LEAD SCORE CASE STUDY

DEVELOPED BY:
ANKUR KUSHWAHA
ANKAN CHATTERJEE
ANKIT ABHISHEK

PROBLEM STATEMENT

- X is an Education company which is selling their online courses to the industrial professionals.
- The company marketing team generated leads from the forms field with the details of Information of the professionals who have browsed for the courses and landed on their websites
- They're getting lots Leads generated at initial stage but only 30% are paying customers or converted to the potential leads so they need help to select the most promising leads.
- The CEO of the company want target lead conversion to be 80% for which they are asking to build a model which provides

1. A higher lead score which have a higher conversion chance
2. A lower lead score have a lower conversion chance.¶

BUSINESS OBJECTIVE

1. 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.
2. A higher score would mean that the lead is hot, i.e. is most likely to convert know as ‘Hot Leads’
3. A lower score would mean that the lead is cold and will mostly not get converted.
4. 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.

STEPS INVOLVED

1. **IMPORTING AND INSPECTING DATASET**
2. **DATA CLEANING**
3. **EDA AND DATA PREPARATION**
4. **MODEL BUILDING**
5. **MODEL EVALUATION**
6. **CONCLUSION**

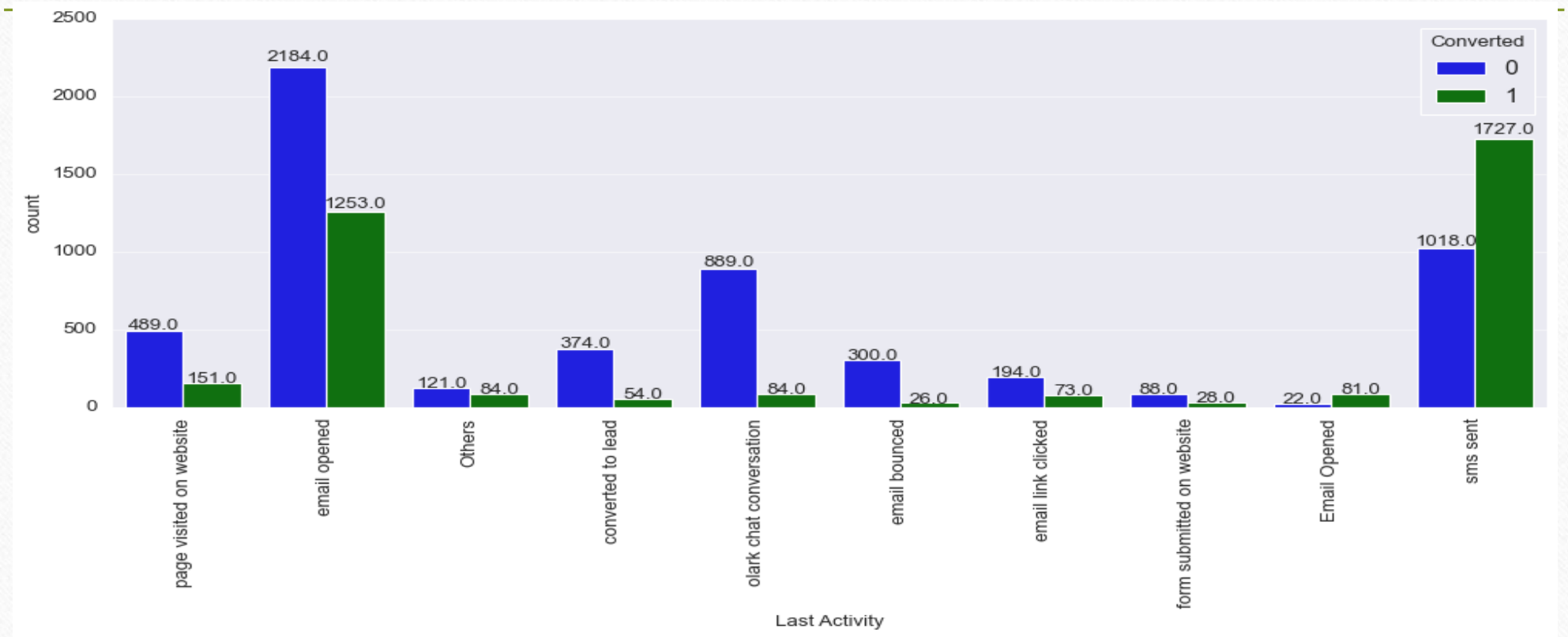
DATA CLEANING

- We removed the unnecessary columns from the data set eliminating the data set with high 40 % of null values
- The columns with skewed values were also removed
- After removing we were left with 13 columns to work ahead

