LEAD SCORE CASE STUDY

Submitted by:

Epsita Bose

Mathangi N

(Batch: C34)

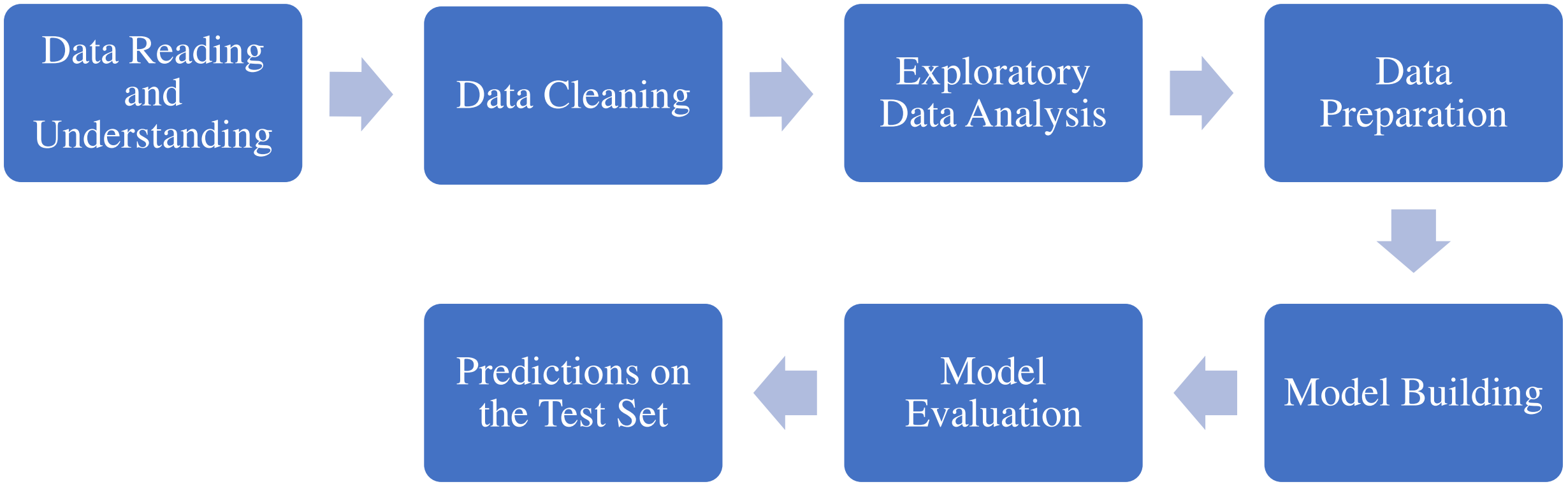
PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites, the company also gets leads through past referrals.
- Leads are acquired through this process, 30% of the leads get converted while most do not.
- 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
- In the middle stage, you need to nurture the potential leads well in order to get a higher lead conversion.

GOALS

- 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

STEPS


Data Reading and Understanding
Data Cleaning
Exploratory Data Analysis
Data Preparation
Model Building
Model Evaluation
Predictions on the Test Set

DATA READING AND UNDERSTANDING

- Total Number of Rows = 37, Total Number of Columns =9240.
- The data contains object, integer and float values.
- There are some null value also presented in some columns
- The description of data is below:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

DATA CLEANING

Find the null value and percentage of it

- There are 17 columns which contain null value

Filter

- Filter null value with 50%

Remove columns with high missing percentage

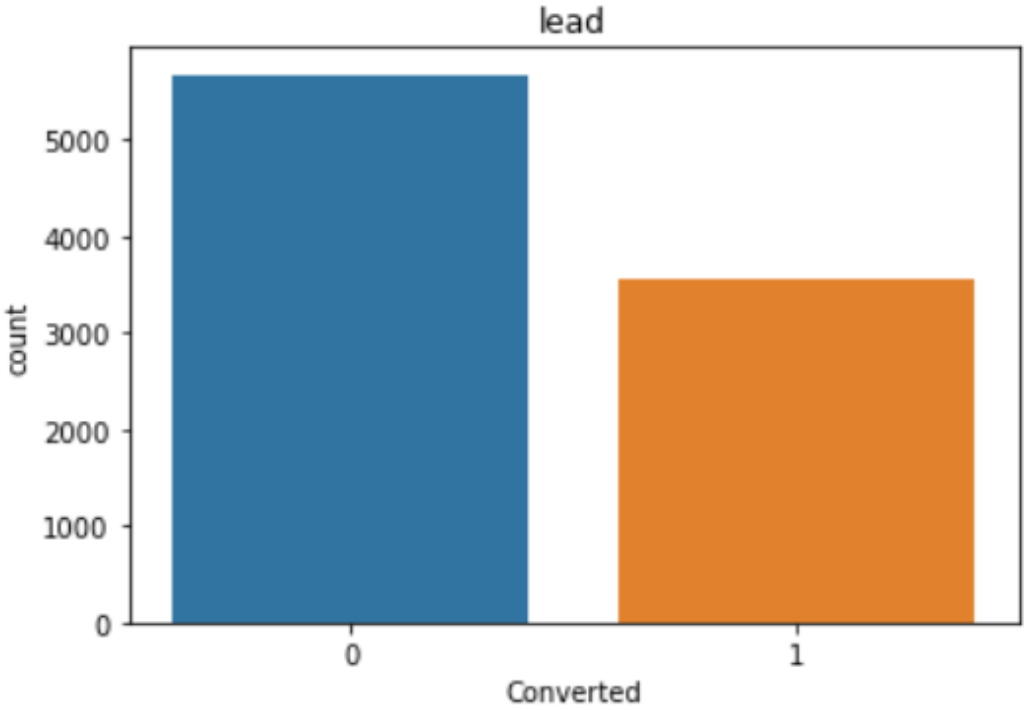
- Remove the null value which containing more than 35%

