LEADS SCORING CASE STUDY

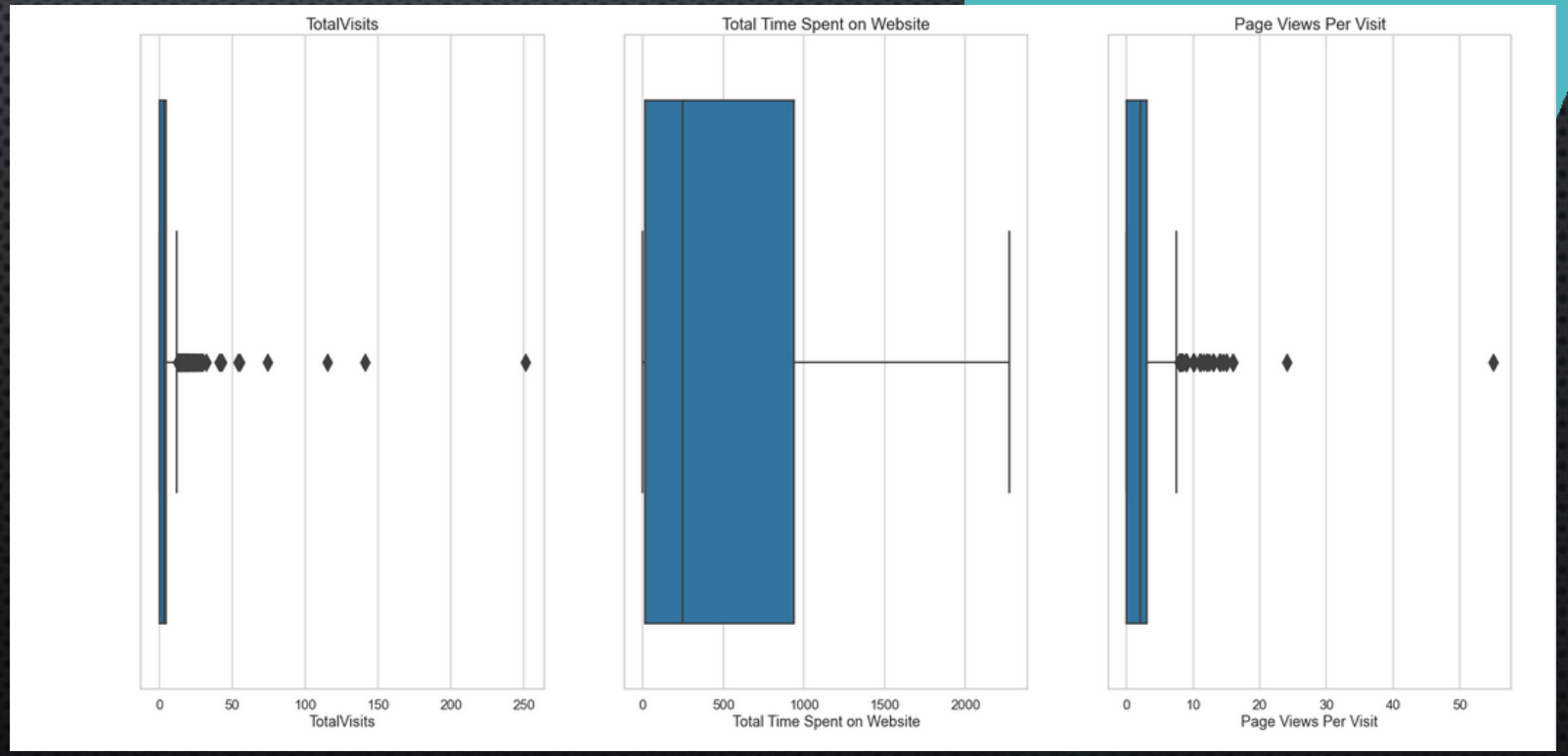
ANN MARY PHILIP
SUMIT KAWAIKAR

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.

