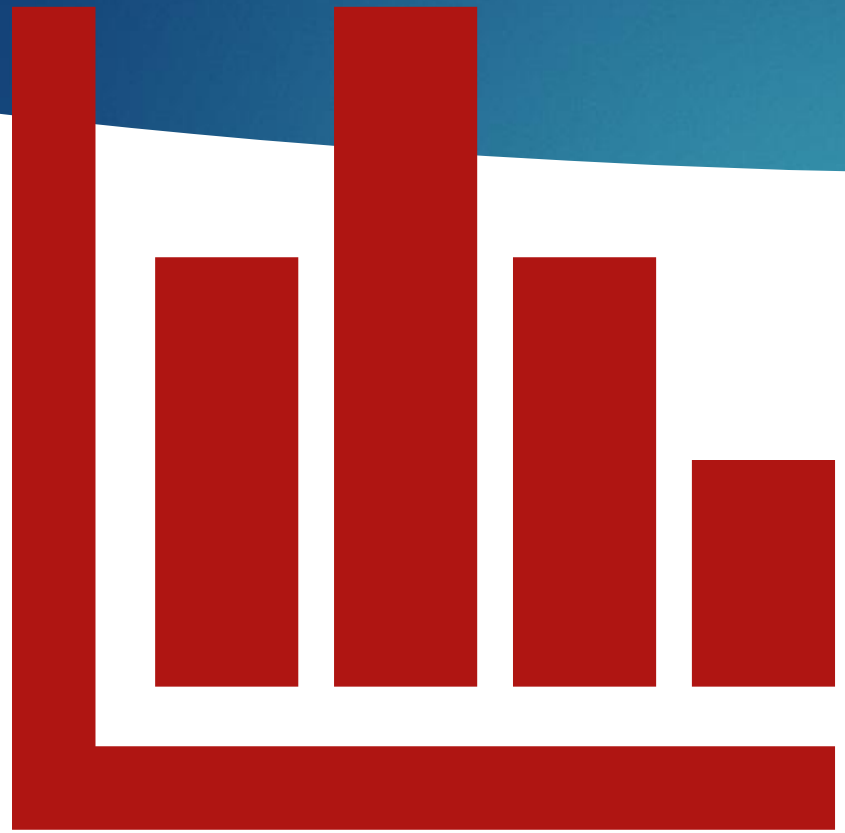


Leads Scoring Case Study

Aditya Kumar & Tanmay Kurmi

Problem statement

Create a model in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%.

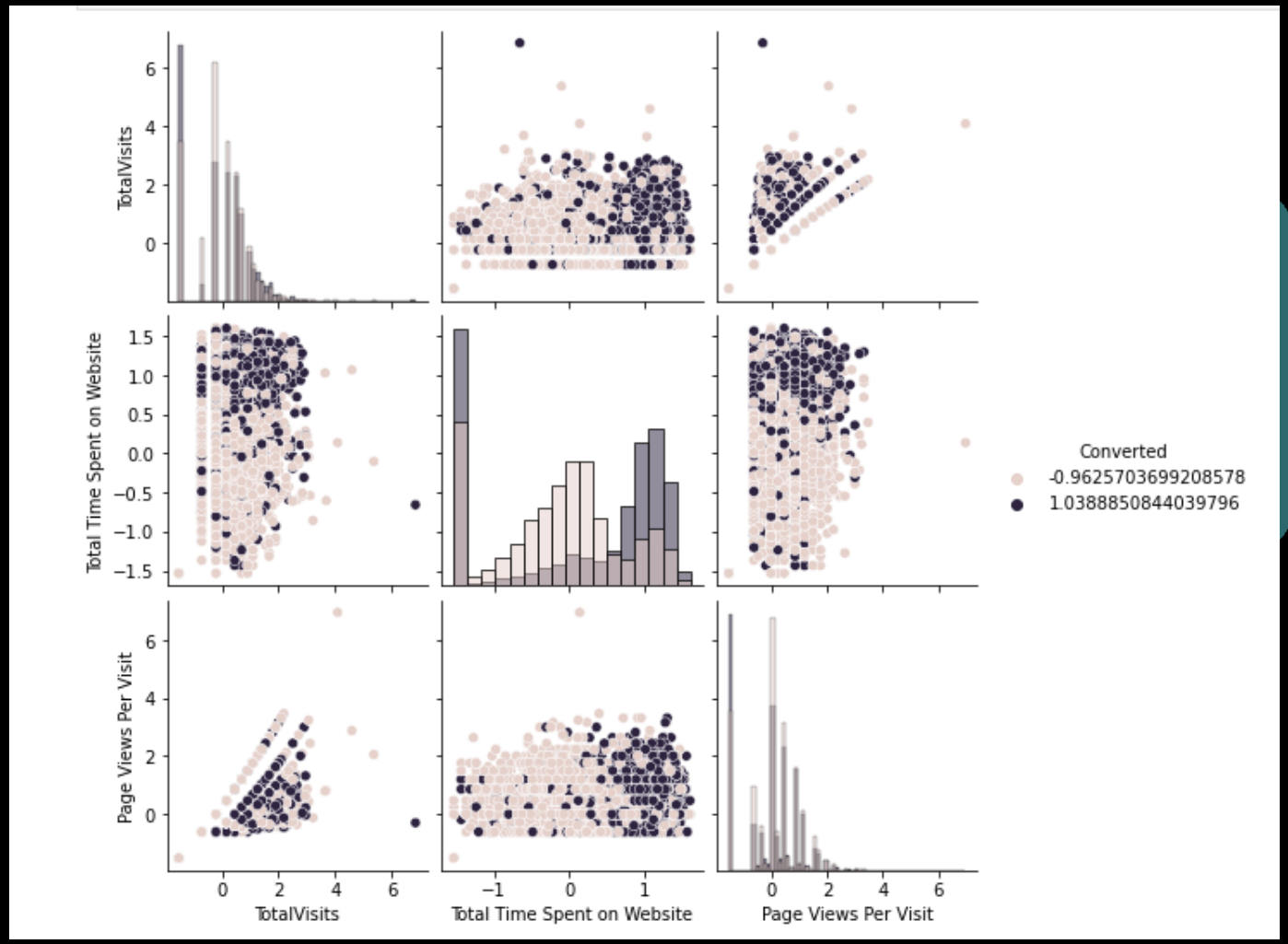
Also the model should be able to adjust if the company's requirement changes in near future.

