LEAD SCORING PROJECT

BY BIBEKANANDAN SAHOO AND RUSHI PANDYA

IMPORTANT OBSERVATIONS FROM EDA

OBSERVATIONS OF BIVARIATE ANALYSIS ON BOOLEAN VARIABLES

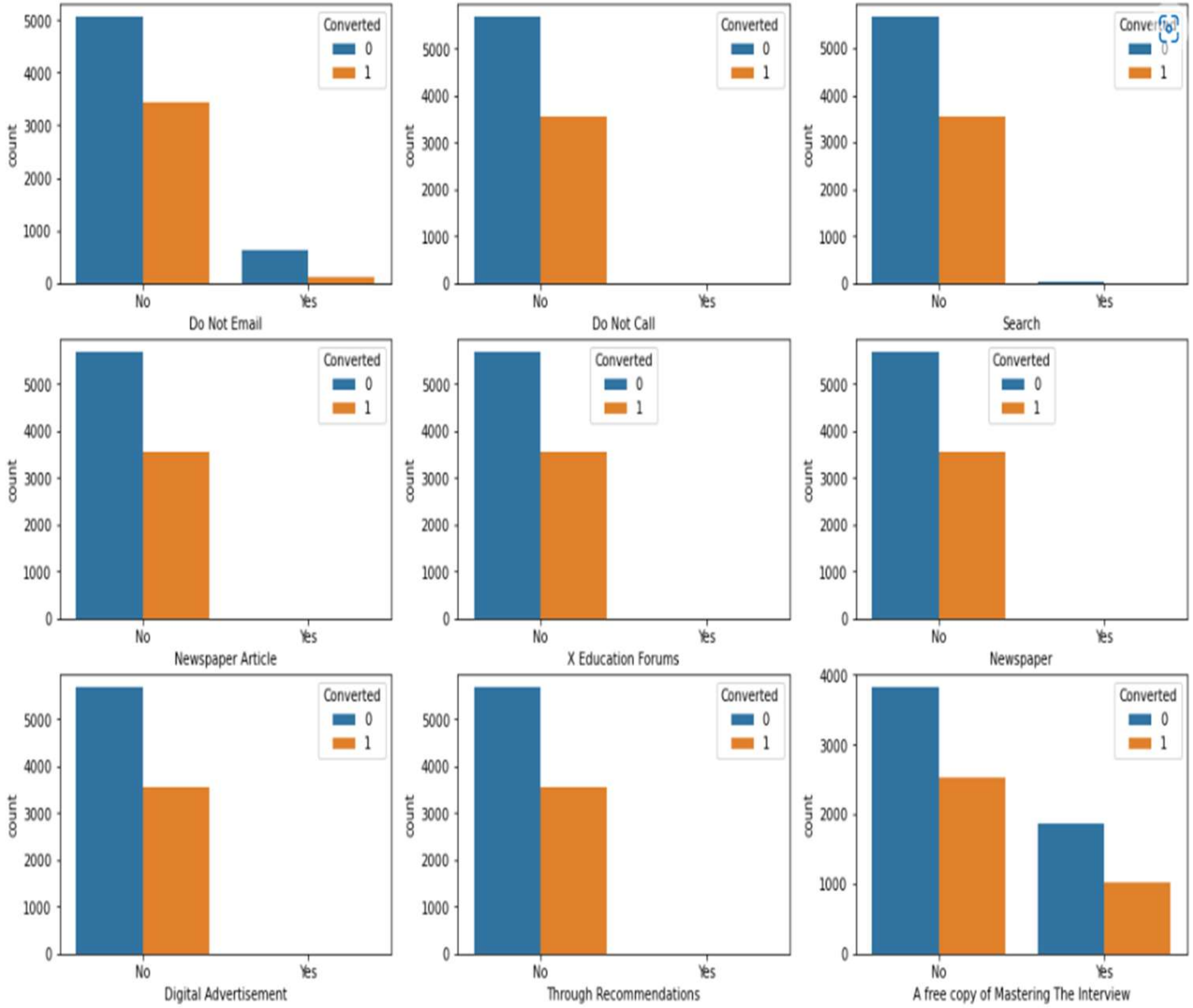
The Obervations here are very general but strengthen the understanding of the data

Some important points are:

The conversion is higher for Leads who said No for:

Do not email & Do not call

Also the conversion rate is higher for those who inquired through digital marketing.