Replacing the null value with relevant values and relevant category

- Replace the null values with 'not provided'
- Some value replaced by mean, median and mode

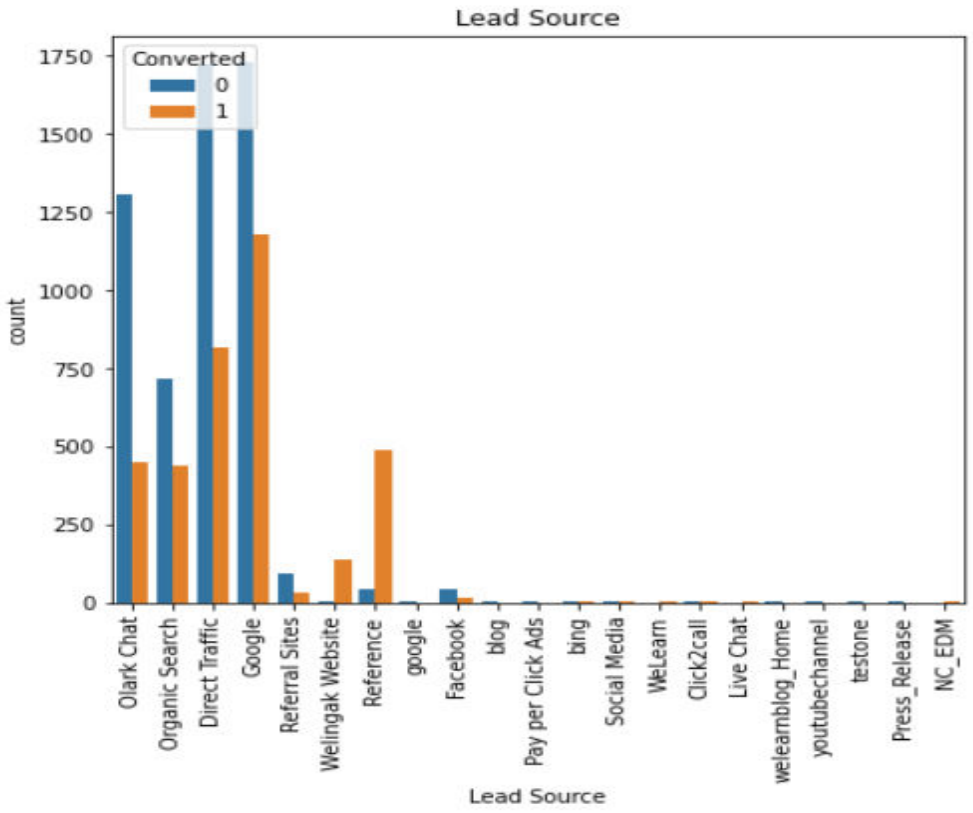
EXPLORATORY DATA ANALYSIS

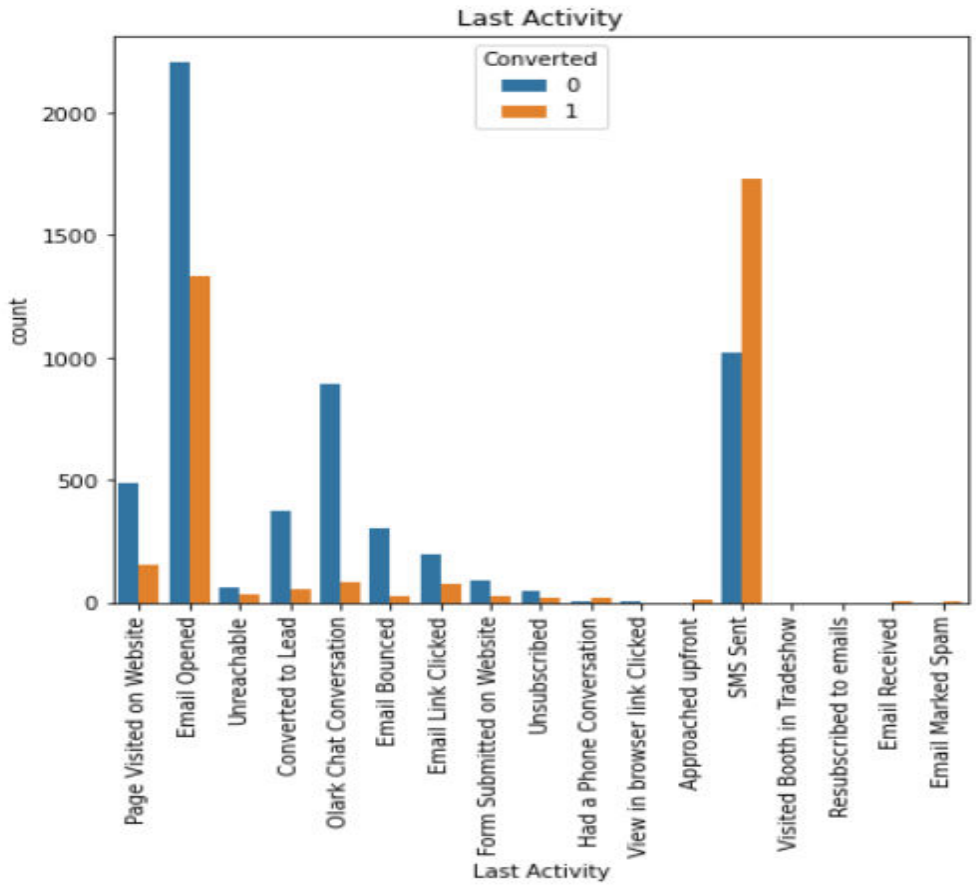
EDA was done by Univariate Analysis, Bivariate Analysis for Categorical and Numerical variable and check relationship.