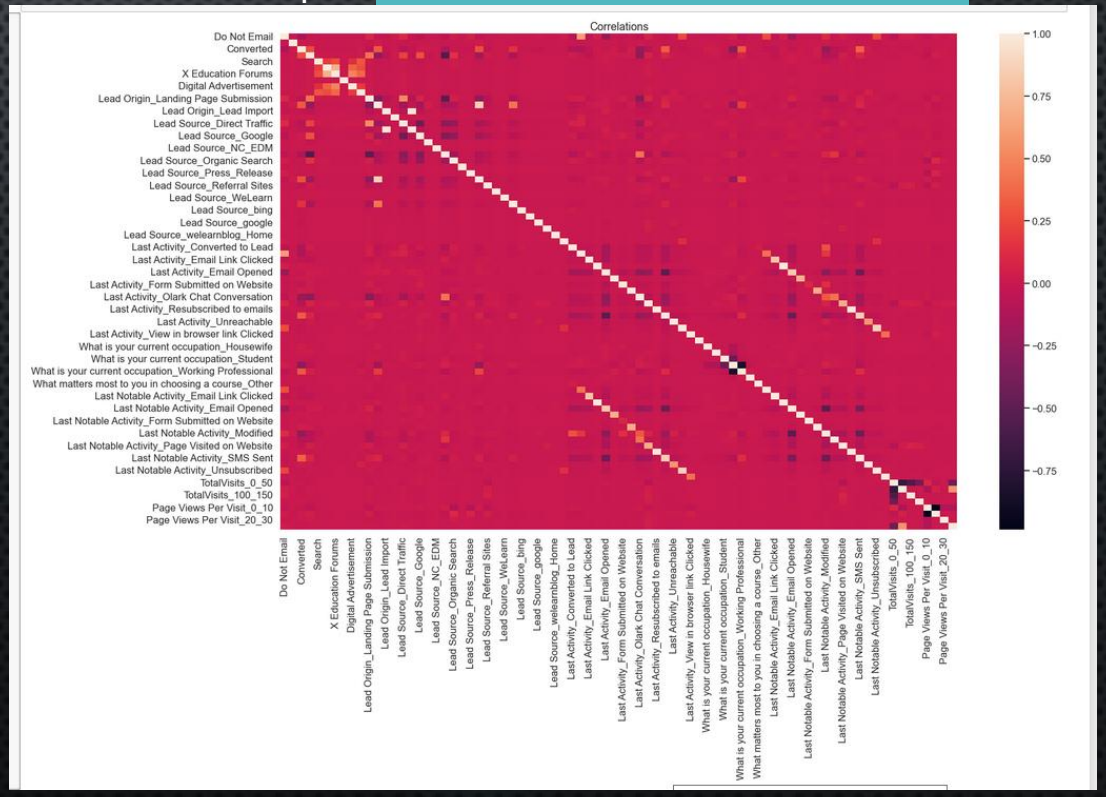
APPROACH OF THE ANALYSIS

- We had started the analysis by cleaning the provided dataset by converting all the binary variables to '0' and '1'.
- Also we have converted multiple categories into dummy variables.
- Then we checked for the outliers of the dataset.The graph depicted represents the outliers in the data set.
- Outliers in the data set can be handled by creating bins.