OBSERVATIONS OF BIVARIATE ANALYSIS ON NUMERICAL VARIABLES

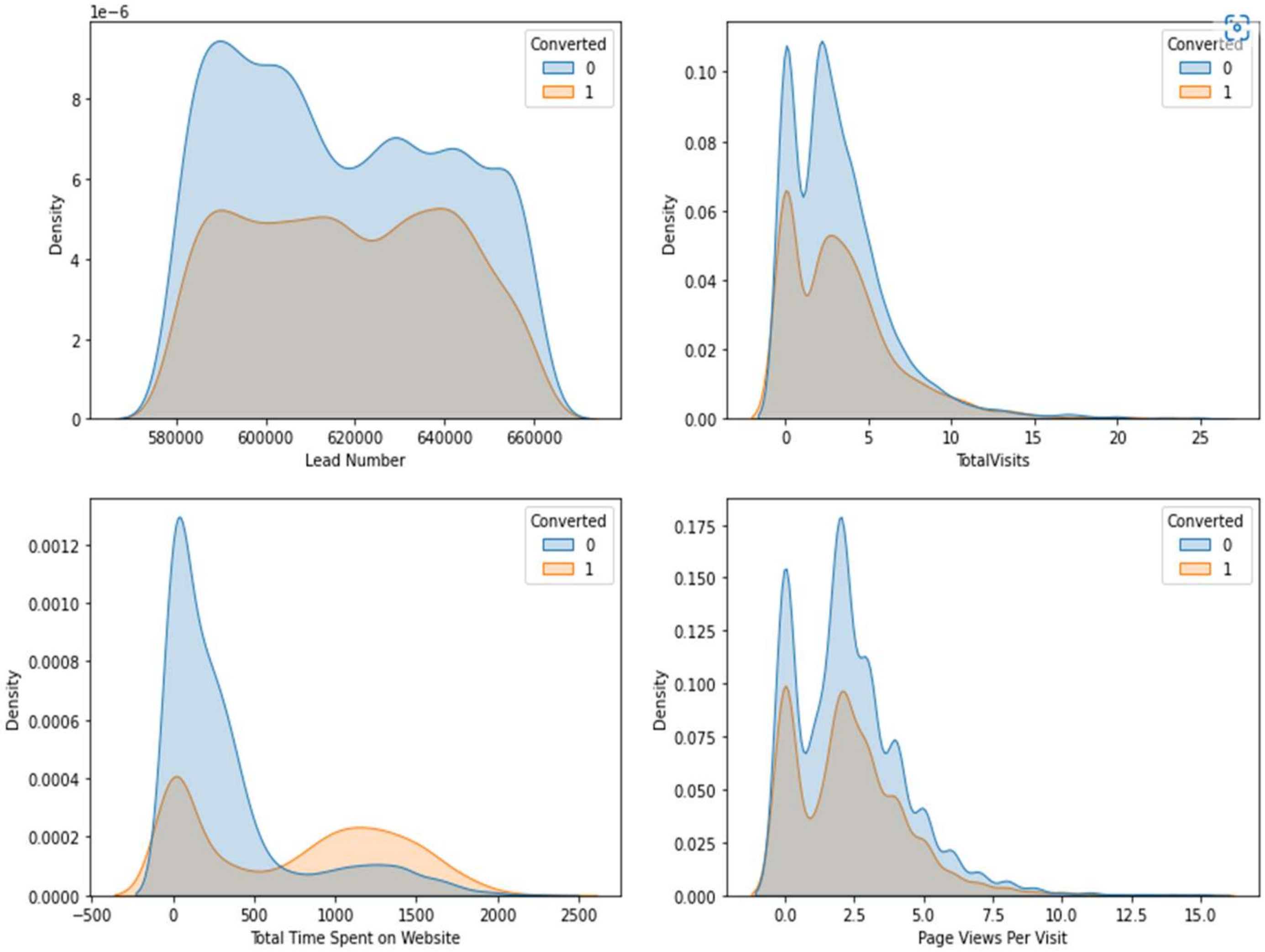
The following can be observed:

The conversion is higher if the following is observed:

1.The total time spent on the website is more.

2.The number of pages visited is more

3.The number of visits to the website is more.



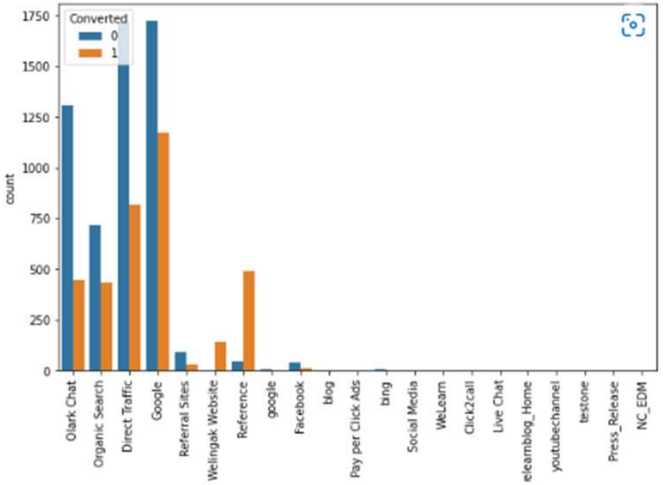
TREATMENT OF OBJECT VARIABLES BEFORE BIVARIATE ANALYSIS TO DERIVE MEANINGFUL INSIGHTS

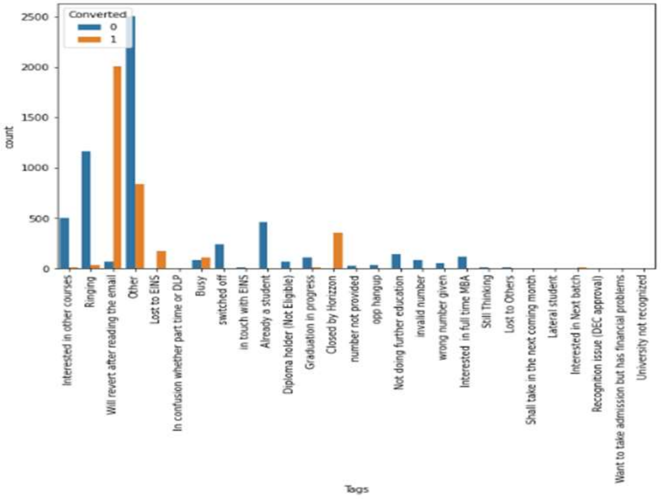
- For each object type variable we compute the percentage of converted in each subcategory
- We club the categories which are lower in percentage to “others” category
- By doing so we can concentrate more on the variables under each sub category under object variables of the dataset which comprise of larger percentages can an significantly impact our decision making process.