The number of non-converted leads are higher.

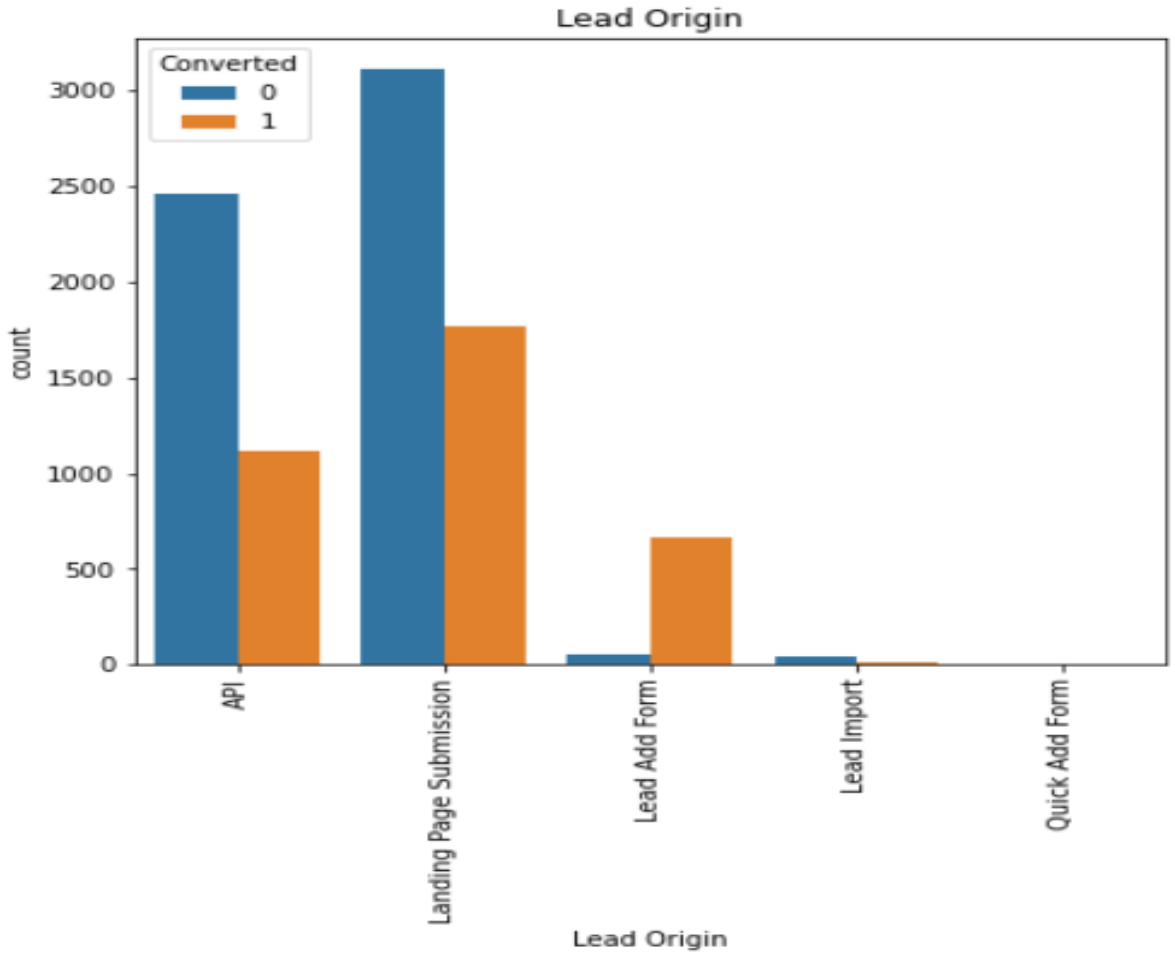
EDA plots for categorical column for those who Converted and those who didn't.





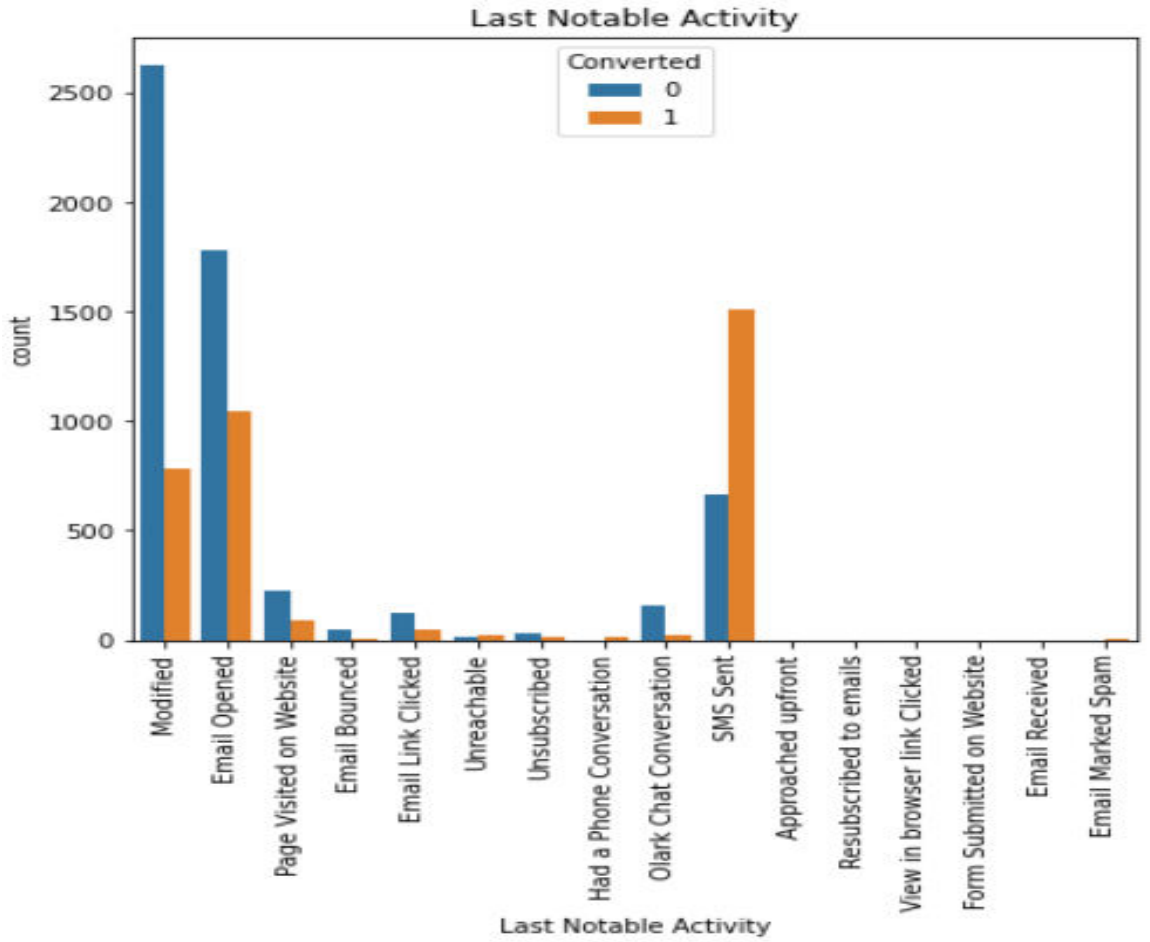
Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category In Last Activity the number of Hot leads is higher in SMS and in EMAIL opened.

