Lead Score Case Study

BY ZISHAN AHMED AND IRFAN MOMIN

Problem Statement

Create a logistic regression model which has high lead score for Higher conversion chance and low score for lower conversion chance. The Ballpark of the target lead conversion rate to be around 80%.

Also model build should be able to adjust if the company's requirement changes in future.

Approach of the Analysis.

- After cleaning the dataset we converted all binary variable to '0' and '1'.

- And then converted categorical variables to dummy variables.

- We checked outliers in the dataset and then plotted the box plot to see the outliers.

- As you can see from the box plot attached on the right side there are outliers for two variables we created bins for these two variables to manage the outliers.
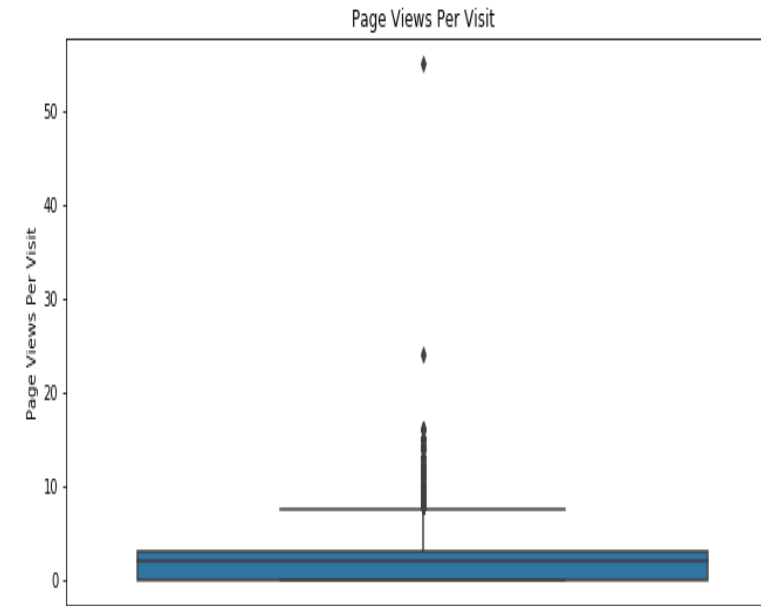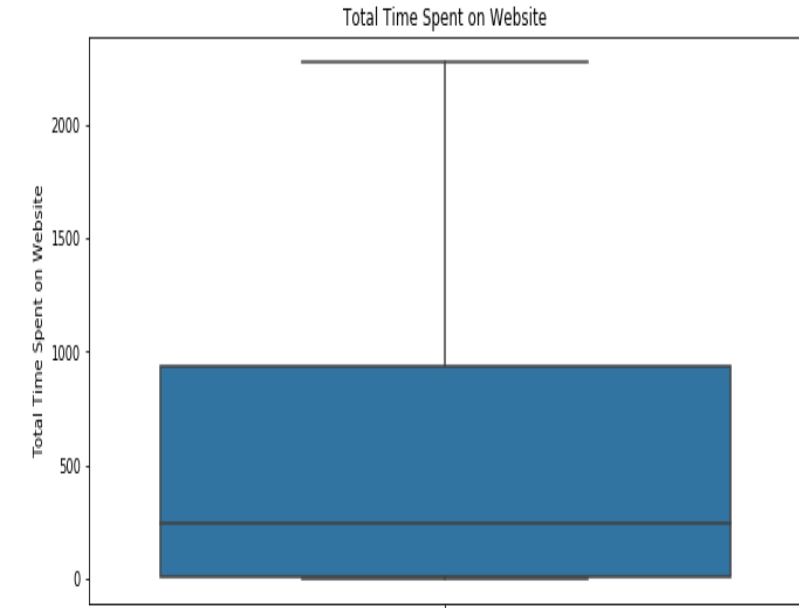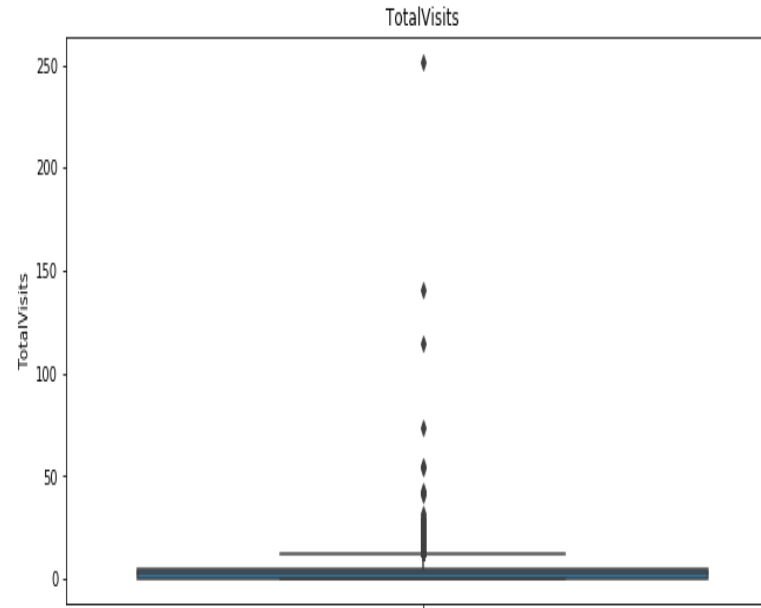
Correlations