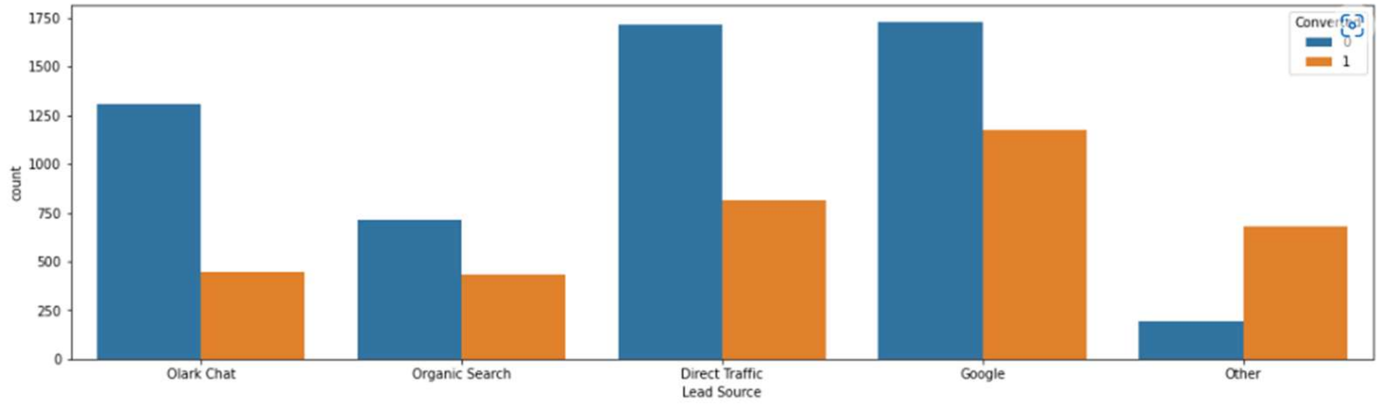
BEFORE AND AFTER TREATMENT OF OBJECT VARIABLES

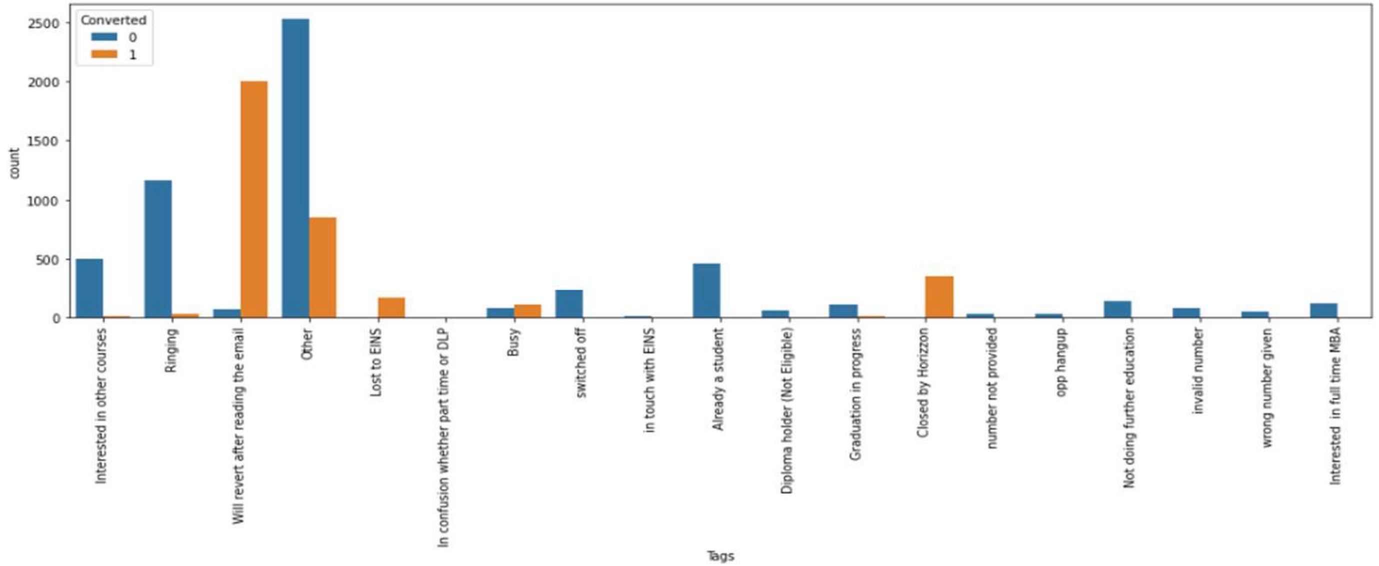
BEFORE CLUBBING SUB CATEGORIES

AFTER CLUBBING THE SUB CATEGORIES



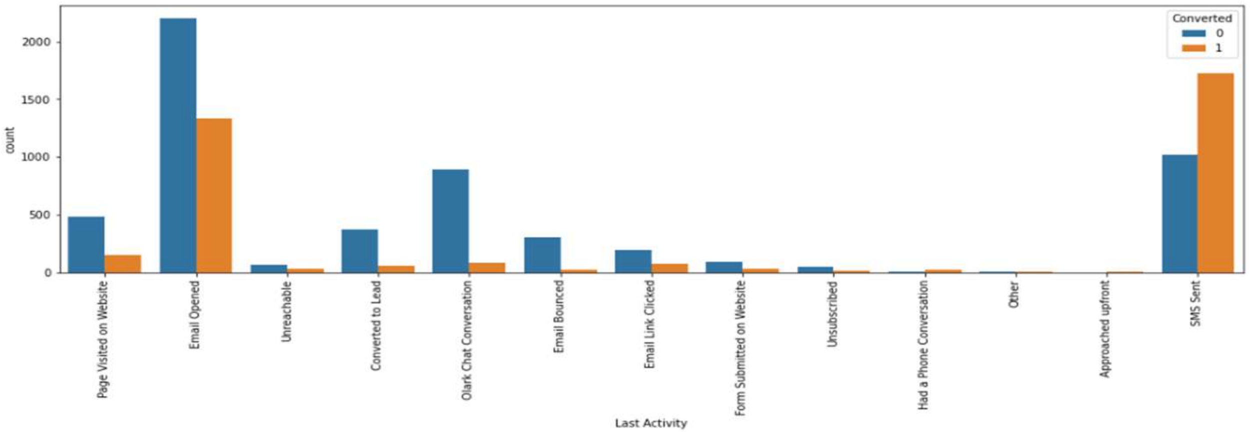


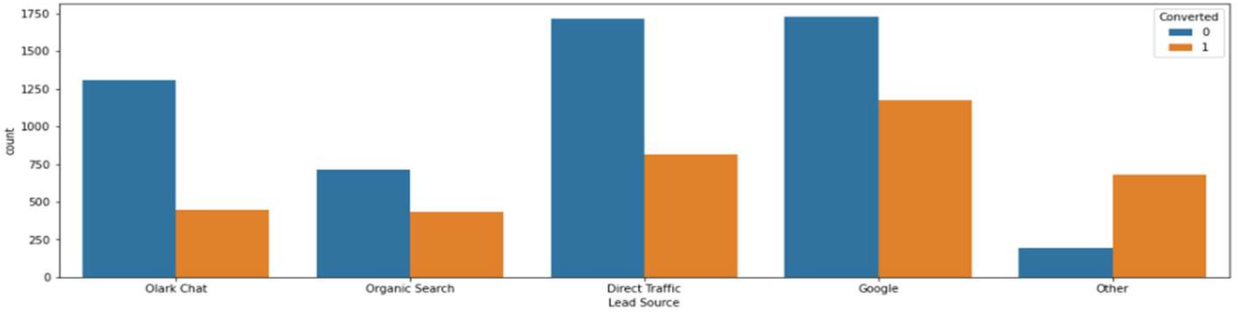


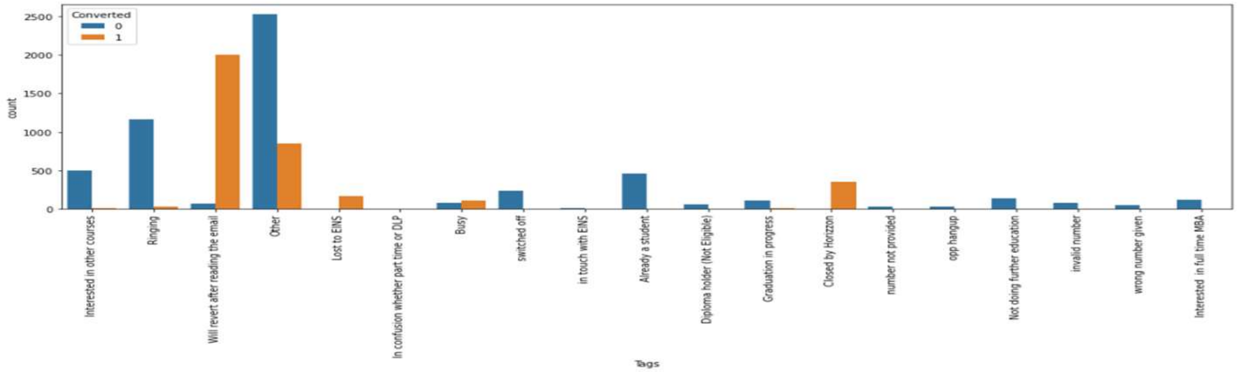


OBSERVATIONS ON ANALYSIS OF OBJECT VARIABLES

SR. NO	OBJECT VARIABLE	OBSERVATION
1.	Lead Score	Highest conversion of leads through olark chats,google & direct traffic lead.
2.	Last Activity	Highest conversion rates for :Email Opened & Olark Chat conversion
3.	Country	Highest conversion is for India
4.	Tags	Highest conversion is for Others and Ringing categories
5.	Last Notable Activity	Highest converstion rate for Modified,Email Opened And SMS Sent.







CONCLUSION FROM EDA

The following variables have been identified as important and should therefore be considered as a part of the model

- Total time spent on the website
- Lead Source
- Last Activity
- Last Notable Activity
- TAGS.

LOGISTIC REGRESSION MODEL

OBSERVATIONS DURING DATA PRE PROCESSING

- A multicollinearity > 80% is observed the following variables and therefore are decided to be dropped from the model:

Lead_Org_Landing Page Submission	OCC_Working Professional	Last Not Act_Email Opened	Last Not Act_SMS Sent	Last Not Act_Unsubscribed

- On performing RFE for 10,15 And 30 variables the following variables appeared to be most important

Lead Source	TAGS	Last Notable Activity	Lead Organisation
Last Activity	Specialization	Occupation	Do Not Mail
Time Spent on Website			