EDA plots for categorical column



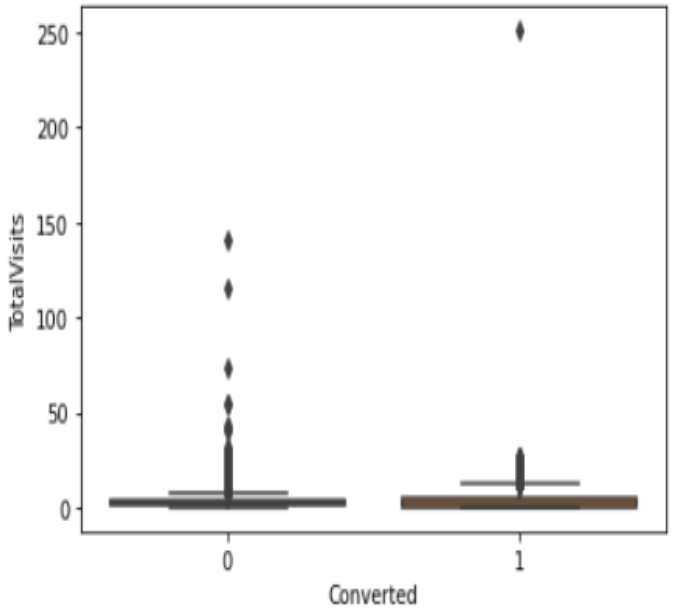
Most of the leads coming at landing page submission, API and lead add form.

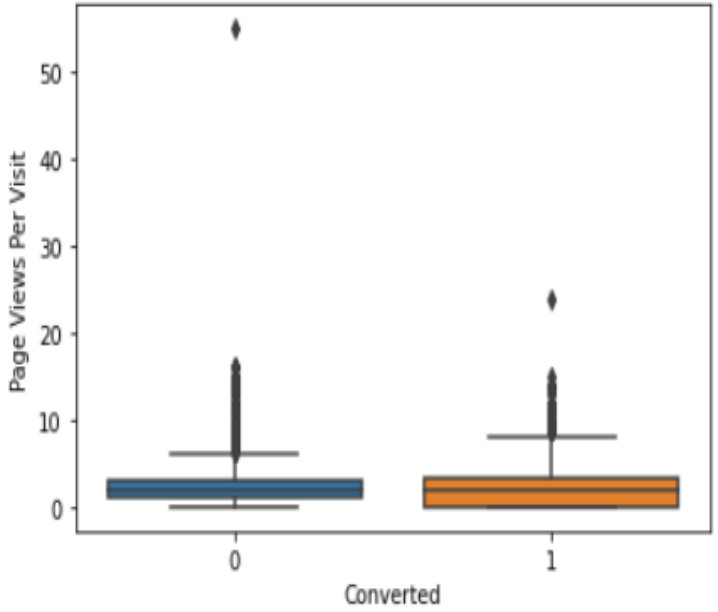
EDA plots for categorical column

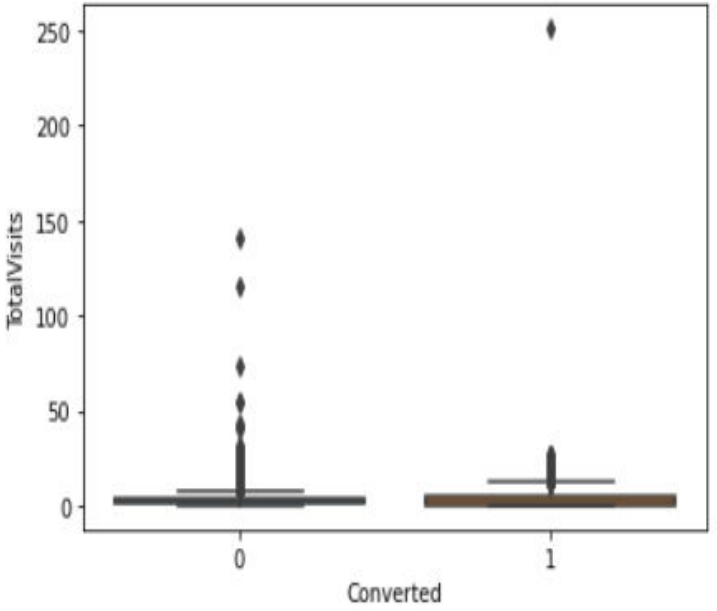


In Last notable activity the number of leads is higher in SMS and in EMAIL

EDA plots for numerical variables for those who Converted and those who didn't.