Approach of the analysis

I. We started our analysis with our cleaned dataset by converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables.

II. Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.

III. Outliers in logistic model is very sensitive hence we need to deal with it without losing our valuable information. This can be achieved by creating bins. Hence, we did it.



Correlation

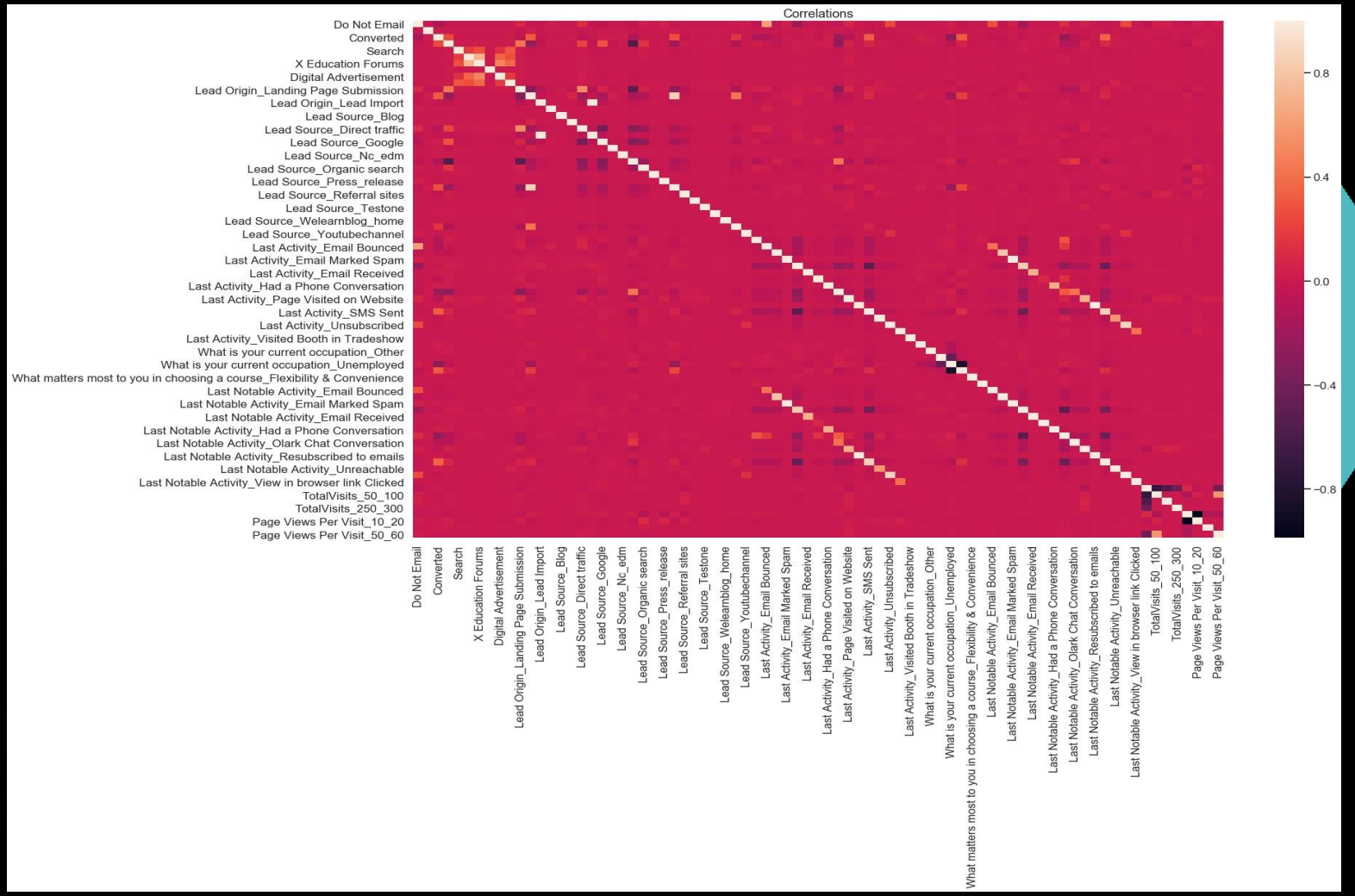
After fixing the outliers and dummy creation we proceed with our next step of analysis which is data preparation.

