




LEAD SCORE CASE STUDY



By
Shreejith S
Sudeep Das
Shubham Tripathi

BUSINESS PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS GOAL

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned 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%

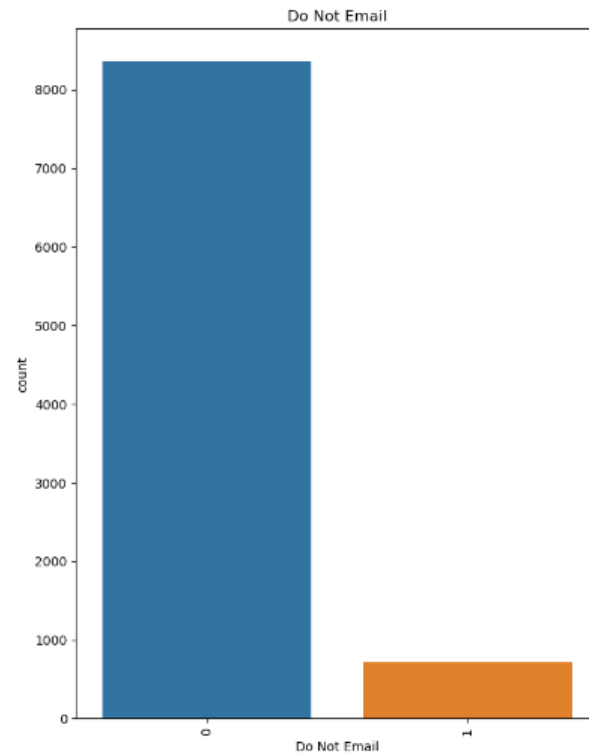
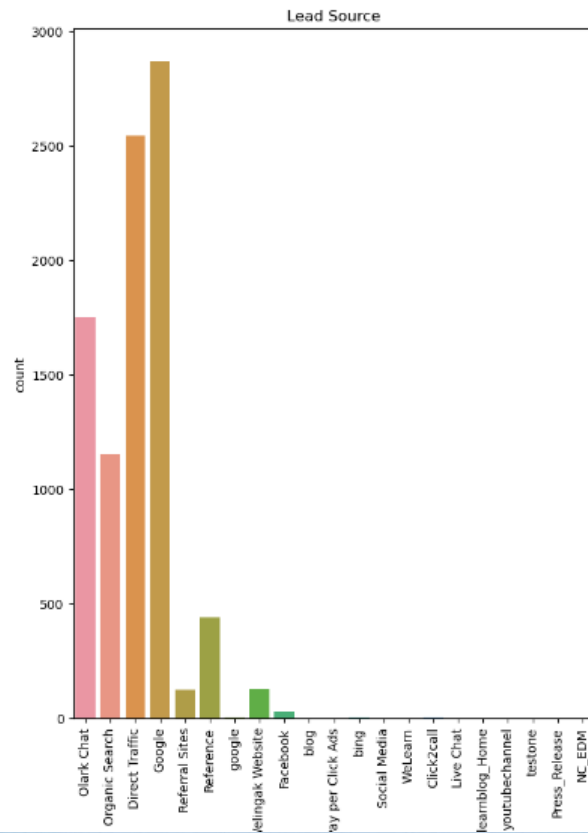
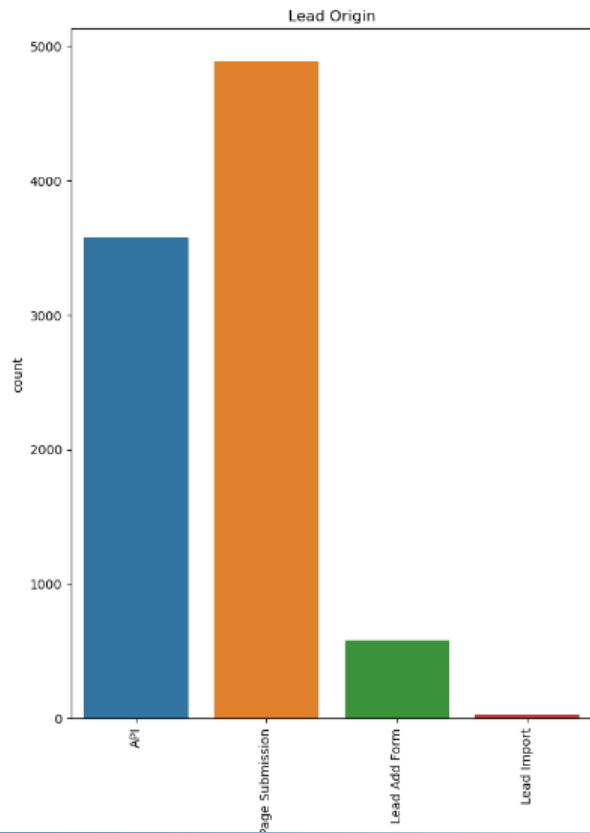
STRATEGY