- We split the dataset into train and test and do standardization on the features.

- Created Heatmap to find the correlations of the variables in our training dataset.

- As you can see from heat map there are two variables which are highly correlated we will be dropping these columns in next steps
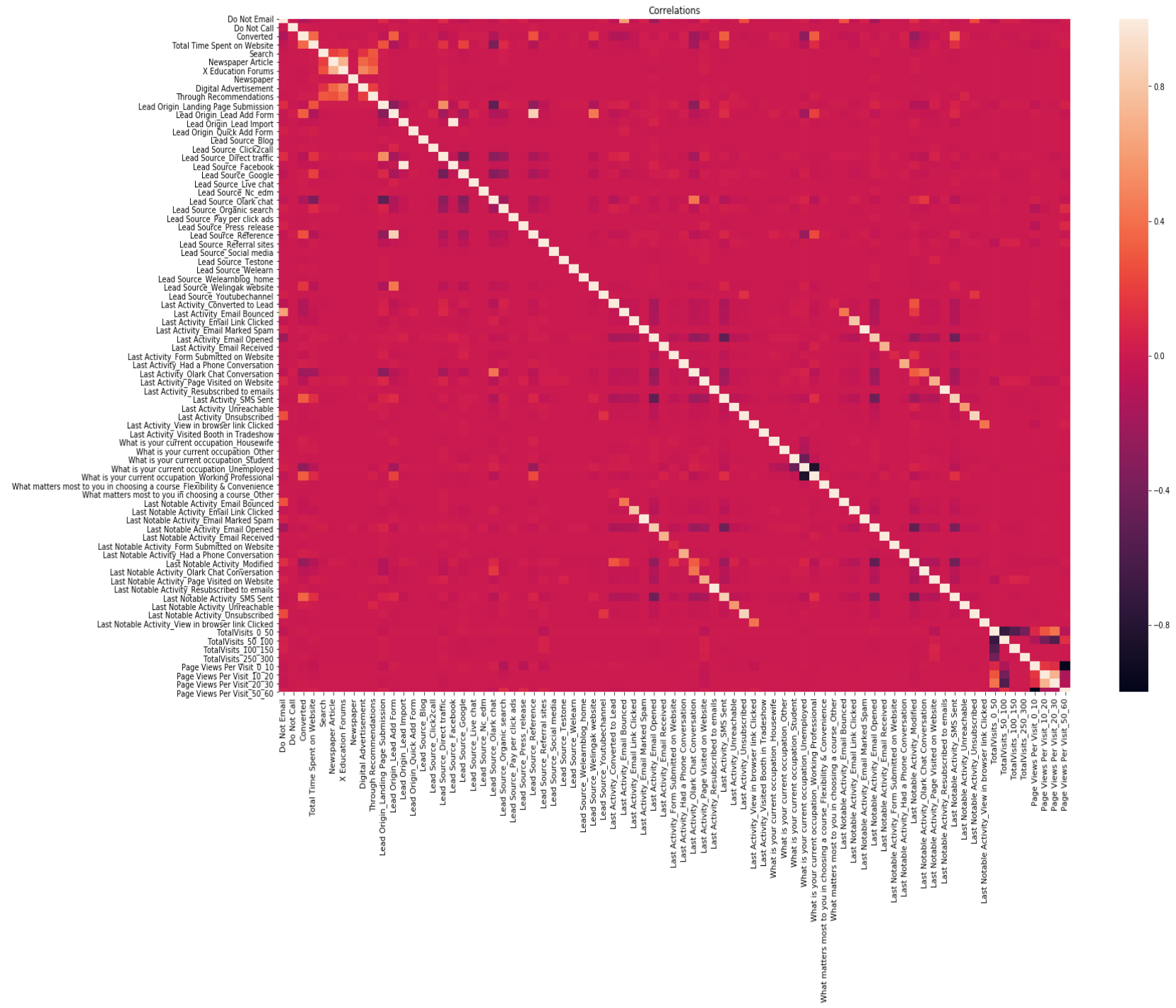
Correlations