a) We split the dataset into train and test set and do standardization on the features.

b) Standardization is required in order to keep all the variables in same scale which will help us in computation in more efficient way.

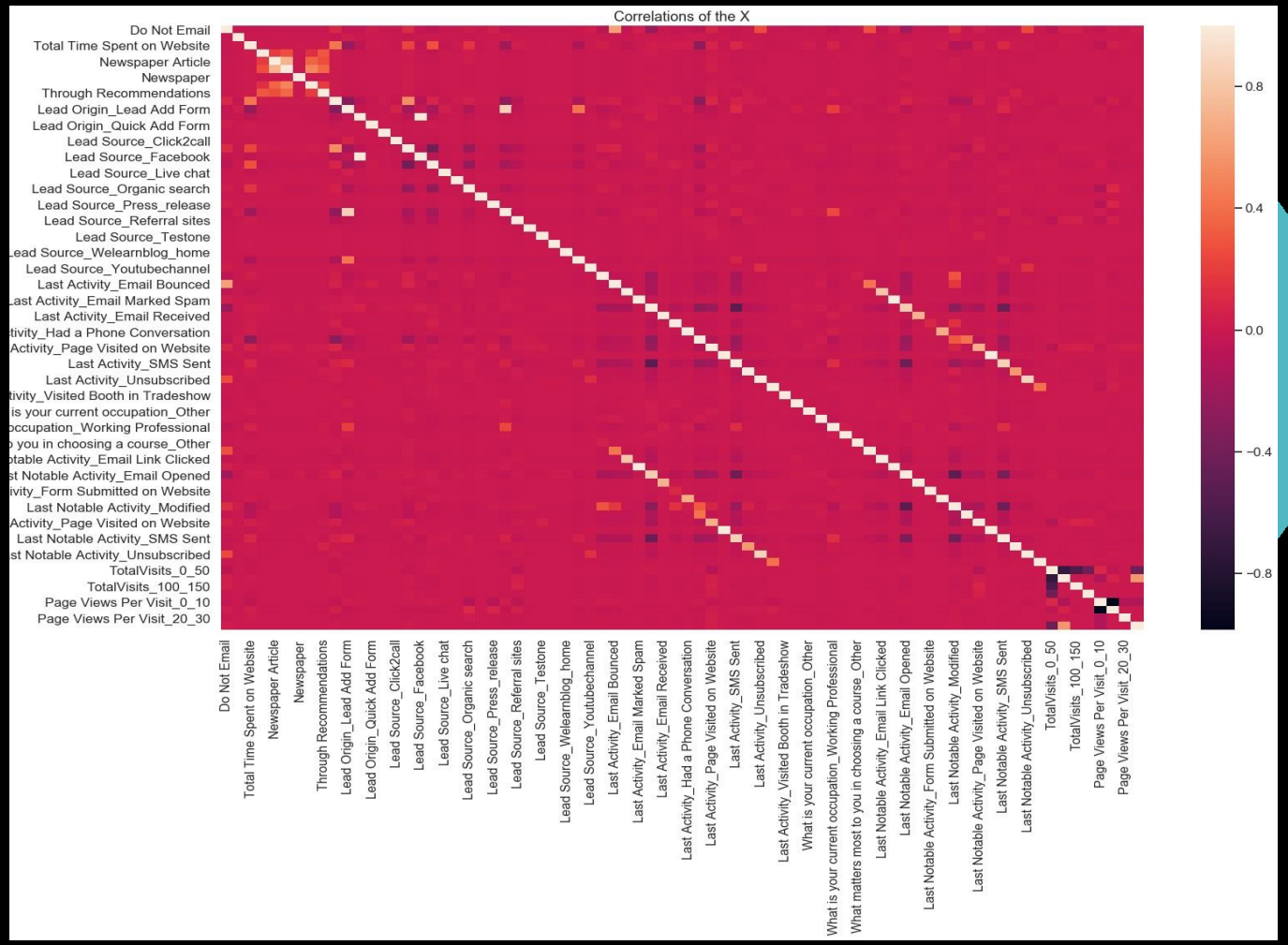
c) Checked the correlation of the dataset. Attached heatmap is showing the correlation of all features present in the dataset.

