Lead Scoring Case Study

Submitted By- Shubhangi Khare

CONTENTS

- Problem Statement
- Problem Approach
- EDA
- Correlation
- Model Evaluation
- Observations
- Conclusion

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%.
- Although 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.

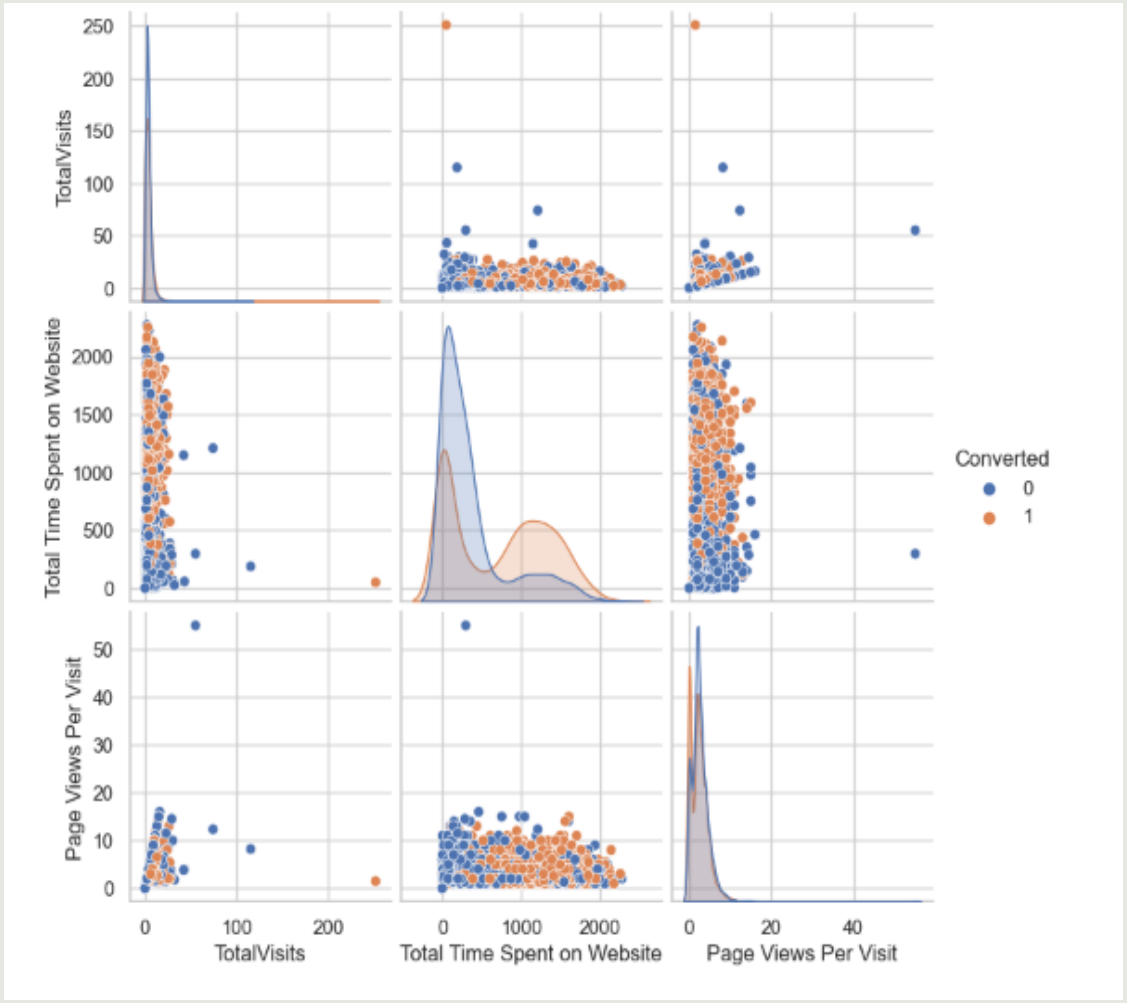
Problem Approach

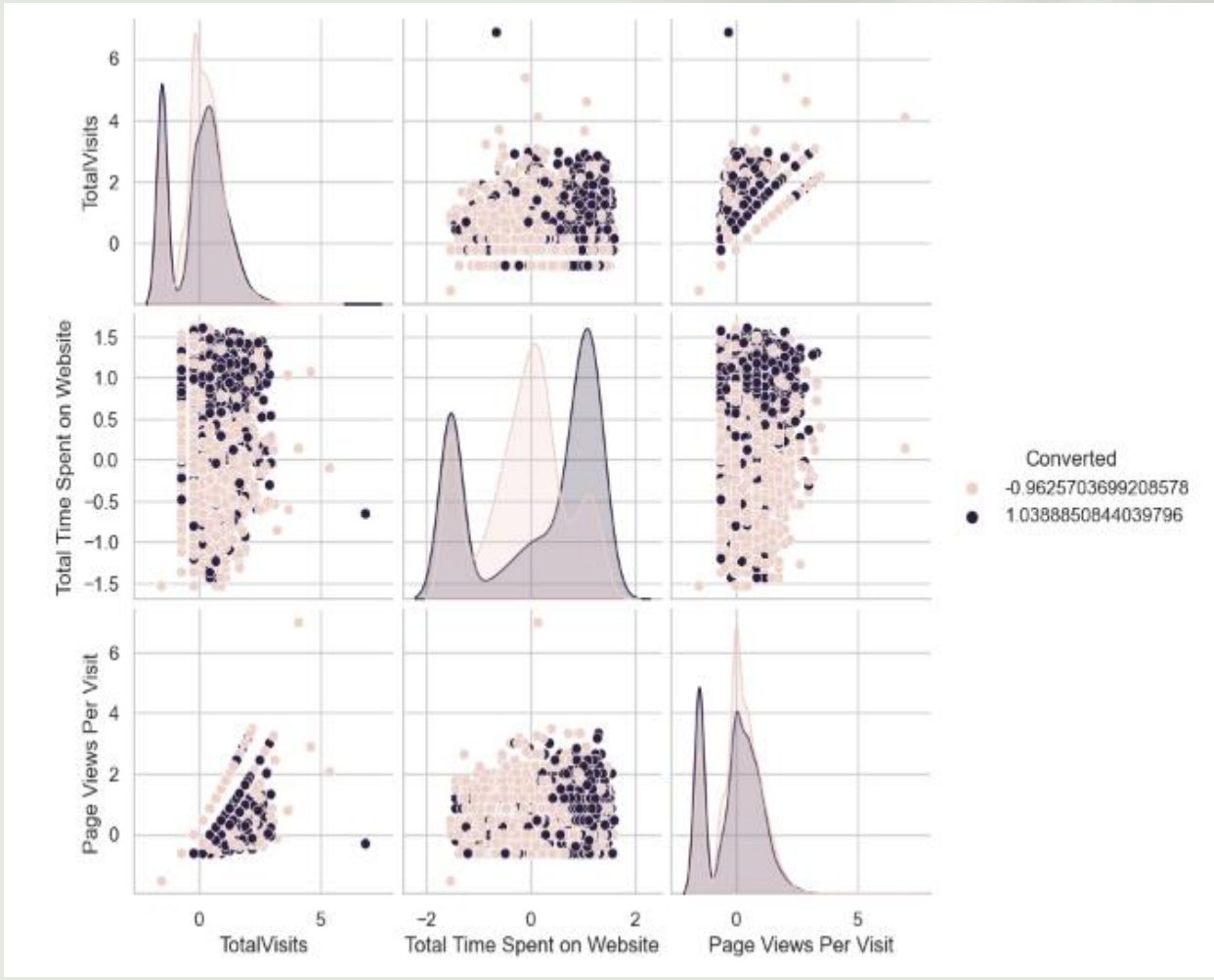
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE,VIFs and P-value)
- Model Evaluation
- Making predictions on test set

EDA- Cleaned Dataset

- Lead Origin
  - Do not Email
  - Total Visits
  - Pages views per visit
  - Country
  - Tags
- Lead Source
  - Converted
  - Total time spent on website
  - Last Activity
  - Specialization
  - Lead Quality

