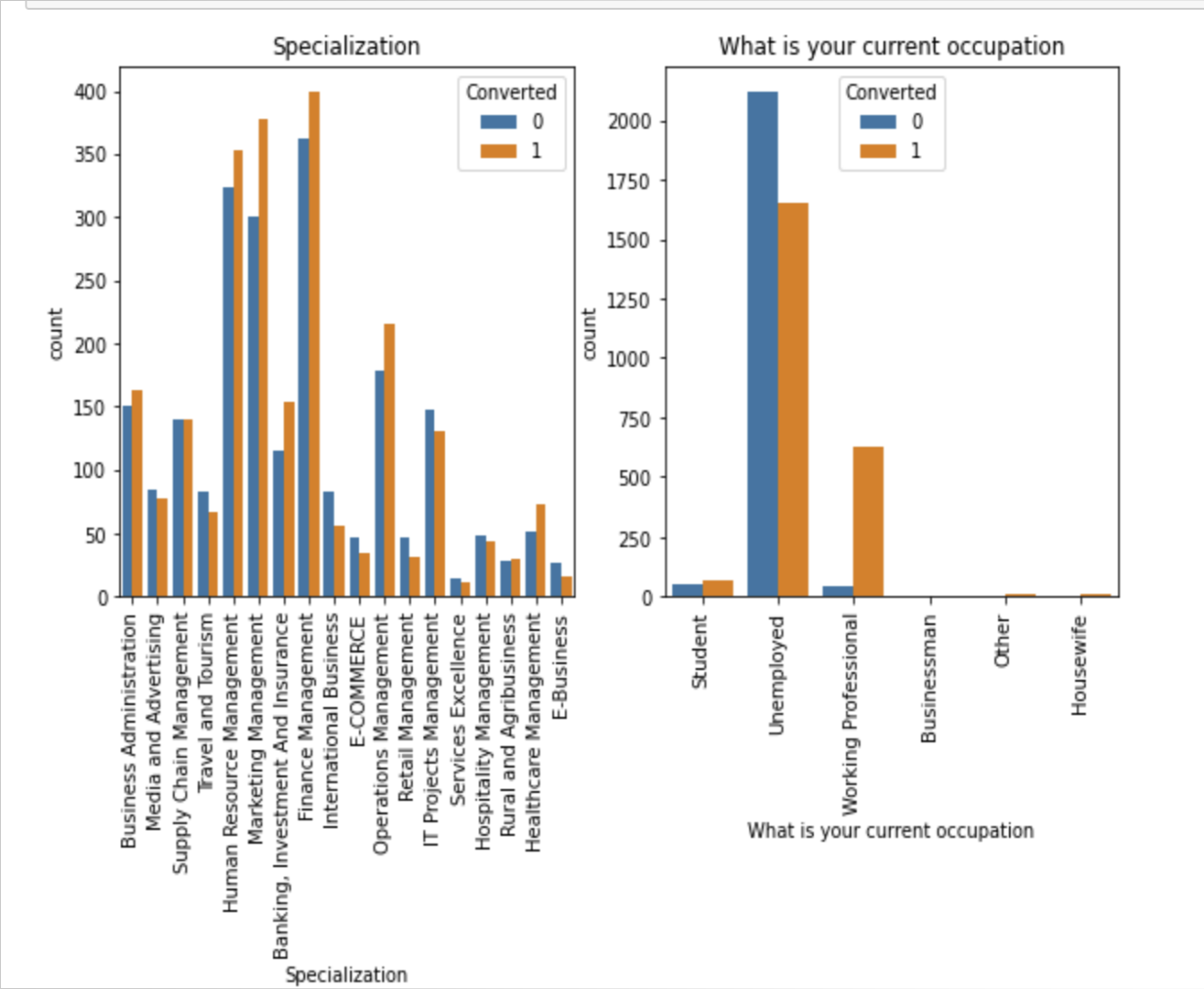
BI VARIATE ANALYSIS

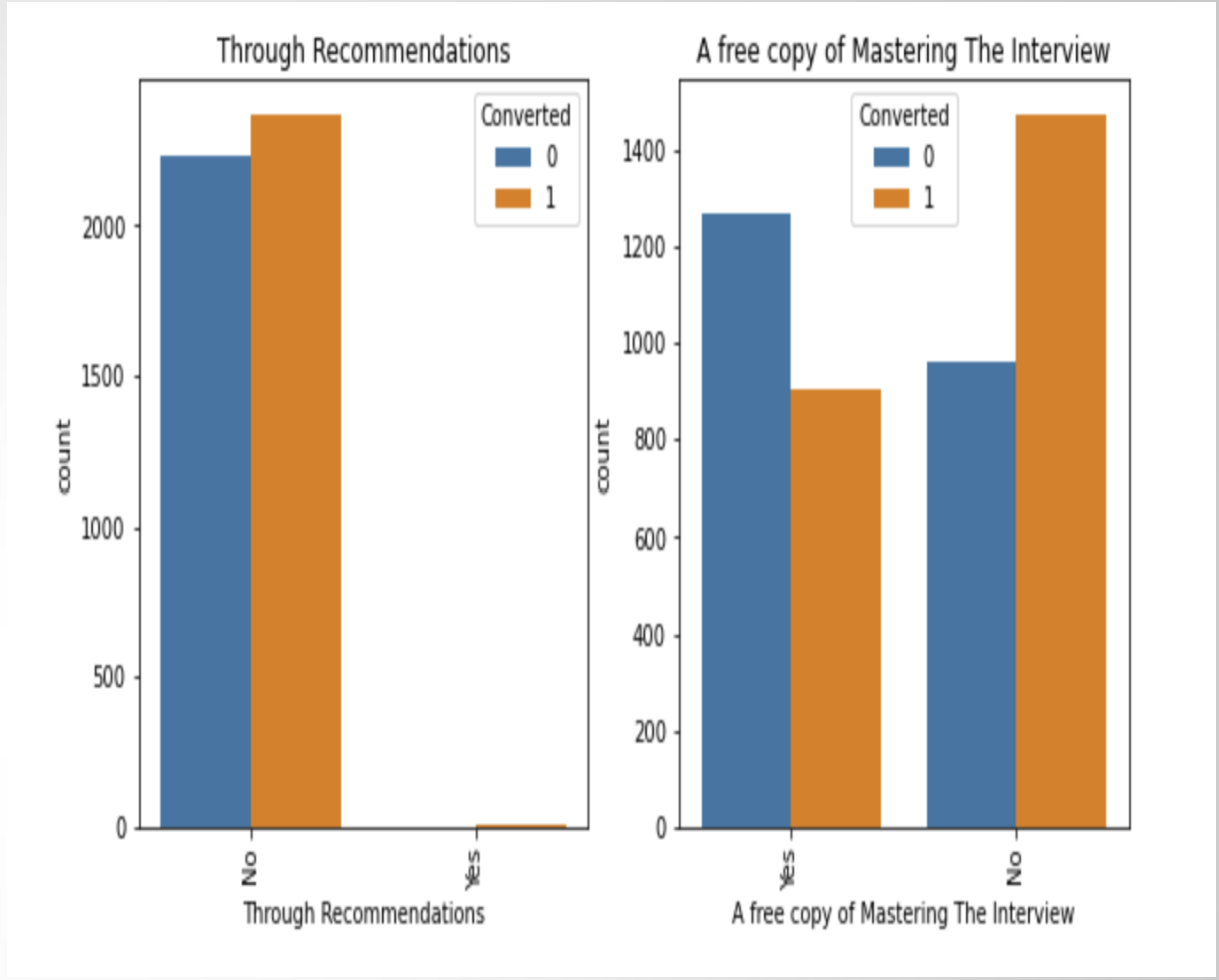






Consumers who have went till the Submission Page of the website or have been through Direct Traffic or have searched for the courses in the Google Search can be considered for potential Leads. Also for the customers who respond back to the email or SMS can be reached out to as well regarding the courses.





- Customers who belong to the Management background such as Finance, Banking , Investment etc. can be reached out to regarding the courses.
- Also people who are unemployed or working professionals can be reached out as well.
- People who are reached out through recommendation and who have requested for a free copys for interview sessions are most like to join the course.

DATA CONVERSION

- Below steps are taken as part of Data Conversion :
- i) Dummy variables has been created for all object type variables.
- ii) Then the dataset is divided into the Train Test dataset and the dataset is scaled using the MinMax Scaler.
- iii) Checking the conversion rate which came to almost 52% and checked the correlation amongst the variables using the Heat plot.