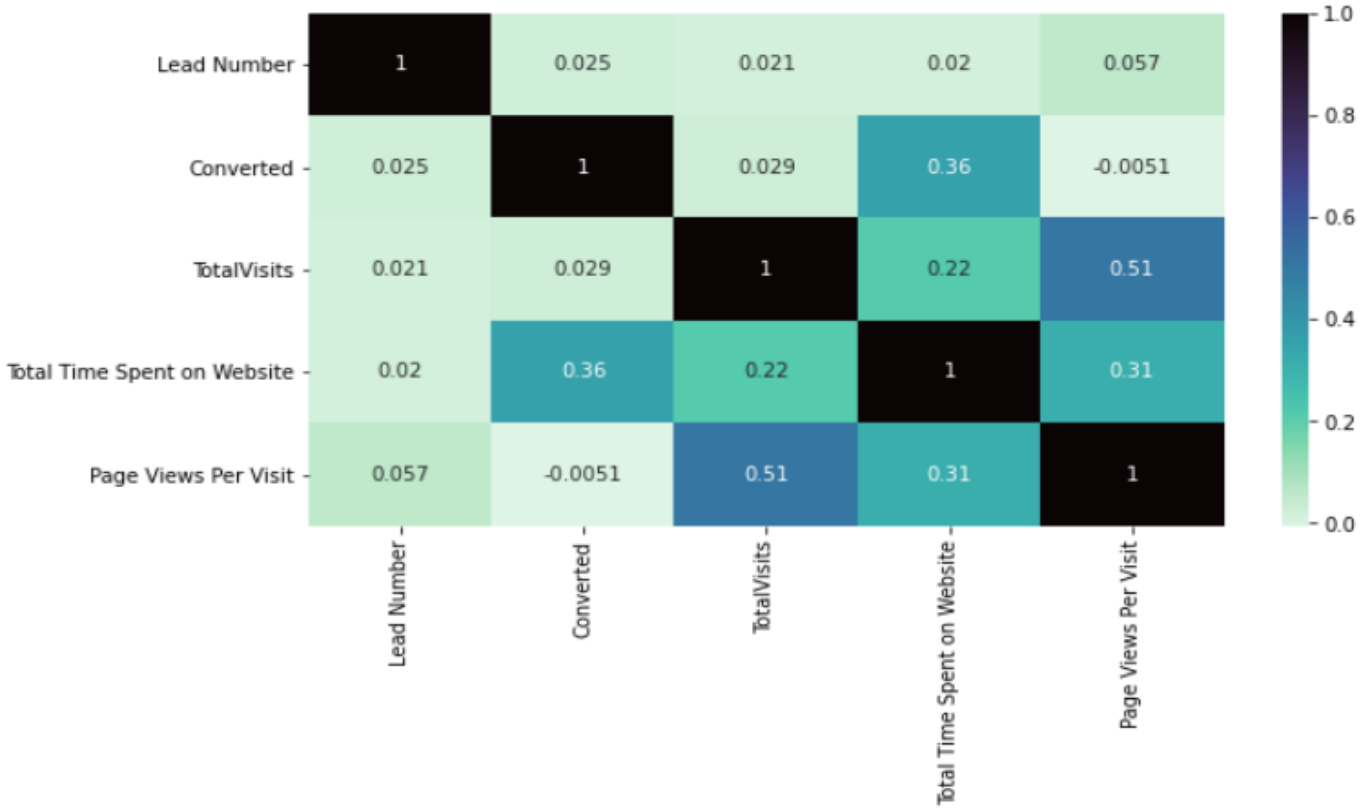




Outliers are present for some columns and leads are less generated

Correlation between the continuous numeric variables in the data

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1.000000	0.025157	0.021366	0.020329	0.057042
Converted	0.025157	1.000000	0.029119	0.362483	-0.005068
TotalVisits	0.021366	0.029119	1.000000	0.217341	0.512214
Total Time Spent on Website	0.020329	0.362483	0.217341	1.000000	0.314266
Page Views Per Visit	0.057042	-0.005068	0.512214	0.314266	1.000000



The correlation between Total Visits and Page Views Per Visit is the highest. There is a positive correlation between all the variables expect between Page Views Per Visit and Converted which has a negative correlation. A positive correlation implies that as the value of one variable increases, the value of the other variable also increases. A negative correlation implies that when the value of one of these variables increases, the value of the other variable decreases

DATA PREPARATION

- Dummies were created for the following variables: 'Lead Origin', 'Lead Source', 'Do Not Email', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'A free copy of Mastering The Interview'.
- The following variables were re-scaled: 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'

MODEL BUILDING

RFE was used for the process of feature selection and initially, 25 variables were considered. After looking at the p-values and the Variance Inflation Factor, the following 13 variables were included in the final model:

- 'Total Time Spent on Website',
- 'Lead Origin_Landing Page Submission',
- 'Lead Origin_Lead Add Form',
- 'Lead Source_Direct Traffic',
- 'Lead Source_Olark Chat',
- 'Lead Source_Referral Sites',
- 'Lead Source_Welingak Website',
- 'Do Not Email_Yes',
- 'Specialization_Select',
- 'Specialization_not provided',
- 'What is your current occupation_Working Professional',
- 'What is your current occupation_not provided',
- 'TotalVisits'

Final Model

The final model has 13 variables all of which have a significant p-value and a VIF score less than 5. This model was used for further analysis.