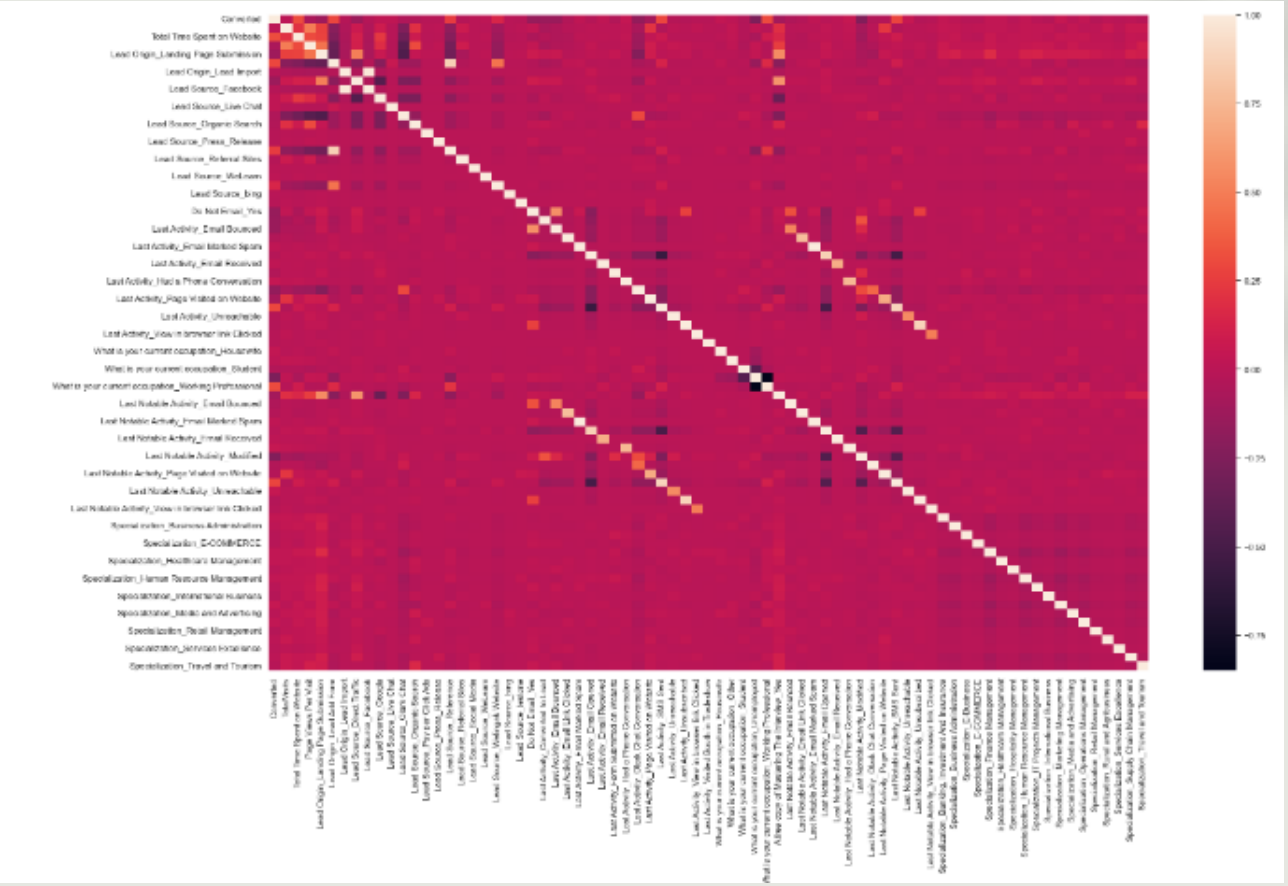
Data preparation and Modelling





Correlation

- From the attached heatmap, Highly correlated attributed create dependency on various independent factors which will give us inappropriate results.
- After dropping those high correlation features, we plotted again the heatmap to check and confirm the same.
- There are still few left but or model is to verify how much they are impacting as from the plot it is not quite understandable which variable is having high correlation.
- Hence, There is no correlation between the Variables.



Model Building

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4445
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2072.8
Date:	Mon, 18 Dec 2023	Deviance:	4145.5
Time:	21:24:07	Pearson chi2:	4.84e+03
No. Iterations:	22	Pseudo R-squ. (CS):	0.3660
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

There are few variables which have p-values greater than 0.05.

VIFs-

VIFs seems to be in a decent range except three variables.
The VIFs are now all less than 5.