d) There are some high correlations in the heatmap which we dropped.



Correlation

- After dropping those high correlations features, we plotted again a heatmap to check and it was confirmed that those highly correlated variables were dropped.
- There are still few left, but we will check them after creating our model to verify how much they are impacting, as from the plot on the right it is not quite understandable which variable is having high correlation.



Building a Model – RFE 1

- We build a model with all the features included and found there were many insignificant variables present in our model.
- We need to drop them, but we can't do it one by one as it is time consuming and not an efficient way to do so.
- Hence, we started with RFE method to deduct those insignificant variables. We choose with RFE count 19 and 15.
- We did two rfe count because we want to find out our final model stability.
- We started creating our model with rfe count 19 and went dropping variables one by one until we reach the point where the model is having all significant variables and low VIF values.
- Now we evaluated our model by first predicting it. We created new dataset with original converted values and the prediction values.

Final model visualization with VIF

	Features	VIF
2	Lead Origin_Lead Add Form	1.40
13	Last Notable Activity_SMS Sent	1.36
3	Lead Source_Direct traffic	1.25
5	Lead Source_Google	1.24
8	Lead Source_Welingak website	1.24
11	What is your current occupation_Working Profes...	1.17
1	Total Time Spent on Website	1.15
0	Do Not Email	1.12
6	Lead Source_Organic search	1.12
9	Last Activity_Converted to Lead	1.10
10	Last Activity_Olark Chat Conversation	1.07
7	Lead Source_Referral sites	1.01
14	Last Notable Activity_Unreachable	1.01
4	Lead Source_Facebook	1.00
12	Last Notable Activity_Had a Phone Conversation	1.00

