Lead Scoring Case Study

Prasun Kumar Palchowdhury

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.

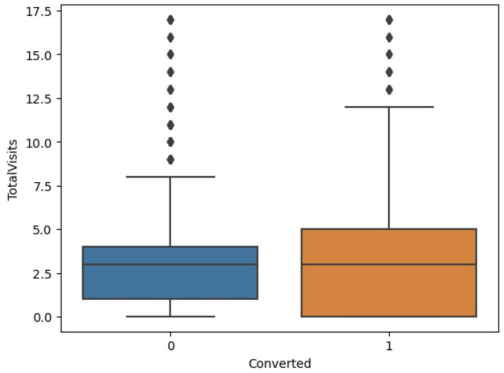
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified 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%.

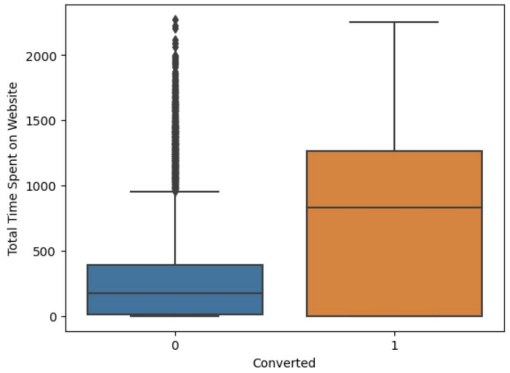
Business Objective

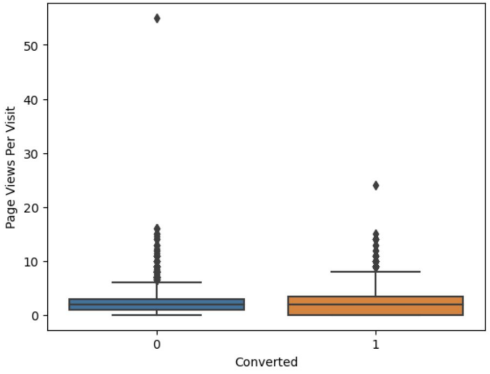
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Outlier Analysis