MODEL BUILDING

- Use RFE for Feature Selection 15 variables are selected as the output variables
- Model is built by removing the variables at each step; the variable are removed whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions are done on test data set with which we got Overall accuracy to be around 81%.

THE FINAL MODEL WITH VIF:

Generalized Linear Model Regression Results

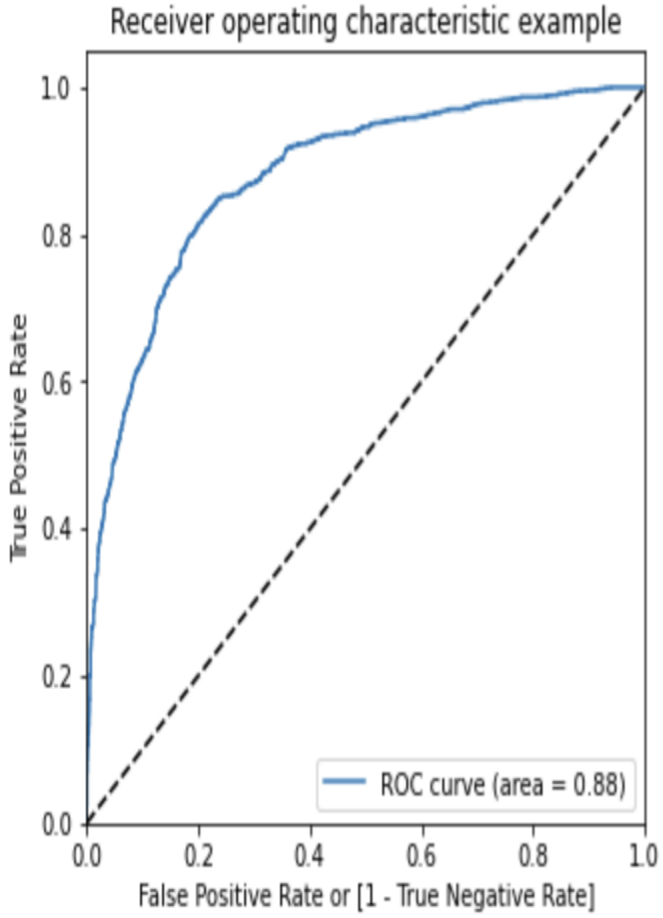
Dep. Variable:	Converted	**No. Observations:**	3084
Model:	GLM	**Df Residuals:**	3073
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1379.8
Date:	Sun, 10 Apr 2022	**Deviance:**	2759.7
Time:	13:15:48	**Pearson chi2:**	3.97e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8241	0.156	-5.271	0.000	-1.131	-0.518
TotalVisits	0.9687	0.283	3.428	0.001	0.415	1.523
Total Time Spent on Website	4.2757	0.210	20.399	0.000	3.865	4.687
Lead Origin_Landing Page Submission	-1.1667	0.143	-8.147	0.000	-1.447	-0.886
Lead Origin_Lead Add Form	3.1464	0.311	10.118	0.000	2.537	3.756
Do Not Email_Yes	-1.7262	0.244	-7.075	0.000	-2.204	-1.248
Last Activity_SMS Sent	0.8834	0.102	8.624	0.000	0.683	1.084
Specialization_Hospitality Management	-1.0514	0.373	-2.817	0.005	-1.783	-0.320
What is your current occupation_Working Professional	2.4695	0.203	12.141	0.000	2.071	2.868
Last Notable Activity_Modified	-0.7955	0.112	-7.103	0.000	-1.015	-0.576
Last Notable Activity_Unsubscribed	1.3856	0.628	2.207	0.027	0.155	2.616

	Features	VIF
2	Lead Origin_Landing Page Submission	3.56
0	TotalVisits	2.64
1	Total Time Spent on Website	2.26
5	Last Activity_SMS Sent	1.59
8	Last Notable Activity_Modified	1.41
7	What is your current occupation_Working Profes...	1.29
3	Lead Origin_Lead Add Form	1.26
4	Do Not Email_Yes	1.20
9	Last Notable Activity_Unsubscribed	1.10
6	Specialization_Hospitality Management	1.02