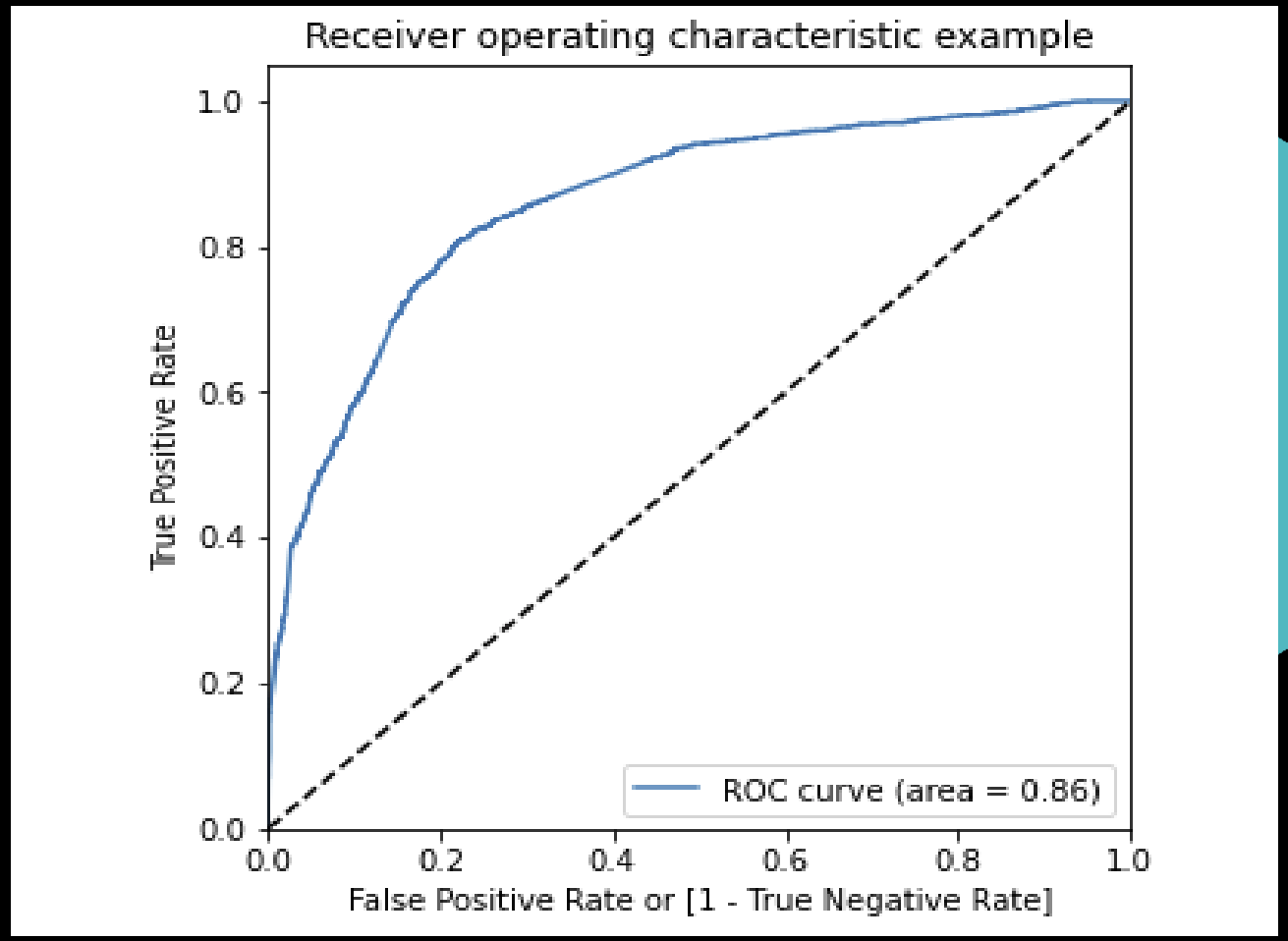
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Gaussian	Df Model:	15
Link Function:	identity	Scale:	0.13856
Method:	IRLS	Log-Likelihood:	-2777.8
Date:	Mon, 26 Aug 2019	Deviance:	893.98
Time:	14:49:33	Pearson chi2:	894.
No. Iterations:	3	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4103	0.013	31.513	0.000	0.385	0.436
Do Not Email	-0.1611	0.018	-9.202	0.000	-0.195	-0.127
Total Time Spent on Website	0.1842	0.005	35.581	0.000	0.174	0.194
Lead Origin_Lead Add Form	0.3895	0.022	17.801	0.000	0.347	0.432
Lead Source_Direct traffic	-0.1921	0.016	-12.210	0.000	-0.223	-0.161
Lead Source_Facebook	-0.1707	0.066	-2.587	0.010	-0.300	-0.041
Lead Source_Google	-0.1295	0.015	-8.615	0.000	-0.159	-0.100
Lead Source_Organic search	-0.1580	0.018	-8.560	0.000	-0.194	-0.122
Lead Source_Referral sites	-0.1777	0.041	-4.369	0.000	-0.257	-0.098
Lead Source_Welingak website	0.1880	0.043	4.332	0.000	0.103	0.273
Last Activity_Converted to Lead	-0.1396	0.023	-6.161	0.000	-0.184	-0.095
Last Activity_Olark Chat Conversation	-0.1703	0.017	-9.895	0.000	-0.204	-0.137
What is your current occupation_Working Professional	0.3445	0.018	19.065	0.000	0.309	0.380
Last Notable Activity_Had a Phone Conversation	0.4780	0.112	4.250	0.000	0.258	0.698
Last Notable Activity_SMS Sent	0.2589	0.011	22.735	0.000	0.237	0.281
Last Notable Activity_Unreachable	0.3063	0.081	3.759	0.000	0.147	0.466

Evaluating the model

- After building the final model making prediction on it(on train set), we created 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 0.86 which is a great score.
- And our graph is leaned towards the left side of the border which means we have good accuracy.