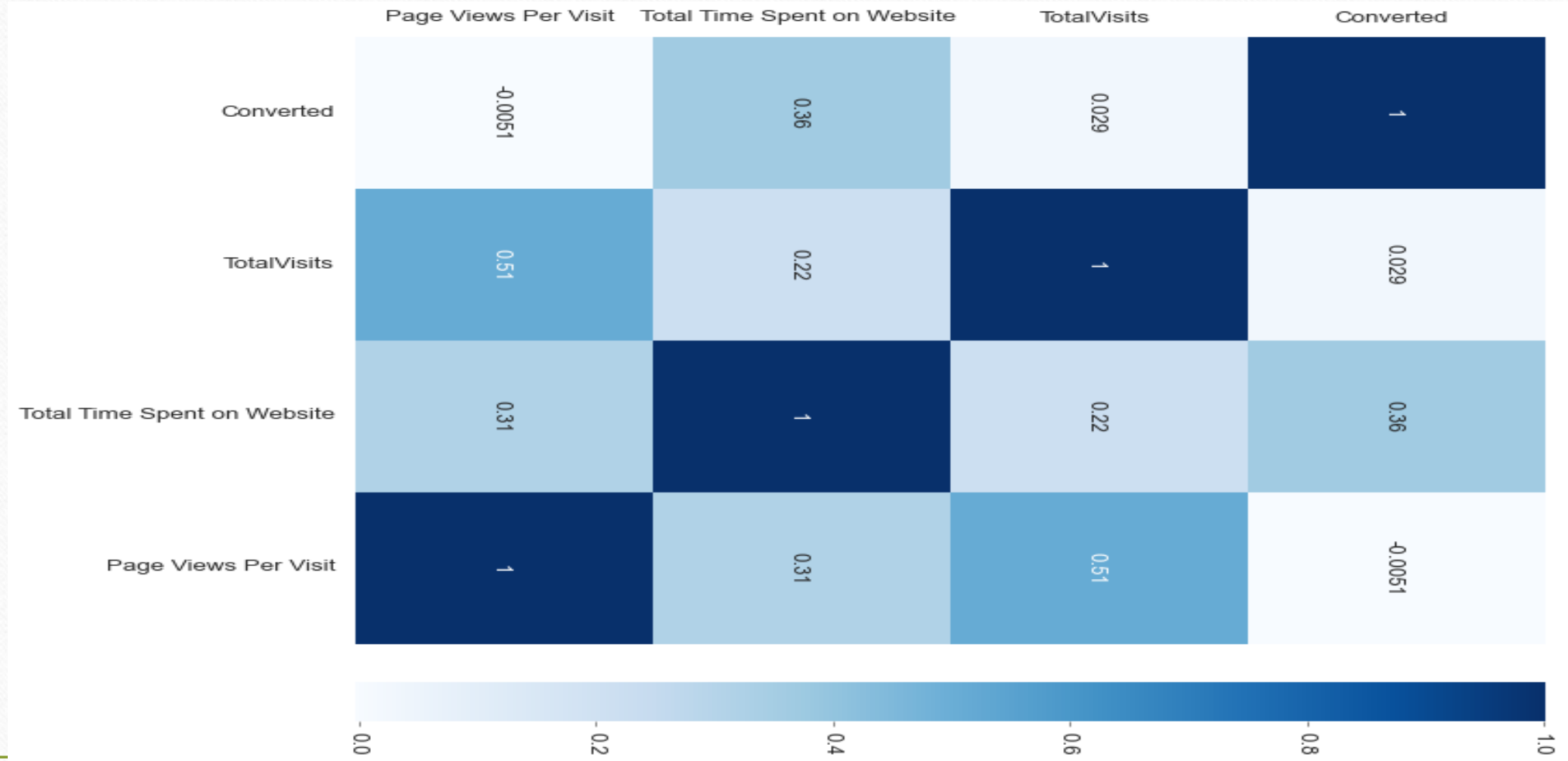
EDA

UNIVARIATE ANALYSIS

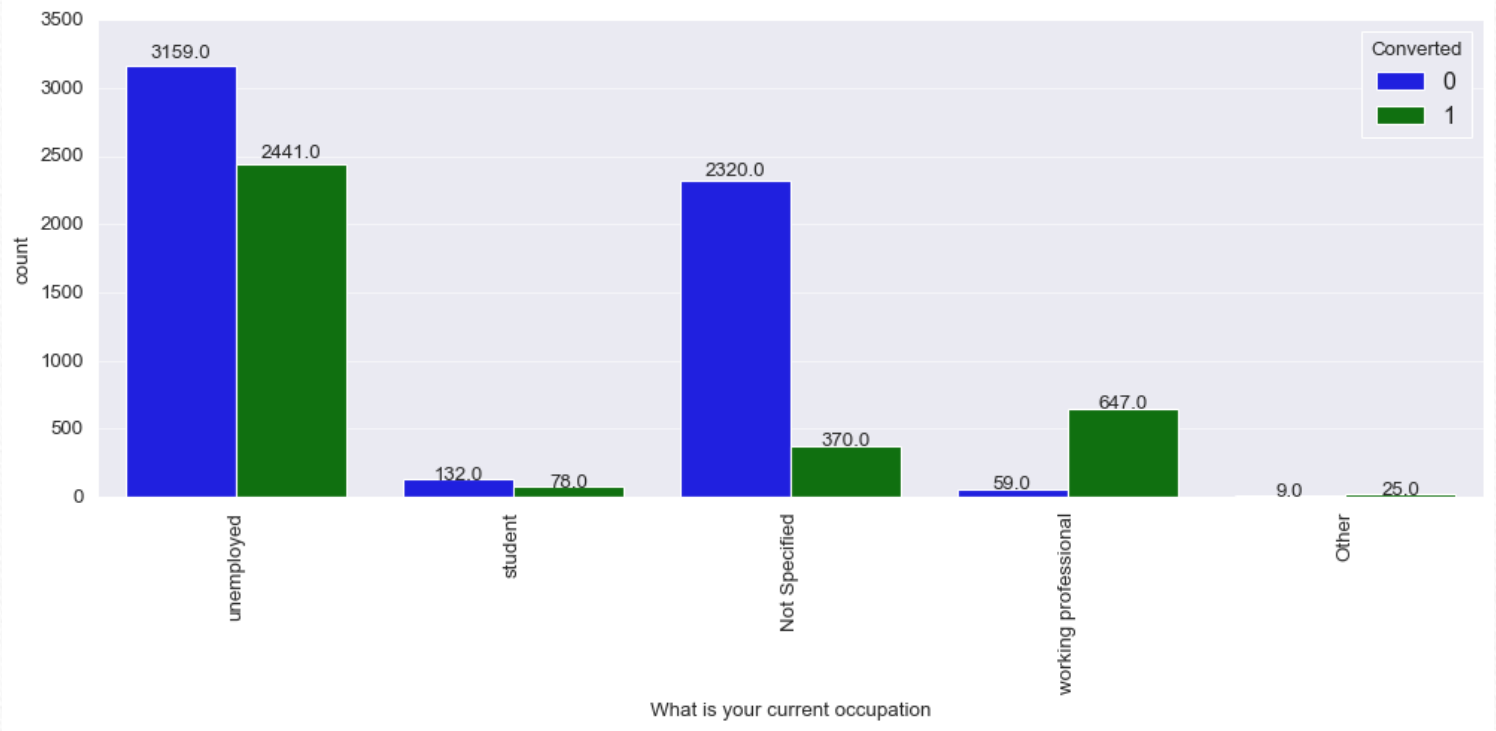
