

Lead Scoring Case Study

Krishna Murthy, Ankit Pandey, Ajinkya Chinchwadkar

Summary

- **Problem Statement**

An education company acquires leads when people provide their contact details after landing on the website or through past referrals. These leads are contacted by sales team through calls, emails. However typical conversion rate is around 30%. It is very low and company wants it to increase to 80% by identifying most potential leads. Company requires to build a model to assign lead score to each of the leads such that customers with higher lead scores have higher conversion chance and vice versa.

- **Machine Learning Problem**

1. Build a logistic regression model to assign a lead score as required.
2. Identify top 3 variables contributing in lead conversion
3. Identify top 3 categorical variables contributing most in lead conversion
4. Good strategy for interns when hired for 2 months for making calls to all potential leads
5. Strategy for reducing useless phone call when target is achieved

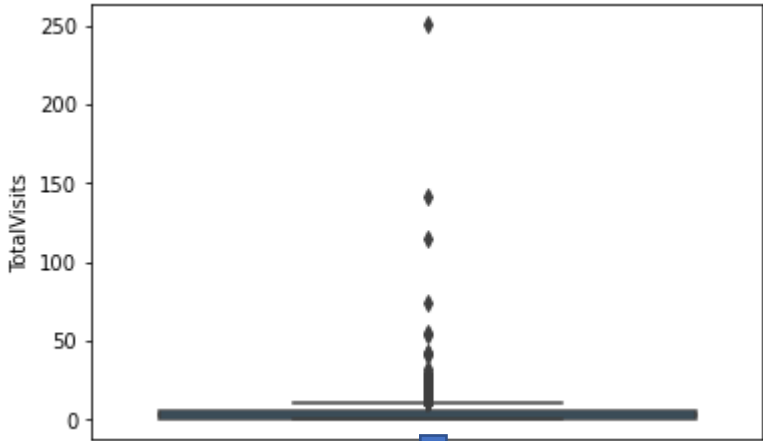
- **Analysis Approach**

EDA, Dummy variables for categorical, Grouping of variables, Train-test split , Logistic regression Model Building, Eliminating variables with P-value,, VIF, Probability Prediction , Model Evaluation

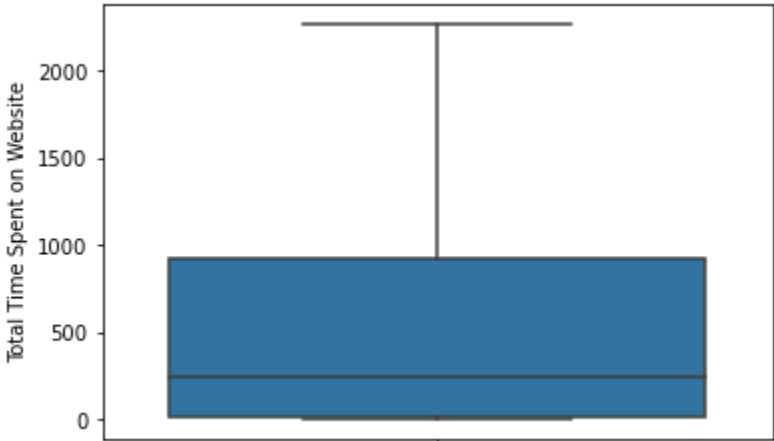
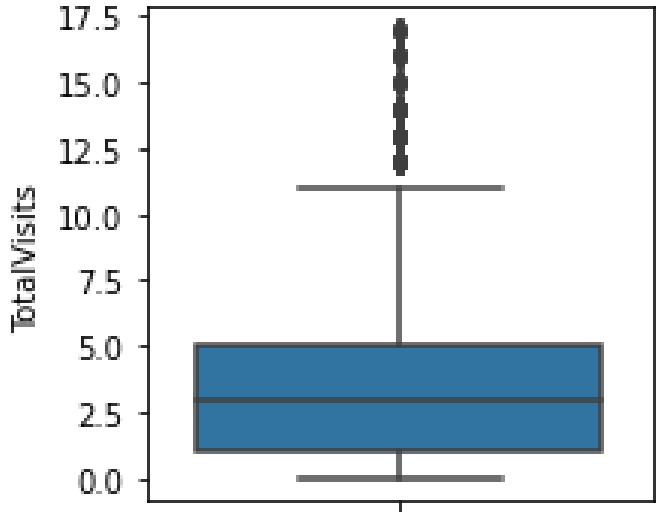
- **Result**

Test set prediction Accuracy is 81%, the precision is about 74% and the recall is about 78% at cut off of 0.40 . So model is able to predict. Hence, it can be useful for business.

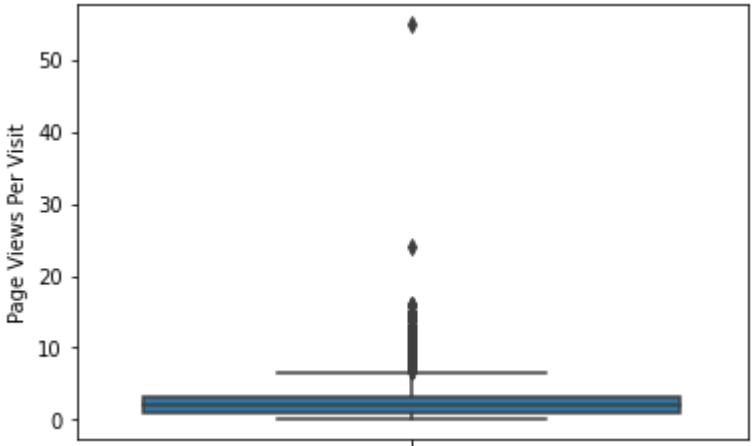
EDA – Outlier Treatment



Removed top and bottom 1% outliers




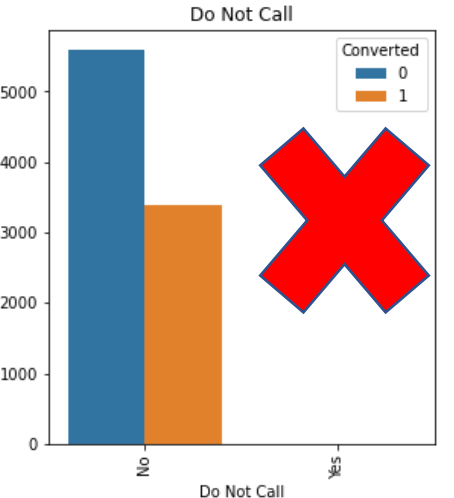
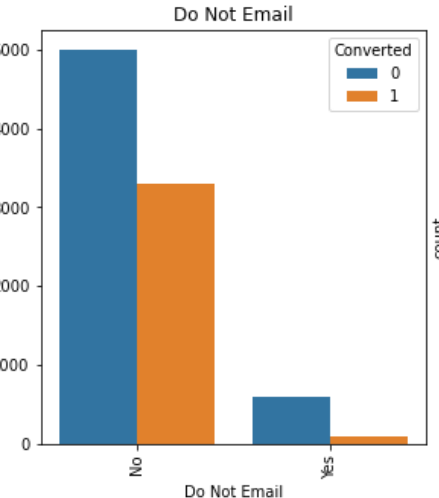
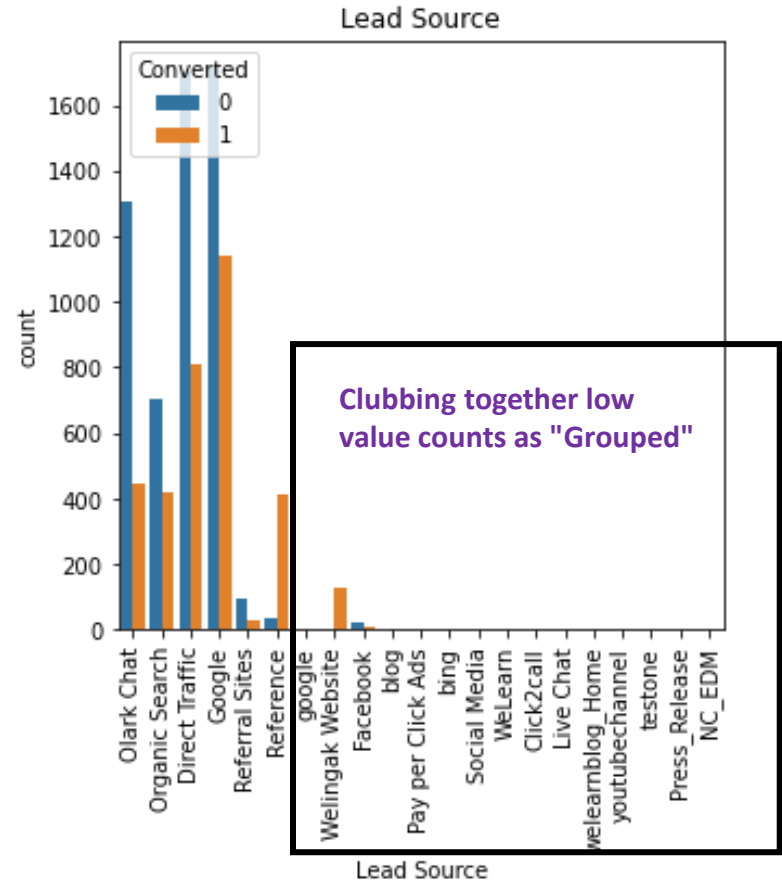
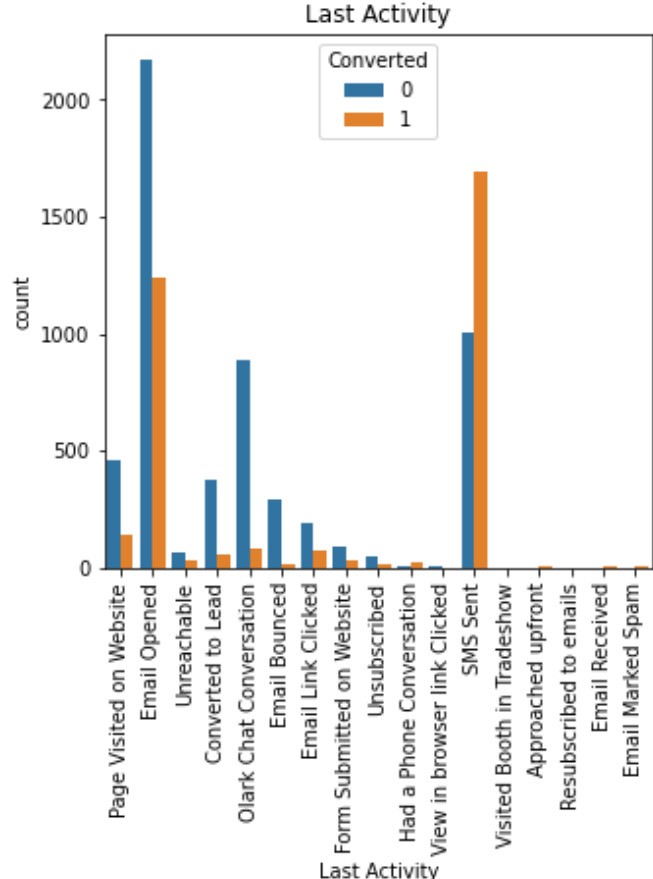
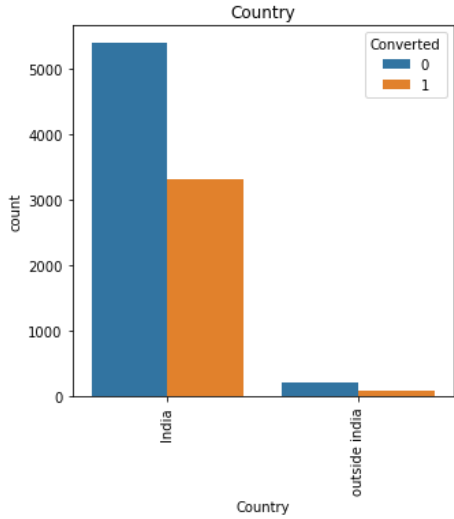
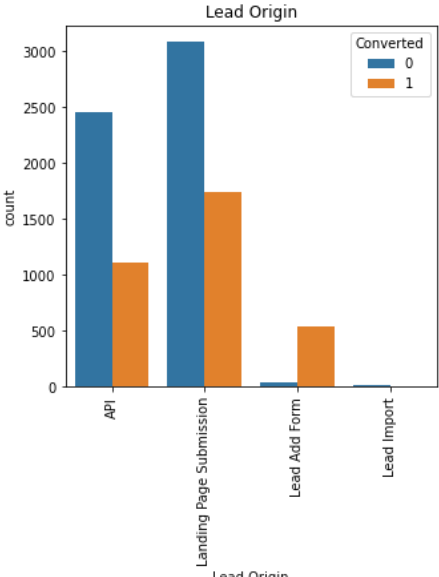
No Outliers



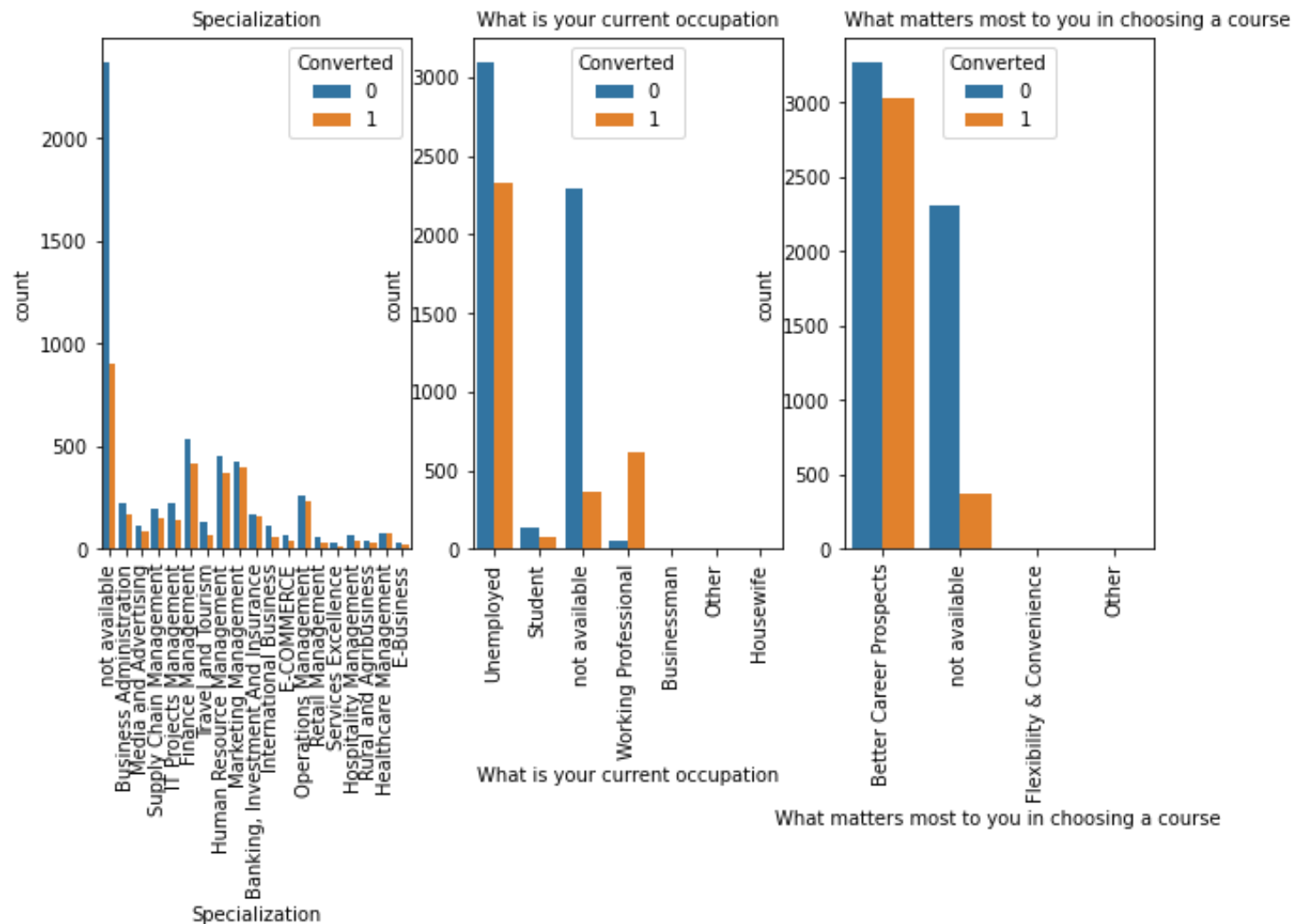
Outliers not too high

Data Visualization

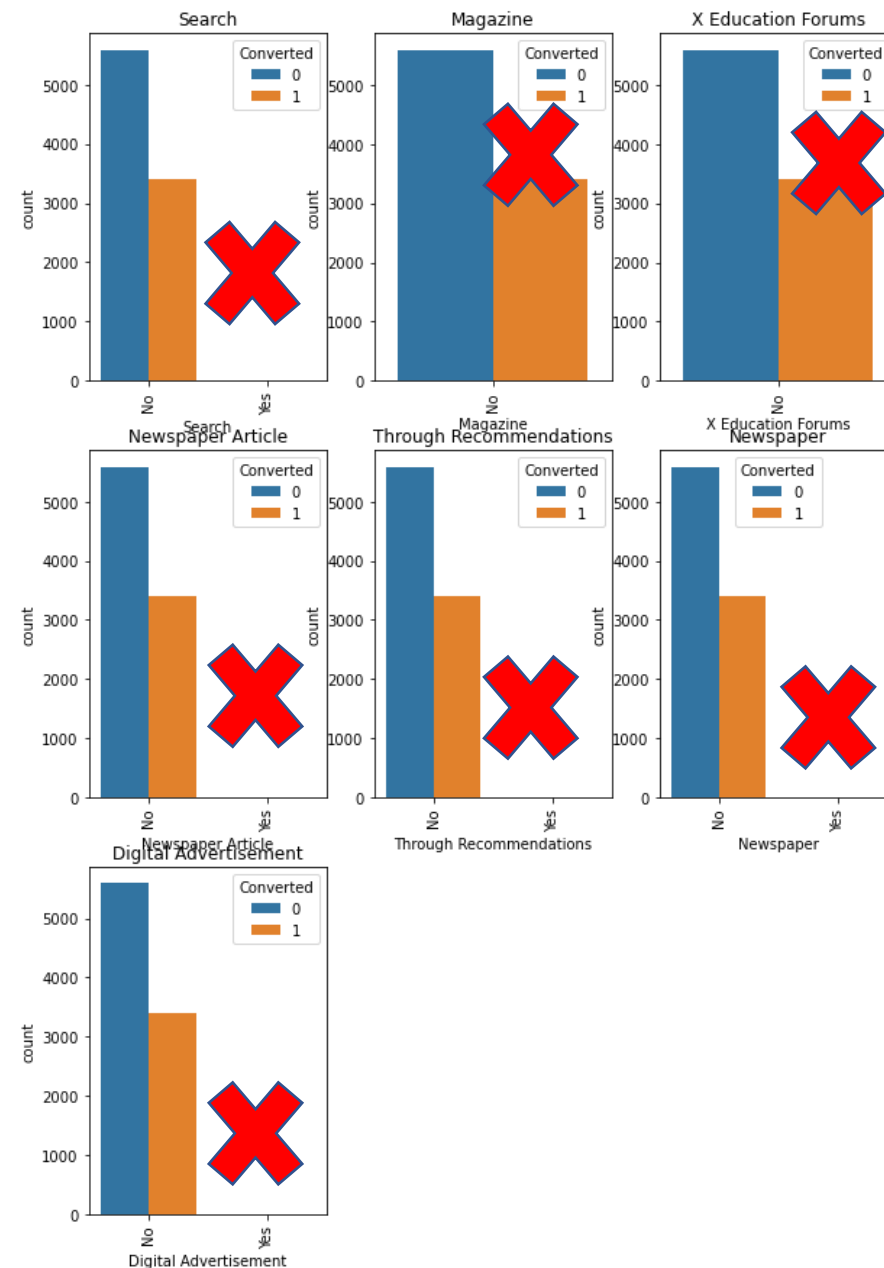
 Dropped because of only one negative feedback



Data Visualisation

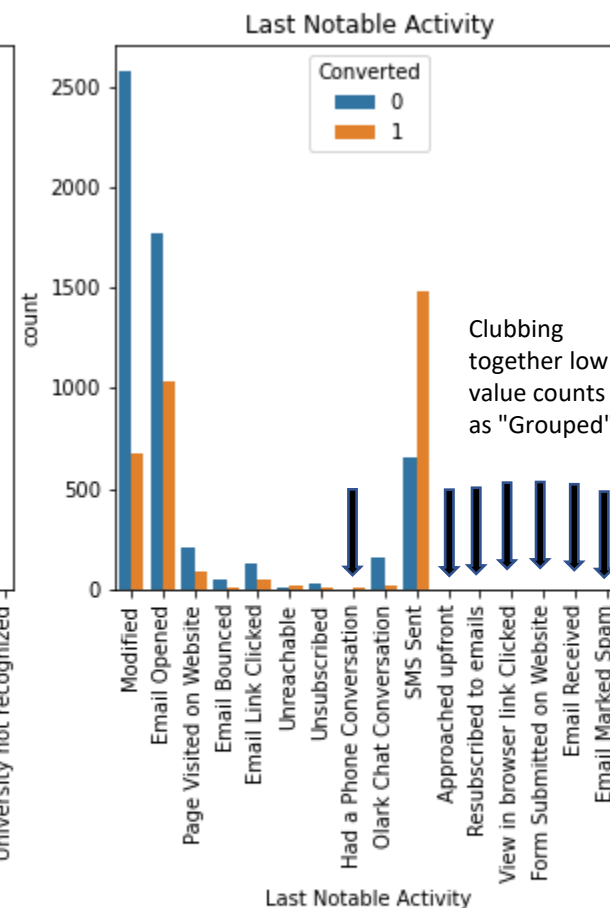
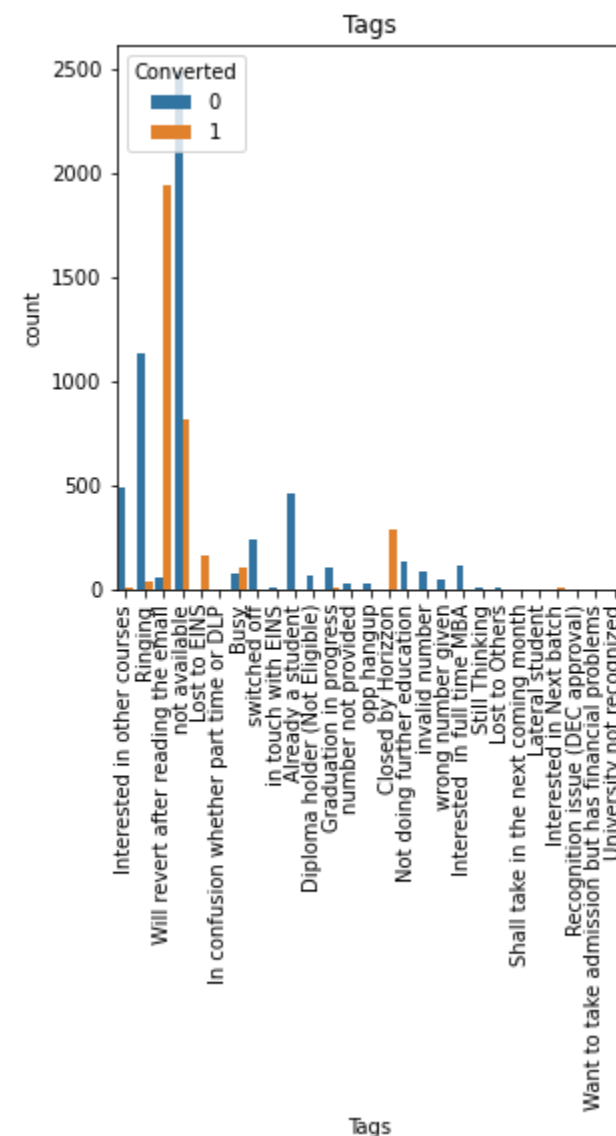
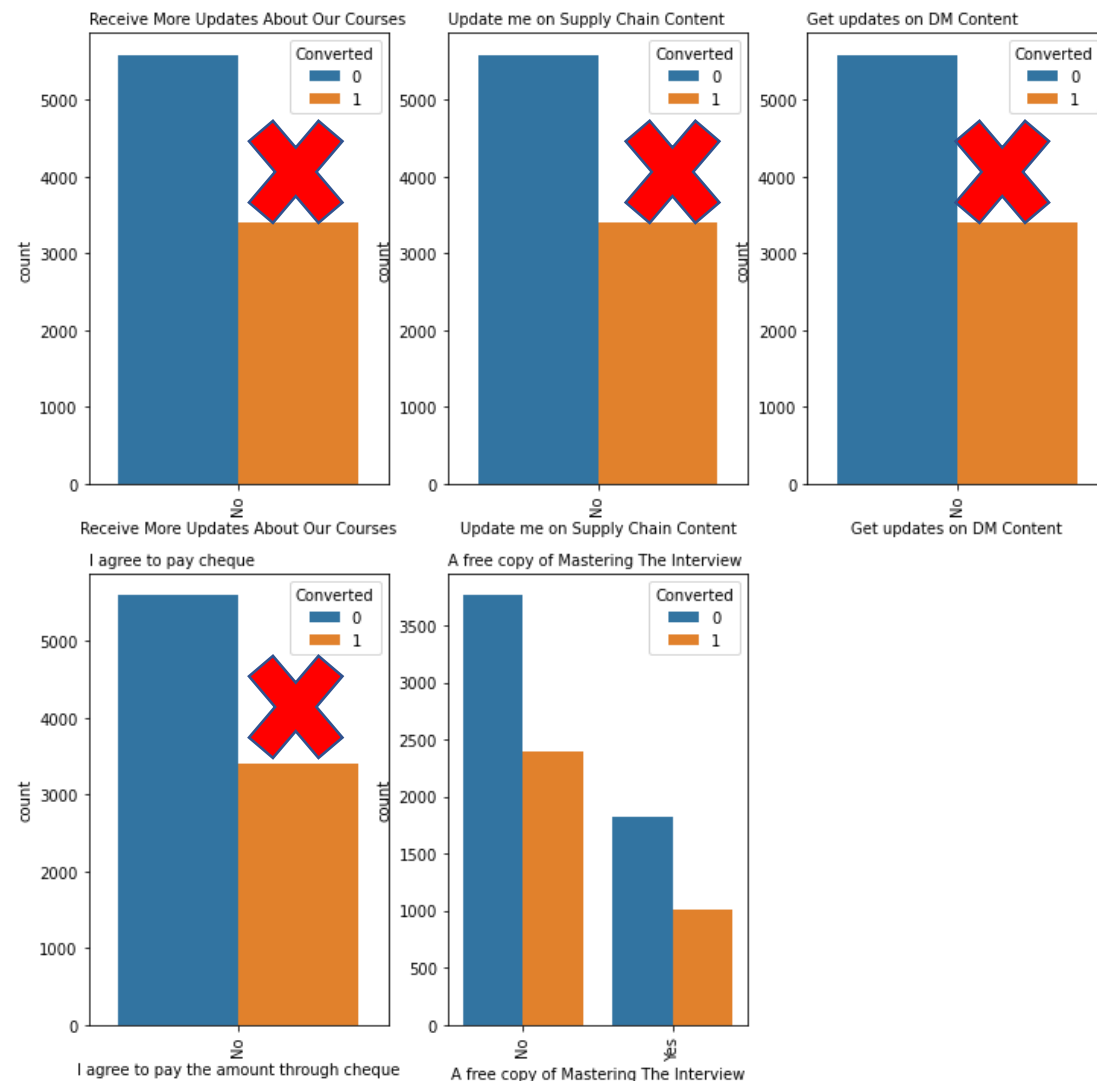


X Dropped because of only one negative feedback

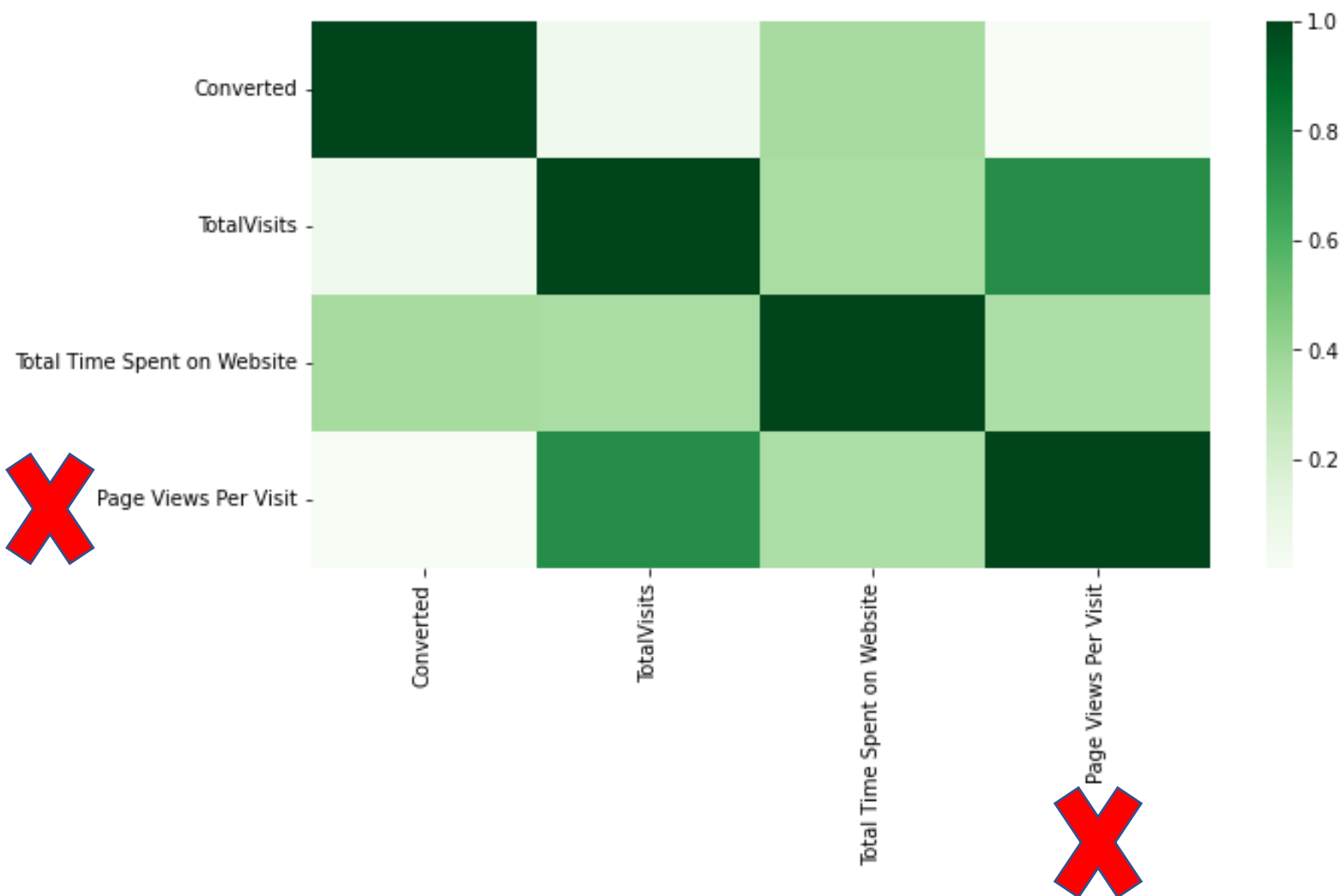


Data Visualisation

✗ Dropped because of only one negative feedback



Heat Map



As we can see page views per visit and Total visit are highly correlated. So we can delete one of them.

Create Dummy Variables

Created dummy variables for below category variables.

'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Specialization', 'What is your current occupation',
'What matters most to you in choosing a course', 'City','A free copy of Mastering The Interview', 'Last Notable Activity'

Train Test Split

Target variable is 'Converted' added to y-axis and added rest of variables as feature variables to x-axis.

MODEL BUILDING

VIF value are seems quiet Ok but the p value of What is your current occupation_Housewife & Last Notable Activity_Grouped is higher side so drop it.

	coef	std err	z	P> z	[0.025	0.975]
const	0.2454	0.148	1.654	0.098	-0.045	0.536
Total Time Spent on Website	4.6536	0.173	26.834	0.000	4.314	4.994
Lead Origin_Landing Page Submission	-0.9751	0.132	-7.376	0.000	-1.234	-0.716
Lead Origin_Lead Add Form	3.4214	0.265	12.931	0.000	2.903	3.940
Specialization_not available	-0.9273	0.129	-7.166	0.000	-1.181	-0.674
Lead Source_Olark Chat	1.1913	0.126	9.427	0.000	0.944	1.439
Lead Source_Welingak Website	2.9433	1.041	2.826	0.005	0.902	4.985
Do Not Email_Yes	-1.4967	0.207	-7.228	0.000	-1.903	-1.091
Last Activity_Converted to Lead	-0.9157	0.213	-4.303	0.000	-1.333	-0.499
Last Activity_Email Bounced	-1.1443	0.393	-2.915	0.004	-1.914	-0.375
Last Activity_Form Submitted on Website	-0.9936	0.377	-2.632	0.008	-1.733	-0.254
Last Activity_Olark Chat Conversation	-1.4266	0.202	-7.060	0.000	-1.823	-1.031
What is your current occupation_Housewife	21.9682	1.76e+04	0.001	0.999	-3.45e+04	3.45e+04
What is your current occupation_Working Professional	2.4070	0.194	12.388	0.000	2.026	2.788
What is your current occupation_not available	-0.9834	0.090	-10.986	0.000	-1.159	-0.808
Last Notable Activity_Email Link Clicked	-1.8793	0.269	-6.991	0.000	-2.406	-1.352
Last Notable Activity_Email Opened	-1.3383	0.090	-14.801	0.000	-1.516	-1.161
Last Notable Activity_Grouped	2.2353	1.187	1.882	0.060	-0.092	4.563
Last Notable Activity_Modified	-1.5064	0.106	-14.229	0.000	-1.714	-1.299
Last Notable Activity_Olark Chat Conversation	-1.0606	0.382	-2.779	0.005	-1.809	-0.313
Last Notable Activity_Page Visited on Website	-1.6800	0.208	-8.096	0.000	-2.087	-1.273

	Features	VIF
1	Lead Origin_Landing Page Submission	3.05
17	Last Notable Activity_Modified	3.04
3	Specialization_not available	2.83
10	Last Activity_Olark Chat Conversation	2.12
4	Lead Source_Olark Chat	2.09
15	Last Notable Activity_Email Opened	1.98
0	Total Time Spent on Website	1.98
6	Do Not Email_Yes	1.92
8	Last Activity_Email Bounced	1.85
13	What is your current occupation_not available	1.64
2	Lead Origin_Lead Add Form	1.54
18	Last Notable Activity_Olark Chat Conversation	1.40
5	Lead Source_Welingak Website	1.39
7	Last Activity_Converted to Lead	1.28
12	What is your current occupation_Working Profes...	1.20
19	Last Notable Activity_Page Visited on Website	1.11
14	Last Notable Activity_Email Link Clicked	1.08
9	Last Activity_Form Submitted on Website	1.07
11	What is your current occupation_Housewife	1.01
16	Last Notable Activity_Grouped	1.01

MODEL BUILDING

After dropping 2 variables (What is your current occupation_Housewife & Last Notable Activity_Grouped) P-value and VIF are OK.

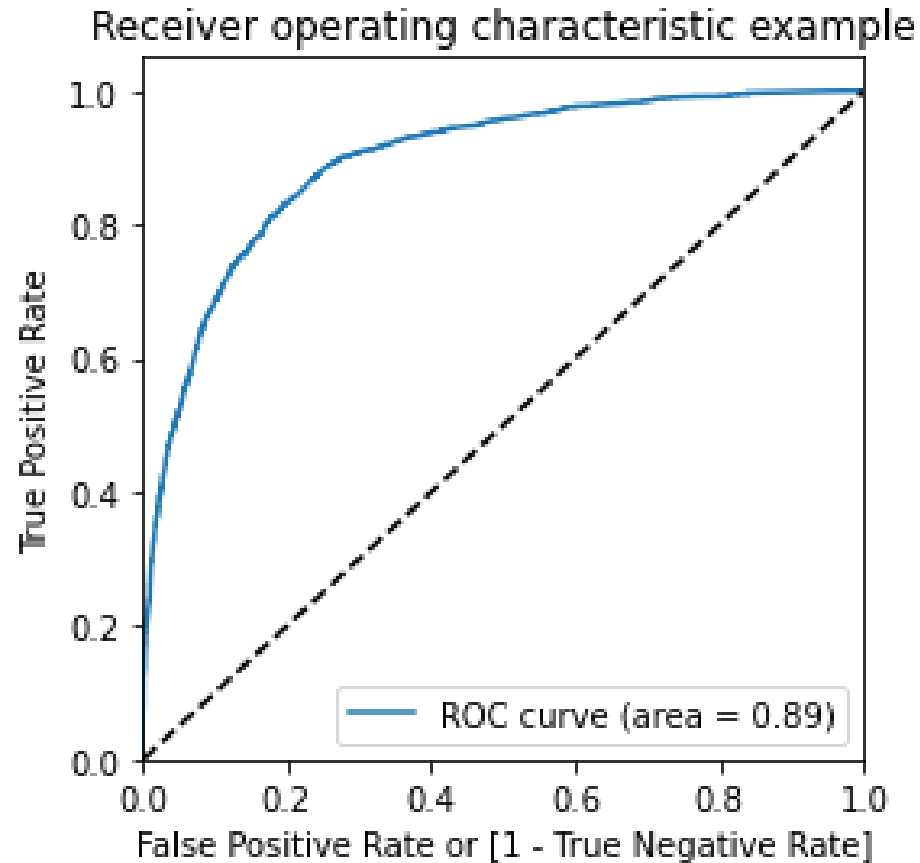
	coef	std err	z	P> z	[0.025	0.975]
const	0.2600	0.148	1.757	0.079	-0.030	0.550
Total Time Spent on Website	4.6521	0.173	26.863	0.000	4.313	4.992
Lead Origin_Landing Page Submission	-0.9767	0.132	-7.399	0.000	-1.235	-0.718
Lead Origin_Lead Add Form	3.4412	0.264	13.017	0.000	2.923	3.959
Specialization_not available	-0.9343	0.129	-7.226	0.000	-1.188	-0.681
Lead Source_Olark Chat	1.1917	0.126	9.435	0.000	0.944	1.439
Lead Source_Welingak Website	2.9229	1.041	2.807	0.005	0.882	4.964
Do Not Email_Yes	-1.4971	0.206	-7.261	0.000	-1.901	-1.093
Last Activity_Converted to Lead	-0.9171	0.213	-4.310	0.000	-1.334	-0.500
Last Activity_Email Bounced	-1.1504	0.392	-2.933	0.003	-1.919	-0.382
Last Activity_Form Submitted on Website	-0.9404	0.375	-2.505	0.012	-1.676	-0.205
Last Activity_Olark Chat Conversation	-1.4271	0.202	-7.062	0.000	-1.823	-1.031
What is your current occupation_Working Professional	2.4029	0.194	12.373	0.000	2.022	2.784
What is your current occupation_not available	-0.9819	0.089	-10.980	0.000	-1.157	-0.807
Last Notable Activity_Email Link Clicked	-1.8815	0.267	-7.037	0.000	-2.405	-1.357
Last Notable Activity_Email Opened	-1.3470	0.090	-14.921	0.000	-1.524	-1.170
Last Notable Activity_Modified	-1.5158	0.106	-14.342	0.000	-1.723	-1.309
Last Notable Activity_Olark Chat Conversation	-1.0698	0.382	-2.803	0.005	-1.818	-0.322
Last Notable Activity_Page Visited on Website	-1.6917	0.207	-8.154	0.000	-2.098	-1.285

	Features	VIF
1	Lead Origin_Landing Page Submission	3.05
15	Last Notable Activity_Modified	3.03
3	Specialization_not available	2.82
10	Last Activity_Olark Chat Conversation	2.12
4	Lead Source_Olark Chat	2.08
14	Last Notable Activity_Email Opened	1.98
0	Total Time Spent on Website	1.97
6	Do Not Email_Yes	1.92
8	Last Activity_Email Bounced	1.85
12	What is your current occupation_not available	1.64
2	Lead Origin_Lead Add Form	1.53
16	Last Notable Activity_Olark Chat Conversation	1.40
5	Lead Source_Welingak Website	1.39
7	Last Activity_Converted to Lead	1.28
11	What is your current occupation_Working Profes...	1.20
17	Last Notable Activity_Page Visited on Website	1.11
13	Last Notable Activity_Email Link Clicked	1.08
9	Last Activity_Form Submitted on Website	1.07

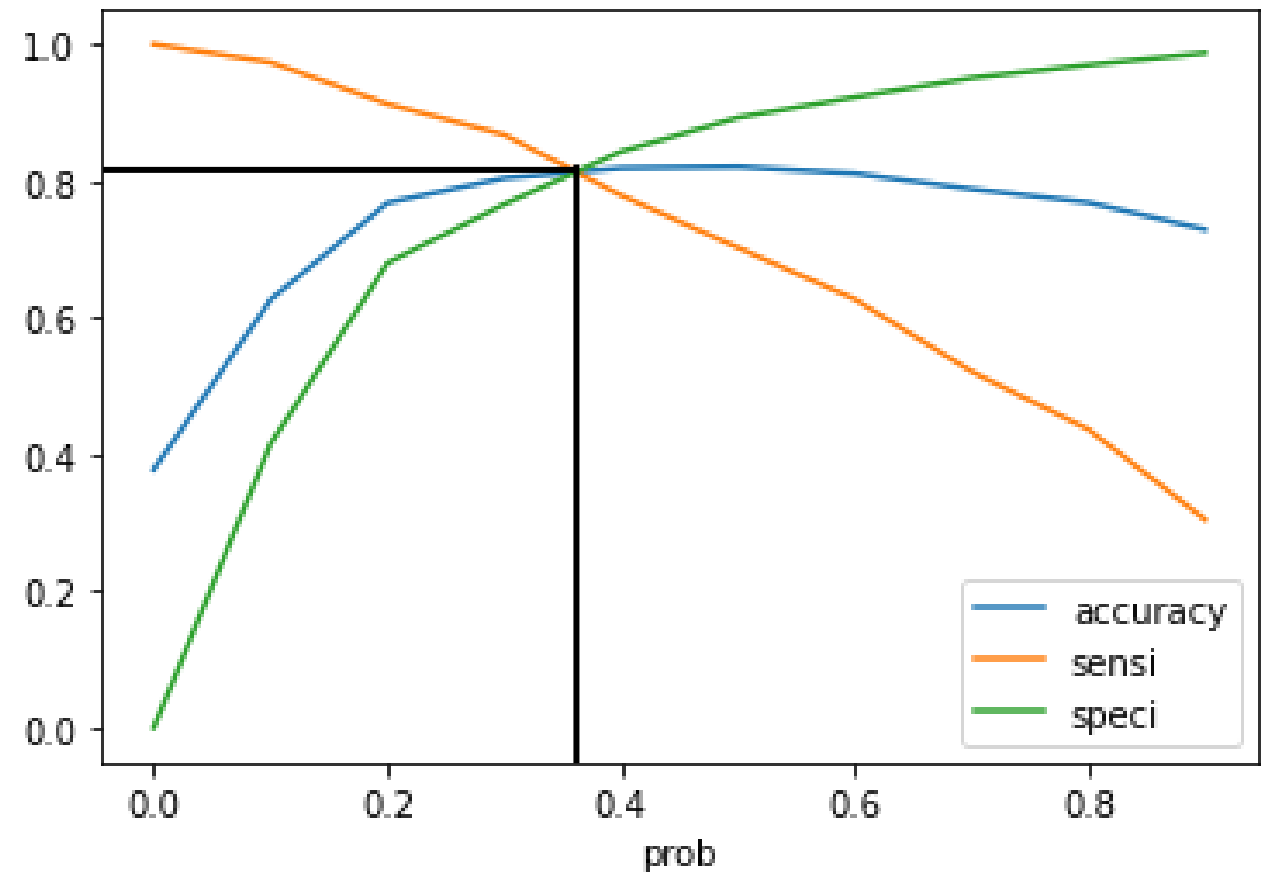
Doing prediction

Accuracy is 82%, the sensitivity is about 70% and the specificity is about 89%. So the readings are quiet good.

The area under ROC curve is 0.89 which is quiet good.



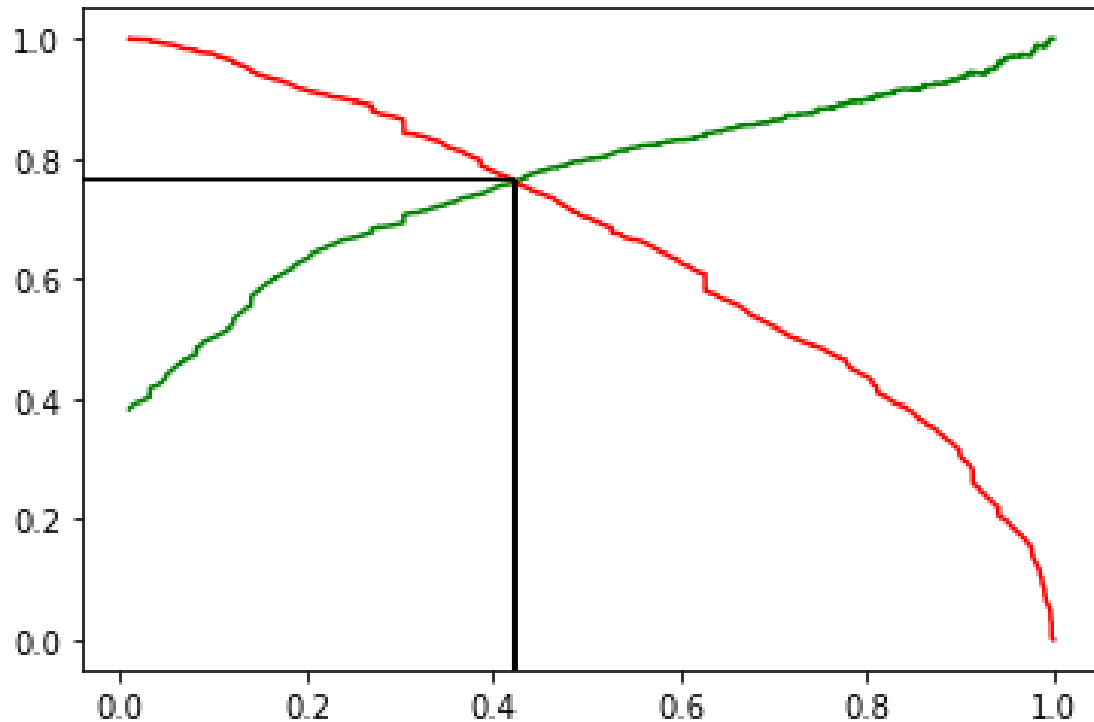
On the present ROC all three i.e accuracy, sensitivity and specificity is around 80%



Checking Test Set

Accuracy is 82%, the precision is about 75% and the recall is about 77% at cut off of 0.41 .

Precision and Recall tradeoff



Confusion Matrix

```
array([[3326,  589],  
       [ 544, 1834]])
```

Test set prediction

Accuracy is 81%, the precision is about 74% and the recall is about 78% at cut off of 0.40 . So model is able to predict. Hence, it can be useful for business.

Confusion Matrix

```
array([[1399,  278],
       [ 223,  798]],
```