- **After dropping the columns we will be plotting heatmap again to check the correlation of the variables as you can see from heatmap the highly correlated variables has been dropped.**

Correlations of X
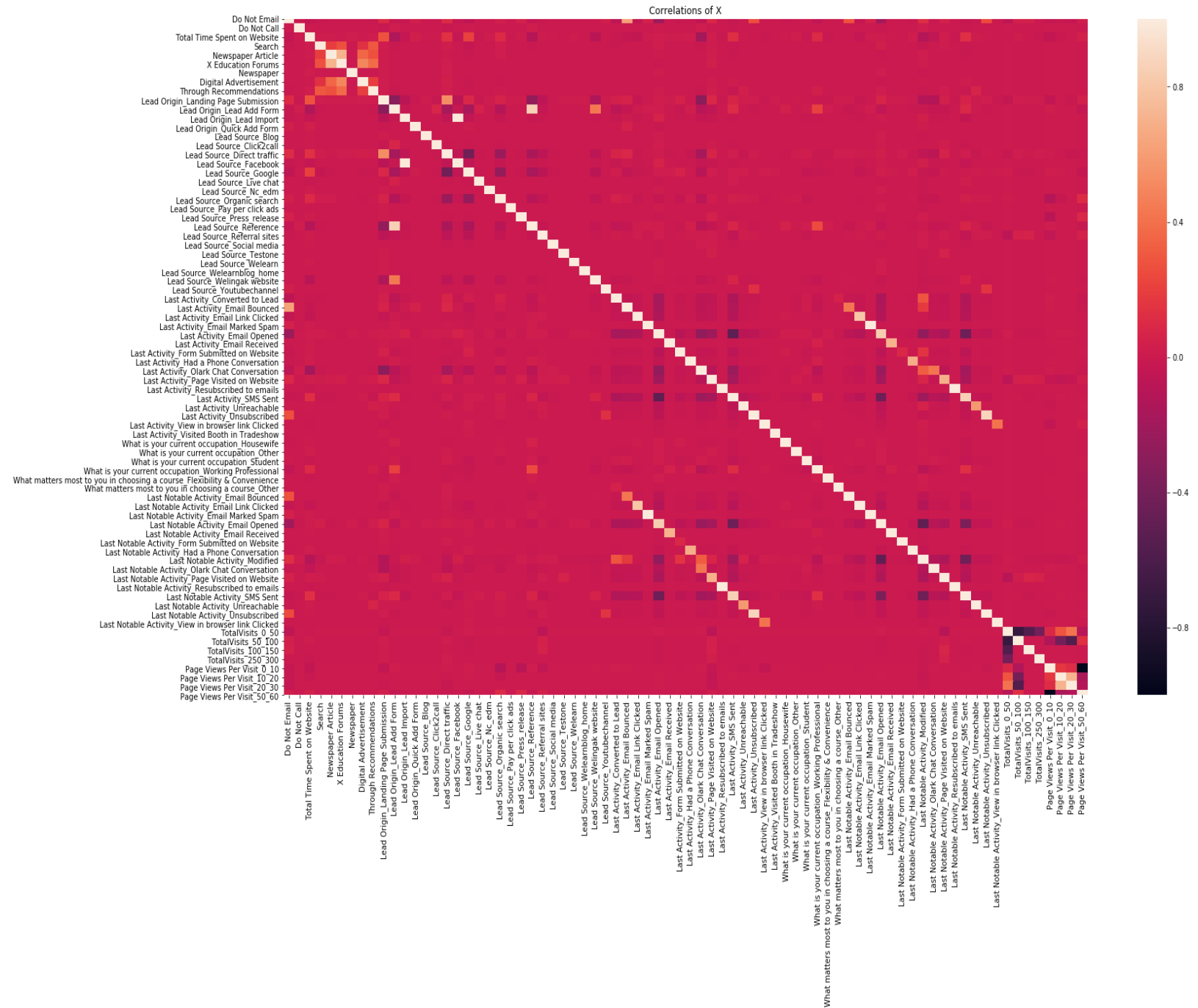
Model Building

- We built a model using all the variables and we found that there are variables which are insignificant and needs to be dropped.

- We used RFE feature selection to drop insignificant variables from the dataset we choose 15 variables in RFE.

- After selecting variables using RFE we created a model and then computed VIF for the model.

- Then we dropped the columns which were having high P values one by one and started re creating the mode until we have our final model.

Final Model and VIF values

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2749.5
Date:	Mon, 12 Apr 2021	Deviance:	5498.9
Time:	00:29:04	Pearson chi2:	7.06e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0088	0.049	-20.477	0.000	-1.105	-0.912
Do Not Email	-1.4008	0.160	-8.740	0.000	-1.715	-1.087
Total Time Spent on Website	0.9429	0.035	27.095	0.000	0.875	1.011
Lead Origin_Lead Add Form	3.1879	0.185	17.260	0.000	2.826	3.550
Lead Source_Direct traffic	-0.5206	0.078	-6.700	0.000	-0.673	-0.368
Lead Source_Welingak website	1.9924	0.743	2.680	0.007	0.535	3.449