- **LAST ACTIVITY**



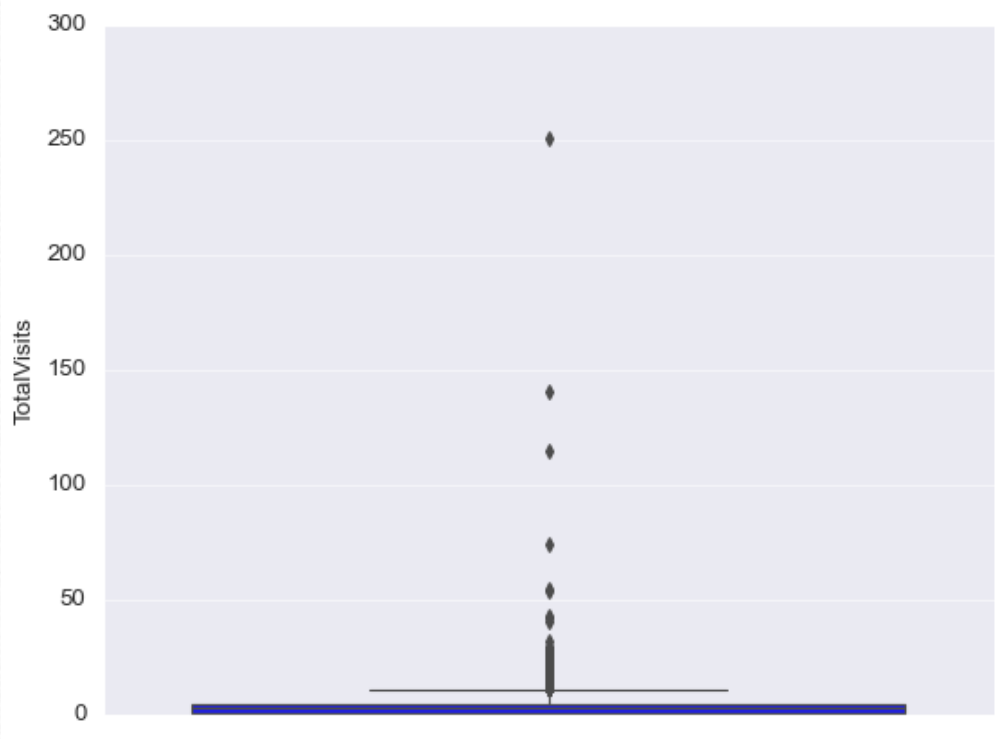
EDA (correlation between