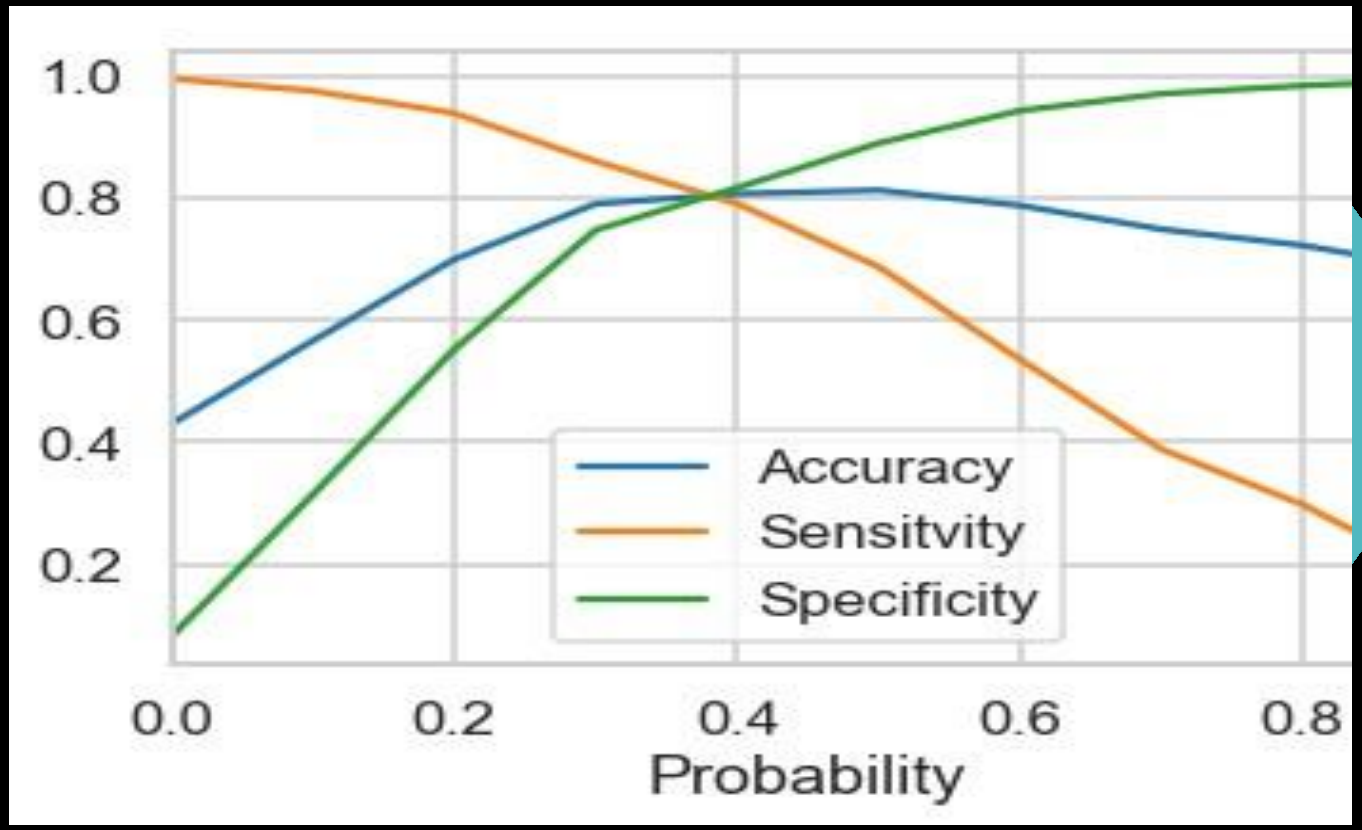


Finding the optimal cutoff point

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.

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.

To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.4 and hence we choose 0.4 as our optimal probability cutoff.