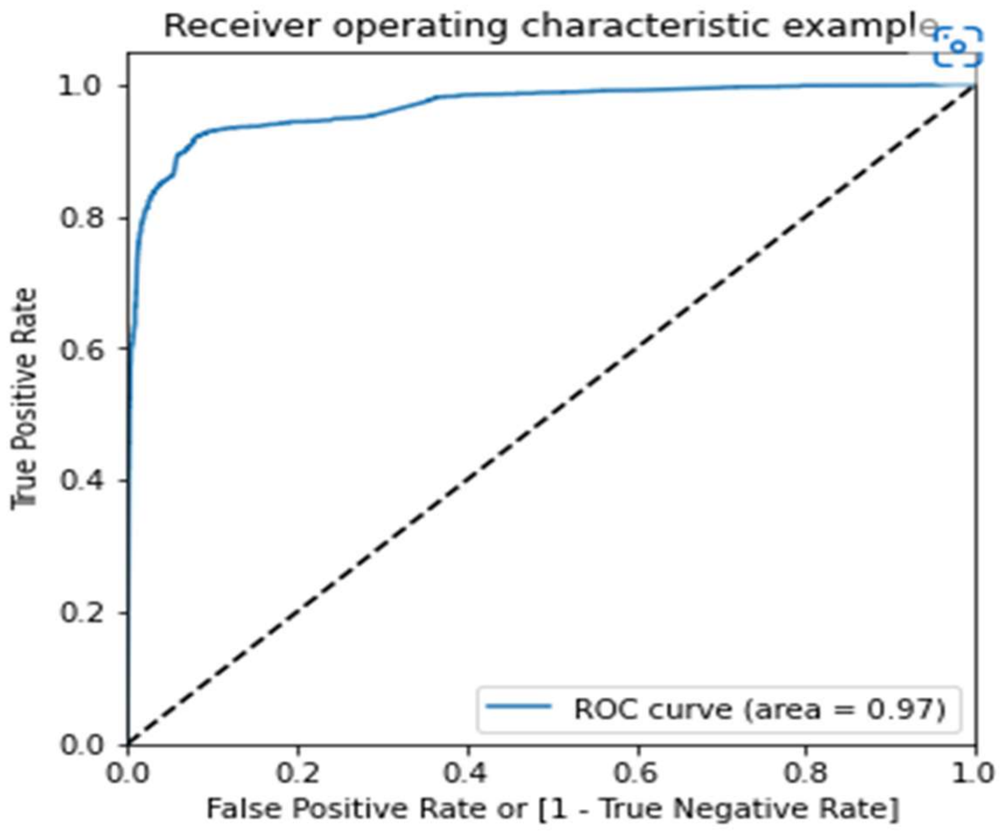
MODEL BUILDING PROCESS

		Model Outputs	p value	VIF	Remarks			Model Outputs	p value	VIF	Remarks			Model Outputs	p value	VIF
MODEL 1	The Model is built considering the 15 variables selected by RFE	Lead Org_Lead Add Form	0.00	1.20	keep	MODEL 2	Making the modification by dropping the variables with p values near 1 from Model 1	Lead Org_Lead Add Form	0	1.2	keep	FINAL MODEL	Adding the variable 'Total time spent on the website' as it was identified as important during EDA	Lead Org_Lead Add Form	0	1.3
		Last Act_SMS Sent	0	1.15	keep			Last Act_SMS Sent	0	1.15	keep			Last Act_SMS Sent	0	1.16
		Tags_Already a student	0	1.08	keep			Tags_Already a student	0	1.07	keep			Tags_Already a student	0	1.07
		Tags_Closed by Horizzon	0	1.22	keep			Tags_Closed by Horizzon	0	1.21	keep			Tags_Closed by Horizzon	0	1.27
		Tags_Interested in full time MBA	0.015	1.02	drop			Tags_Interested in other courses	0	1.09	keep			Tags_Interested in other courses	0	1.09
		Tags_Interested in other courses	0	1.11	keep			Tags_Lost to EINS	0	1.03	keep			Tags_Lost to EINS	0	1.05
		Tags_Lost to EINS	0	1.04	keep			Tags_Ringing	0	1.15	keep			Tags_Ringing	0	1.15
		Tags_Not doing further education	0.99	1.03	drop			Tags_Will revert after reading the email	0	1.33	keep			Tags_Will revert after reading the email	0	1.46
		Tags_Ringing	0	1.17	keep			Tags_Invalid number	0.001	1.01	keep			Tags_Invalid number	0.001	1.01
		Tags_Will revert after reading the mail	0	1.36	keep			Tags_Switched off	0	1.04	keep			Tags_Switched off	0	1.04
		Tags_Invalid number	0	1.01	keep			Last Not Act_Modified	0	1.16	keep			Last Not Act_Modified	0	1.16
		Tags_Number not provided	1	1.01	drop									Total time spent on the website	0	1.21
		Tags_Switched off	0	1.04	keep											
		Tags_Wrong number given	0.99	1.01	drop											
		Last Not Act_Modified	0	1.16	keep											