numerical variables)

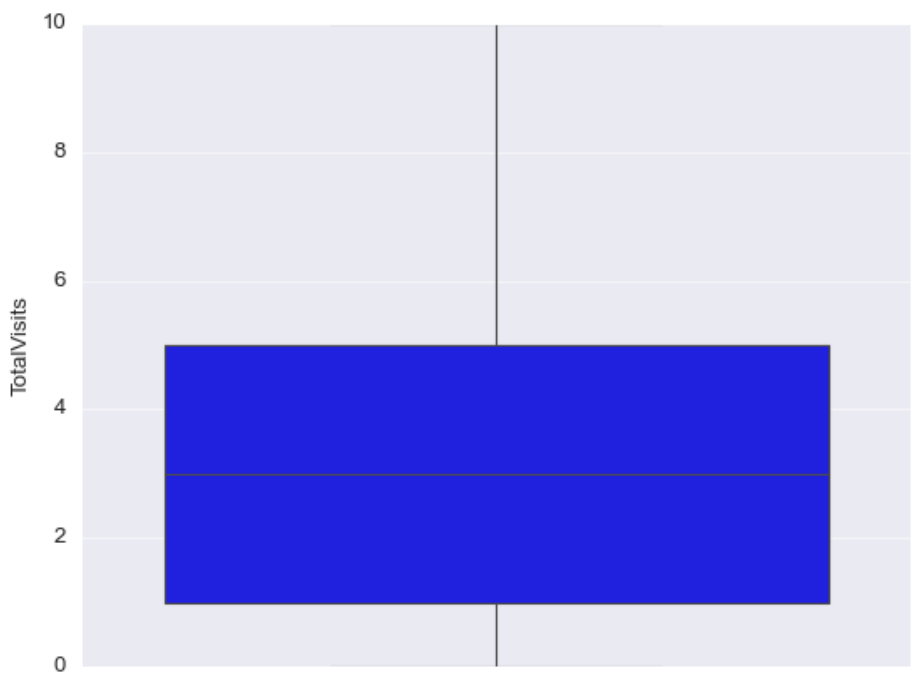


EDA
UNIVARIATE ANALYSIS (OCCUPATION)