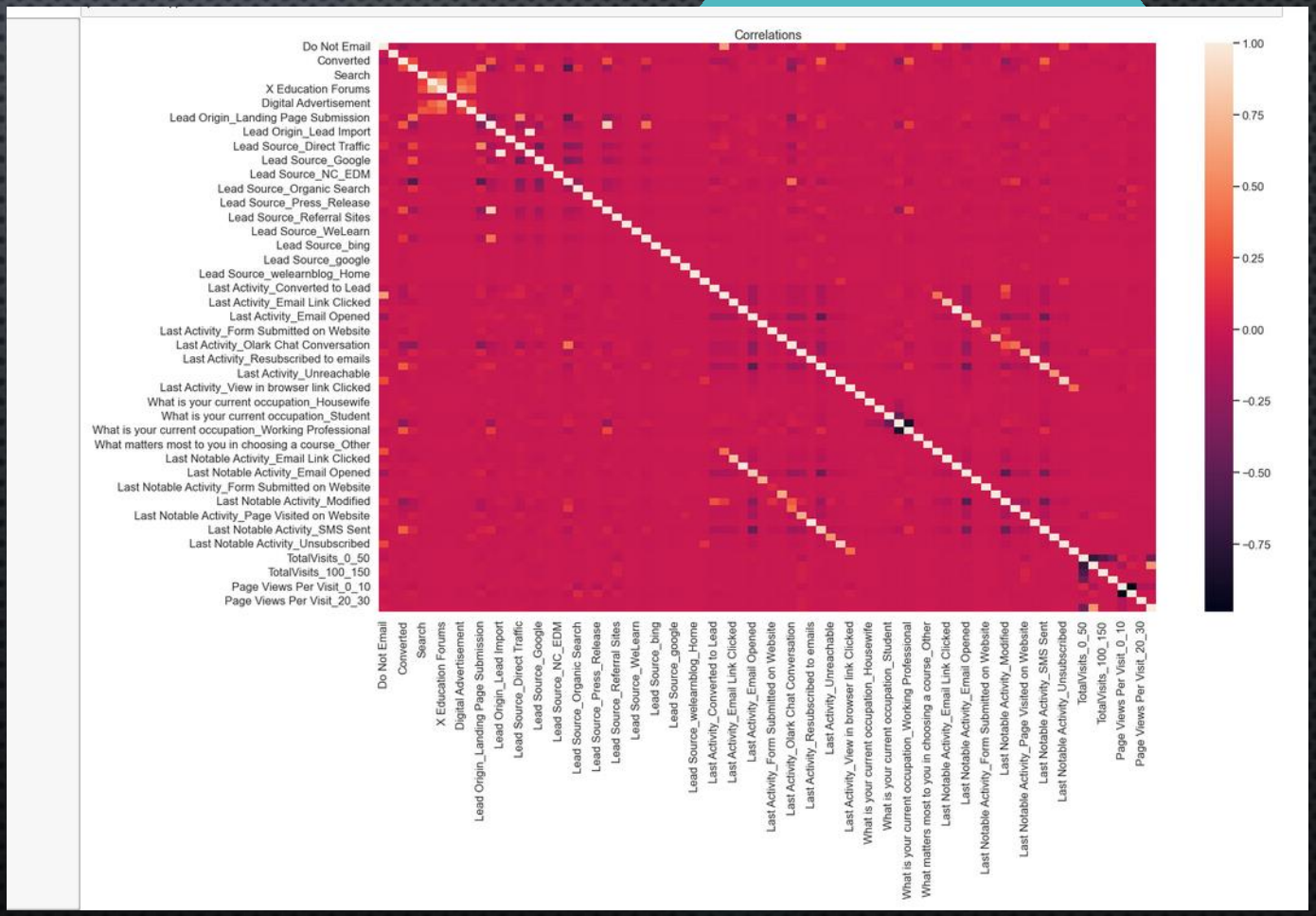
CORRELATION

- After correcting the outliers in the data set and the dummy creation, we proceed with data preparation.
- Then We split the dataset into train and test sets and standardized the features available. To k eep all the variables on the same scale and make computations more effective, standardization is necessary.
- Next we checked the correlation of the given dataset.
- Attached heatmap showing the correlation of all features in the dataset.
- As we can see that there are some high correlations in the heatmap which we have dropped in the dataset.



CORRELATION

- After dropping the attributes with high correlations features, we have plotted again a heatmap to check.
- The heat map has been shown here, and it was confirmed that those highly correlated variables were dropped.