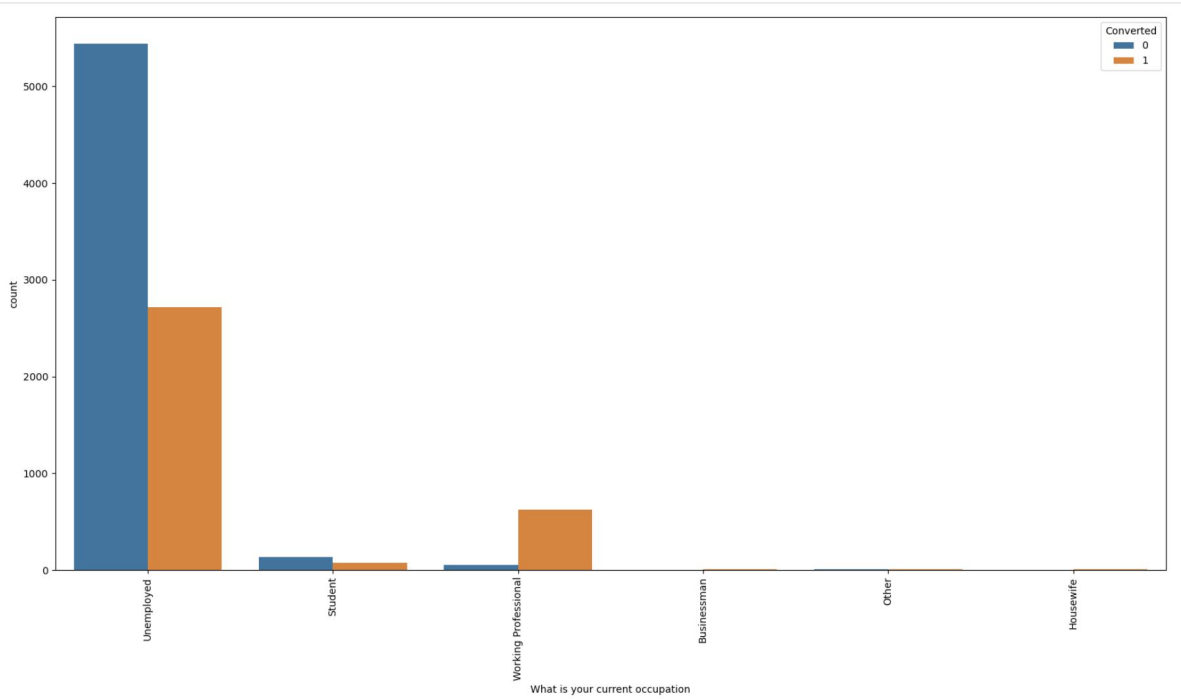






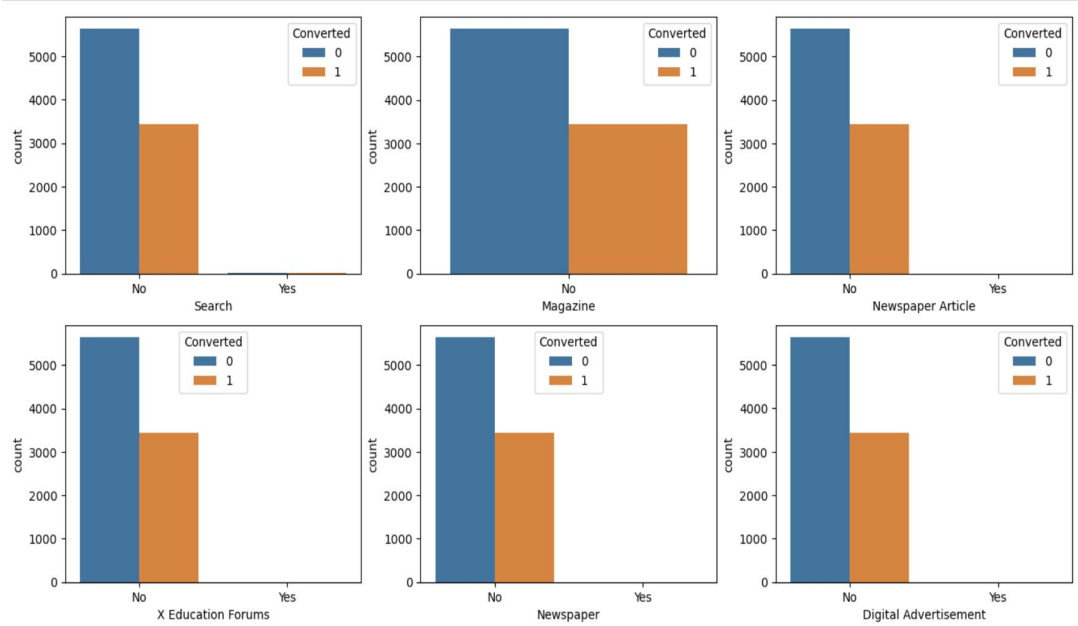
- There are outliers in TotalVisits, Total Time Spent on Website and Page Views Per Visit columns
- To eliminate the outliers first did a 99th percentile and then 95th percentile analysis. Considered 95th percentile data to be the final choice

Professional Attributes Analysis



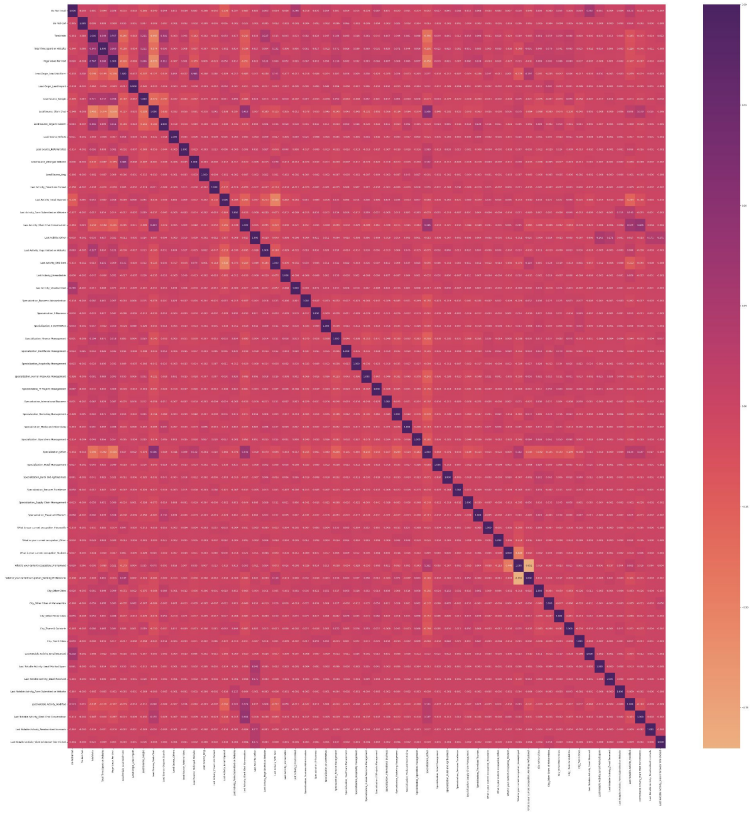
- Working professionals are more inclined towards taking a course than students
- Also, unemployed leads have shown interest in courses by X Education but their conversion rate is not that great

