Leads Scoring Case Study

By: Aniruddha Nath

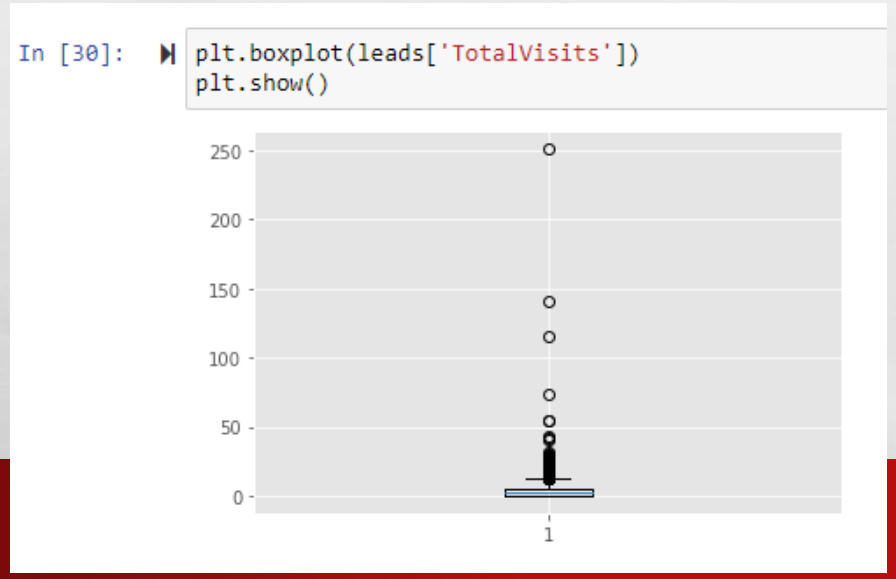
PROBLEM STATEMENT

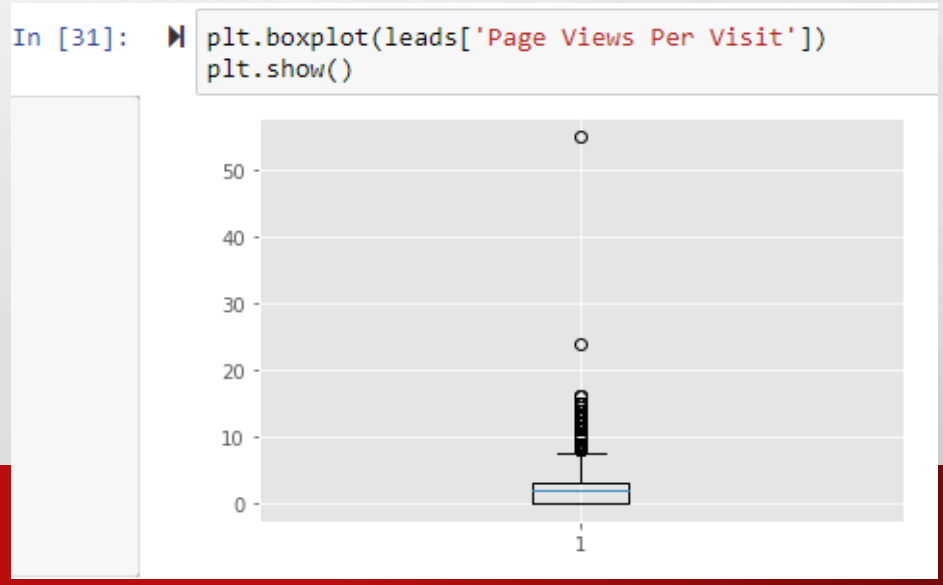
X EDUCATION SELLS ONLINE COURSES TO INDUSTRY PROFESSIONALS. THE TYPICAL LEAD CONVERSION RATE AT X EDUCATION IS AROUND 30%. THEY HAVE APPOINTED YOU TO HELP THEM SELECT THE MOST PROMISING LEADS. YOU NEED TO ASSIGN A LEAD SCORE TO EACH OF THE LEADS SO THAT CUSTOMERS WITH HIGHER LEAD SCORE HAVE A HIGHER CONVERSION CHANCE.

THE CEO, IN PARTICULAR, HAS GIVEN A BALLPARK OF THE TARGET LEAD CONVERSION RATE TO BE AROUND 80%.