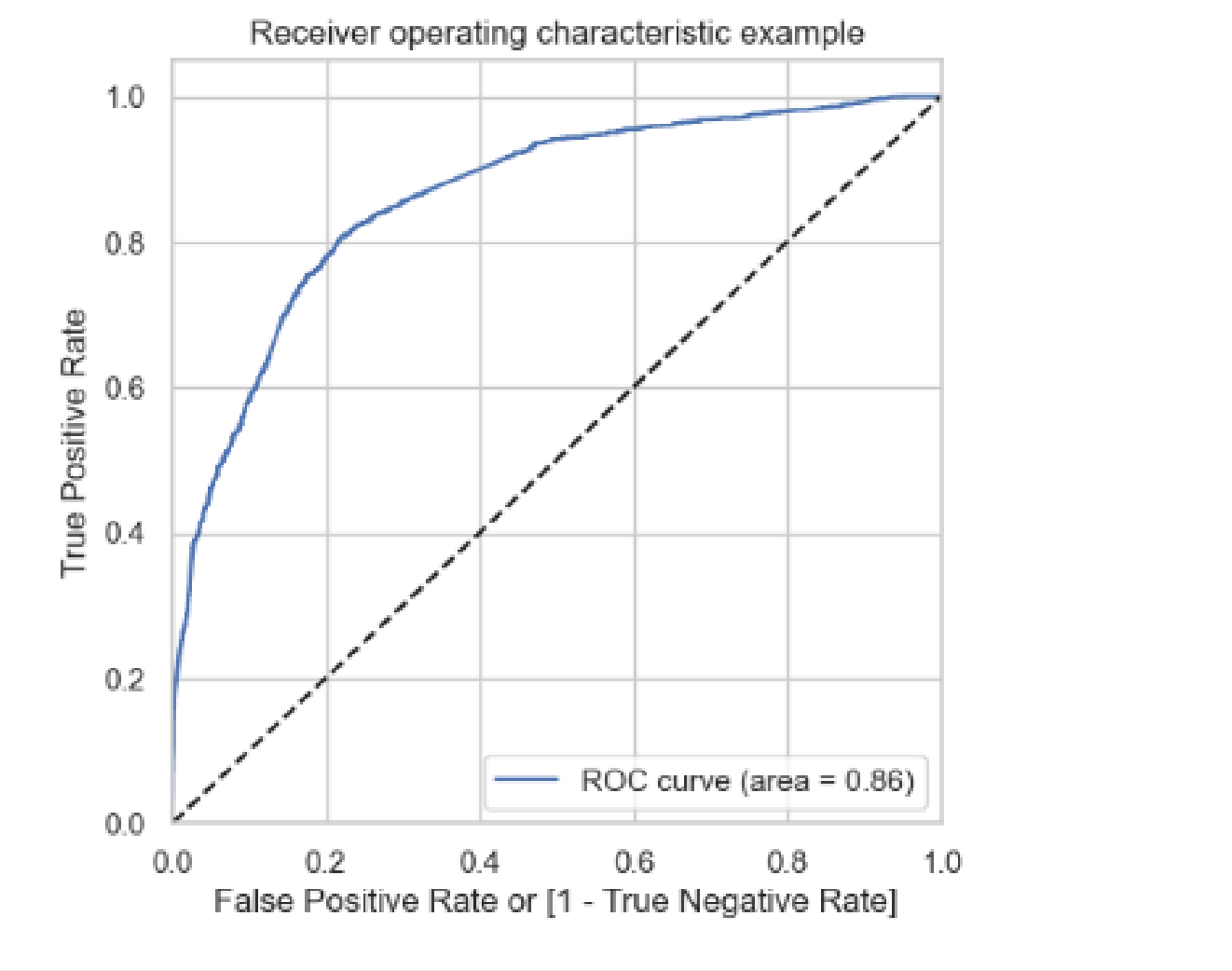
	Features	VIF
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Professional	1.56
3	Lead Source_Olark Chat	1.44
6	Do Not Email_Yes	1.09
10	What is your current occupation_Student	1.09
9	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Unreachable	1.01