Advertisement Mechanics Analysis



- None of the advertising mechanism has been effective so far
- X Education might need to focus on other means of reaching out to leads, such as Email, SMS even phone calls

Correlation Analysis



- There were good amount of correlation among multiple variables
- Some of the variables were removed manually
- Further were deleted based on RFE analysis

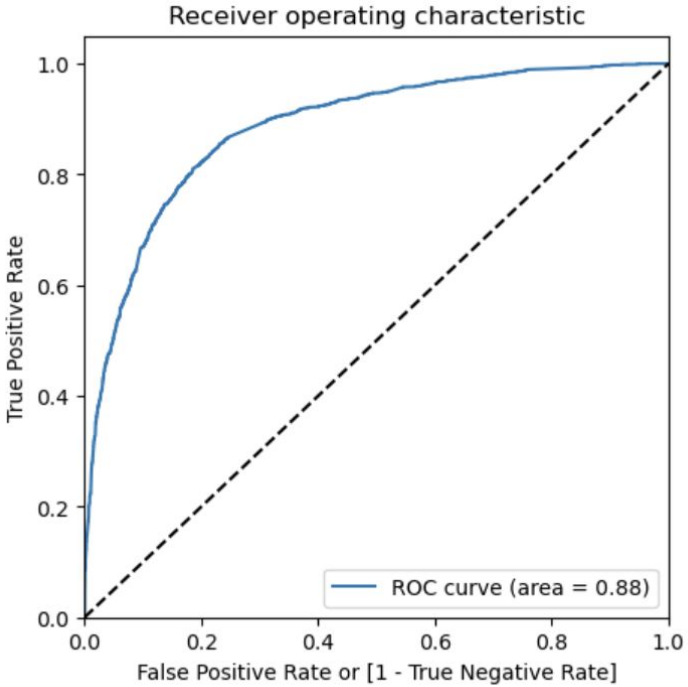
Model Selection

	Features	VIF
10	Specialization_Other	2.23
5	Lead Source_Olark Chat	2.17
6	Last Activity_Olark Chat Conversation	1.59
12	Last Notable Activity_Modified	1.58
4	Lead Source_Google	1.42
8	Last Activity_SMS Sent	1.37
1	Total Time Spent on Website	1.30
2	Lead Origin_Lead Add Form	1.29
0	Do Not Email	1.18
11	What is your current occupation_Working Profes...	1.16
9	Last Activity_Unsubscribed	1.08
3	Lead Origin_Lead Import	1.01
7	Last Activity_Other	1.01

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2788	0.069	-18.551	0.000	-1.414	-1.144
Do Not Email	-1.6224	0.186	-8.741	0.000	-1.986	-1.259
Total Time Spent on Website	1.1198	0.040	27.684	0.000	1.041	1.199
Lead Origin_Lead Add Form	4.5723	0.218	21.007	0.000	4.146	4.999
Lead Origin_Lead Import	1.7279	0.462	3.737	0.000	0.822	2.634
Lead Source_Google	0.3836	0.080	4.817	0.000	0.228	0.540
Lead Source_Olark Chat	1.6500	0.126	13.098	0.000	1.403	1.897
Last Activity_Olark Chat Conversation	-0.9116	0.168	-5.417	0.000	-1.241	-0.582
Last Activity_Other	2.2247	0.463	4.806	0.000	1.317	3.132
Last Activity_SMS Sent	1.2838	0.075	17.218	0.000	1.138	1.430
Last Activity_Unsubscribed	1.3689	0.476	2.875	0.004	0.436	2.302
Specialization_Other	-0.4097	0.088	-4.652	0.000	-0.582	-0.237
What is your current occupation_Working Professional	2.6714	0.191	14.002	0.000	2.297	3.045
Last Notable Activity_Modified	-0.8920	0.080	-11.091	0.000	-1.050	-0.734