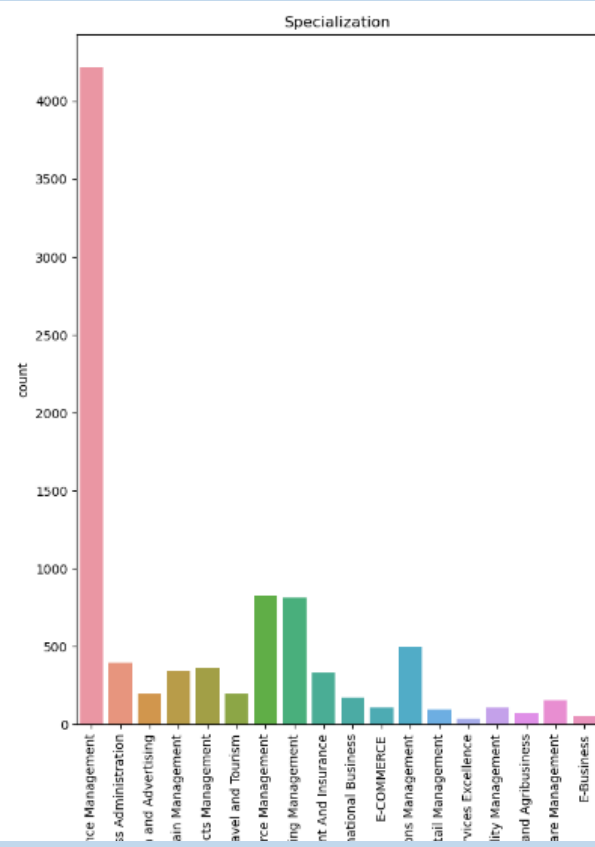
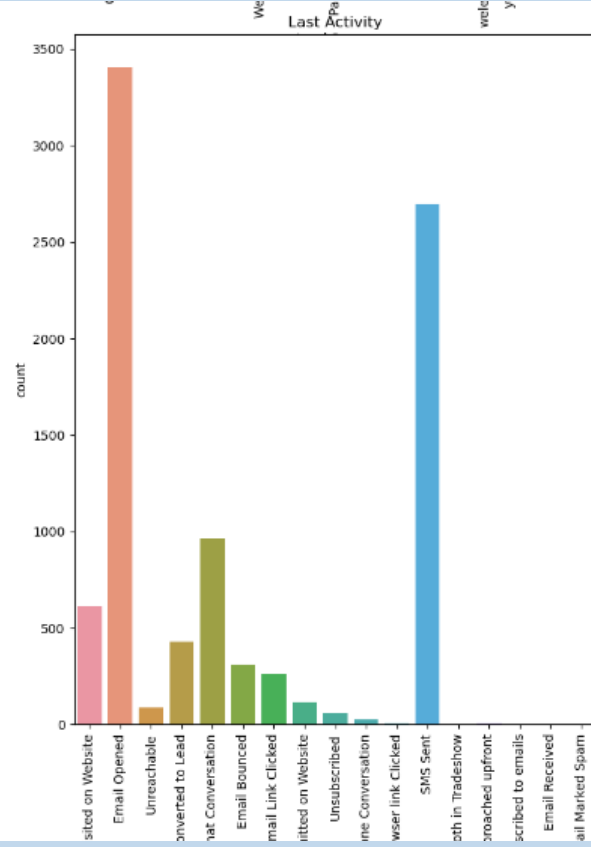
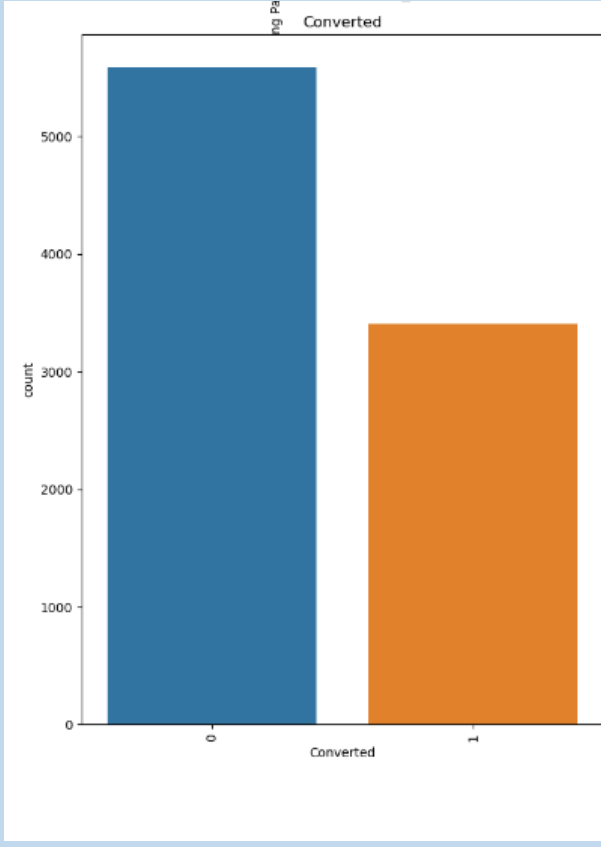
- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculating Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model on Test data

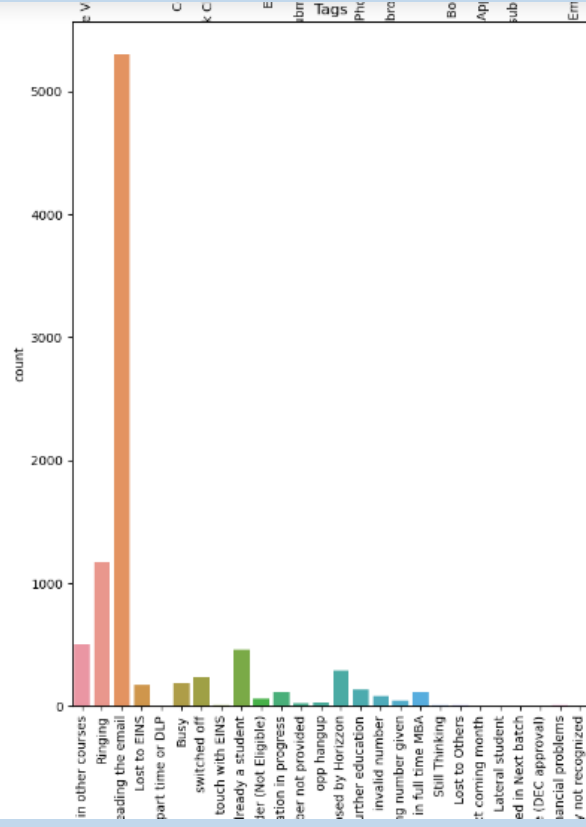
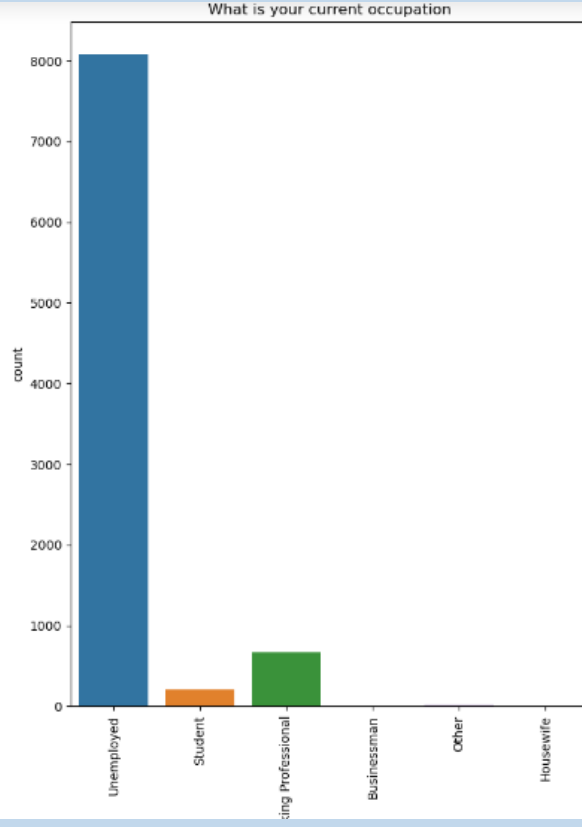