	Features	VIF
1	Lead Origin_Landing Page Submission	3.46
9	Specialization_not provided	2.90
11	What is your current occupation_not provided	2.74
0	Total Time Spent on Website	1.96
3	Lead Source_Direct Traffic	1.92
4	Lead Source_Olark Chat	1.86
12	TotalVisits	1.64
8	Specialization_Select	1.59
2	Lead Origin_Lead Add Form	1.44
6	Lead Source_Welingak Website	1.30
10	What is your current occupation_Working Profes...	1.21
7	Do Not Email_Yes	1.11
5	Lead Source_Referral Sites	1.05

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2862.5
Date:	Sat, 08 Jan 2022	**Deviance:**	5725.0
Time:	18:32:58	**Pearson chi2:**	7.75e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0717	0.134	-7.968	0.000	-1.335	-0.808
Total Time Spent on Website	4.3814	0.155	28.234	0.000	4.077	4.686
Lead Origin_Landing Page Submission	-0.5766	0.130	-4.442	0.000	-0.831	-0.322
Lead Origin_Lead Add Form	3.2788	0.214	15.350	0.000	2.860	3.697
Lead Source_Direct Traffic	-0.3161	0.085	-3.732	0.000	-0.482	-0.150
Lead Source_Olark Chat	0.9169	0.118	7.790	0.000	0.686	1.148
Lead Source_Referral Sites	-0.7325	0.353	-2.072	0.038	-1.425	-0.040
Lead Source_Welingak Website	3.1196	1.027	3.037	0.002	1.106	5.133
Do Not Email_Yes	-1.2998	0.159	-8.183	0.000	-1.611	-0.989
Specialization_Select	-0.6892	0.123	-5.594	0.000	-0.931	-0.448
Specialization_not provided	-0.9347	0.178	-5.240	0.000	-1.284	-0.585
What is your current occupation_Working Professional	2.3366	0.179	13.048	0.000	1.986	2.688
What is your current occupation_not provided	-1.1838	0.109	-10.876	0.000	-1.397	-0.970
TotalVisits	4.8465	1.937	2.501	0.012	1.049	8.644

MODEL EVALUATION

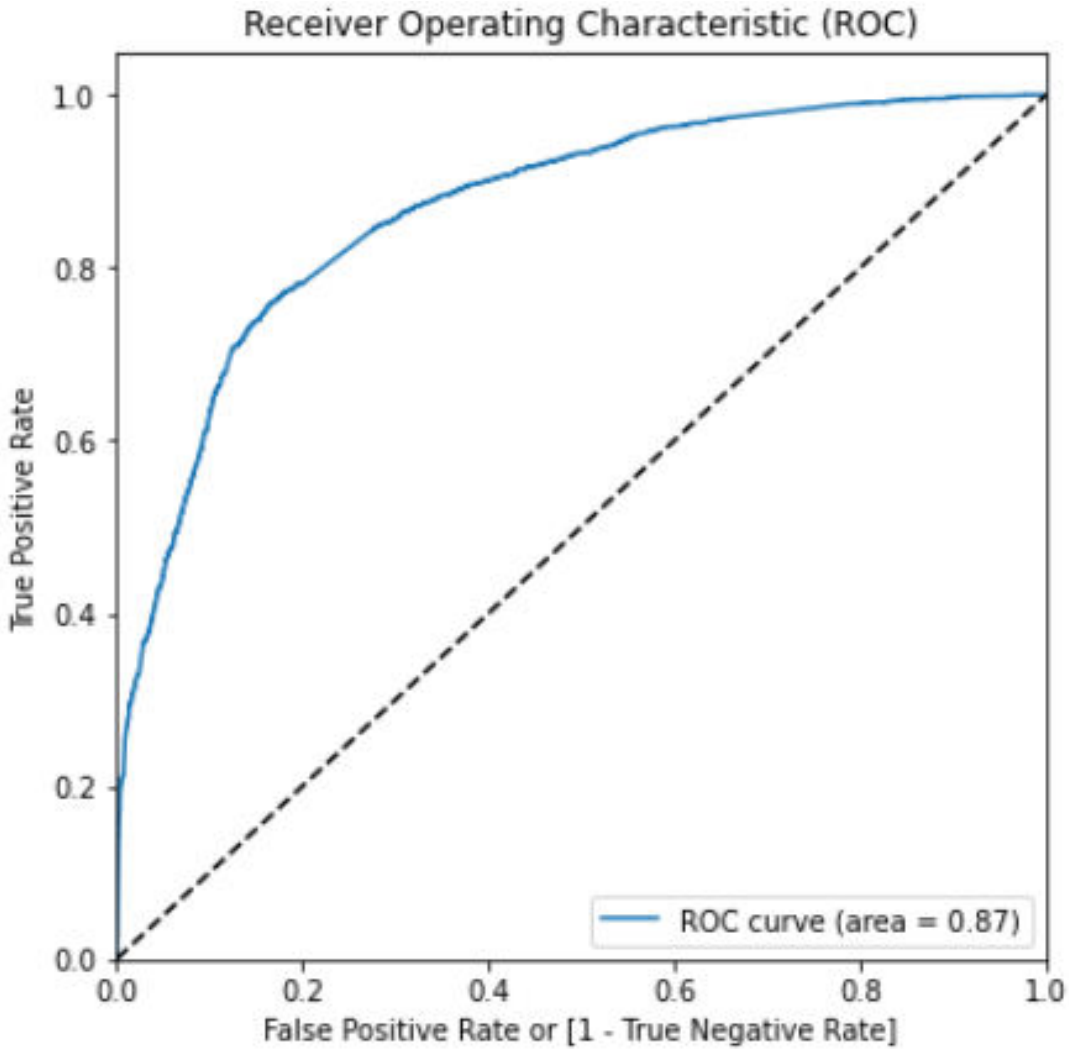
- Confusion Matrix:

Predicted	Not converted	Converted
Actual		
Not converted	3531	424
Converted	864	1649

- Accuracy on train dataset is 80%
- The sensitivity value is 0.66
- The specificity value is 0.89

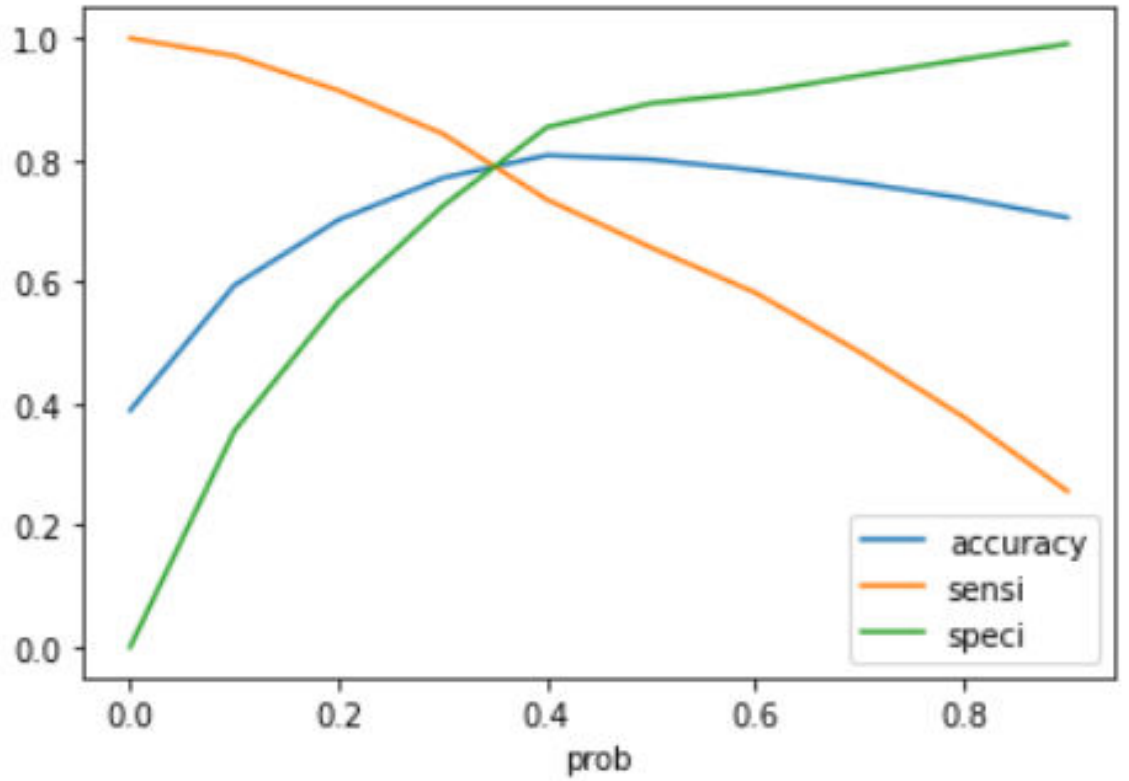
ROC Curve

A good classifier is that the curve must be as close to the boundary as possible. Here, the curve is far from the diagonal line is towards the upper-left corner. The area under the curve of the ROC is 0.87 which is a pretty good value and hence the model at hand looks like a good model.