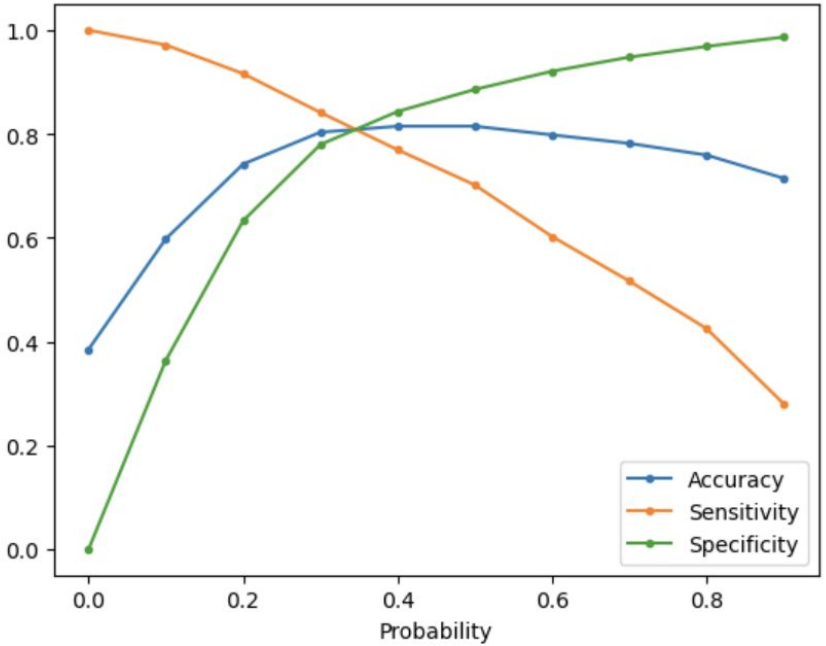
By doing VIF and P-Value analysis we have decided the final set of features

Model evaluation



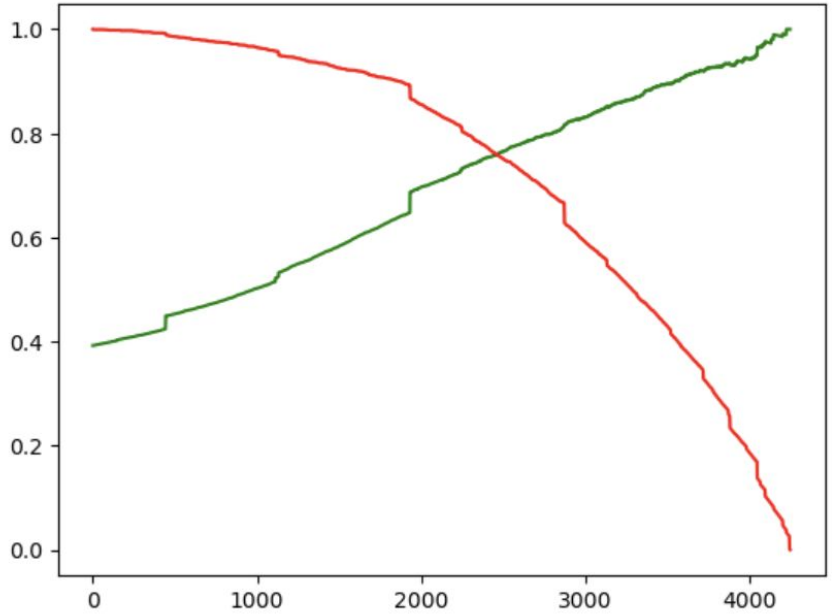
- After the final model building, ROC curve shows that area score is 0.88 which is a great score for prediction
- Also, the ROC curve is leaned toward the bottom left part which means it has a very good accuracy

Optimal Cut-off



- The optimal cut-off to be found at 0.35 which is around 35%
- This is used for calculating lead score on train and test data

Precision & Recall trade off



- Precision and recall trade off shown on the train dataset

Conclusion

1. Accuracy, Precision, Recall, Sensitivity and Specificity seems to be in acceptable region
2. This model would meet the business requirement as mentioned earlier.
3. Leads who have origin of `Add Form` and `Lead Import` should be contacted
4. Leads whose sources are `Olark Chat` and `Google` should be contacted
5. Leads who are currently `working professionals` they must be contacted
6. Leads who have done any activities such as `SMS Sent`, `Olark Chat`, `Modified` should be contacted
7. Leads who spend more time on the website should be contacted
8. Overall when the leads have engagement with the platform or the sales representative, they should be contacted.