Last Activity_Converted to Lead	-1.3398	0.212	-6.312	0.000	-1.756	-0.924
Last Activity_Olark Chat Conversation	-1.0501	0.159	-6.592	0.000	-1.362	-0.738
What is your current occupation_Working Professional	2.7875	0.187	14.900	0.000	2.421	3.154
Last Notable Activity_Had a Phone Conversation	3.2848	1.101	2.984	0.003	1.127	5.443
Last Notable Activity_SMS Sent	1.4738	0.078	18.917	0.000	1.321	1.627
Last Notable Activity_Unreachable	1.6261	0.516	3.153	0.002	0.615	2.637

|	features	VIF
2	Lead Origin_Lead Add Form	1.39
4	Lead Source_Welingak website	1.24
3	Lead Source_Direct traffic	1.19
9	Last Notable Activity_SMS Sent	1.19
7	What is your current occupation_Working Profes...	1.16
1	Total Time Spent on Website	1.10
0	Do Not Email	1.07
6	Last Activity_Olark Chat Conversation	1.04
5	Last Activity_Converted to Lead	1.03
8	Last Notable Activity_Had a Phone Conversation	1.00
10	Last Notable Activity_Unreachable	1.00

Plotting ROC Curve

- We will be plotting ROC curve to calculate Area Under Curve to see stability of our model.

- As you can see from graph on right side you can see we have Area Under Curve of about 87%.

- And also you can see that the graph is more leaning towards the left side of the borer which means we have good accuracy.

Receiver operating characteristic
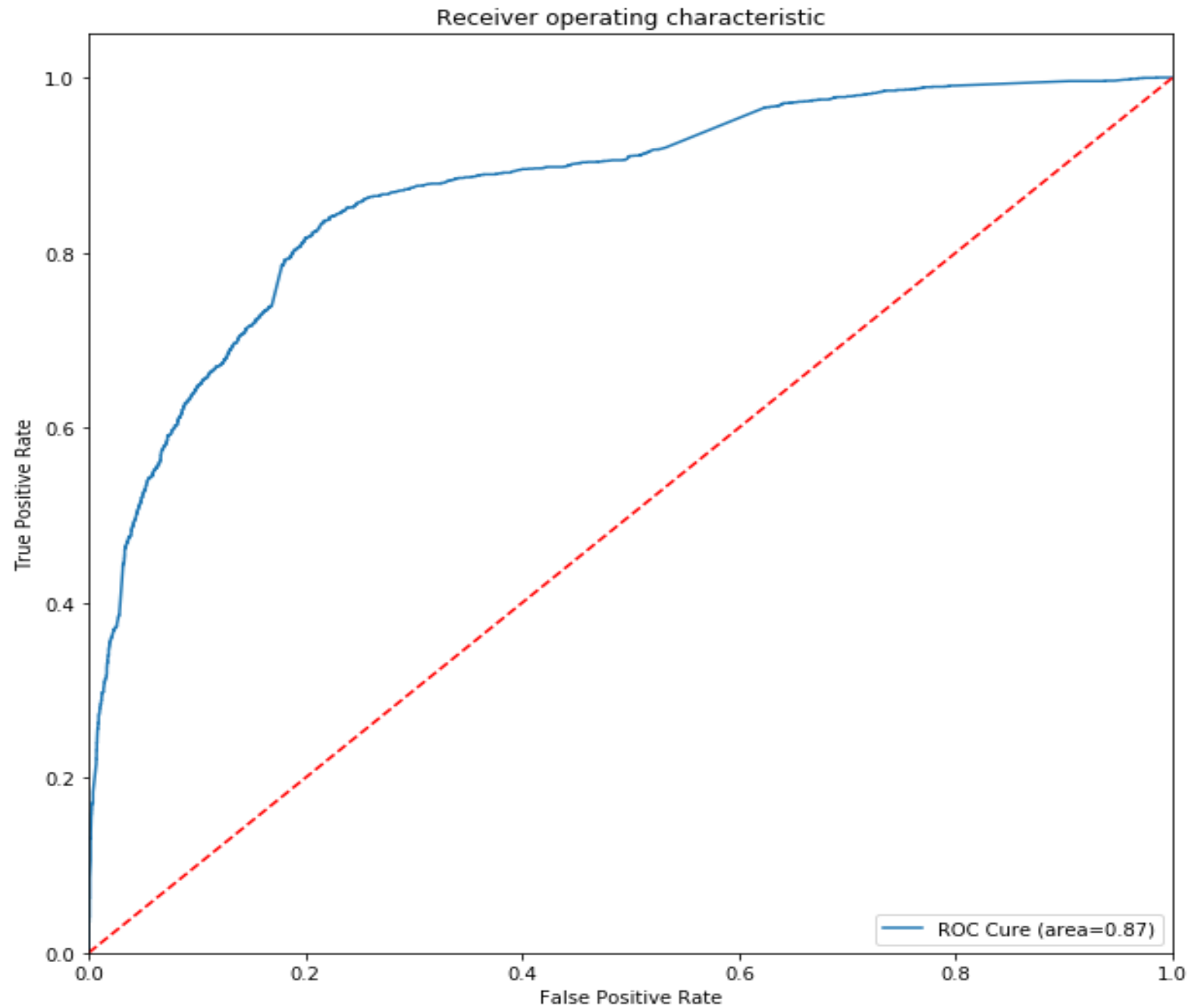
Finding Optimal cutoff point

- Once we are done with plotting ROC curve in next step we will be calculating Accuracy, Specificity and Sensitivity for each point to choose Optimal Cutoff Point.

- As you can see from the graph on right side we can say that optimal cutoff point is 0.4.
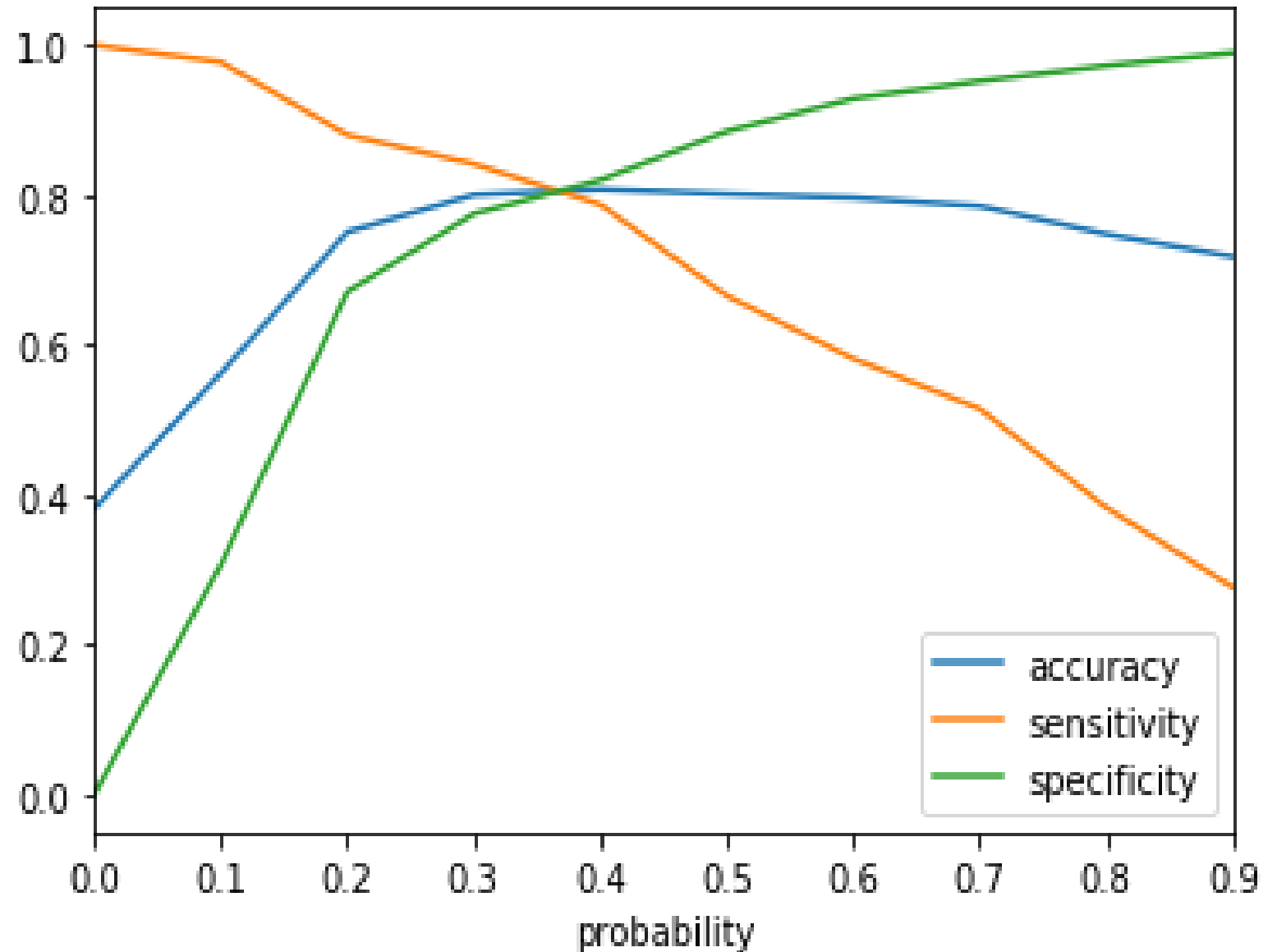
Precision and Recall

- We will be creating a new column 'Predicted' in our final dataset using the optimal cutoff point.

- Next step we will be calculating Precision and Recall. Both Precision and Recall plays important role in building our model more business oriented and also tell how our model is behaving.

- We got Precision and Recall values for model as below:
  - Precision : 0.72
  - Recall : 0.78

Precision Vs Recall Trade off

- Graph was plotted to check the trade off between Precision and Recall.

- As you can see from the graph on right side the meeting point for Precision vs Recall tradeoff is approximately 0.42