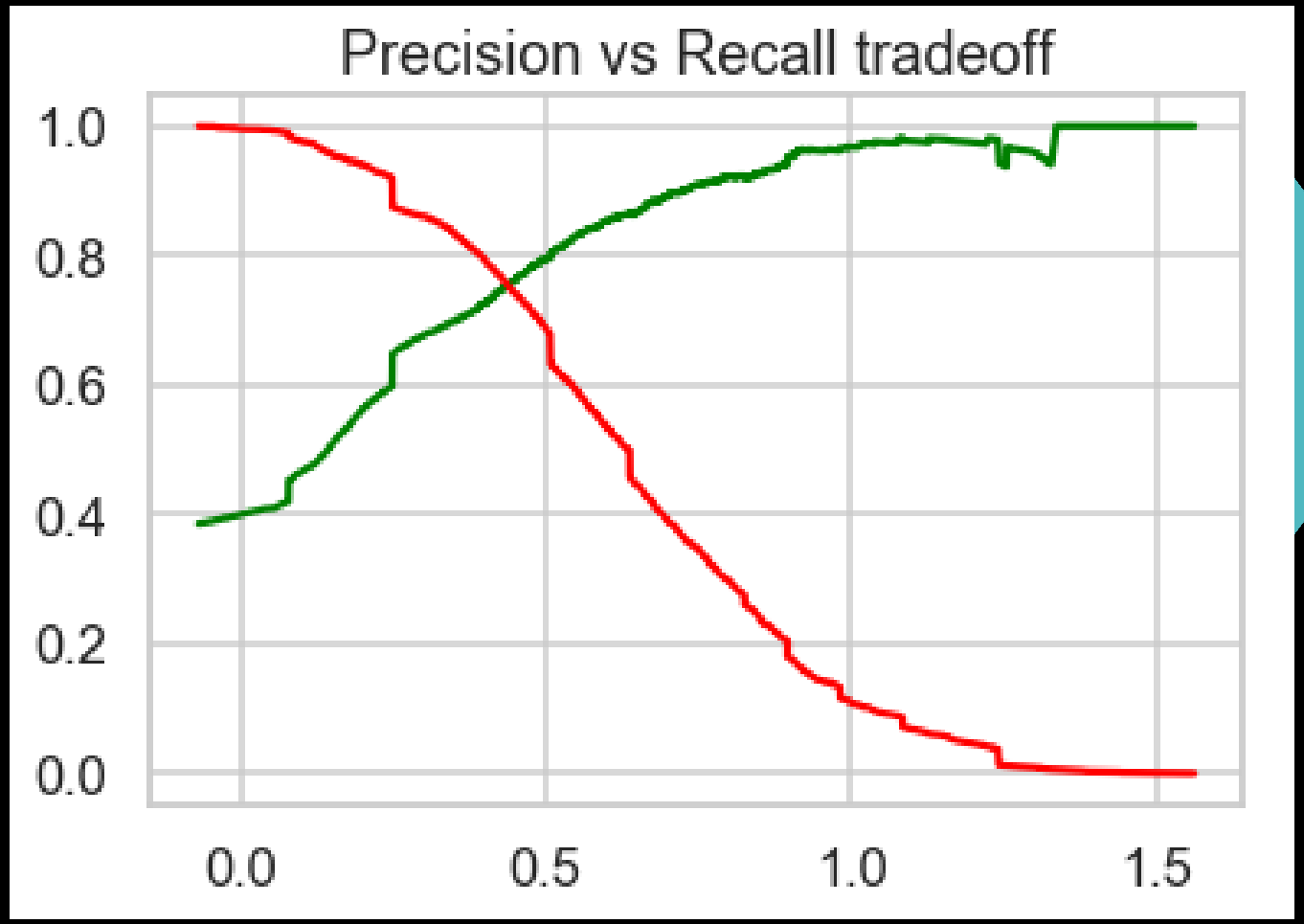


Precision and Recall

- We used this cutoff point to create a new column in our final dataset for predicting the outcomes.
- After this we did another type of evaluation which is by checking Precision and Recall
- As we all know, Precision and Recall plays very important role in build our model more business oriented and it also tells how our model behaves.
- Hence, we evaluated the precision and recall for this model and found the score as 0.73 for precision and 0.79 for recall.
- Now, recall our business objective - the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.
- i.e We get more relevant results - as many as hot lead customers from our model .

Precision and Recall tradeoff

- We created a graph which will show us the tradeoff 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.

