Leads Scoring Case Study

By - Chayan Banerjee and Pulkit Arya

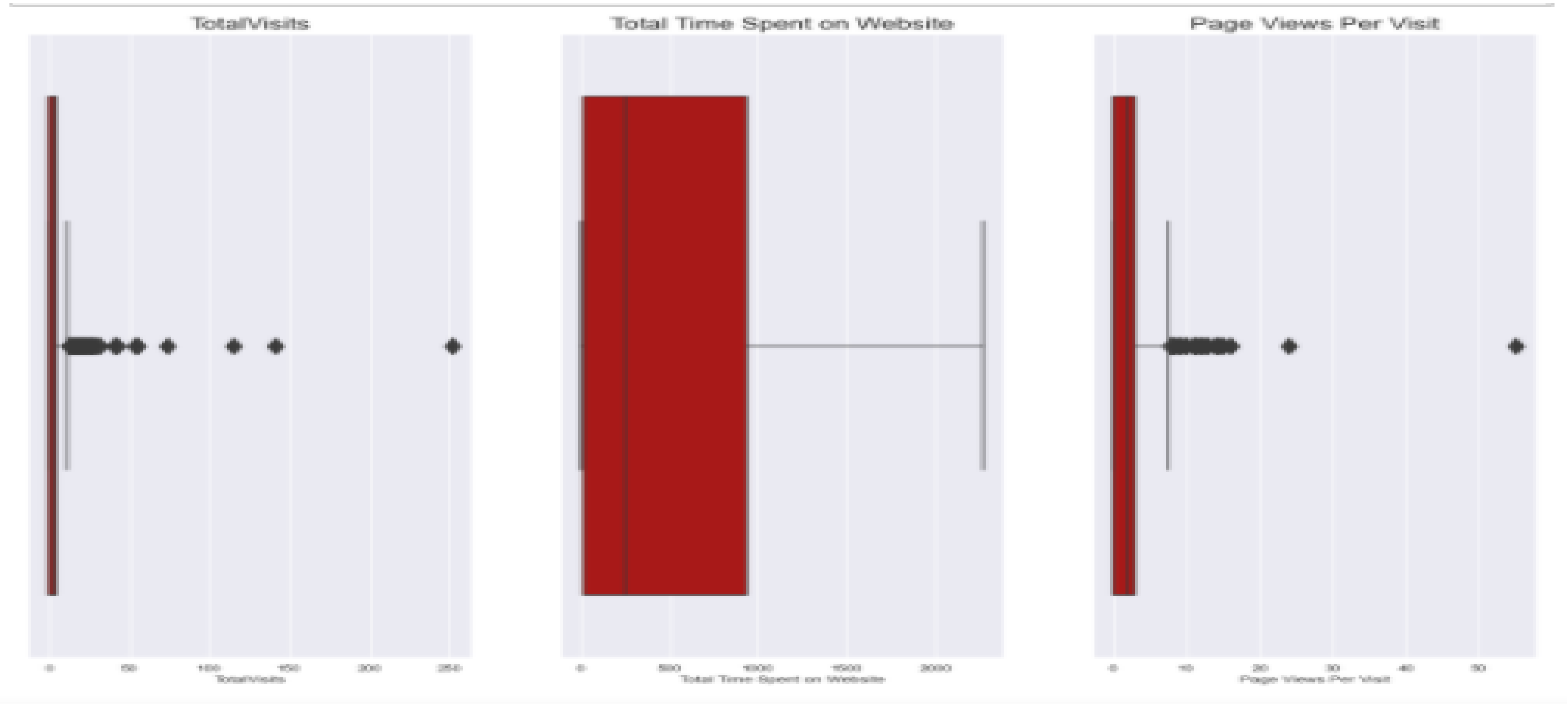
Problem statement

Create a model such that the customers have higher conversion chance

How Did We Do It?

- We cleaned the data
- Created dummies and checked and visualised the outliers

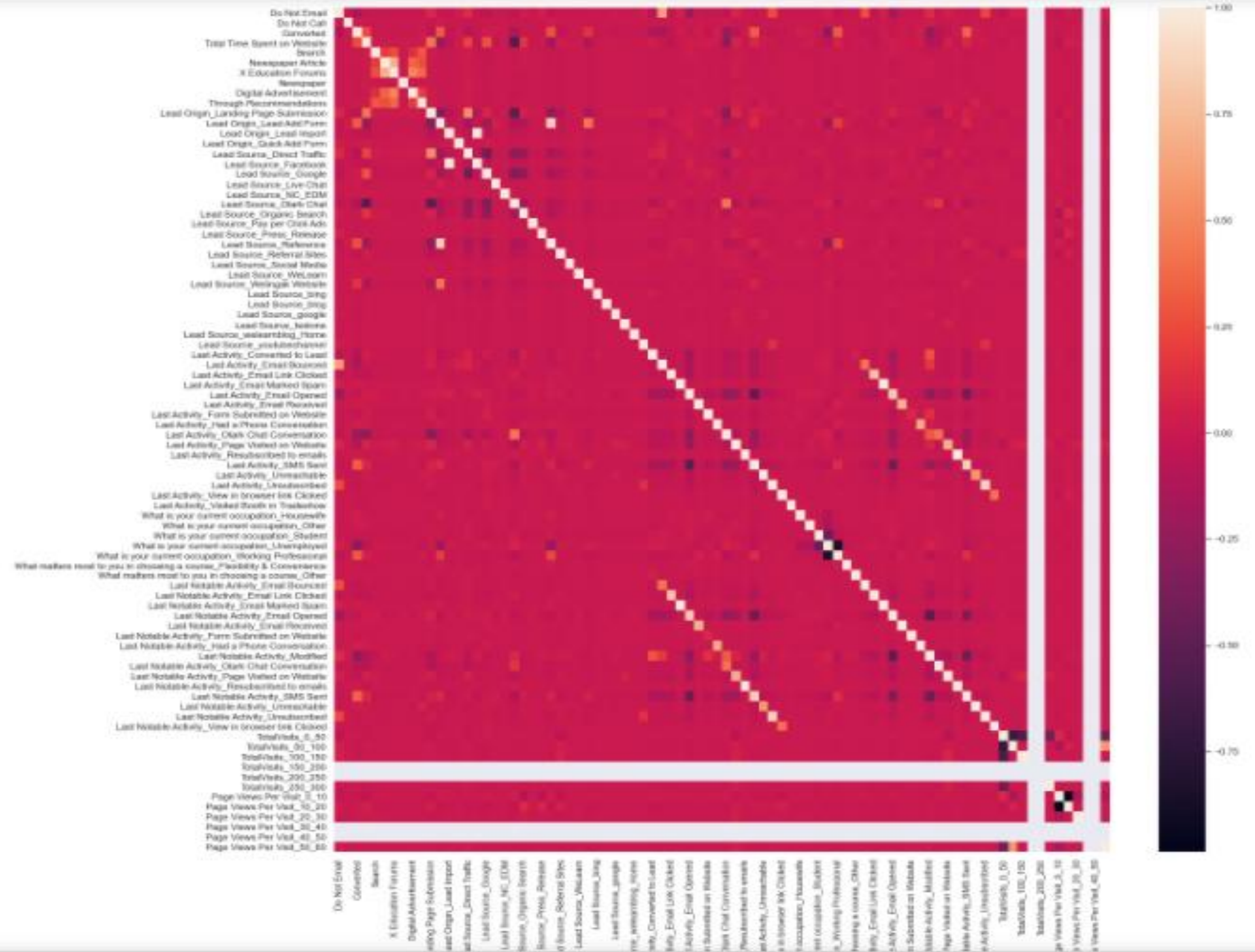
OUTLIERS:



Checked for Co-relation

An important step in data preparation, We split the dataset into train and test set and did standardization on some features.

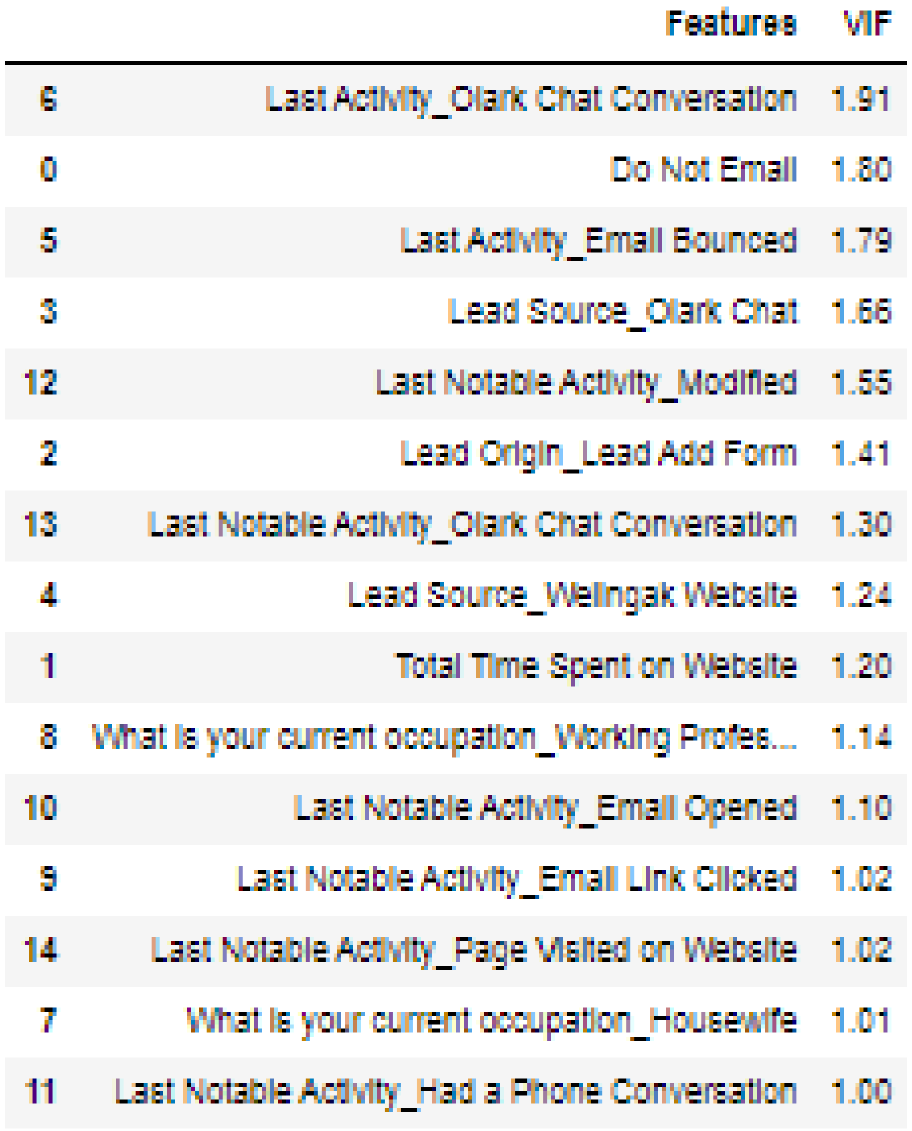
HEATMAP:-



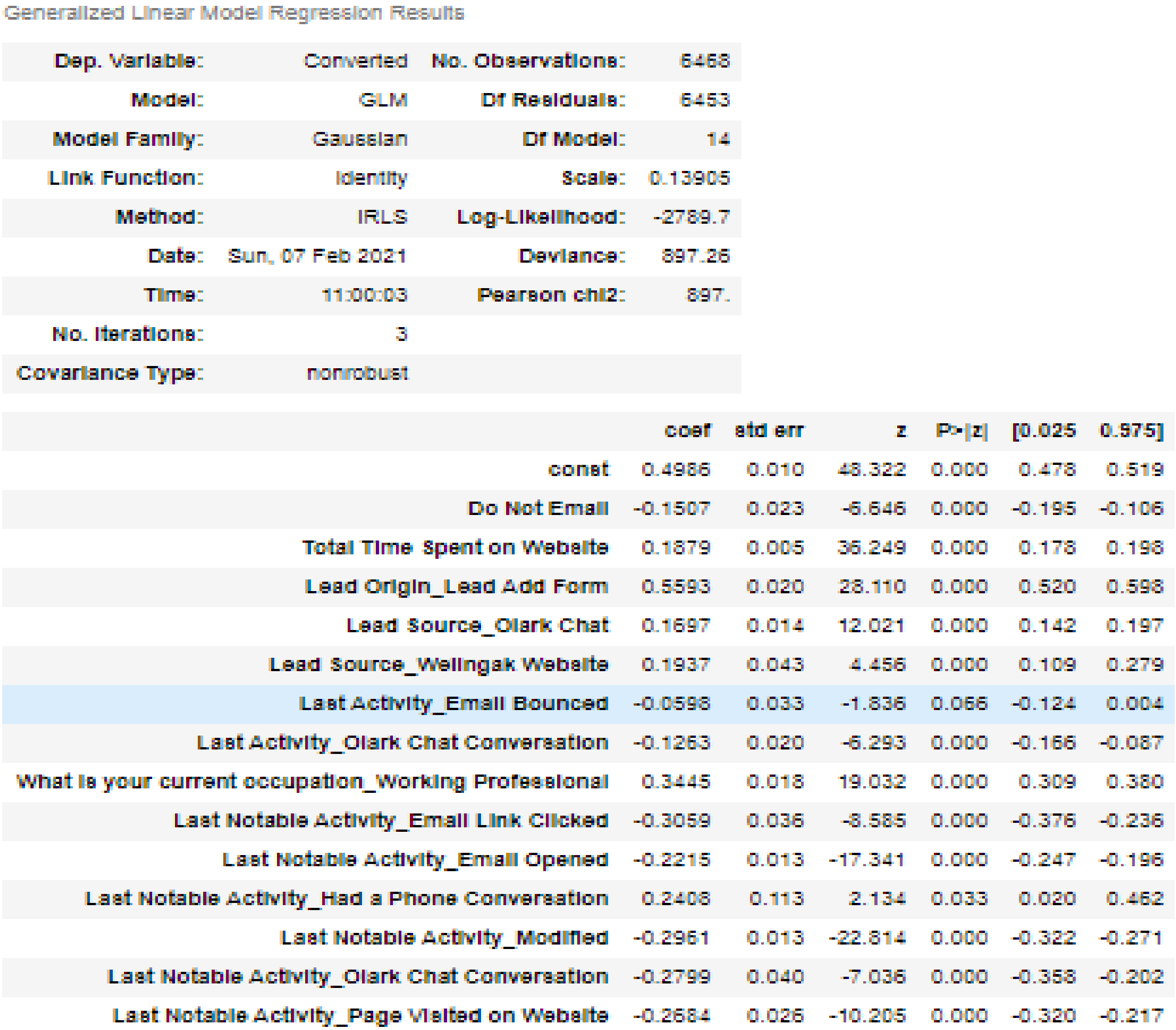
Our Model :-

- We started creating our model with rfe count 15 and went dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.
- we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

Our model visualization with VIF



	Features	VIF
6	Last Activity_Olark Chat Conversation	1.91
0	Do Not Email	1.80
5	Last Activity_Email Bounced	1.79
3	Lead Source_Olark Chat	1.66
12	Last Notable Activity_Modified	1.55
2	Lead Origin_Lead Add Form	1.41
13	Last Notable Activity_Olark Chat Conversation	1.30
4	Lead Source_Welingak Website	1.24
1	Total Time Spent on Website	1.20
8	What is your current occupation_Working Profes...	1.14
10	Last Notable Activity_Email Opened	1.10
9	Last Notable Activity_Email Link Clicked	1.02
14	Last Notable Activity_Page Visited on Website	1.02
7	What is your current occupation_Housewife	1.01
11	Last Notable Activity_Had a Phone Conversation	1.00



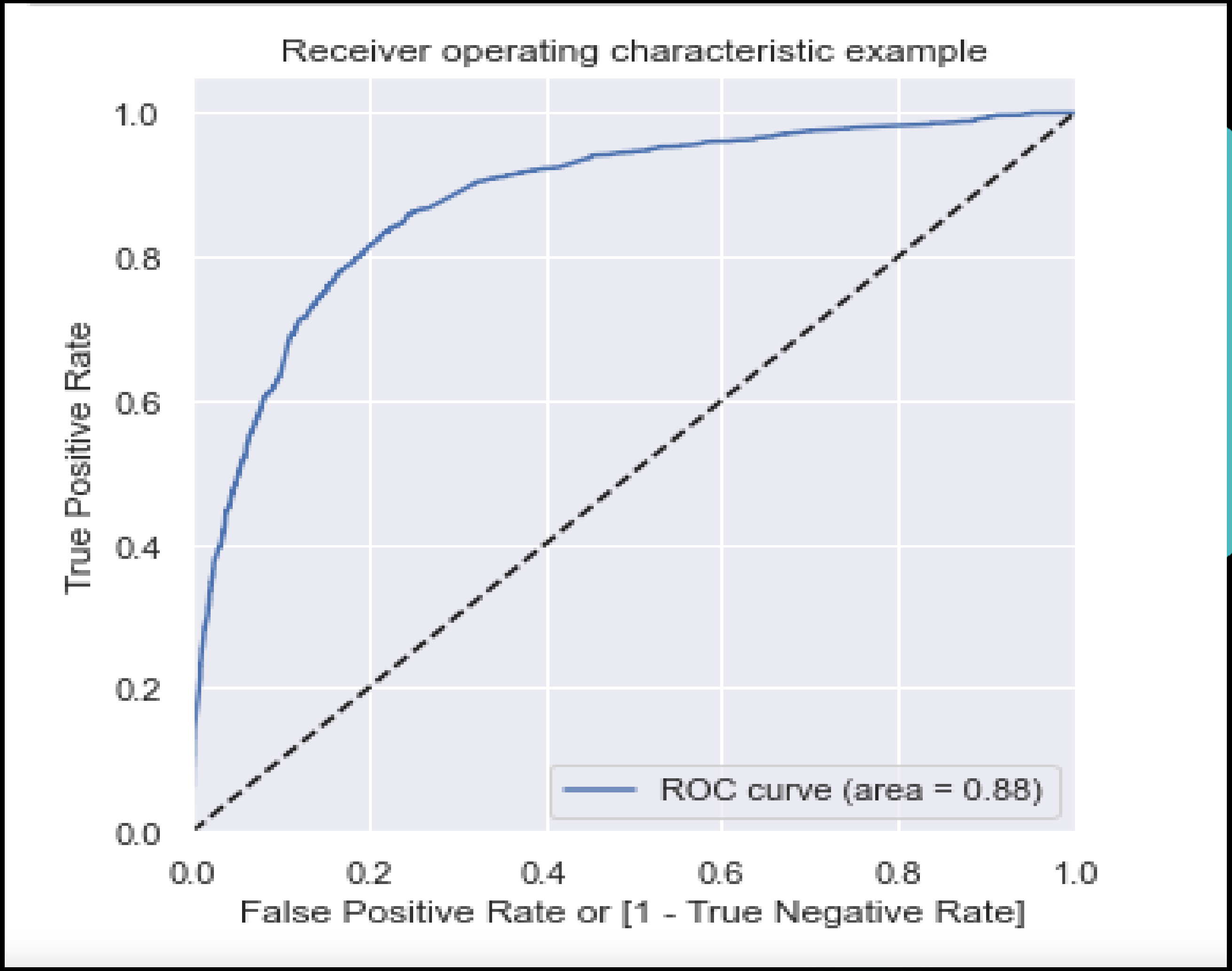
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6453
Model Family:	Gaussian	Df Model:	14
Link Function:	identity	Scale:	0.13905
Method:	IRLS	Log-Likelihood:	-2789.7
Date:	Sun, 07 Feb 2021	Deviance:	897.26
Time:	11:00:03	Pearson chi2:	897.
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4986	0.010	48.322	0.000	0.478	0.519
Do Not Email	-0.1507	0.023	-6.646	0.000	-0.195	-0.106
Total Time Spent on Website	0.1879	0.005	36.249	0.000	0.178	0.198
Lead Origin_Lead Add Form	0.5593	0.020	28.110	0.000	0.520	0.598
Lead Source_Olark Chat	0.1697	0.014	12.021	0.000	0.142	0.197
Lead Source_Welingak Website	0.1937	0.043	4.456	0.000	0.109	0.279
Last Activity_Email Bounced	-0.0598	0.033	-1.836	0.066	-0.124	0.004
Last Activity_Olark Chat Conversation	-0.1263	0.020	-6.293	0.000	-0.166	-0.087
What is your current occupation_Working Professional	0.3445	0.018	19.032	0.000	0.309	0.380
Last Notable Activity_Email Link Clicked	-0.3059	0.036	-8.585	0.000	-0.376	-0.236
Last Notable Activity_Email Opened	-0.2215	0.013	-17.341	0.000	-0.247	-0.196
Last Notable Activity_Had a Phone Conversation	0.2408	0.113	2.134	0.033	0.020	0.462
Last Notable Activity_Modified	-0.2961	0.013	-22.814	0.000	-0.322	-0.271
Last Notable Activity_Olark Chat Conversation	-0.2799	0.040	-7.036	0.000	-0.358	-0.202
Last Notable Activity_Page Visited on Website	-0.2684	0.026	-10.205	0.000	-0.320	-0.217

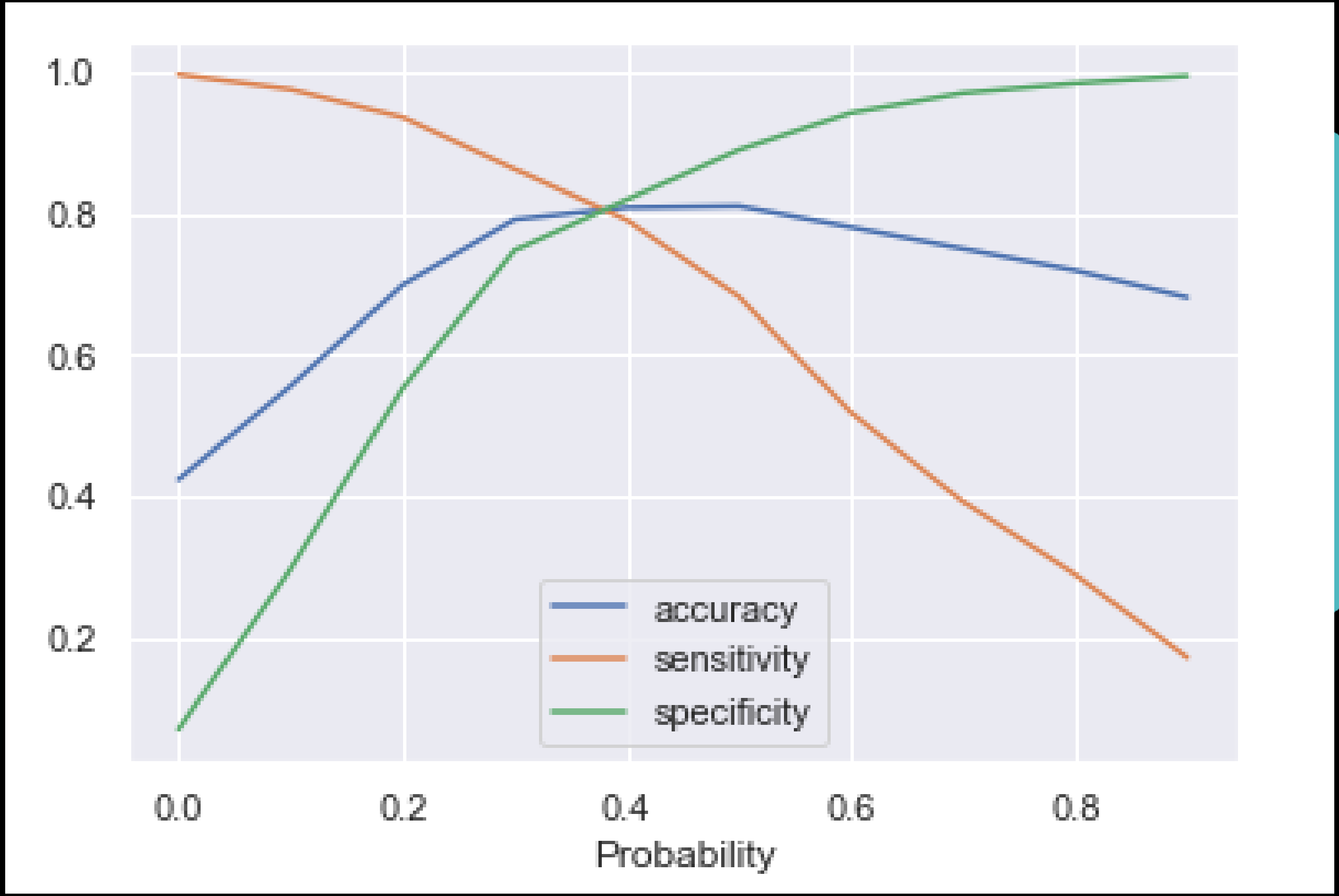
ROC Curve

- And our graph is leaned towards the left side of the border which means we have good accuracy.



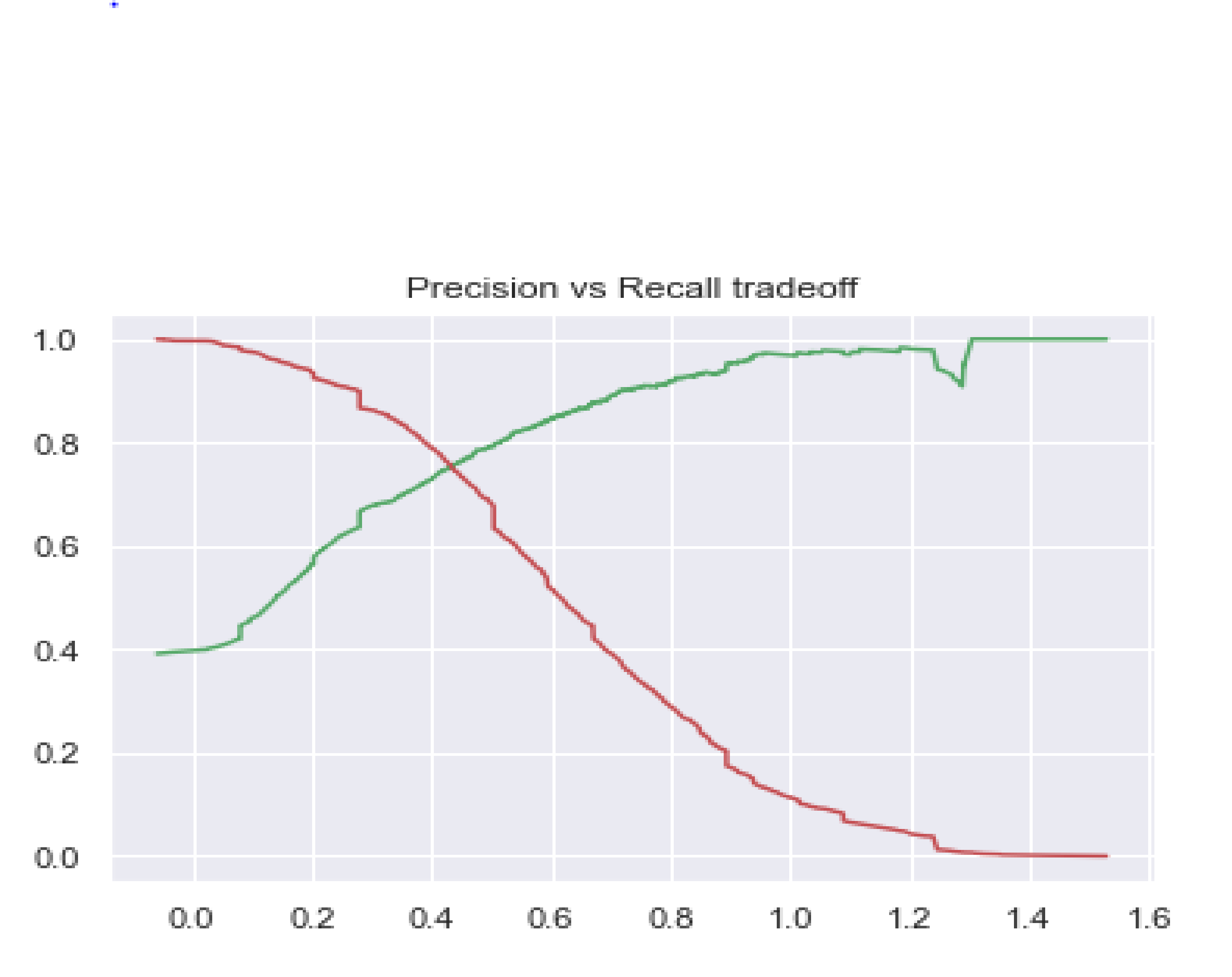
Finding the optimal cutoff point

- Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff.



Precision and Recall tradeoff

- We created a graph which will show us the tradeoff between Precision and recall.



Conclusion

- We have high recall score than precision score
- Important features are:
  - a) Last Notable Activity_Page Visited on Website
  - b) Last Notable Activity_Email Opened
  - c) Last Notable Activity_Email Link Clicked

Thank You