USING THE ROC TO FIND THE BEST CUT OFF

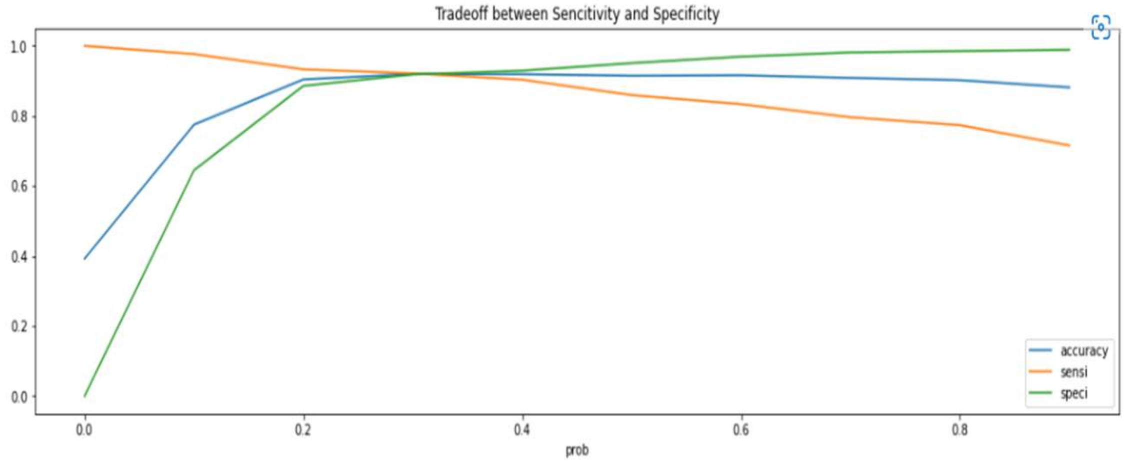
The area under the curve is 97%

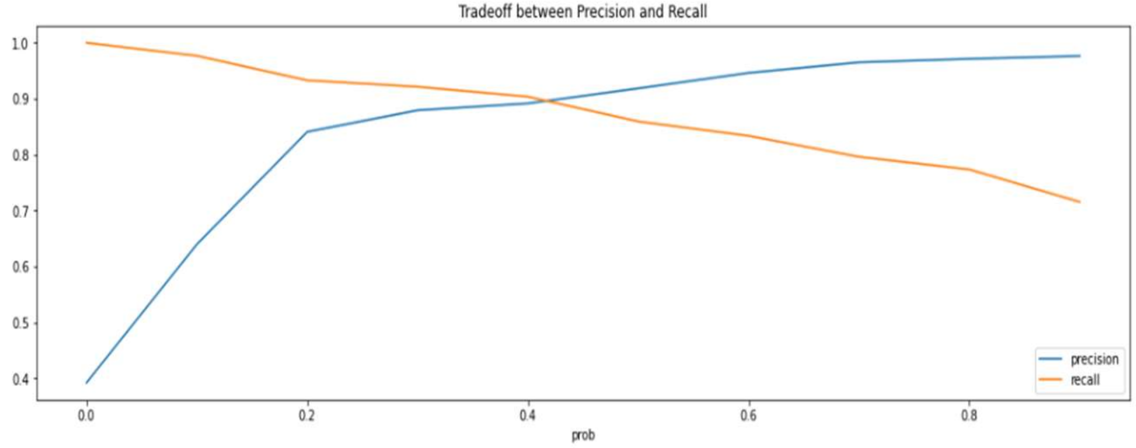
This means that the model is highly capable of distinguishing between the classes



DETERMINING THE SENSITIVITY OF THE MODEL

By observing the tradeoff between sensitivity and specificity we can conclude that the cutoff can be 27%





MODEL SUMMARY

The most important features of the model are:

Tags_Closed by Horizzon	Tags_Lost to EINS	Tags_Will revert after reading the email	Lead Org_Lead Add Form	Last Act_ SMS Sent	Total Time spent on website

The CUTOFF is decided at 0.27 taking into consideration ROC and the Tradeoffs
The 0.27 CUTOFF indicates that if the probability of a lead is > 0.27 the probability of conversion is high

Sr.No	Decision Matrix Parameters		Remarks
1	Accuracy	92%	This implies that the rate at which the model correctly identifies converted and not converted leads.
2	Sensitivity	92%	This implies how well we have identified the converted leads as hot leads
3	Specificity	91%	This implies how well we have been able to identify not converted leads
4	Precision	86%	This Implies how well we predicted the hot leads and how many were actually converted