APPROACH OF THE ANALYSIS

- FIRSTLY, WE CLEANED THE DATASET WITH MISSING VALUES AND ALSO CONVERTED THE BINARY VARIABLES TO ‘0’ AND ‘1’. MULTIPLE CATEGORIES WERE CONVERTED INTO DUMMY VARIABLES
- CHECKING THE OUTLIERS OF THE DATASET AND FOUND THAT TWO OF THE VARIABLES HAD OUTLIERS BUT THOSE VARIABLES COULDN’T BE DROPPED SO WE BINNED THOSE VARIABLES AS WE DIDN’T WANTED TO LOSE THE USEFUL INFORMATION THAT IT WAS PROVIDING TO US. WE USED BOXPLOT TO CHECK FOR THE SAME.

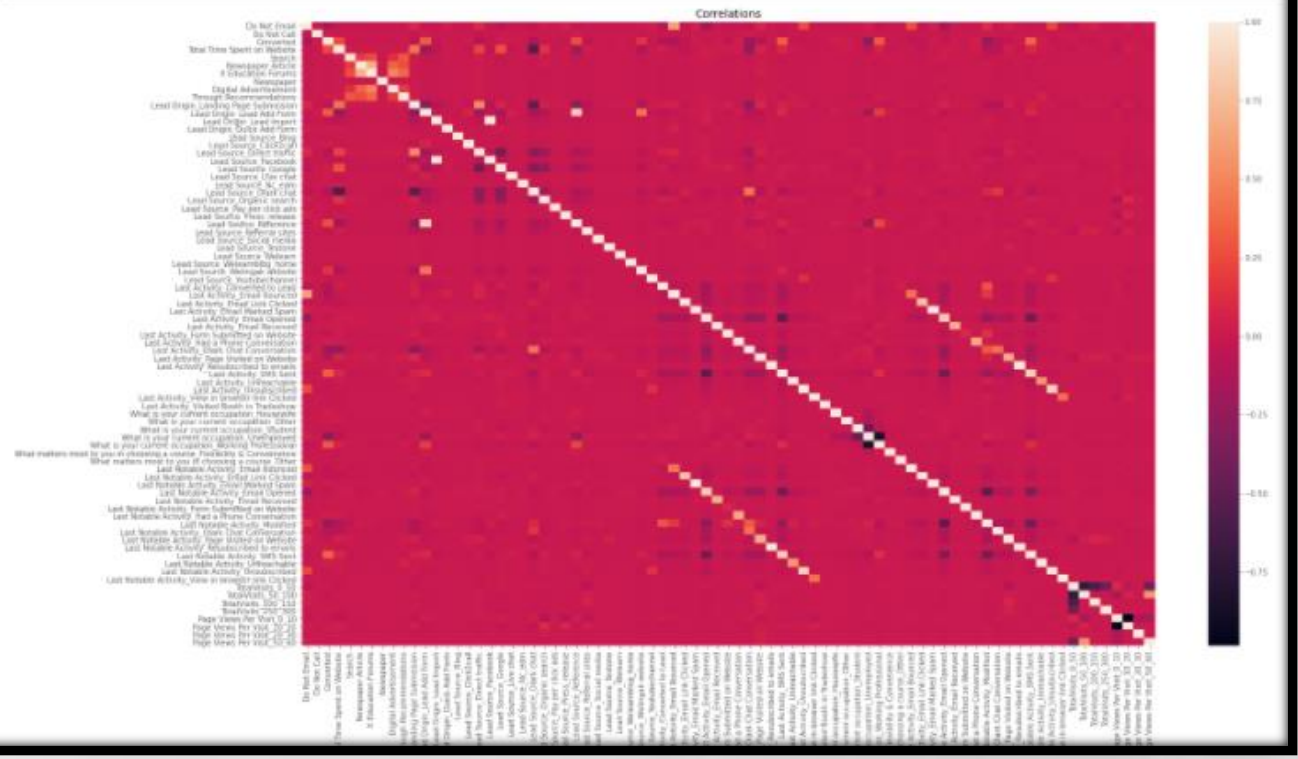




CORRELATION

- We prepare the data by splitting it into train and test set and use feature standardization
- Standardization is required to keep all the variables in same scale which will help us in computation in more efficient way.
- Using heatmap checked the correlation for the data.
- Variables with high correlation are then dropped and then we again plot the heatmap leaving some of them which we will verify later while building the model.

```
In [41]:
# setting the figure size
plt.figure(figsize=(25,15))
# setting the title
plt.title('Correlations')
# Plotting a heatmap
sns.heatmap(leads.corr(method='spearman'))
plt.show()
```