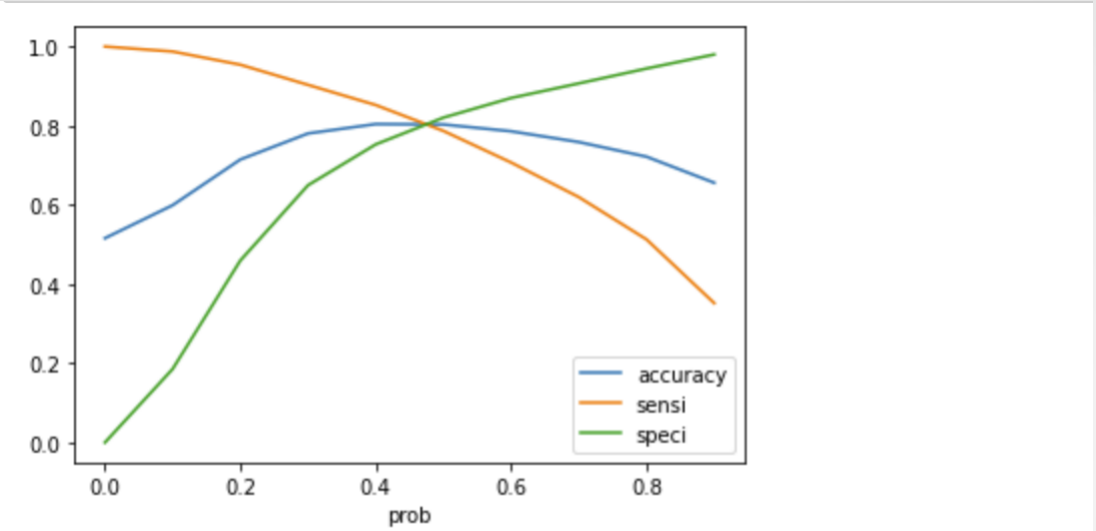
EVALUATING THE MODEL:

- After building the final model making prediction on it(on train set), we created ROC curve to find the model stability with auc score(area under the curve).
- As we can see from the graph the area score is 0.88 which is a great score and our graph is leaned towards the left side of the border which means we have good accuracy.



The area under the curve of the ROC is 0.88 which is quite good. It seems to have a good model

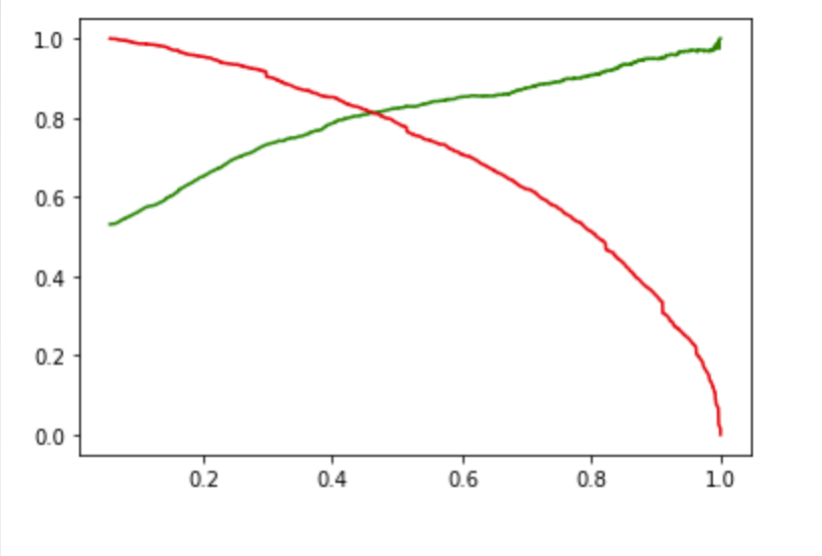
THE OPTIMAL CUTOFF POINT



From the curve, 0.45 is the optimum point to take it as a cutoff probability.

- We have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.
- We found that on 0.4 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected. To verify the same we have plotted the above graph (line plot) and we stand corrected that the meeting point is close to 0.4 and hence we choose **0.4** as our optimal probability cutoff.

PRECISION AND RECALL TRADEOFF



We created a graph which will show us the tradeoff between Precision and recall and we found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.5.

PREDICTION ON TEST SET

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision and recall as got the below findings:
- Accuracy : 78.8%
- Sensitivity : 81.3%
- Specificity : 76%
- Precision : 78.3%
- Recall : 81.3%
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

CONCLUSION

The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps.

Means the recall is having high score value than precision which is acceptable. In business terms, this model has an ability to adjust with the company's future requirements.

This concludes that the model is in stable state.

Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are :

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional