Lead Scoring Case Study

By

Lokesh Bathula

Pankhudi Bhavate

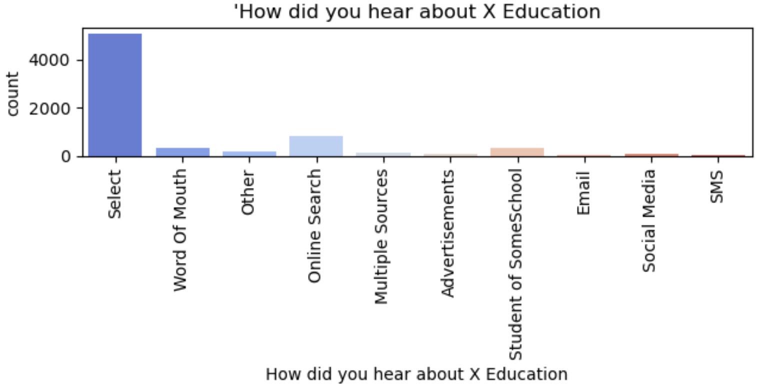
Pavan Kumar C S

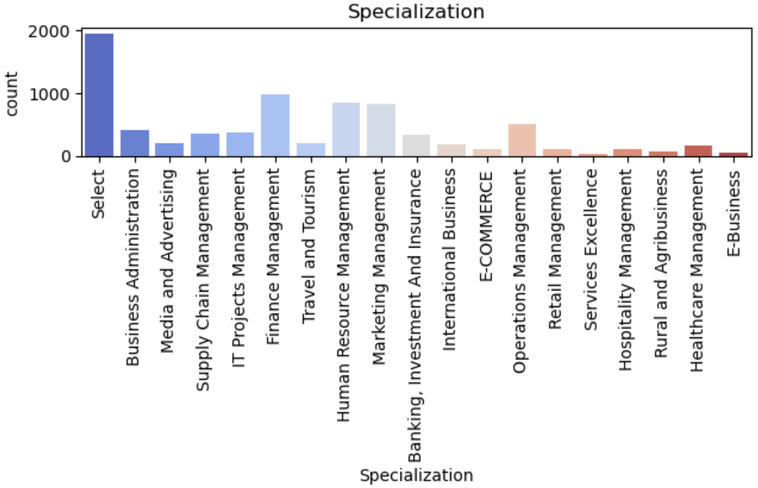
- **Problem Statement** :
  - X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
  - Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. 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 you 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%

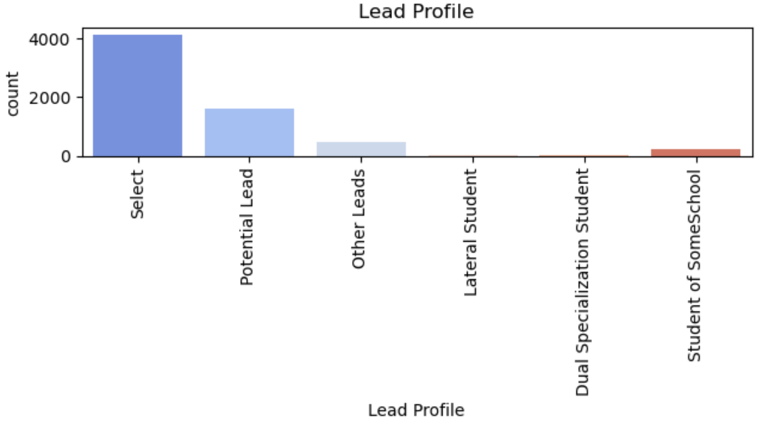
Approach

- The problem is a classification problem. So we will use logistic regression to assign the user to either converted or non-converted category.
- EDA is performed on the data set.
- Feature selection is performed by eliminating the feature with high p value and VIF. This is resulted in 3 different models.
- The model with VIF and p values below the desired values is selected
- The model is evaluated for accuracy, sensitivity and specificity.
- Precision and recall from the confusion matrix.

Visualising the columns that has select values

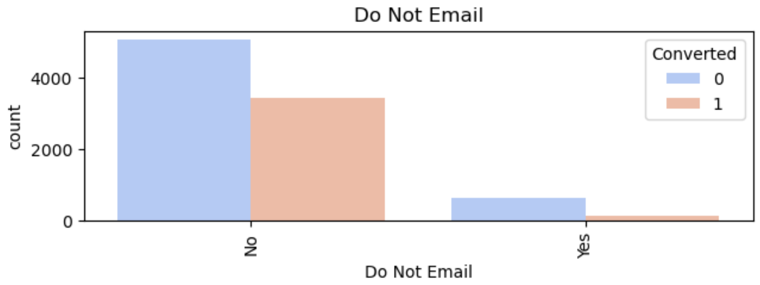


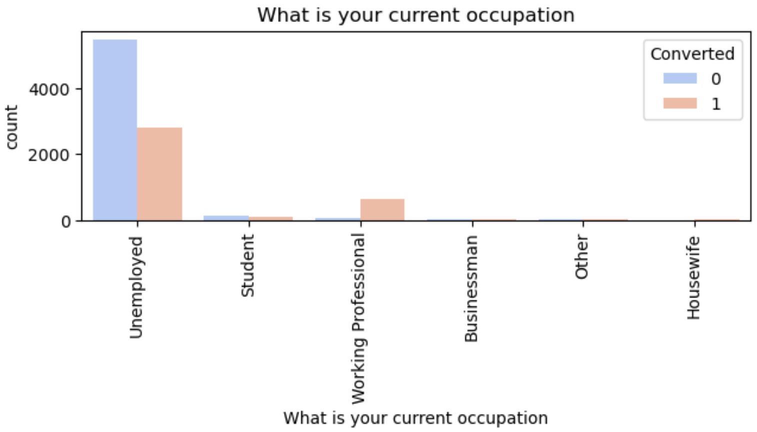


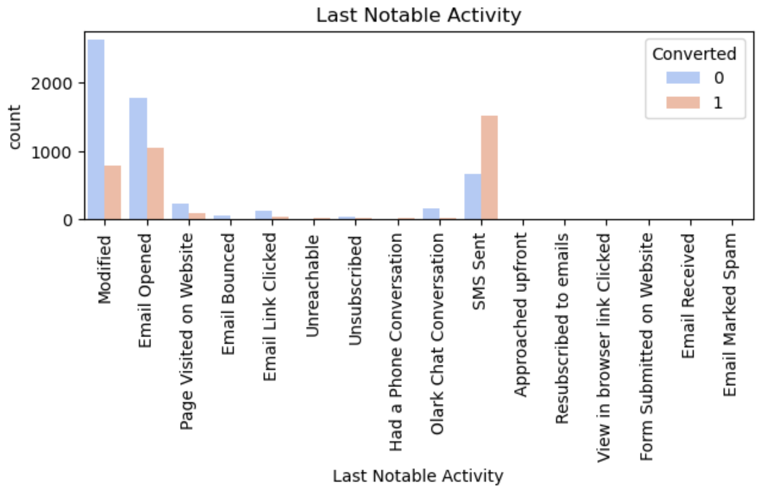


The columns which has the select values are nothing but those who didn't select that option for that particular column

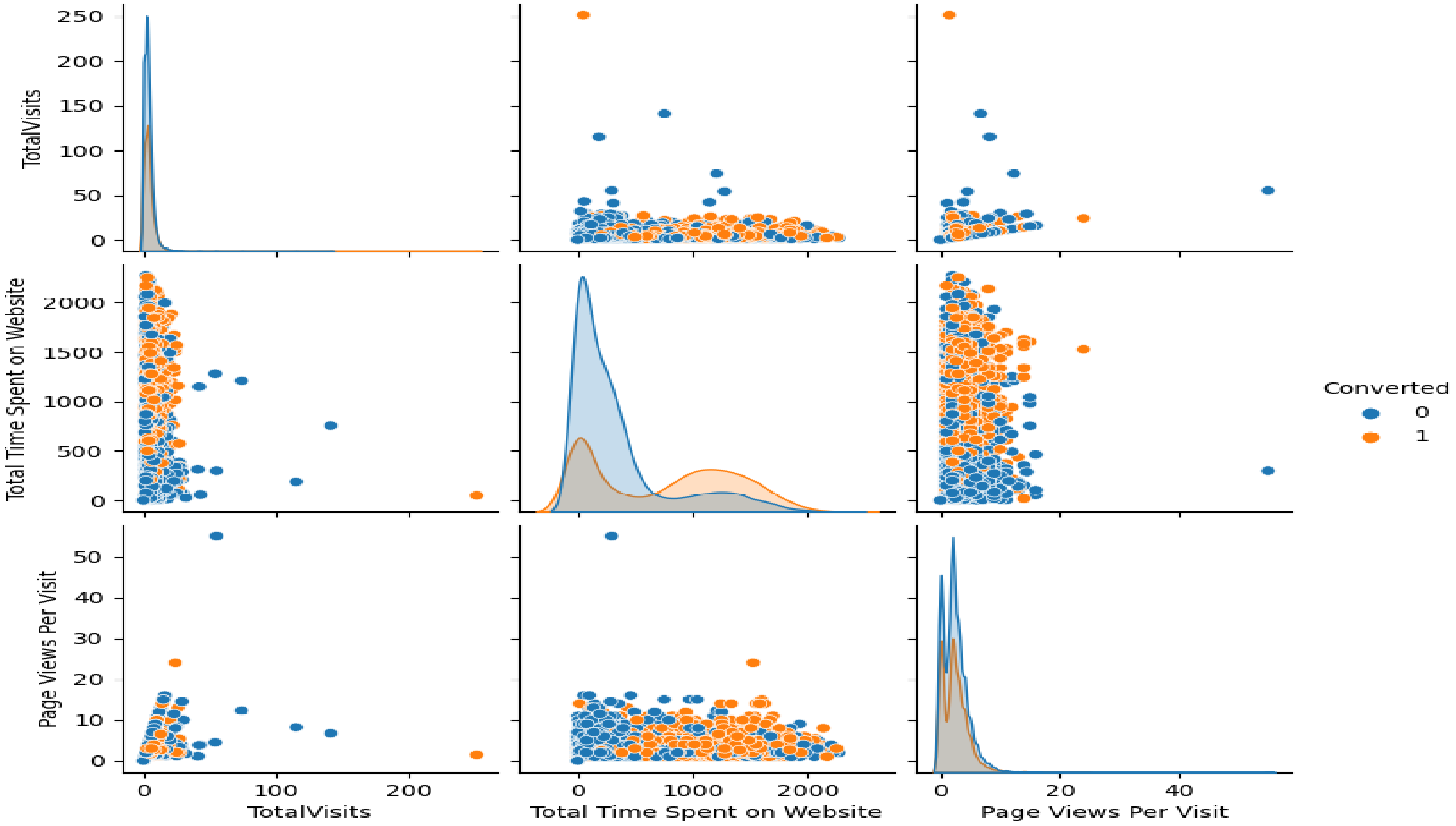
Visualising for categorical variables





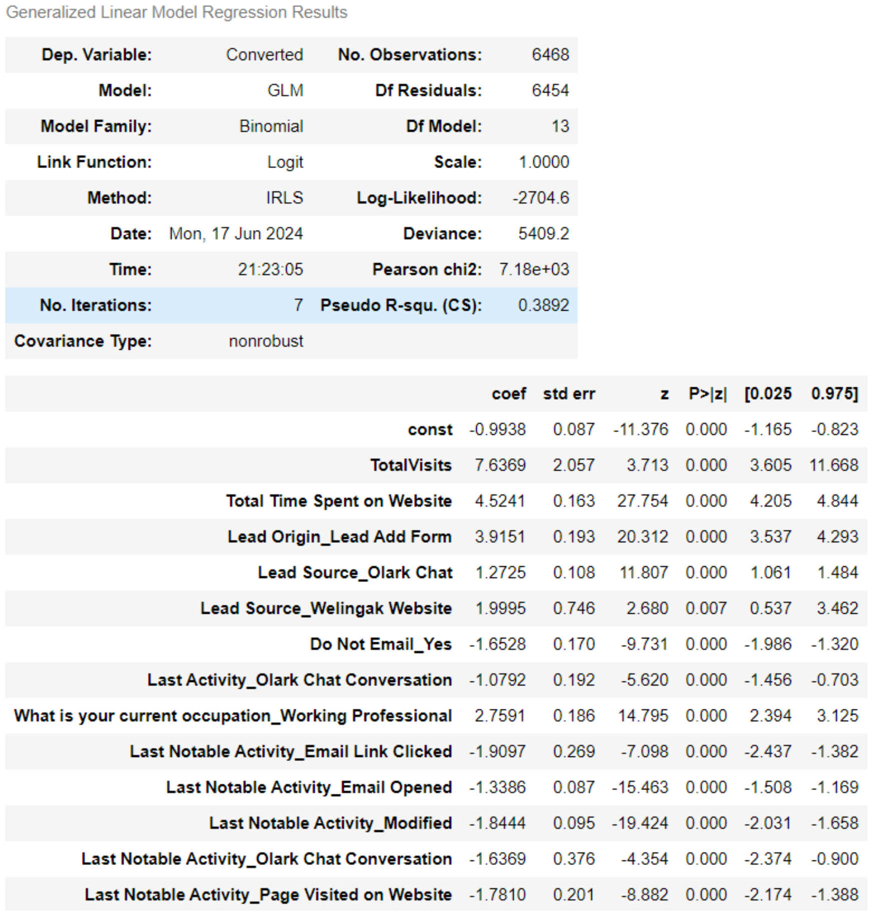


Visulaising for numerical variables



Model selection

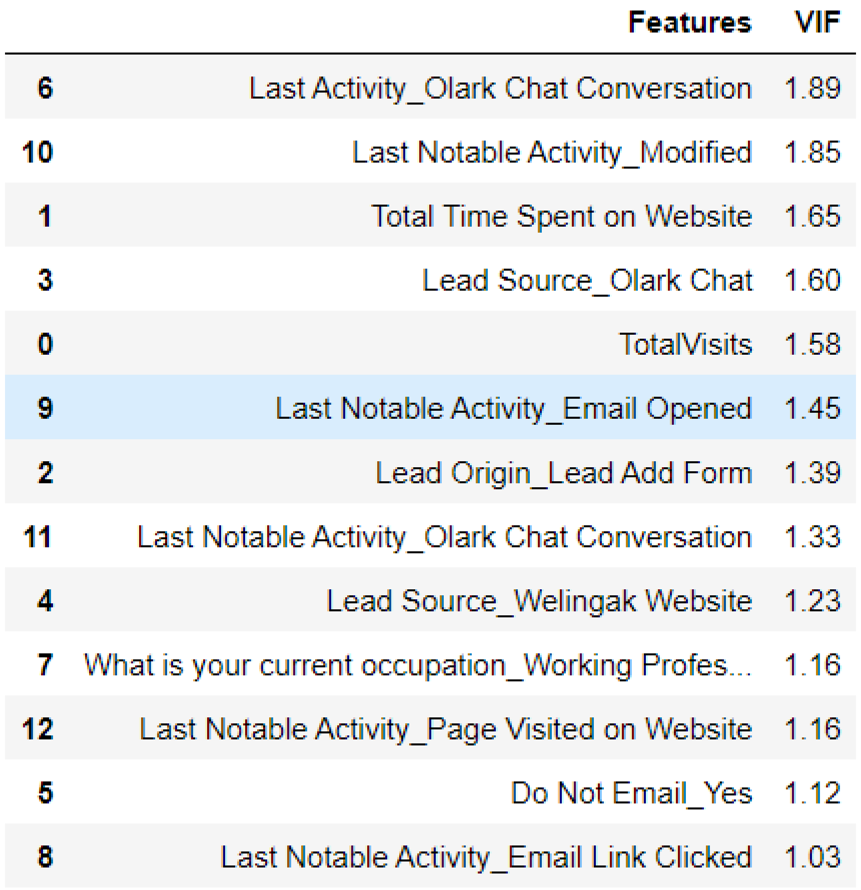
The selected model has VIF and p values below the desired values



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2704.6
Date:	Mon, 17 Jun 2024	**Deviance:**	5409.2
Time:	21:23:05	**Pearson chi2:**	7.18e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3892
Covariance Type:	nonrobust		

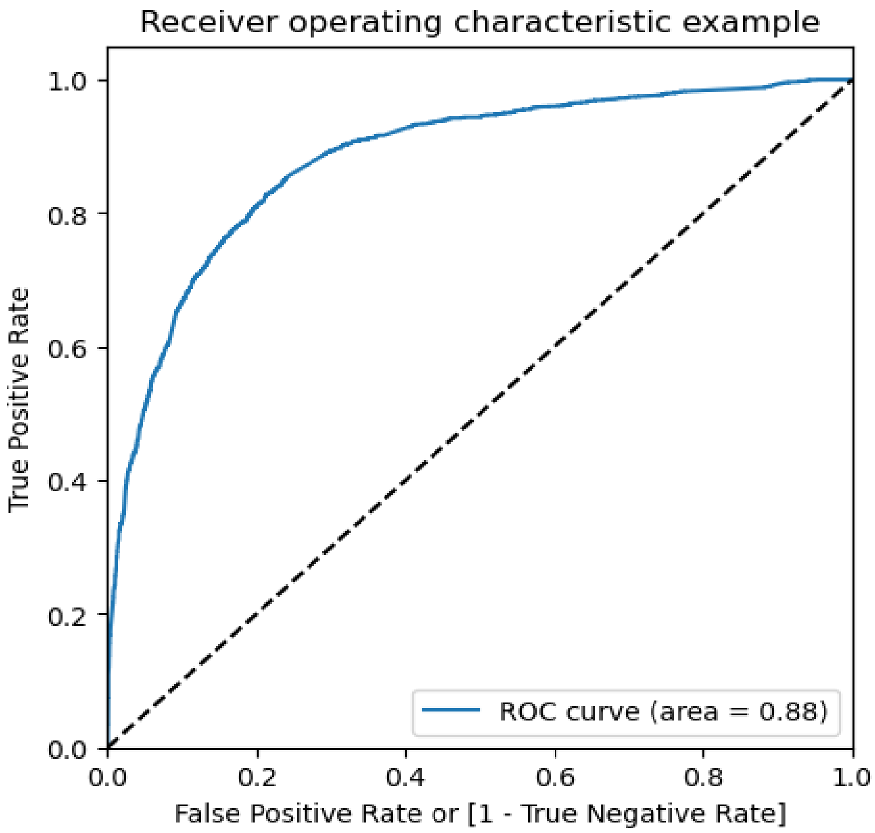
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9938	0.087	-11.376	0.000	-1.165	-0.823
TotalVisits	7.6369	2.057	3.713	0.000	3.605	11.668
Total Time Spent on Website	4.5241	0.163	27.754	0.000	4.205	4.844
Lead Origin_Lead Add Form	3.9151	0.193	20.312	0.000	3.537	4.293
Lead Source_Olark Chat	1.2725	0.108	11.807	0.000	1.061	1.484
Lead Source_Welingak Website	1.9995	0.746	2.680	0.007	0.537	3.462
Do Not Email_Yes	-1.6528	0.170	-9.731	0.000	-1.986	-1.320
Last Activity_Olark Chat Conversation	-1.0792	0.192	-5.620	0.000	-1.456	-0.703
What is your current occupation_Working Professional	2.7591	0.186	14.795	0.000	2.394	3.125
Last Notable Activity_Email Link Clicked	-1.9097	0.269	-7.098	0.000	-2.437	-1.382
Last Notable Activity_Email Opened	-1.3386	0.087	-15.463	0.000	-1.508	-1.169
Last Notable Activity_Modified	-1.8444	0.095	-19.424	0.000	-2.031	-1.658
Last Notable Activity_Olark Chat Conversation	-1.6369	0.376	-4.354	0.000	-2.374	-0.900
Last Notable Activity_Page Visited on Website	-1.7810	0.201	-8.882	0.000	-2.174	-1.388



	Features	VIF
6	Last Activity_Olark Chat Conversation	1.89
10	Last Notable Activity_Modified	1.85
1	Total Time Spent on Website	1.65
3	Lead Source_Olark Chat	1.60
0	TotalVisits	1.58
9	Last Notable Activity_Email Opened	1.45
2	Lead Origin_Lead Add Form	1.39
11	Last Notable Activity_Olark Chat Conversation	1.33
4	Lead Source_Welingak Website	1.23
7	What is your current occupation_Working Profes...	1.16
12	Last Notable Activity_Page Visited on Website	1.16
5	Do Not Email_Yes	1.12
8	Last Notable Activity_Email Link Clicked	1.03

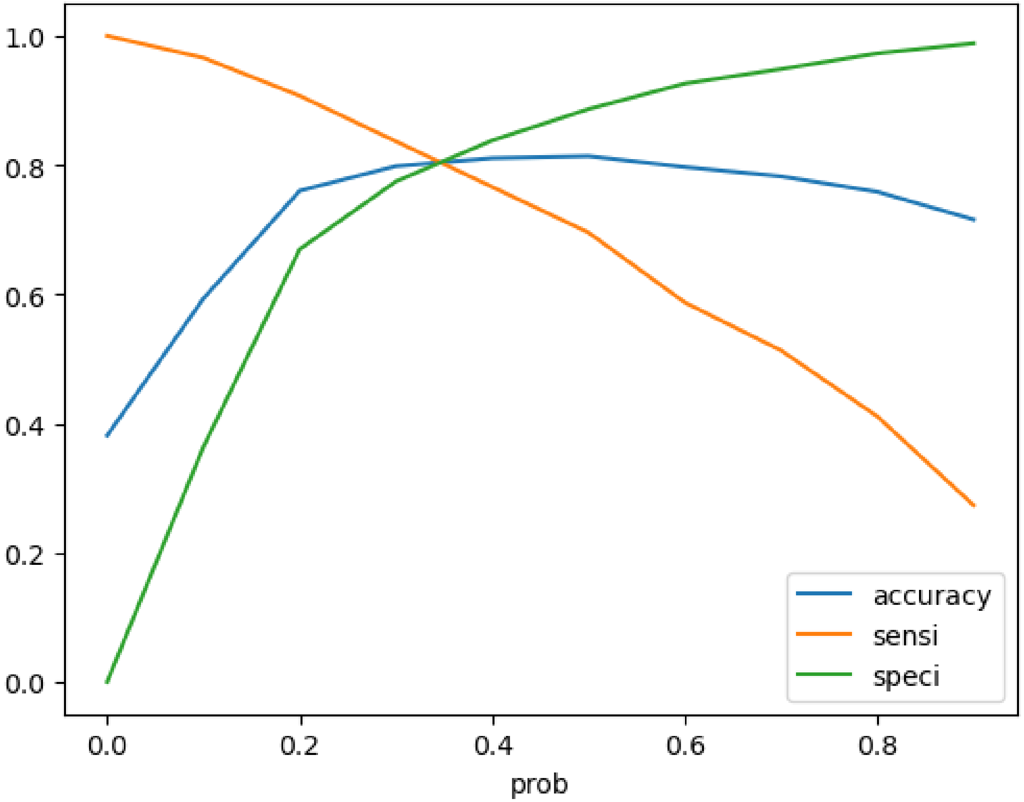
ROC curve

The area under Roc is 0.88 which is good



Determine cutoff

0.37 seems to be the cutoff



Results

- Train data
  - Accuracy : 81%
  - Sensitivity : 78%
  - Specificity : 81%
  - Precision : 75
  - Recall : 75

- Test Data
  - Accuracy :81%
  - Sensitivity : 78%
  - Specificity : 82%
  - Precision:77
  - Recall:77

Conclusion

- Optimal cut off was selected based on the Sensitivity and Specificity of the model.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 78% and 82% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are
  a. TotalVisits
  b. Total Time Spent on Website
  c. Lead Origin_Lead Add Form
- overall this model seems to be good.