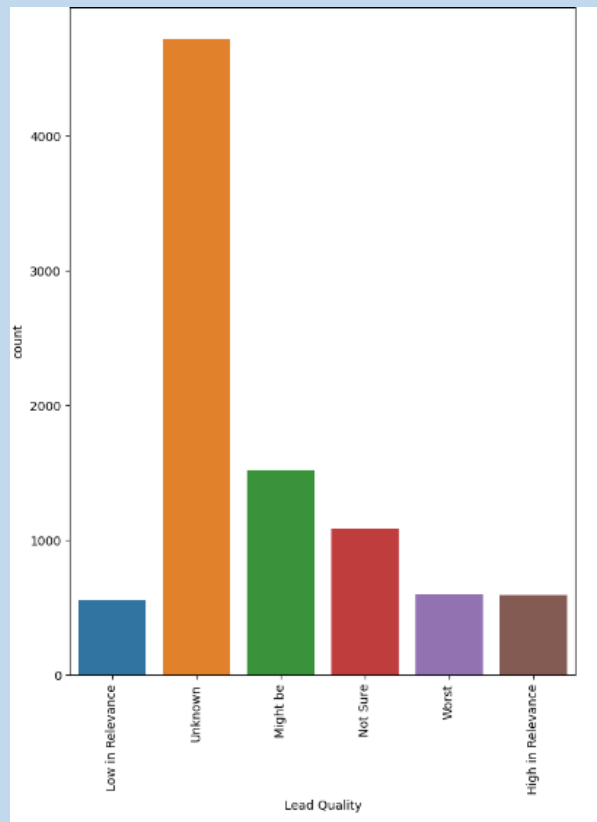
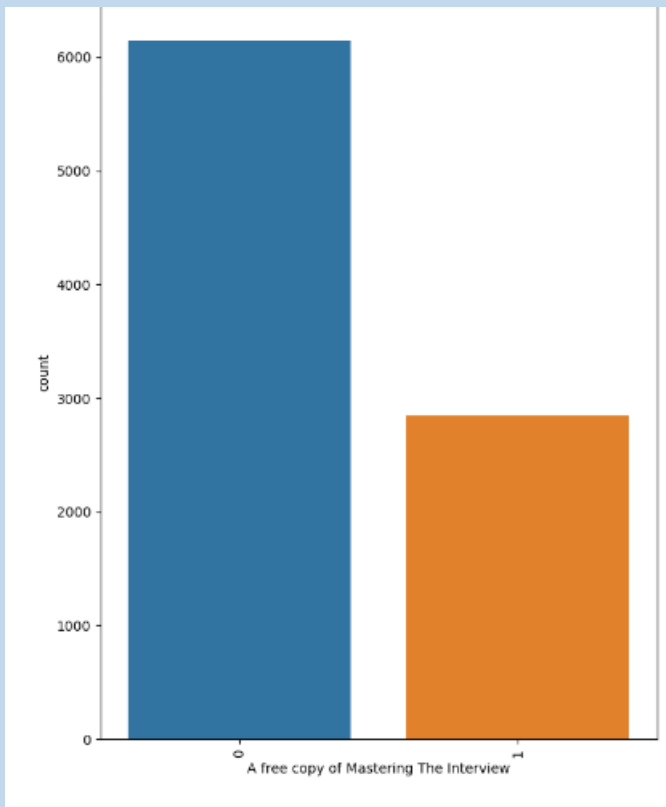
UNIVARIATE ANALYSIS