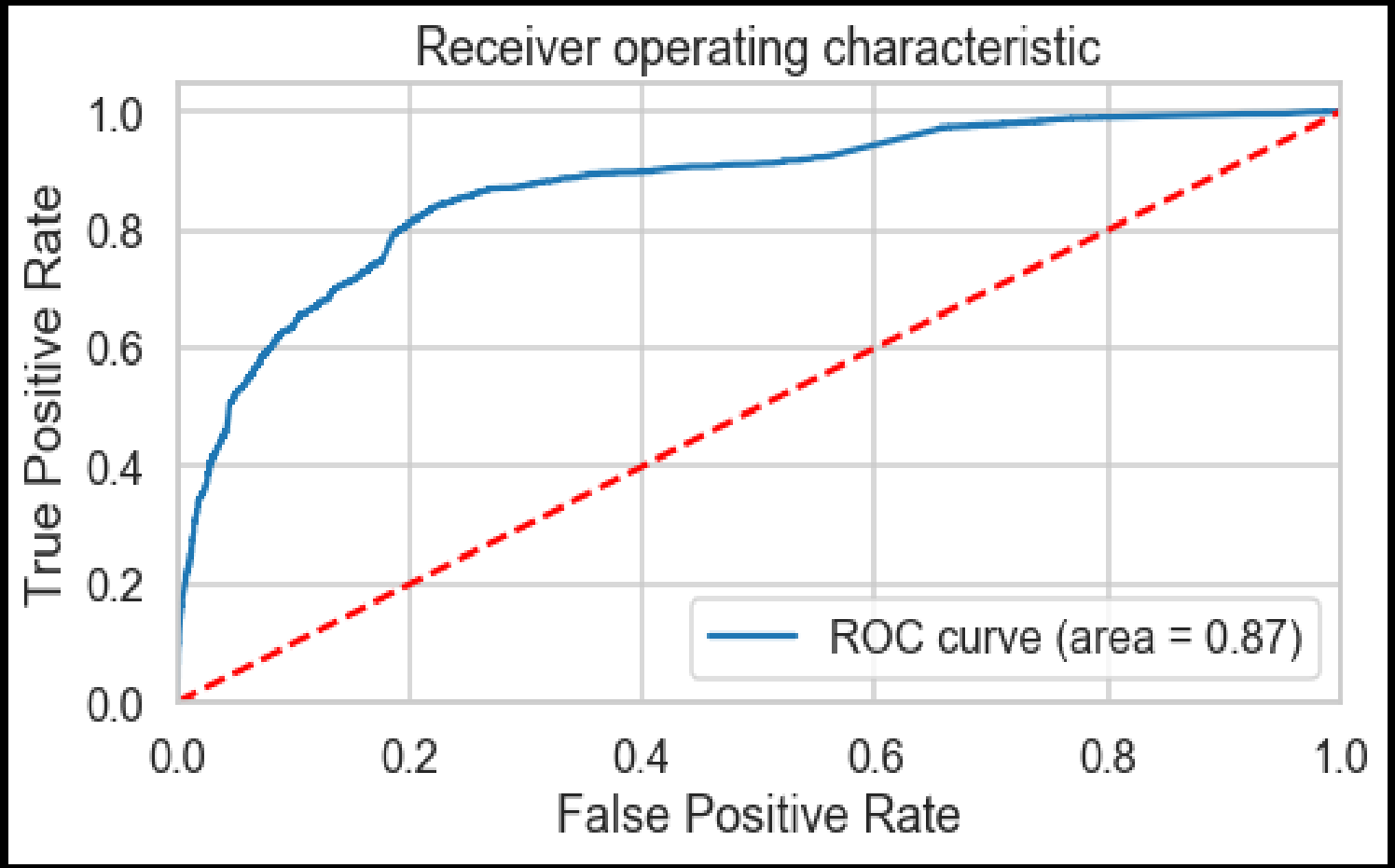


With RFE 2

- After completing our model evaluation from rfe 1, we proceeded with our second rfe method with count 15.
- We did that same steps as were mentioned in rfe 1, like creating a model and checking the insignificant values and VIFs and dropping those and running again until we reach our model with no insignificant variables and low VIFs.
- Ultimately, we found out last final model with all significant values and low VIFs.
- We predicted the final model in train set and created a new dataset with original converted values and prediction values.
- After this want to verify which final model is the best – one that was created with 19 variables or the one created with 15 variables.

RFE 1 vs RFE 2

- We want to choose our final model for test dataset prediction and in order to do that we plotted ROC curve for the RFE 2 model and compared these two graphs
- Attached graph plotted for the RFE 2 on the right.
- What we found was the auc score(area under the curve) in rfe 2 was 0.87 which was less than auc score generated in rfe 1.
- As we all know that the auc score shows the model accuracy and stability, we found that the final model created by RFE 1 is more stable and accurate than the final model created by RFE 2.



Prediction on test set

- Before predicting on test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set and the new predictions values were saved in new dataframe.
- After this we did model evaluation i.e. finding the accuracy, precision and recall.
- The accuracy score we found was 0.82, precision 0.76 and recall 0.79 approximately.
- This shows that our test prediction is having accuracy , precision and recall score in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of converted, low the lead score lower the chance of getting converted.

Conclusion

Valuable Insights -

- The Accuracy, Precision and Recall/Sensitivity are showing promising scores in test set which is as expected after looking the same in train set evaluation steps. Means the recall is having high score value than precision which is acceptable for business needs.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

a) **Last Notable Activity_Had a Phone Conversation**
b) **Lead Origin_Lead Add Form and**
c) **What is your current occupation_Working Professional**

Thank You