OUTLIER TREATMENT
(TOTAL VISITS)
LEFT – BEFORE TREATMENT
RIGHT- AFTER TREATMENT



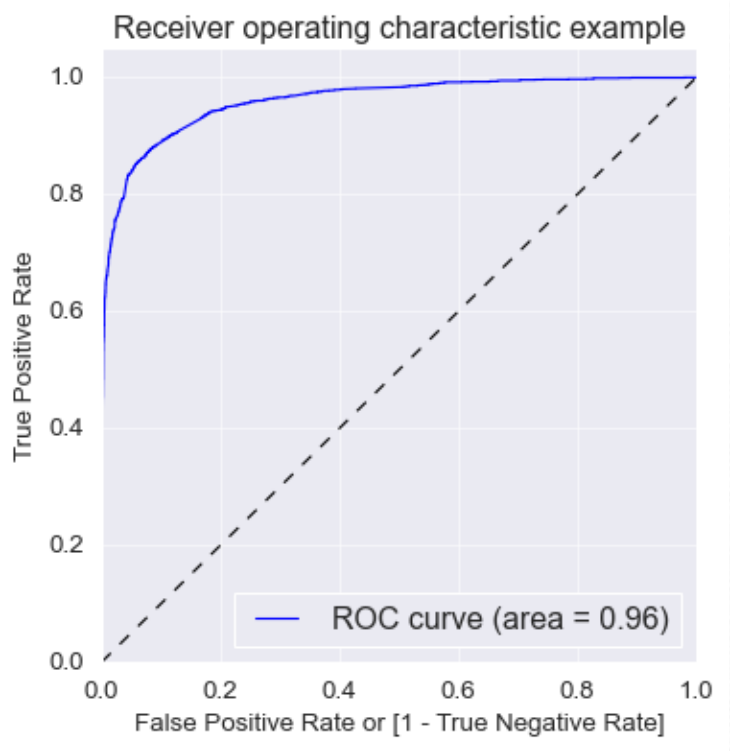


DATA PREPARATION

- Dummy variables were created for the categorical columns
- Scaling of the data was done
- The test train split on the data was carried out
- Features selection was done on the data set with the help of RFE
- VIF was also used to check the optimal features
- p-value 0.05 and RFE < 5