BUILDING A MODEL RFE-1

- We build a model with all the features included and found there were many attributes which has a greater p value (insignificant variables) present in our model.
- We need to drop all those attributes and make our model more refined
- Hence to eliminate the unwanted attirbutes, we started with the RFE method to deduct those insignificant variables. We choose with RFE count 19 and 15.
- We did with two rfe count because we want to find out our final model stability and both our model was in equal stability.
- We started creating our model with the rfe count 19 and then went to 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

Out[579]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Gaussian	**Df Model:**	11
Link Function:	identity	**Scale:**	0.13958
Method:	IRLS	**Log-Likelihood:**	-2803.7
Date:	Sun, 16 Oct 2022	**Deviance:**	901.16
Time:	21:42:19	**Pearson chi2:**	901.
No. Iterations:	3		
Covariance Type:	nonrobust		

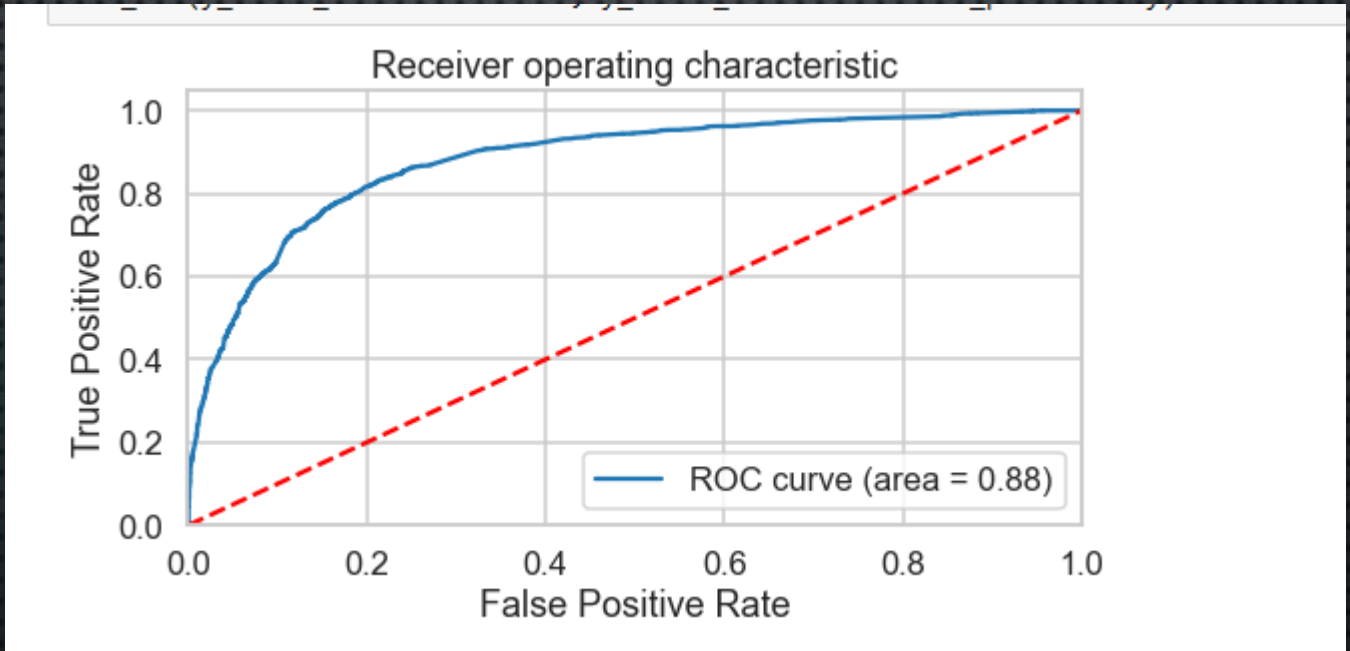
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5034	0.010	48.992	0.000	0.483	0.524
Do Not Email	-0.1770	0.018	-9.916	0.000	-0.212	-0.142
Total Time Spent on Website	0.1875	0.005	36.118	0.000	0.177	0.198
Lead Origin_Lead Add Form	0.5946	0.018	32.621	0.000	0.559	0.630
Lead Source_Olark Chat	0.1681	0.014	11.898	0.000	0.140	0.196
Last Activity_Olark Chat Conversation	-0.1225	0.020	-6.123	0.000	-0.162	-0.083
What is your current occupation_Working Professional	0.3346	0.018	18.604	0.000	0.299	0.370
Last Notable Activity_Email Link Clicked	-0.3097	0.036	-8.676	0.000	-0.380	-0.240
Last Notable Activity_Email Opened	-0.2256	0.013	-17.677	0.000	-0.251	-0.201
Last Notable Activity_Modified	-0.3030	0.013	-23.613	0.000	-0.328	-0.278
Last Notable Activity_Olark Chat Conversation	-0.2871	0.040	-7.212	0.000	-0.365	-0.209
Last Notable Activity_Page Visited on Website	-0.2692	0.026	-10.237	0.000	-0.321	-0.218