BUILDING A MODEL RFE-I

- WE BUILD MODEL WITH ALL THE FEATURES INCLUDED AND THERE ARE MANY INSIGNIFICANT VARIABLES WHICH WE CAN SEE FROM HIGH P-VALUE WHICH NEEDS TO BE DROPPED.
- WE USE RFE METHOD TO DROP THE VARIABLES BY KEEPING THE COUNT 19 AND 15.
- WE DID TWO RFE COUNT AS 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.
- EVALUATING OUR MODEL BY PREDICTING IT AND THEN CREATED NEW DATASET WITH ORIGINAL CONVERTED VALUES AND THE PREDICTION VALUES.

FINAL MODEL VISUALIZATION WITH VIF

	Features	VIF
13	Last Notable Activity_Modified	2.50
9	Last Activity_Olark Chat Conversation	1.80
4	Lead Source_Google	1.77
12	Last Notable Activity_Email Opened	1.73
3	Lead Source_Direct traffic	1.72
2	Lead Origin_Lead Add Form	1.46
14	Last Notable Activity_Olark Chat Conversation	1.35
5	Lead Source_Organic search	1.29
7	Lead Source_Welingak website	1.24
8	Last Activity_Converted to Lead	1.24
1	Total Time Spent on Website	1.22
0	Do Not Email	1.19
10	What is your current occupation_Working Profes...	1.16
15	Last Notable Activity_Page Visited on Website	1.10
6	Lead Source_Referral sites	1.04
11	Last Notable Activity_Email Link Clicked	1.04

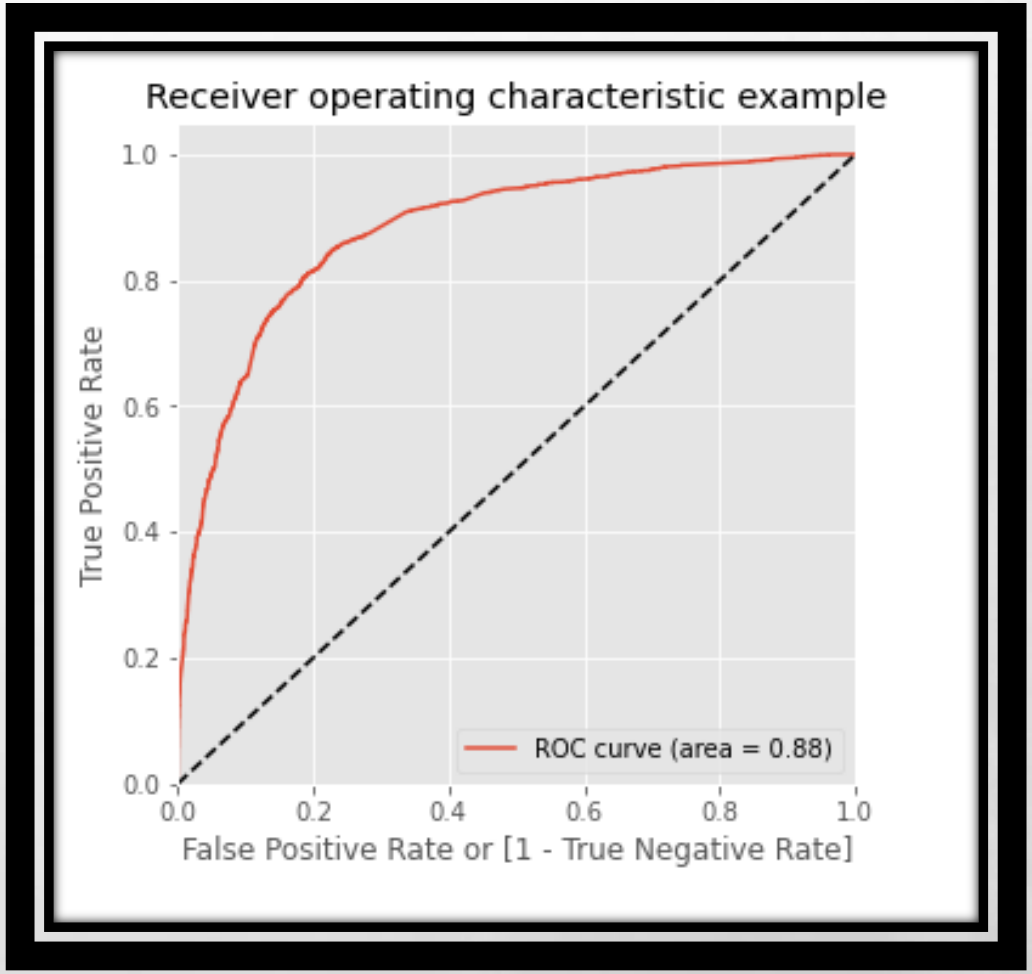
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6451
Model Family:	Gaussian	**Df Model:**	16
Link Function:	identity	**Scale:**	0.13845
Method:	IRLS	**Log-Likelihood:**	-2774.8
Date:	Sun, 16 Jan 2022	**Deviance:**	893.15
Time:	23:44:10	**Pearson chi2:**	893.
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6580	0.015	43.910	0.000	0.629	0.687
Do Not Email	-0.1779	0.018	-9.838	0.000	-0.213	-0.142
Total Time Spent on Website	0.1841	0.005	35.533	0.000	0.174	0.194
Lead Origin_Lead Add Form	0.4044	0.022	18.505	0.000	0.362	0.447
Lead Source_Direct traffic	-0.1854	0.016	-11.897	0.000	-0.216	-0.155
Lead Source_Google	-0.1234	0.015	-8.287	0.000	-0.153	-0.094
Lead Source_Organic search	-0.1502	0.018	-8.196	0.000	-0.186	-0.114
Lead Source_Referral sites	-0.1714	0.041	-4.220	0.000	-0.251	-0.092
Lead Source_Welingak website	0.1864	0.043	4.296	0.000	0.101	0.271
Last Activity_Converted to Lead	-0.1065	0.024	-4.428	0.000	-0.154	-0.059
Last Activity_Olark Chat Conversation	-0.1352	0.020	-6.706	0.000	-0.175	-0.096
What is your current occupation_Working Professional	0.3441	0.018	19.045	0.000	0.309	0.379
Last Notable Activity_Email Link Clicked	-0.3028	0.036	-8.509	0.000	-0.373	-0.233
Last Notable Activity_Email Opened	-0.2268	0.013	-17.828	0.000	-0.252	-0.202
Last Notable Activity_Modified	-0.2873	0.013	-21.482	0.000	-0.314	-0.261
Last Notable Activity_Olark Chat Conversation	-0.2707	0.040	-6.792	0.000	-0.349	-0.193
Last Notable Activity_Page Visited on Website	-0.2647	0.026	-10.099	0.000	-0.316	-0.213

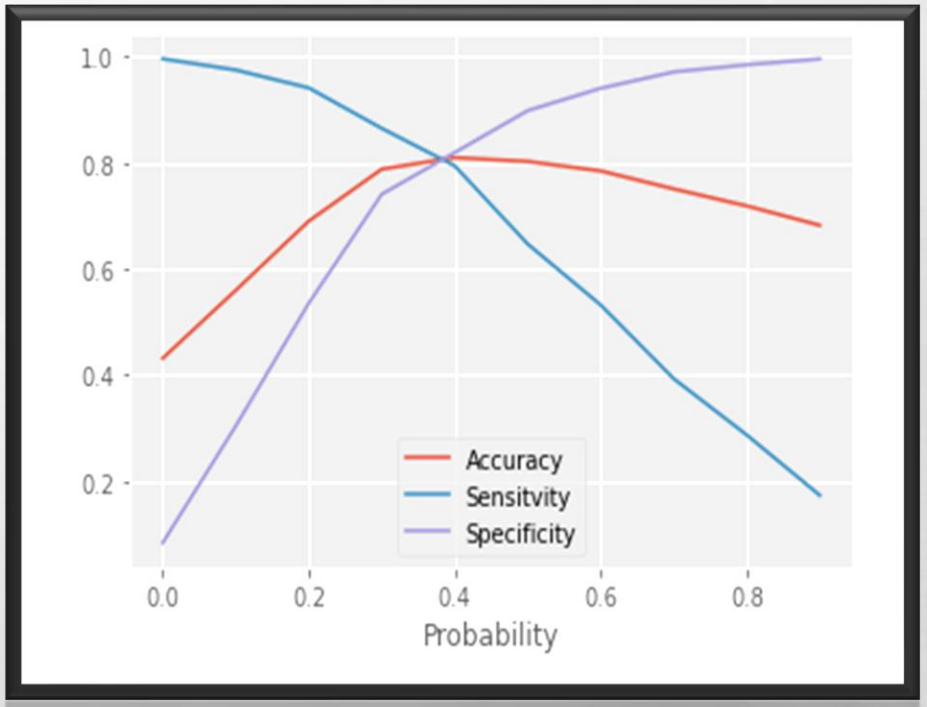
EVALUATING THE MODEL

- 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 0.88 WHICH IS A GREAT SCORE.
- AND LOOKING AT GRAPH WE CAN SAY THAT IT HAS 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