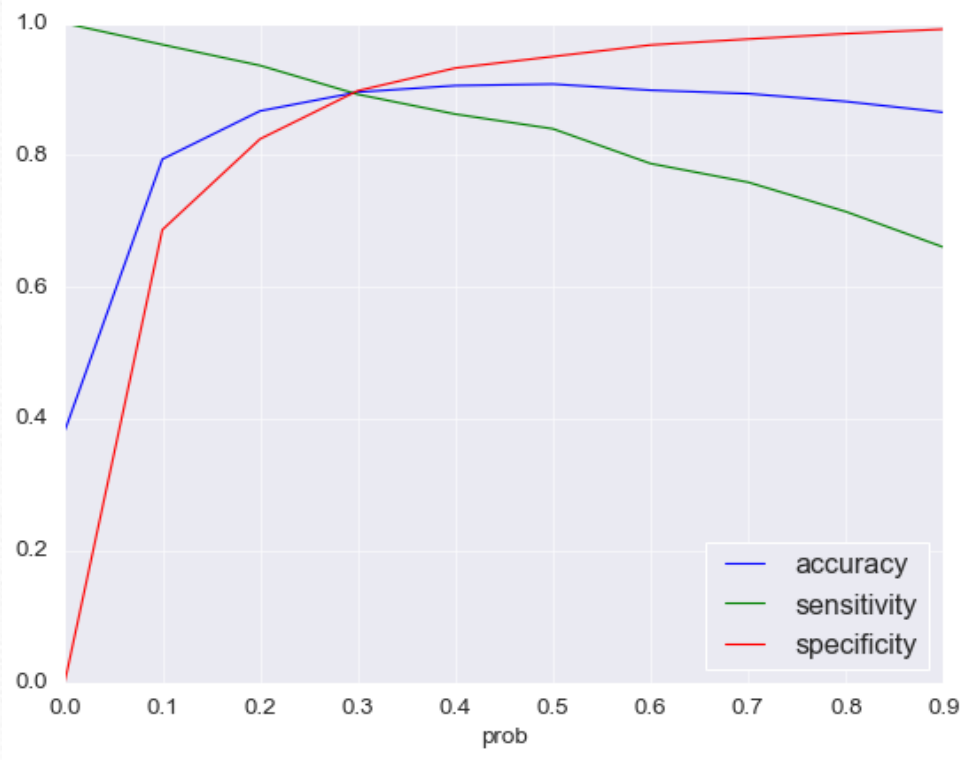
MODEL EVALUATION

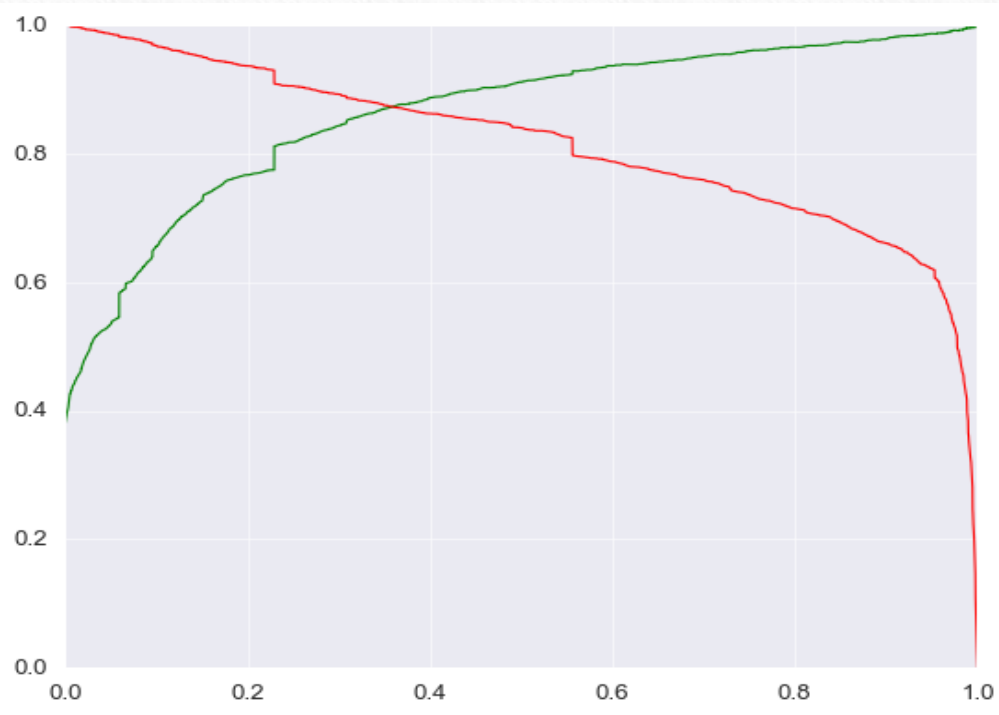
- **ROC Curve**



1. **Accuracy – 89.54%**
2. **Sensitivity – 89.21%**
3. **Specificity – 89.75%**
4. **Precision-84.29%**
5. **Recall – 89.21%**

Model evaluation Train





MODEL EVALUATION TEST

Accuracy - 89.61%

Sensitivity – 89.68%

Specificity – 89.56%

Precision – 84.87%

Recall – 89.68%

	Converted	Prospect ID	Converted_prob	Final_Predicted
0	1	4269	0.709067	1
1	1	2376	0.994489	1
2	1	7766	0.894005	1
3	0	9199	0.001614	0
4	1	4359	0.977198	1

Hot leads based on score

	const	Lead Score	Lead Number
2	1.0	99.0	660727
4	1.0	95.0	660681
6	1.0	98.0	660673
10	1.0	99.0	660608
12	1.0	98.0	660562

FEATURE SELECTED

Tags_closed by horizzon	6.685087
Tags_will revert after reading the email	4.223903
Lead Source_welingak website	3.573060
Last_Notable_Activity_had a phone conversation	2.665803
Lead Origin_lead add form	2.080225
Lead_Activity_sms sent	1.956086
Lead Source_olark chat	1.316219
Total Time Spent on Website	1.090308
Occupation_working professional	0.856425
Lead_Activity_email opened	0.518418
Lead_Activity_email bounced	-0.798591
Last_Notable_Activity_olark chat conversation	-0.913706
Last_Notable_Activity_modified	-1.033250
Tags_interested in other courses	-1.768735
const	-2.082527
Tags_ringing	-3.149726
Tags_already a student	-3.645769

RECOMMENDATIONS

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

- Tags_closed by horizzon
- Tags_will revert after reading the email
- Lead Source_welingak website

- The evaluation matrics are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.
- a. The model achieved a sensitivity of 89.21% in the train set and 89.68% in the test set, using a cut-off value of 0.345.
- b. The model also achieved an accuracy of 89.61%, which is in line with the study's objectives.The CEO of X Education had set a target sensitivity of around 80%.

THANK YOU

HOPE OUR RECOMMENDATION HELPS YOU