Categorical variables







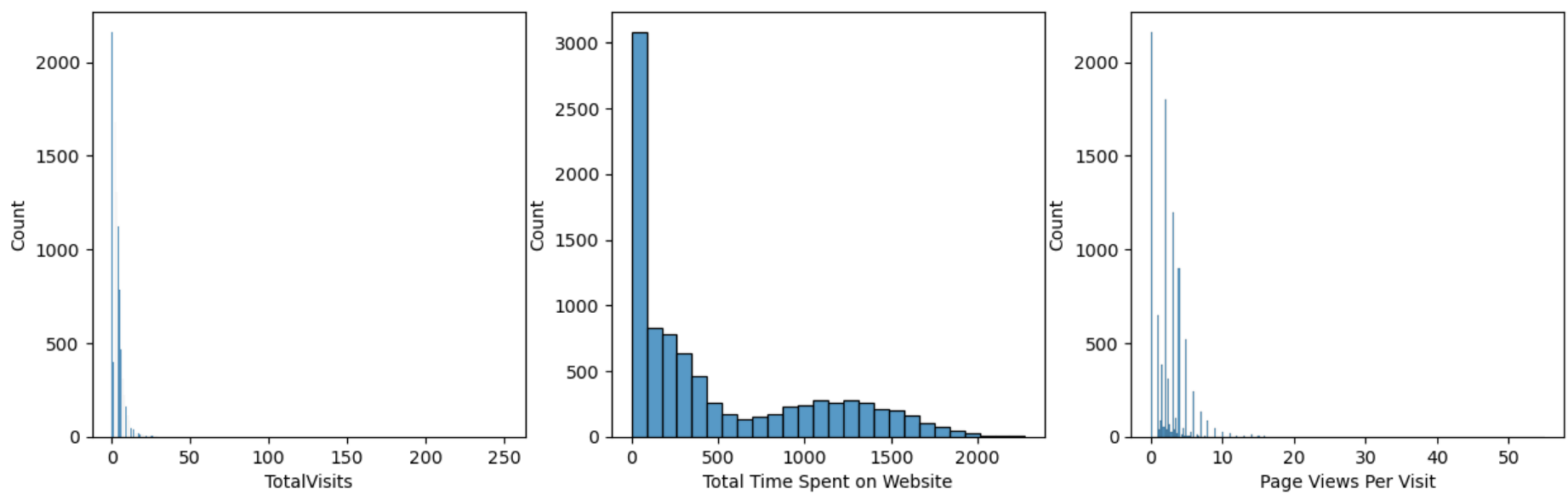


UNIVARIATE ANALYSIS

- Most of the lead source was found to be Google and naturally have originated from Landing page submission.
- Most of the leads want to get email's from the education institution.
- Number of leads converting to customers are significantly lesser compared to the ones rejecting the course.
- A large number of clients have tagged that they will revert back after reading the mail.
- Most of the leads belong to Finance management specialization and rest were specialized in the field of Resource management and Marketing management.
- Most of the enquiries are being done by unemployed people.
- Surprisingly large number of clients don't even want the free copy of Mastering the interview.
- Most of the lead quality is unknown which isn't desirable.

UNIVARIATE ANALYSIS

Numerical variables

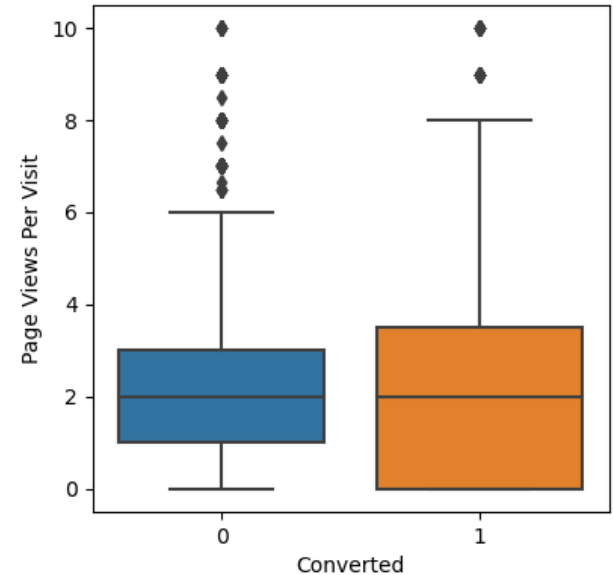
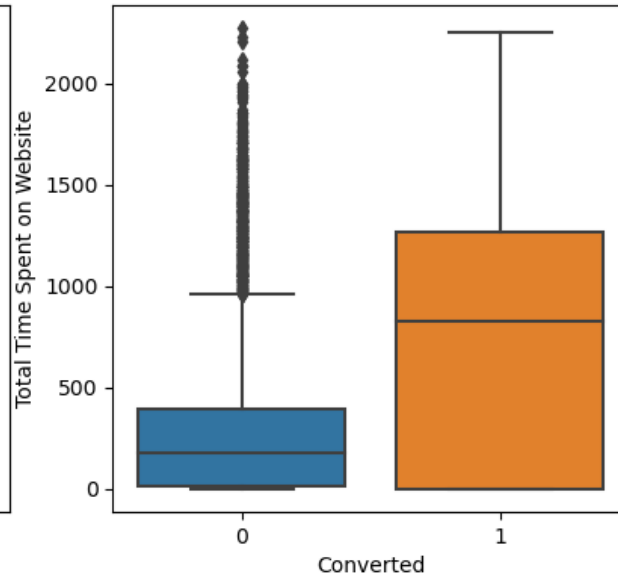
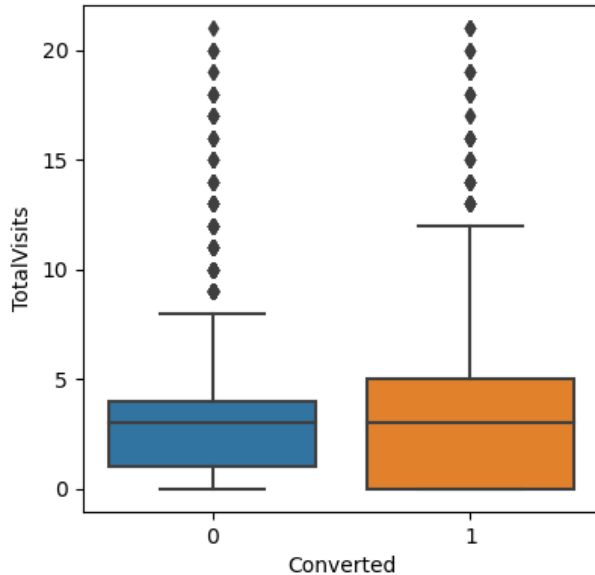


Numerical variables

- We see that most of the clients would visit the website less than 20 times and had spent a very short duration on the website about 5 minutes or lesser.
- Most of the people who visited the website had gone through less than 5 pages.