Out[581]:

	Features	VIF
4	Last Activity_Olark Chat Conversation	1.89
3	Lead Source_Olark Chat	1.65
8	Last Notable Activity_Modified	1.51
9	Last Notable Activity_Olark Chat Conversation	1.30
1	Total Time Spent on Website	1.20
2	Lead Origin_Lead Add Form	1.16
5	What is your current occupation_Working Profes...	1.12
0	Do Not Email	1.11
7	Last Notable Activity_Email Opened	1.10
6	Last Notable Activity_Email Link Clicked	1.02
10	Last Notable Activity_Page Visited on Website	1.02

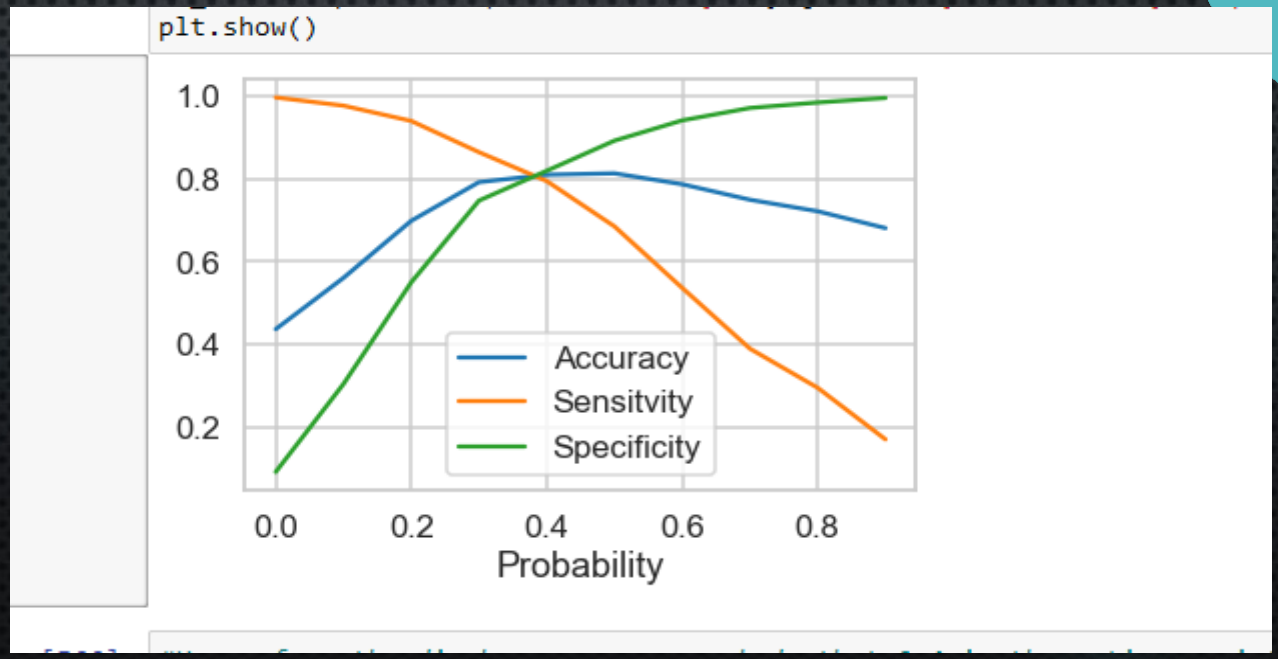
EVALUATING THE MODEL

- After building the final model prediction on it(on train set), we created the ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted, the area score is 0.88
- Our graph is leaned towards the left side of the border which means we have good accuracy.