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4446
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2073.2
Date:	Mon, 18 Dec 2023	Deviance:	4146.5
Time:	21:28:48	Pearson chi2:	4.82e+03
No. Iterations:	22	Pseudo R-squ. (CS):	0.3658
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0057	0.600	-1.677	0.094	-2.181	0.170
TotalVisits	11.3428	2.682	4.229	0.000	6.086	16.599
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	4.2084	0.259	16.277	0.000	3.702	4.715
Lead Source_Olark Chat	1.4583	0.122	11.960	0.000	1.219	1.697
Lead Source_Welingak Website	2.1557	1.037	2.079	0.038	0.124	4.188
Do Not Email_Yes	-1.5036	0.193	-7.779	0.000	-1.882	-1.125
Last Activity_Had a Phone Conversation	1.0398	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6511	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1537	0.630	-1.830	0.067	-2.389	0.082
What is your current occupation_Unemployed	-1.3401	0.594	-2.255	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2748	0.623	2.046	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1934	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7872	0.807	3.454	0.001	1.205	4.369

The variables Lead Profile_Dual Specialization student also needs to be dropped.

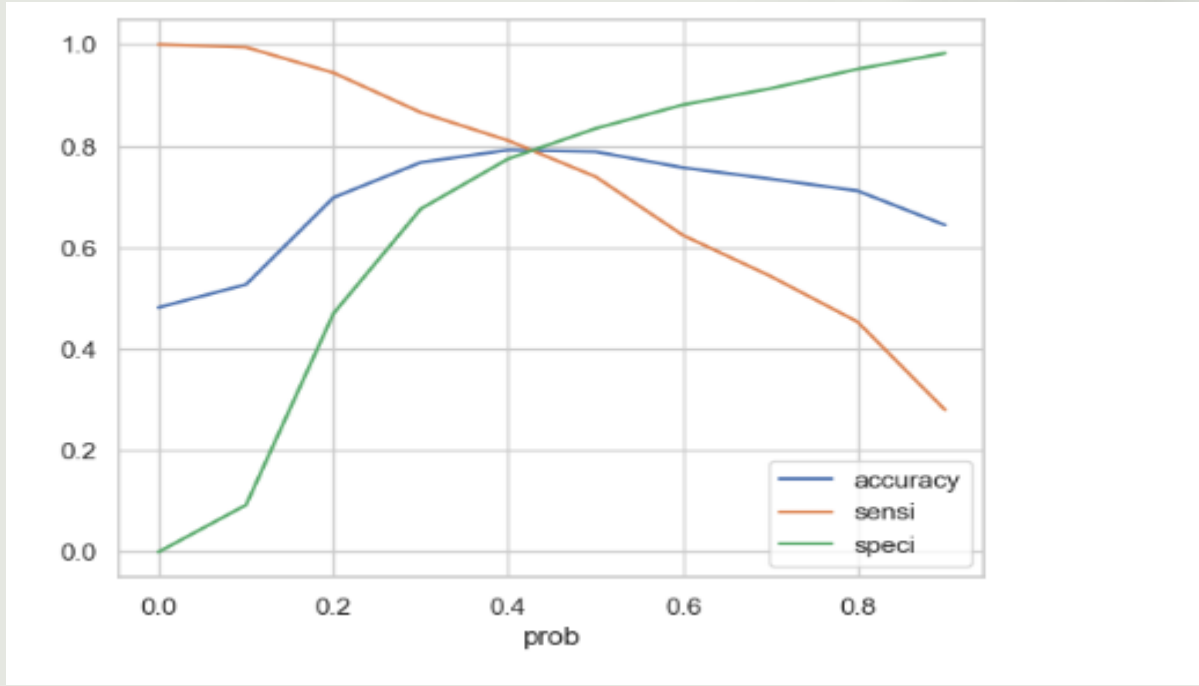
Model Evaluation



- The area under the curve of the ROC is 0.86 which is quite good.
- It seems that we have a good model.

Check Accuracy, Sensitivity and Specificity.

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699



- Around 0.42 we can get the optimal values of the three metrics.
- 0.42 is our cutoff and 42% probability of hot leads.

Observations

- Train Data
  - Accuracy- 78.8%
  - Sensitivity- 73.9%
  - Specificity- 83.4%
- Test Data
  - Accuracy-78.4%
  - Sensitivity- 77.9%
  - Specificity- 78.9%

Final Feature list

- Lead origin
- Lead Source
- Total time spent on website
- Specialization
- Lead Origin Landing page submission
- What is your current occupation
- Working profession
- Do not Email

Conclusion

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
  First, sort out the best prospects from the leads you have generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
  Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.
  Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.