BI-VARIATE ANALYSIS

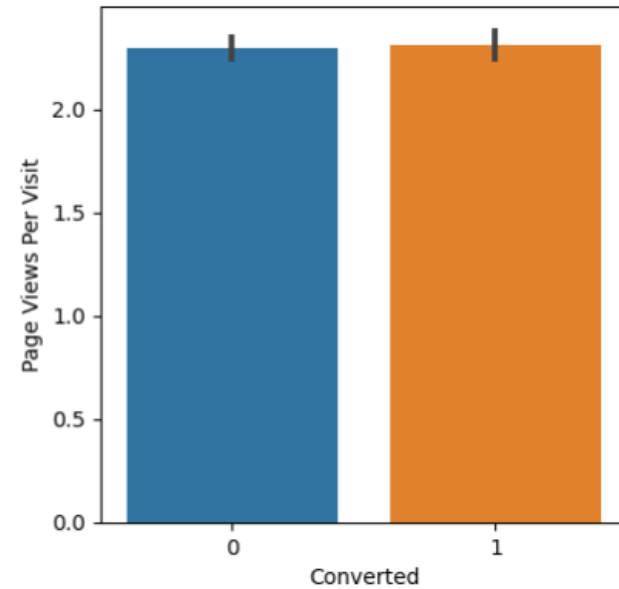
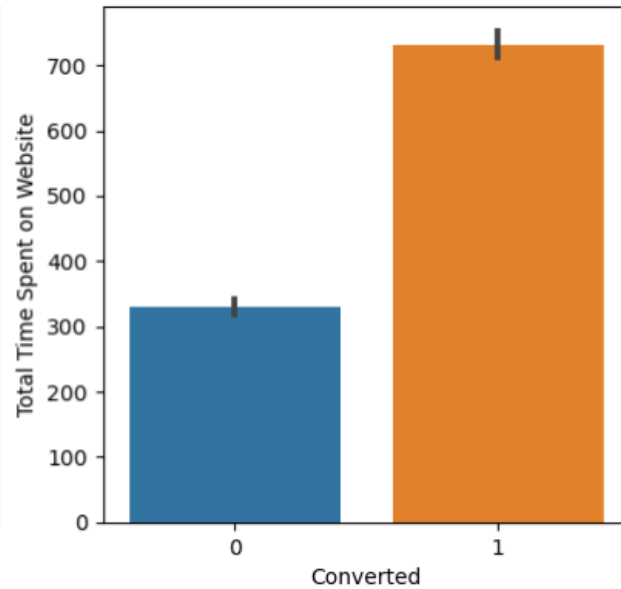
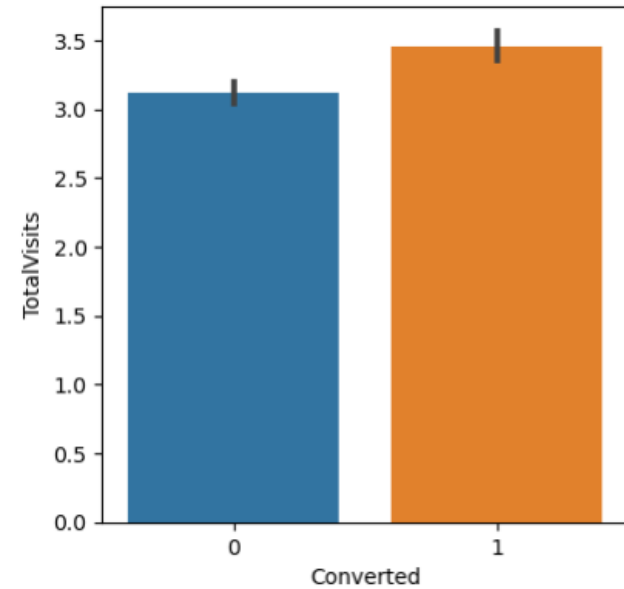
Target v/s numerical variables



BI-VARIATE ANALYSIS

- We see that range of Total visits among the leads which got converted is wider compared to the non converted ones.
- The median of the leads who took the course with respect to Total time spent on the website is higher compared to the ones who rejected the course.
- The median looks similar for converted and not converted leads in terms of pages viewed per visit but the range for converted leads is wider.

BI-VARIATE ANALYSIS

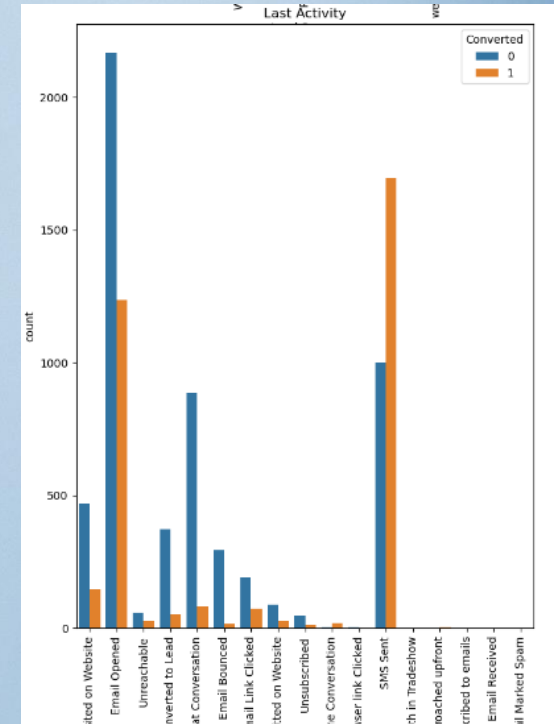
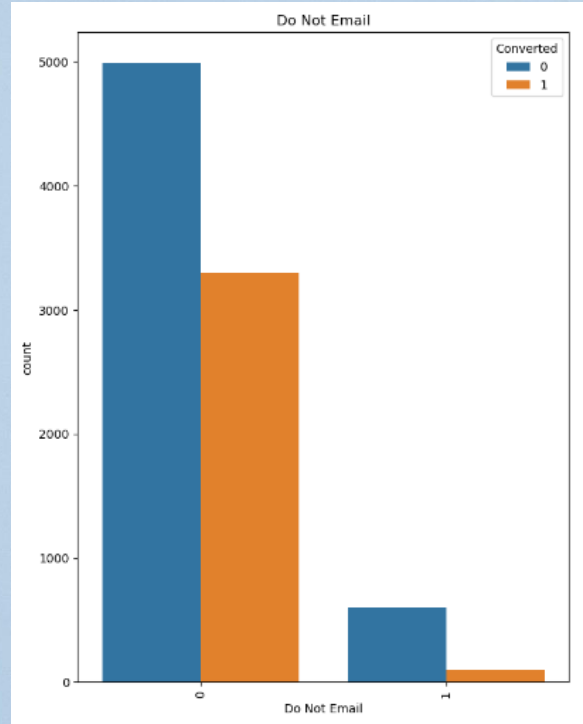
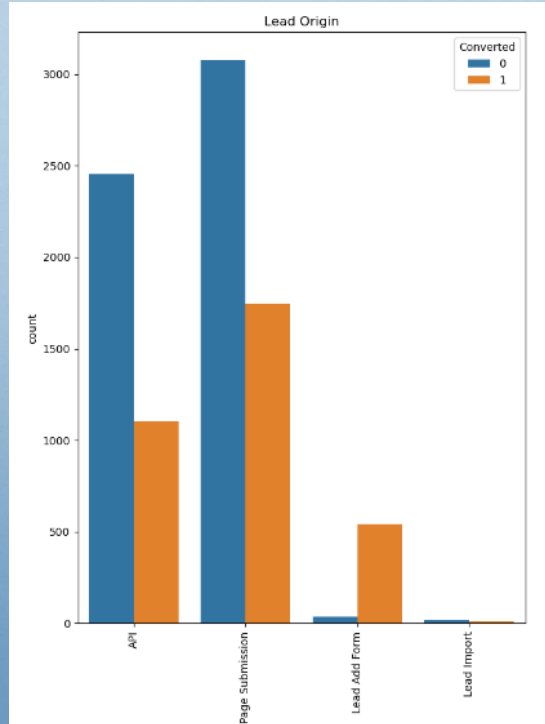


