Lead Scoring Case Study

Submitted By

Anshuman Barthakur

Pradeep Yadav

Usha Rana

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.

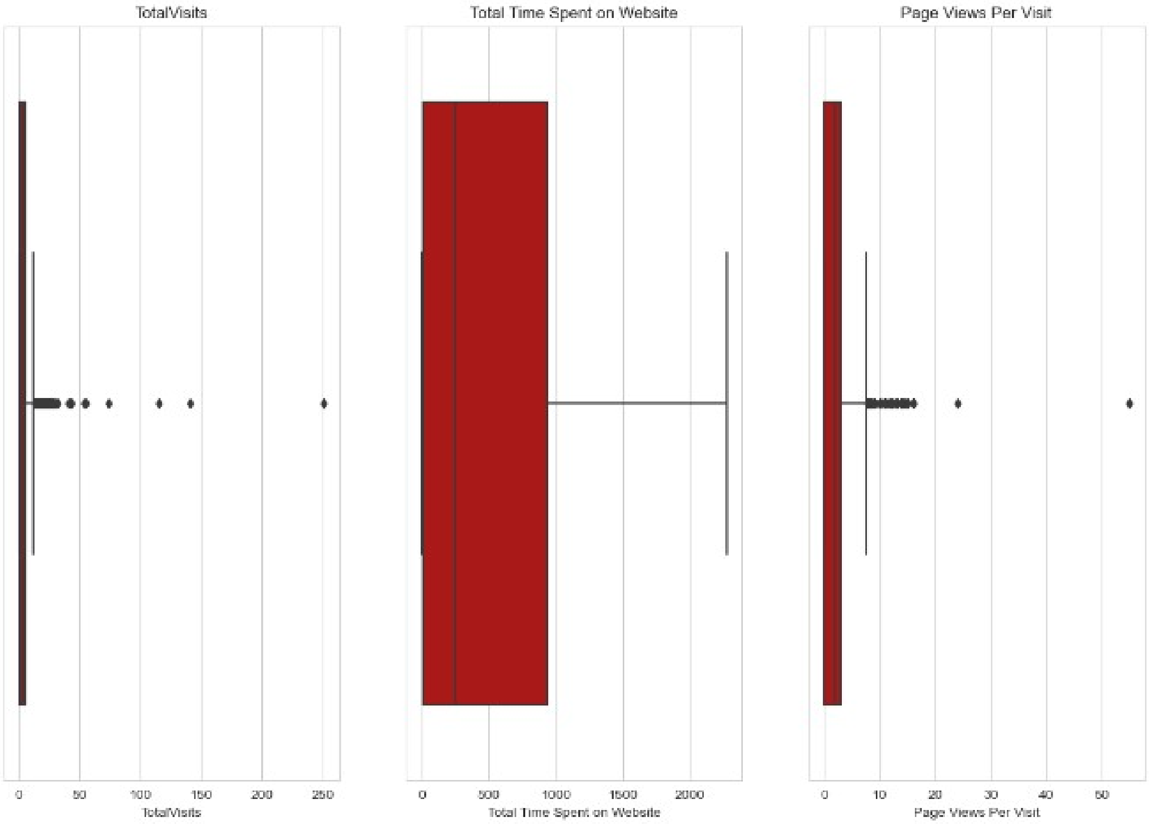
The company markets its courses on several websites and search engines like google. Once these people land on the website, they might browse the course or fill up a form or the course or watch some videos. When these people fill up a form for providing their email address or phone number, they are classified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

The company requires us to build a model where in we need to assign a lead score 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.

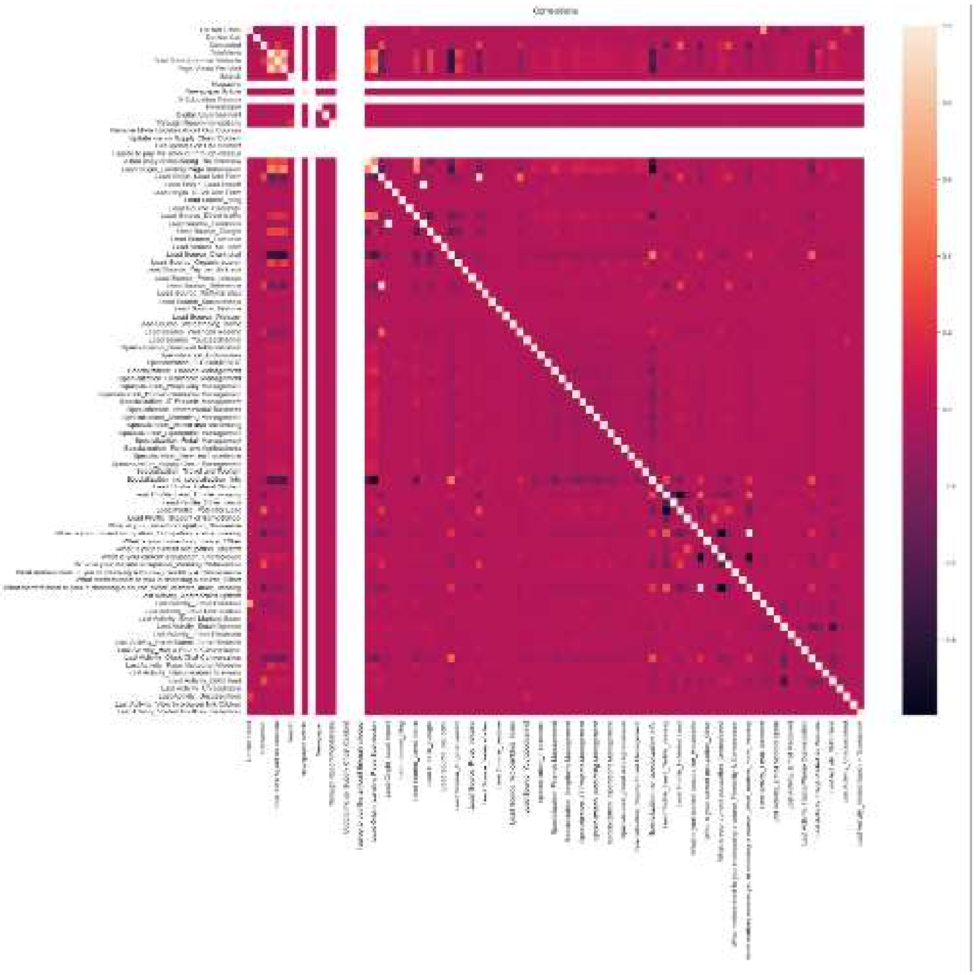
Outlier analysis

- There are outliers in "TotalVisit" colums and "Page Views Per Visit" column.
- To treat them we have to do 0.99-0.1% analysis to get rid of the outliers.



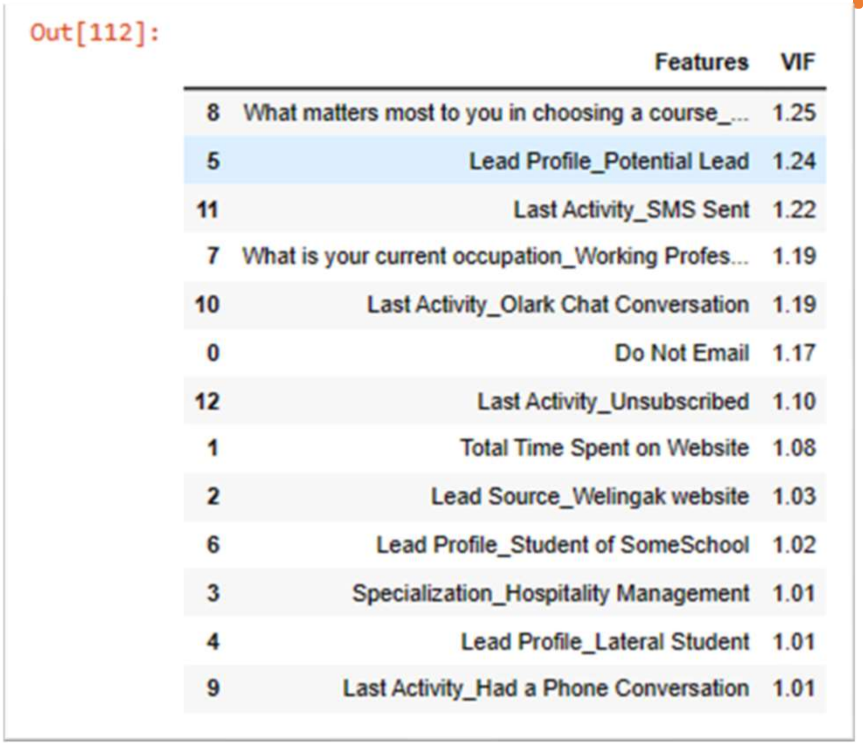
CORRELATION

- From the attached heatmap, we observe that there are many correlated attributes that needs to be removed.
- Highly correlated attributes create dependency on various independent factors which will gives us inappropriate results.



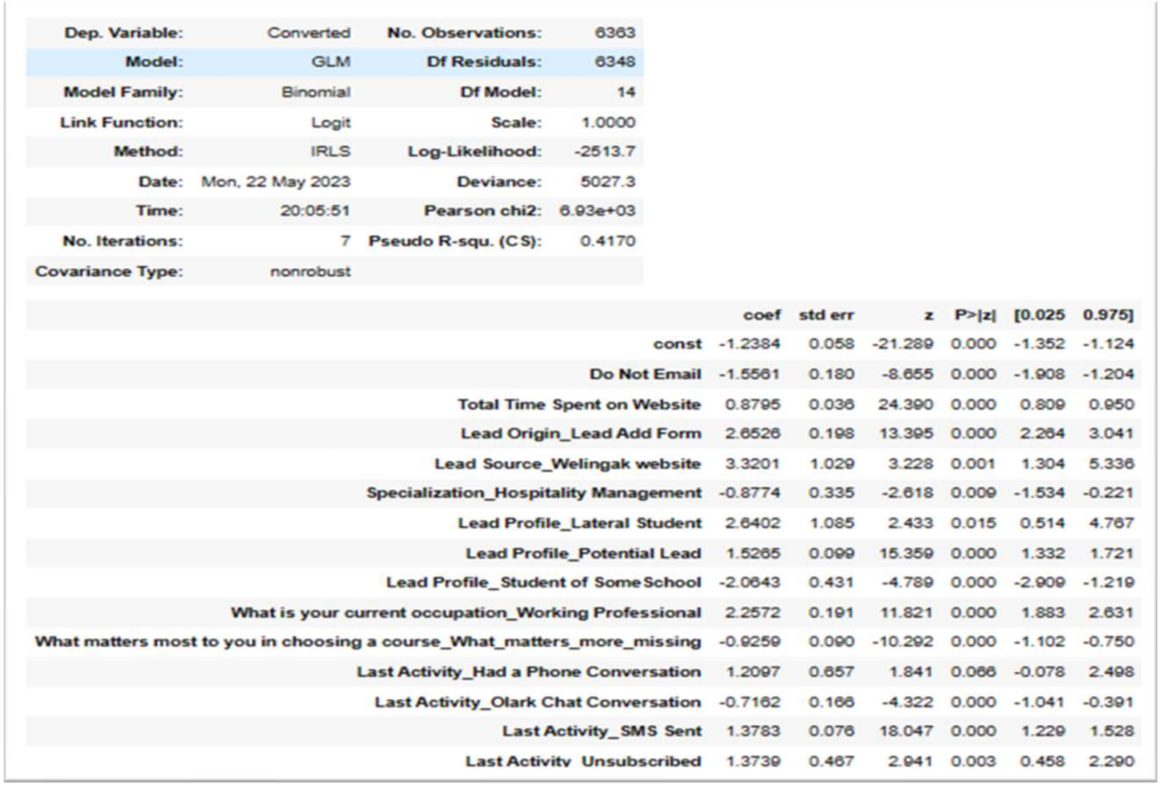
MODEL BUILDING

- With the help of RFE, we can identify the insignificant variables present in our model.



Out[112]:

	Features	VIF
8	What matters most to you in choosing a course_...	1.25
5	Lead Profile_Potential Lead	1.24
11	Last Activity_SMS Sent	1.22
7	What is your current occupation_Working Profes...	1.19
10	Last Activity_Olark Chat Conversation	1.19
0	Do Not Email	1.17
12	Last Activity_Unsubscribed	1.10
1	Total Time Spent on Website	1.08
2	Lead Source_Welingak website	1.03
6	Lead Profile_Student of SomeSchool	1.02
3	Specialization_Hospitality Management	1.01
4	Lead Profile_Lateral Student	1.01
9	Last Activity_Had a Phone Conversation	1.01

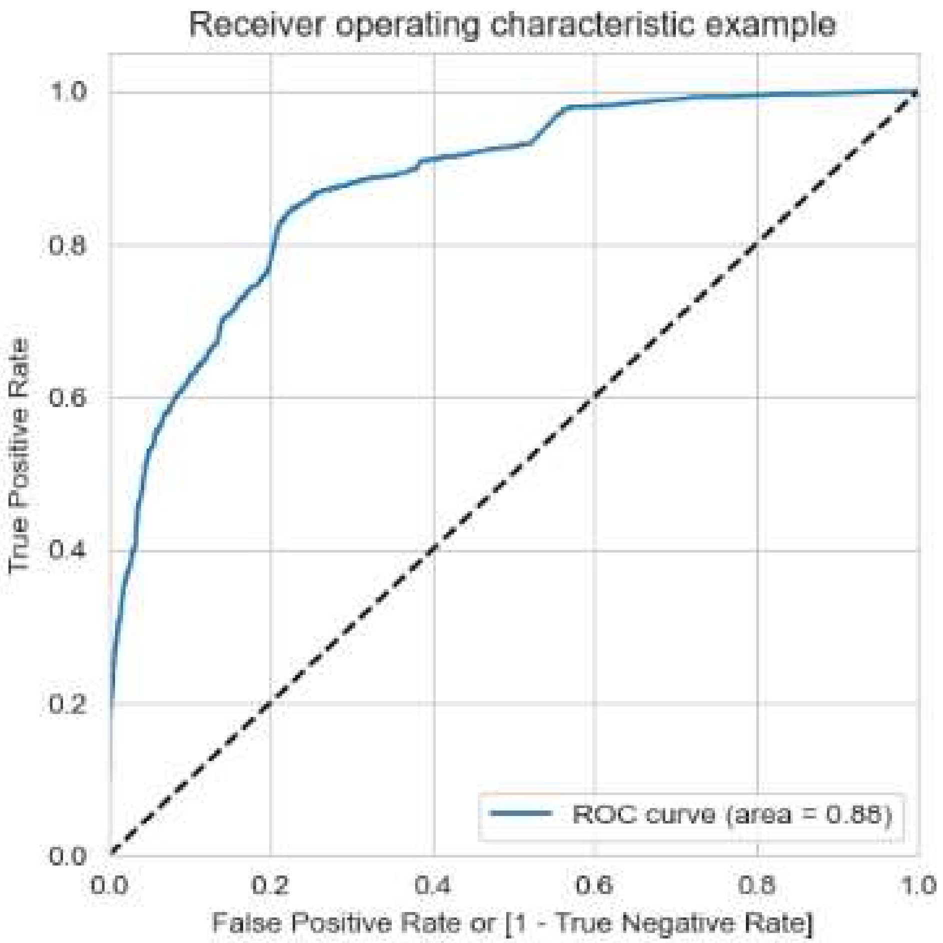


Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6348
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2513.7
Date:	Mon, 22 May 2023	Deviance:	5027.3
Time:	20:05:51	Pearson chi2:	6.93e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4170
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2384	0.058	-21.289	0.000	-1.352	-1.124
Do Not Email	-1.5561	0.180	-8.655	0.000	-1.908	-1.204
Total Time Spent on Website	0.8795	0.036	24.390	0.000	0.809	0.950
Lead Origin_Lead Add Form	2.6526	0.198	13.395	0.000	2.264	3.041
Lead Source_Welingak website	3.3201	1.029	3.228	0.001	1.304	5.336
Specialization_Hospitality Management	-0.8774	0.335	-2.618	0.009	-1.534	-0.221
Lead Profile_Lateral Student	2.6402	1.085	2.433	0.015	0.514	4.767
Lead Profile_Potential Lead	1.5265	0.099	15.359	0.000	1.332	1.721
Lead Profile_Student of SomeSchool	-2.0643	0.431	-4.789	0.000	-2.909	-1.219
What is your current occupation_Working Professional	2.2572	0.191	11.821	0.000	1.883	2.631
What matters most to you in choosing a course_What_matters_more_missing	-0.9259	0.090	-10.292	0.000	-1.102	-0.750
Last Activity_Had a Phone Conversation	1.2097	0.657	1.841	0.066	-0.078	2.498
Last Activity_Olark Chat Conversation	-0.7162	0.166	-4.322	0.000	-1.041	-0.391
Last Activity_SMS Sent	1.3783	0.076	18.047	0.000	1.229	1.528
Last Activity_Unsubscribed	1.3739	0.467	2.941	0.003	0.458	2.290

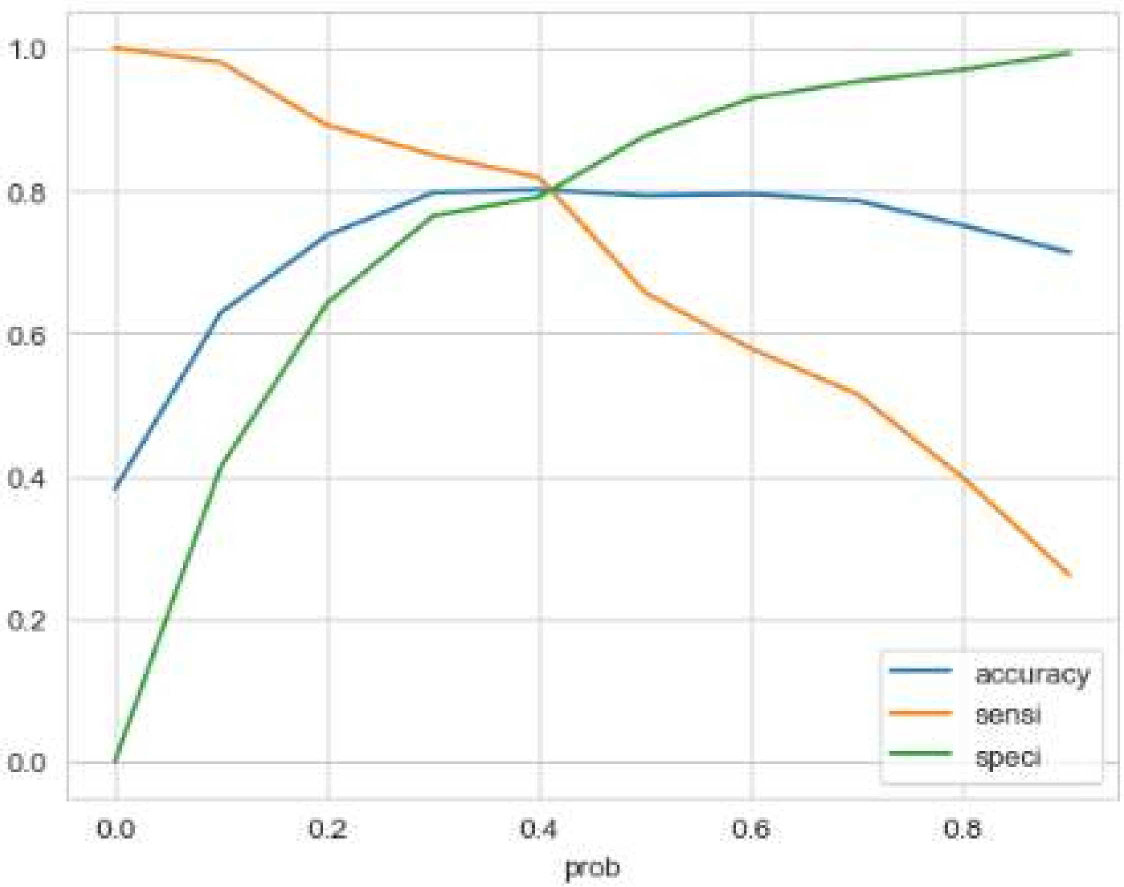
EVALUATING THE MODEL

- After building the final model making prediction on it(on train set), we creat ROC curve to find the model stability with AUC score(area under the curve) As we can see from the graph plotted on the right side, the area score is 0.88 which is a great score.
- And our graph is leaned towards the left side of the boarder which means we have good accuracy.



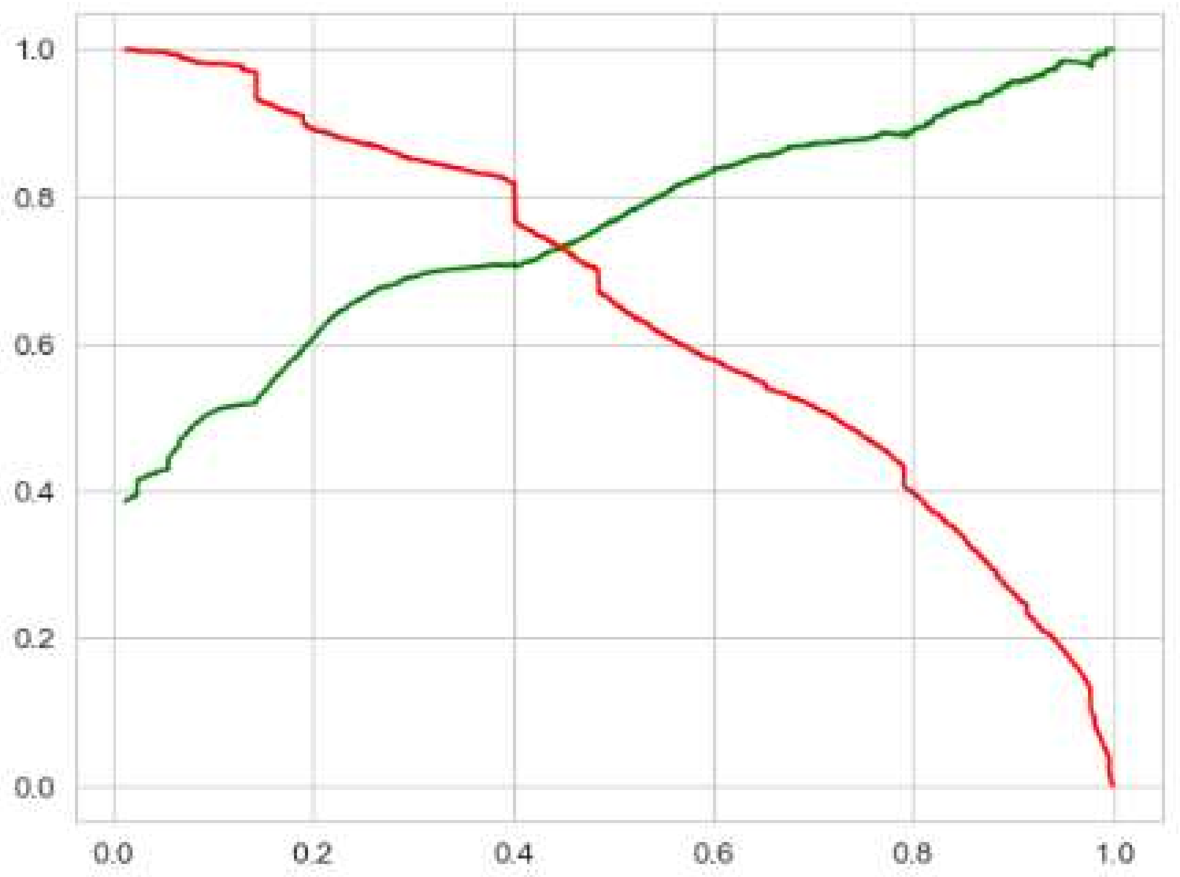
FINDING THE OPTIMAL CUT OFF POINT

- We found that on 0.4 point 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 of between Precision and recall.
- We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.5.



CONCLUSION

- The accuracy, Precision and Recall score we got from the test data are in the acceptable region.
- In business terms, this model has an ability to adjust with the company’s requirements in coming future.
- Important features responsible for good conversion rate or the one’s which contributes more towards the probability of a lead getting converted are:

- Do Not Email
- Total Time Spent on websites
- Lead Source_Welingak website
- Specialization_Hospitality Management
- Lead Profile Lateral student
- Lead Profile Student of Some School
- What is your current occupation_Working professional
- What matters most to you in choosing a course_what_matters_more_missing
- Last Activity_Had a phone Conversation
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- Last Activity_Unsubscribed