Precision Vs Recall tradeoff
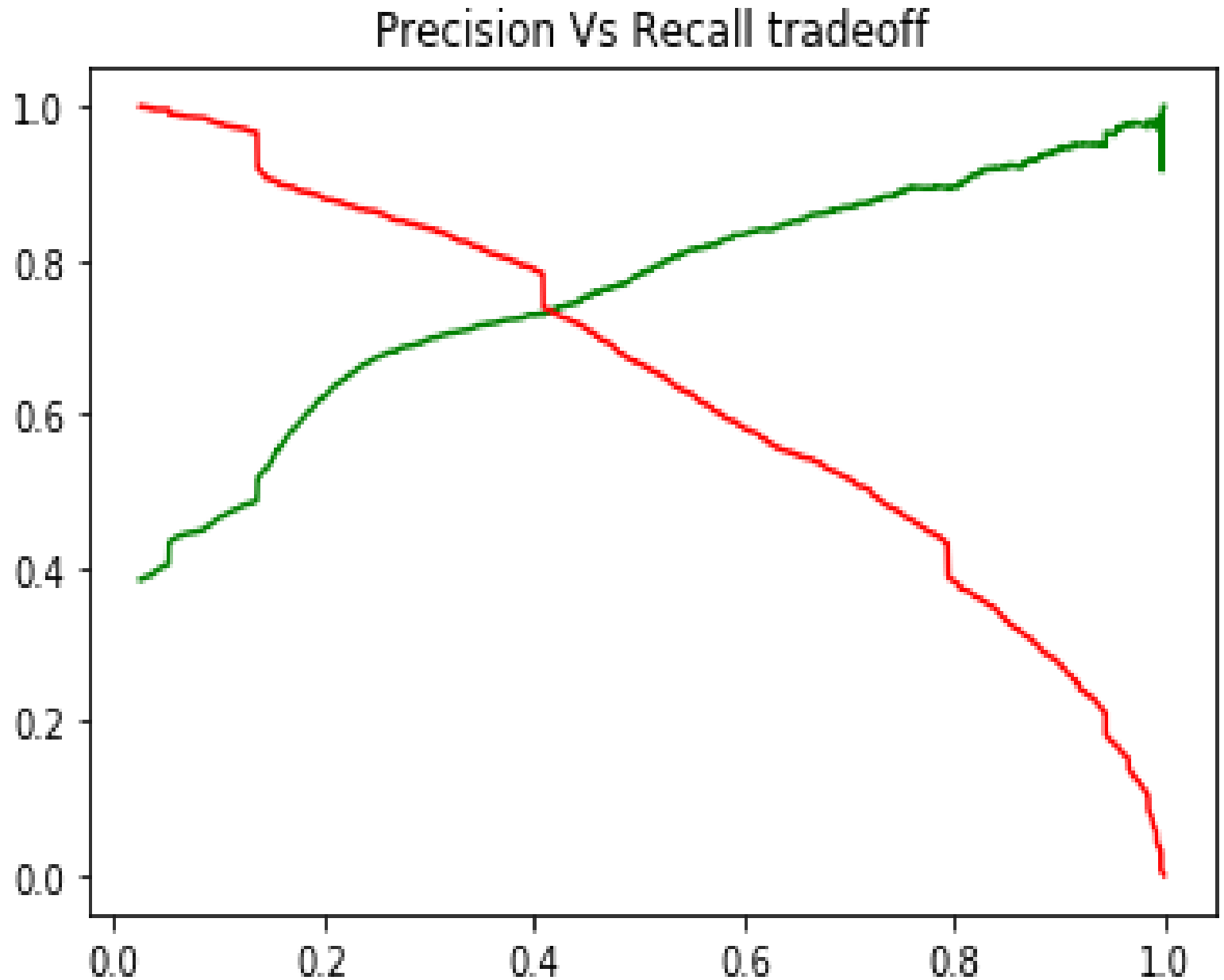
Making Prediction on Test Set

- Now we will be making predictions on Test Dataset.

- We standardized the test dataset and then created columns 'Lead Score' and 'Final_Predicted'.

- Then we will be calculating Accuracy, Sensitivity and Specificity for the Test Dataset.
  - Accuracy : 0.79

- And we also calculated Precision and Recall for the Test Dataset as well.

- We can see that test prediction is having accuracy, precision and recall in acceptable range.

- We can conclude that our model is stable with good accuracy, precision and recall.

- Lead Score is created on test dataset to identify hot leads – High the Lead Score Higher chance of getting converted and low the lead score lower chance of getting converted.

Conclusion

- Both train and test set have accuracy, recall in acceptable range.

- This model has ability to adjust in future with company requirements.

- Top feature for good conversion rate are as below:
  - Last Notable Activity_Had a Phone Conversation.
  - Lead Origin_Lead Add Form.
  - What is your current occupation_Working Professional

Thank You