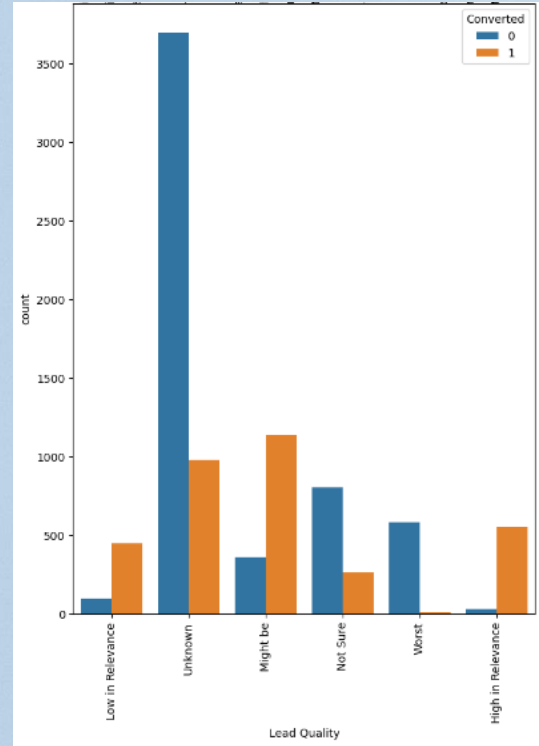
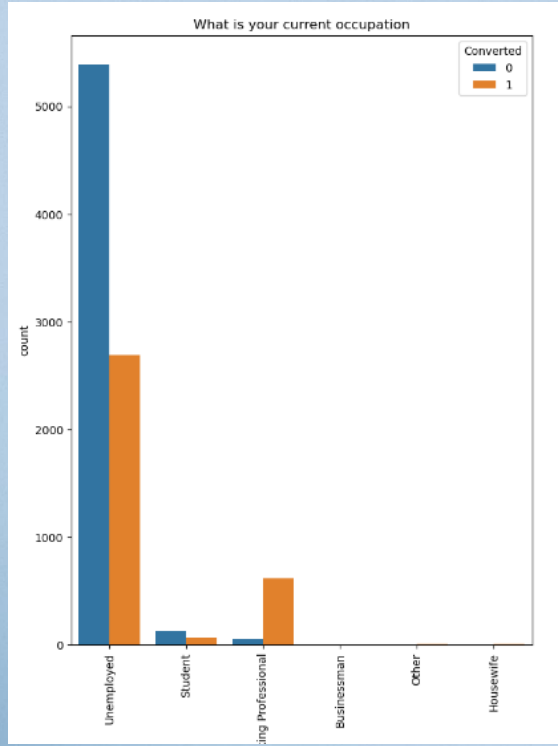
BI-VARIATE ANALYSIS

- People who spent more time on the website tend to opt the course.
- We see that number of pages viewed per visit doesn't have much influence on whether the lead gets converted or not, but if the total visits to the website are more then there are more chances of the lead getting converted as they might be visiting because they are interested in the course.