FINDING THE OPTIMAL CUTOFF POINT

- We then built a range of points for which we'll determine each point's accuracy, sensitivity, and specificity before analysing which point to select as the probability cutoff.
- We discovered that the accuracy, sensitivity, and specificity scores all fall inside a narrow range at 0.4, making it the best spot to choose.We put this in a graph to validate our conclusion.
- The line plot is on the right. We stand corrected that the meeting point is near to 0.4, thus we select 0.4 as our ideal probability cutoff.

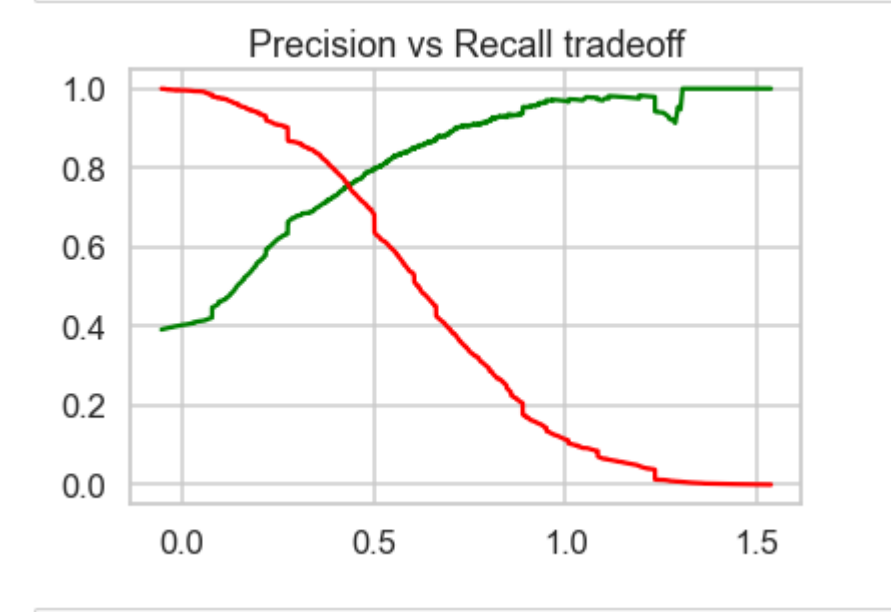


PRECISION AND RECALL

- In order to predict the outcomes, we added a new column to our final dataset using this cutoff point. Next, we performed another sort of evaluation by examining Precision and Recall.
- As a result, we assessed the model's precision and recall and discovered that the scores were 0.72 for precision and 0.79 for recall.
- Now, if we 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.
- Hence we have got more relevant results - as many as hot lead customers from our model .

PRECISION AND RECALL TRADEOFF

- We then created a graph which will show us the tradeoff between Precision and recall.
- We found out that there is a trade off between Precision and Recall and the meeting point is approximately at 0.5.