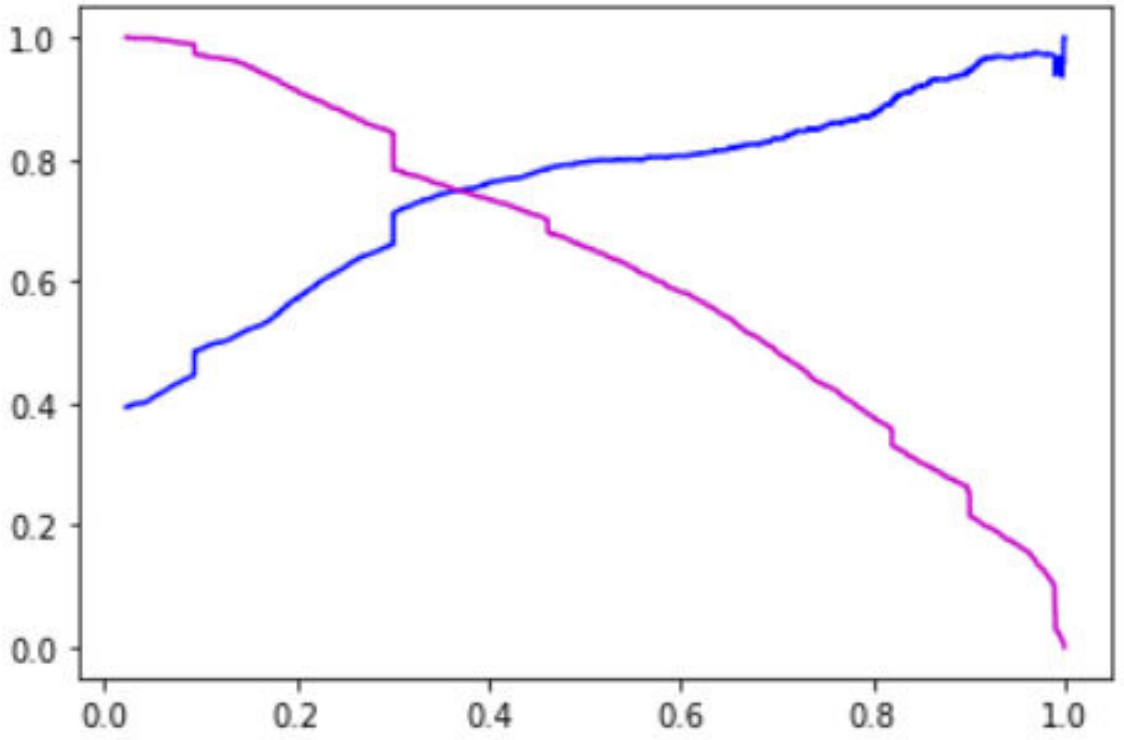
Finding Optimal Cut-Off

- The Accuracy, Sensitivity and specificity intersect at 0.39. Hence, at around 0.39,optimal values of the three metrics can be obtained.
- Hence it was used as the cutoff.
- The accuracy value is 0.81 which is a pretty good value, sensitivity value is 0.74 and specificity value is 0.85.



Precision and Recall

The precision score is 0.79 and the recall score is 0.65. The precision and recall intersects at 0.4. Using 0.4 as the cut-off, The accuracy value is 0.81 which is a pretty good value, sensitivity value is 0.73 and the specificity value is 0.85.

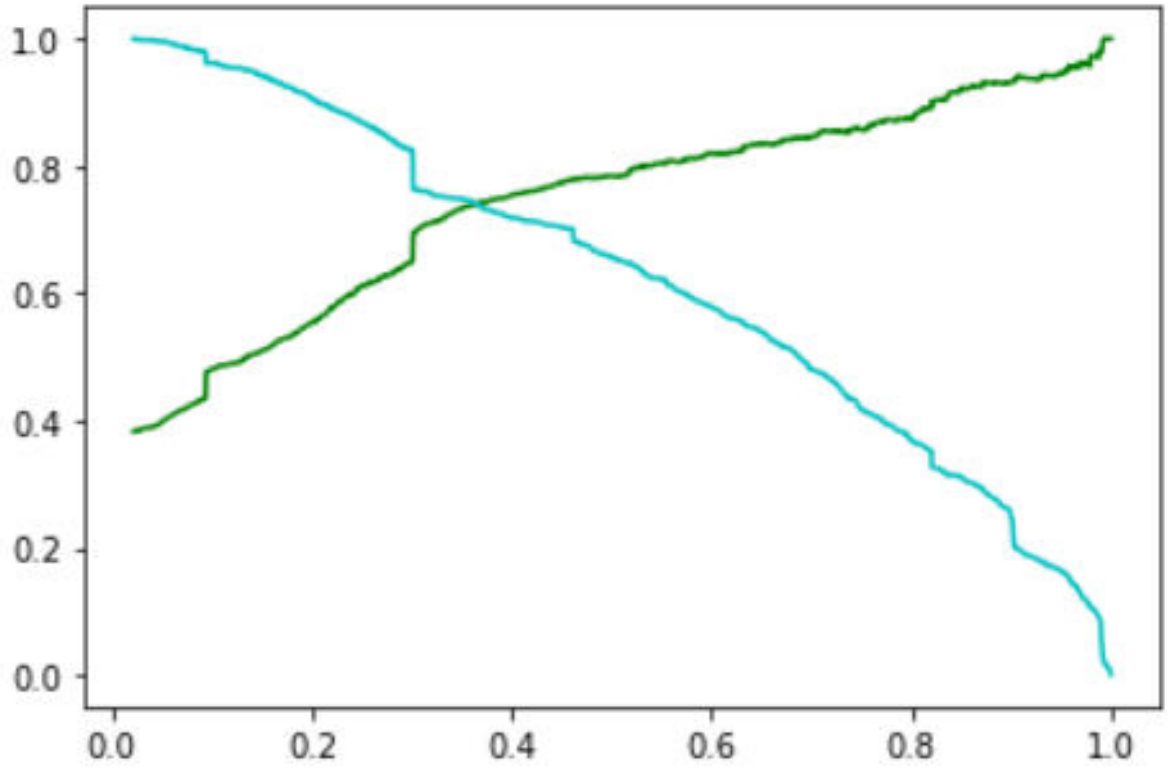


PREDICTION ON THE TEST SET

- Conversion rate on test set is 72%.
- The accuracy value on the test set is 0.81 which is a good value. The sensitivity value is 0.72 and specificity value is 0.86.

- **Precision and Recall on the test data**

The precision score is 0.76 and the recall score is 0.72. The precision and recall intersects at 0.39



CONCLUSION

The Accuracy, Sensitivity and Specificity values of the test set are around 81%, 72% and 86% which are approximately close to the values obtained from the train set implying that the model is good. The variable TotalVisits has the highest coefficient value implying that this variable has the highest significant positive influence on whether the lead will be converted or not. The variable total time spent on the website has a high coefficient value implying that it has the highest positive influence on whether the lead will be converted or not. Leads originating from add form also has a high positive impact on whether the lead will be converted or not. The lead source being Welingak Website also has a positive impact on whether a particular lead converts. Working professionals seem to have a high positive impact on whether a particular lead converts. The lead source being Olark Chat also has a positive impact on whether a particular lead converts. Variables that have a negative impact on whether a lead will be converted or not are the leads originating from landing page submission, direct traffic, Specialization being Select, don't email being yes, Lead Source being Referral Sites, specializations and occupations that are not provided and the last notable activity that is modified. Since the conversion rate is 74% on the train test and 72% on the test set which is a significant increase in the rate of conversion as compared to the conversion rate of 39%, the company X Education can look into the impact that these variables have on the conversion of leads and work towards improving the conversion rate by focusing on leads with a higher lead score.

THANK YOU