BI-VARIATE ANALYSIS

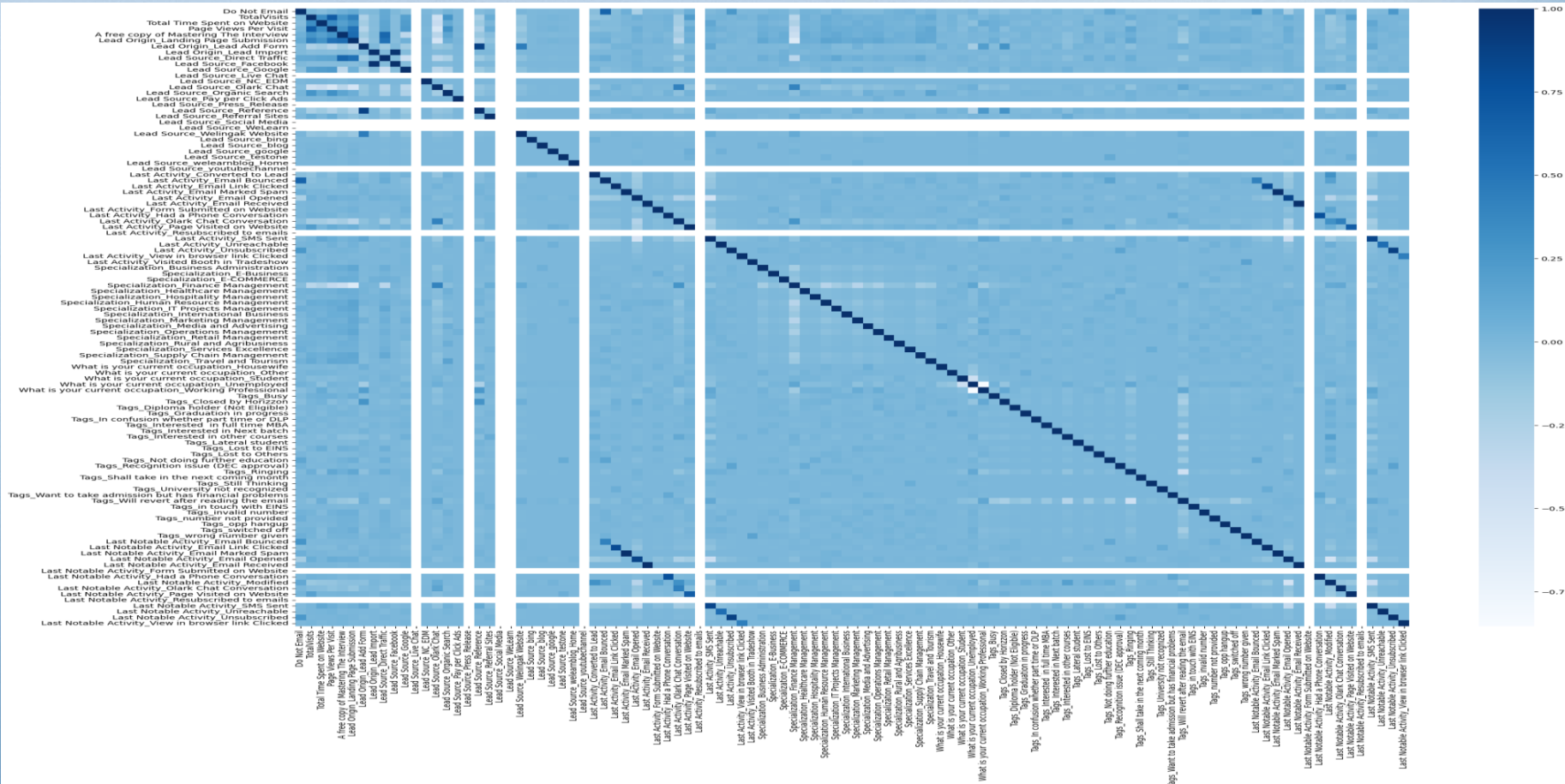
Target v/s categorical variables





- Among the people who filled the lead add form, a good proportion of them have opted the course , but as a whole the people who filled the form are less compared to API's or Landing Page.
- Emails don't seem to show any positive affect on lead conversion.
- There are more chances of a lead opting the course when the lead had sent a SMS as last activity.
- Most of the people enquiring about the courses were unemployed and a large majority of them didn't opt for the course.
- We see that most of the leads whose relevance(quality) was unknown didn't opt for the course.

CORRELATION



MODEL BUILDING

Dep. Variable:	Converted	No. Observations:	6292
Model:	GLM	Df Residuals:	6278
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1334.6
Date:	Sun, 18 Aug 2024	Deviance:	2669.2
Time:	19:29:26	Pearson chi2:	6.77e+04
No. Iterations:	9	Pseudo R-squ. (CS):	0.5943
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-2.3422	0.164	-14.270	0.000	-2.664	-2.021
Total Time Spent on Website	3.2004	0.211	15.162	0.000	2.787	3.614
Lead Origin_Lead Add Form	1.7442	0.461	3.784	0.000	0.841	2.648
Lead Source_Welingak Website	3.1142	0.864	3.603	0.000	1.420	4.808
Last Activity_SMS Sent	1.8415	0.109	16.888	0.000	1.628	2.055
Tags_Closed by Horizzon	10.1493	1.073	9.456	0.000	8.046	12.253
Tags_Lost to EINS	9.6154	0.660	14.578	0.000	8.323	10.908
Tags_Ringing	-2.5564	0.274	-9.321	0.000	-3.094	-2.019
Tags_Will revert after reading the email	4.5759	0.210	21.752	0.000	4.164	4.988
Tags_invalid number	-2.4462	1.069	-2.287	0.022	-4.542	-0.350
Tags_switched off	-2.7671	0.549	-5.042	0.000	-3.843	-1.691
Last Notable Activity_Modified	-1.6861	0.120	-14.095	0.000	-1.921	-1.452
Lead Quality_Unknown	-4.1683	0.167	-24.923	0.000	-4.496	-3.841
Lead Quality_Worst	-3.3024	0.685	-4.820	0.000	-4.645	-1.959

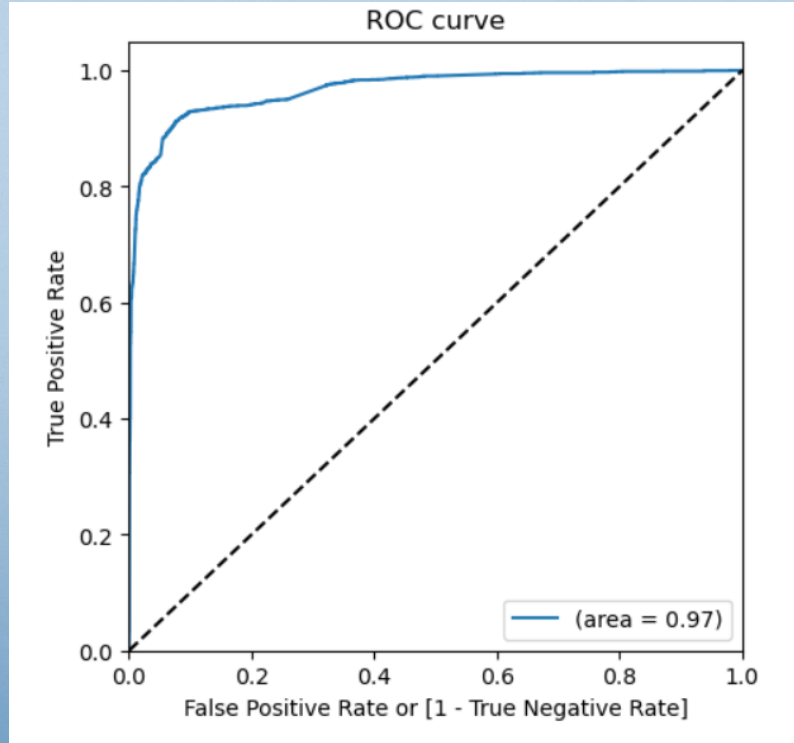
With the help of RFE, we identify top 15 features initially and then build the model.

Out of those 15 features we dropped the features 'Tags_number not provided' and 'Tags_wrong number given' as their p-value was higher than 0.05.