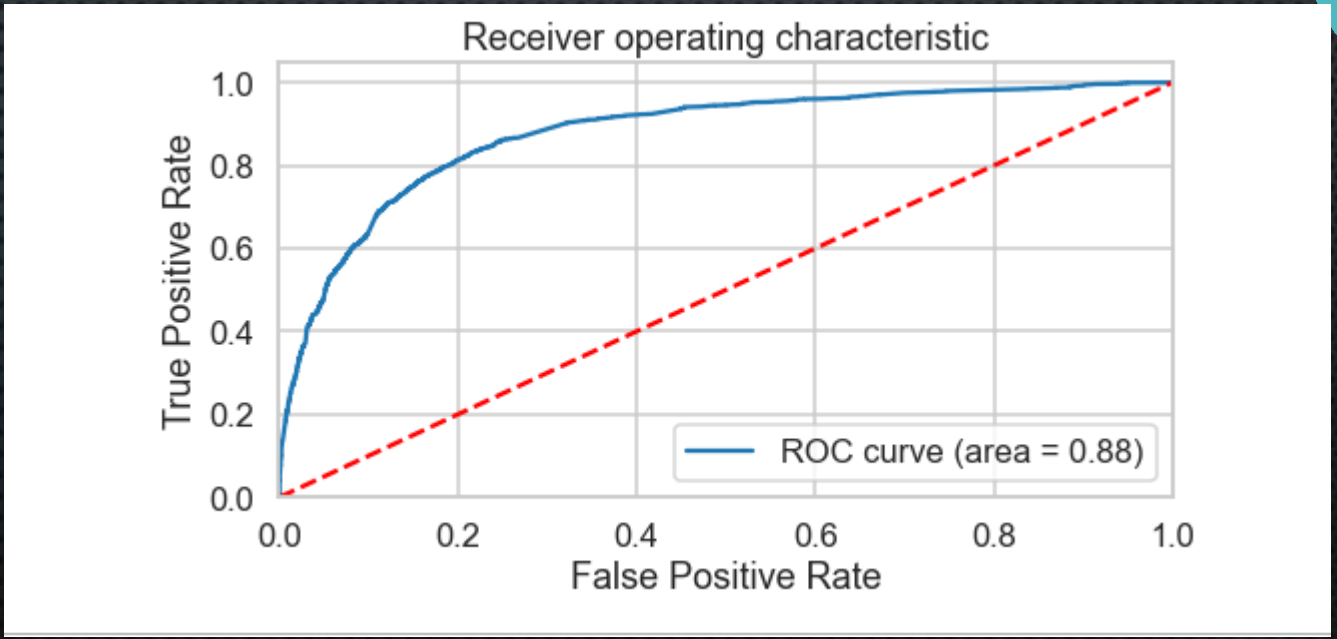
BUILDING THE MODEL WITH RFE-2

- We continued with our second rfe technique with count 15 after finishing our model evaluation from rfe 1.We followed the same procedures as those described in rfe 1:
- We created a model, checked it for insignificant values and VIFs, removed those, and then ran the model again until it had no insignificant variables and had low VIFs.
- Finally, a model with all significant values and a low VIF was discovered.
- We generated a new dataset using the original converted values and the final model predictions after predicting it in the train set.
- After that, you should determine which of the two final models—the one developed with 19 variables or the one created with 15 variables—is the best.

RFE 1 VS RFE 2

We want to choose our final model for the test dataset prediction and in order to do that we plotted the ROC curve for the RFE 2 model and compared these two graphs

- Attached graph plotted for the RFE 2.
- What we found was the auc score(area under the curve) in rfe 2 was 0.88 which was equal to the 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 and RFE 2 is equally stable and accurate



PREDICTION ON TEST SET

- We must standardise the test set and ensure that the exact identical columns are present in our final train dataset before making predictions on the test set.
- We began predicting the test set after completing the a fore mentioned phase, and the new predictions values were saved in a new dataframe.
- Following this, we evaluated the model by determining its accuracy, precision, and recall. We discovered accuracy scores of 0.82, precision scores of 0.76, and recall scores of roughly 0.79.This demonstrates that our test prediction has appropriate levels of accuracy, precision, and recall.
- This demonstrates the stability, accuracy, and recall/sensitivity of our model.
- 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 -

- After examining the same in train set evaluation stages, the Accuracy, Precision, and Recall/Sensitivity are exhibiting promising scores in the test set, which is expected. This indicates that the recall has a high score value compared to the precision, which is suitable for business demands.
- In business terms, this model has the ability to adjust to the company's requirements in the coming future.
- This concludes that the model is in stable state.

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

Last Notable Activity_Had a Phone Conversation

Lead Origin_Lead Add Form and

What is your current occupation_Working Professional

THANK YOU