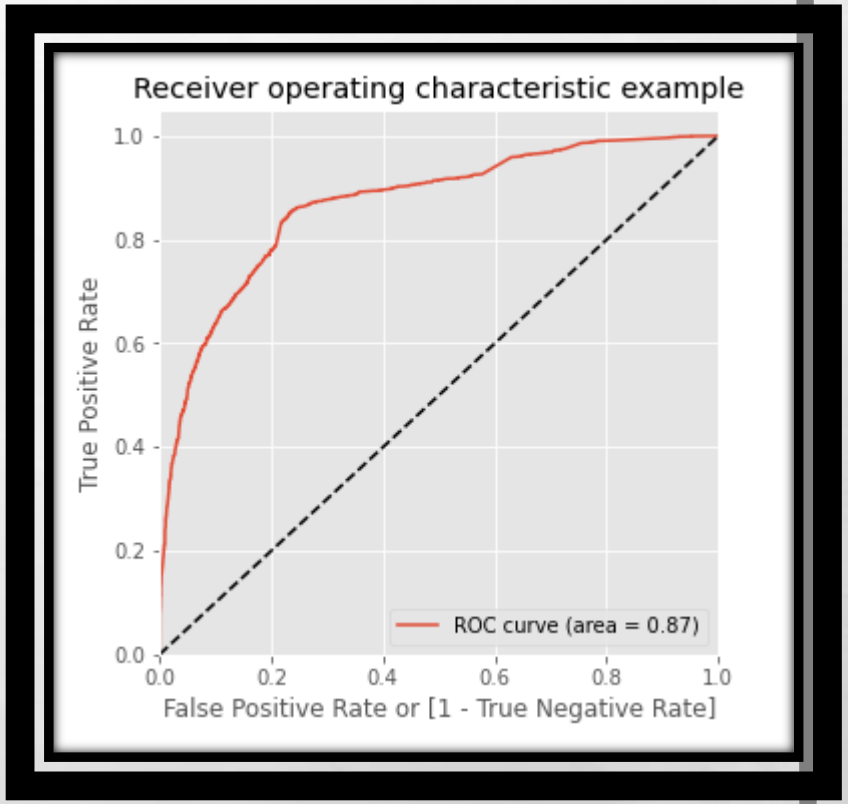
- WE USED THIS CUTOFF POINT TO CREATE A NEW COLUMN IN OUR FINAL DATASET FOR PREDICTING THE OUTCOMES.
- AFTER THIS WE DID ANOTHER TYPE OF EVALUATION WHICH IS BY CHECKING PRECISION AND RECALL
- AS WE ALL KNOW, PRECISION AND RECALL PLAYS VERY IMPORTANT ROLE IN BUILD OUR MODEL MORE BUSINESS ORIENTED AND 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.
- WE GET MORE RELEVANT RESULTS -AS MANY AS HOT LEAD CUSTOMERS FROM OUR MODEL .

BUILDING A MODEL RFE-II

- 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-I VS RFE-II

- SO, WE WANT TO CHOOSE OUR FINAL MODEL FOR THE PREDICTION OF TEST DATASET AND SO WE PLOT ROC CURVE FOR RFE-II MODEL.
- WE FIND THAT AREA UNDER CURVE(AUC) IN RFE-II IS 0.87 WHICH IS LESS THAN WHAT IS GENERATED BY RFE-I.
- 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

- WE HAVE HIGHER RECALL SCORE THAN PRECISION SCORE.
- THE ACCURACY, PRECISION AND RECALL ARE IN ACCEPTABLE RANGE.
- THE MODEL HAS THE ABILITY TO ADJUST TO COMPANY'S UPCOMING REQUIREMENTS.
- 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 :

  I) LAST NOTABLE ACTIVITY_HAD A PHONE CONVERSATION

  II) LEAD ORIGIN_LEAD ADD FORM AND

  III) WHAT IS YOUR CURRENT OCCUPATION_WORKING PROFESSIONAL

THANK YOU!