Checking the VIF values and we would be removing the variable having VIF value higher than 5. But as none of the variables are having $VIF > 5$ we will keep the variables.

	variables	values
0	const	11.71
8	Tags_Will revert after reading the email	2.42
7	Tags_Ringing	1.80
2	Lead Origin_Lead Add Form	1.58
13	Lead Quality_Worst	1.45
12	Lead Quality_Unknown	1.38
5	Tags_Closed by Horizon	1.36
3	Lead Source_Welingak Website	1.31
1	Total Time Spent on Website	1.19
4	Last Activity_SMS Sent	1.17
10	Tags_switched off	1.17
11	Last Notable Activity_Modified	1.14
6	Tags_Lost to EINS	1.13
9	Tags_invalid number	1.05

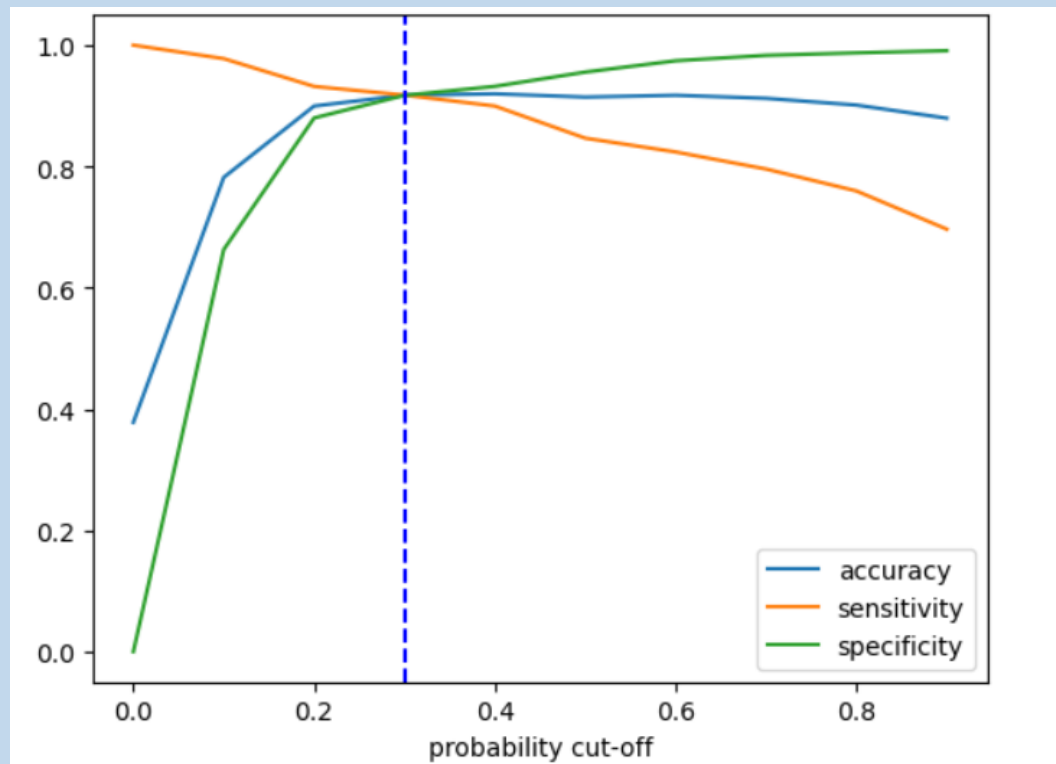
EVALUATION



After building the final model and 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 plotted on the left side, the area score is 0.97 which is a great score.

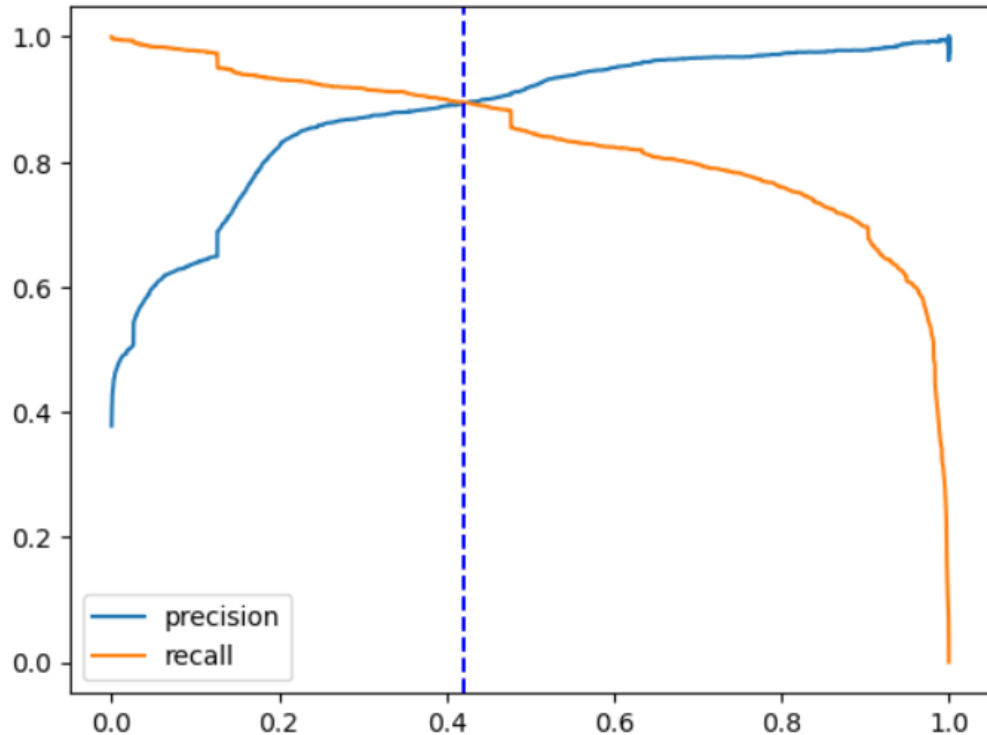
And our graph is more towards the left side of the border which means we have good accuracy.

CUT OFF POINT



We found that at 0.3 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.

PRECISION AND RECALL TRADE OFF POINT



We created a graph which will show us the trade off between Precision and recall. We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.42.

CONCLUSION

The model has achieved commendable performance across key metrics, with its accuracy, precision, and recall scores falling well within the desired range.

Furthermore, the model exhibits significant flexibility, ensuring it can seamlessly integrate and evolve in line with our company's future strategic objectives.

The factors that hold the greatest sway over potential buyers, listed in order of importance, are:

1. Tags_Closed by Horizon.
2. Tags_Lost to EINS.
3. Tags_Will revert after reading the email
4. Total Time Spent on Website
5. Lead Source_Welingak Website

Given these factors, X Education is well-positioned to capitalize on a significant opportunity, by strategically leveraging these insights. X Education has the potential to effectively convert a substantial portion of potential buyers into course purchasers, thereby increasing their market share and revenue streams.



THANK YOU