LEAD SCORING CASE STUDY

Submitted By:
Abhishek De
Rahul Roy

BUSINESS PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS OBJECTIVE

The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

OUTLIERS ANALYSIS

• There are outliers in 'Total Visits' column and 'Page Views Per Visit' column.

• To treat them we have to do **0.99-0.1%** analysis to get rid of the outliers.
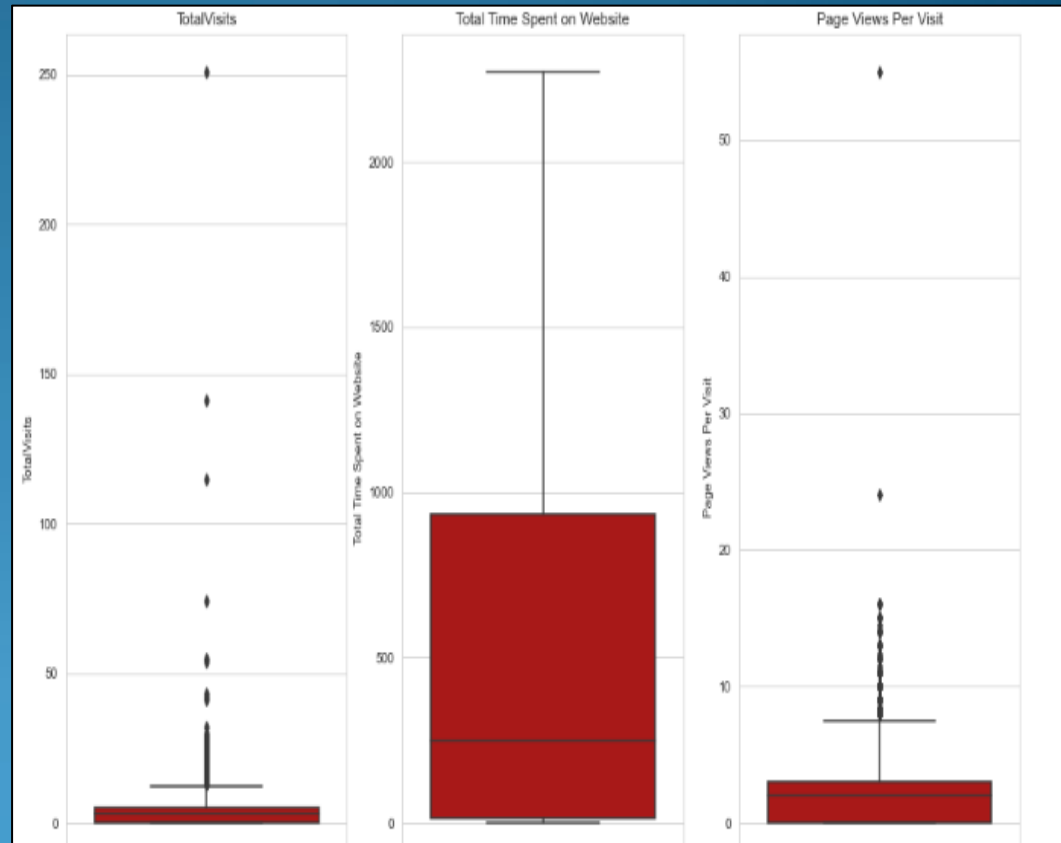
CORRELATION

•From the attached heatmap, we observe that there are many correlated attributes that needs to be removed.

• Highly correlated attributes create dependency on various independent factors which will give us inappropriate results.
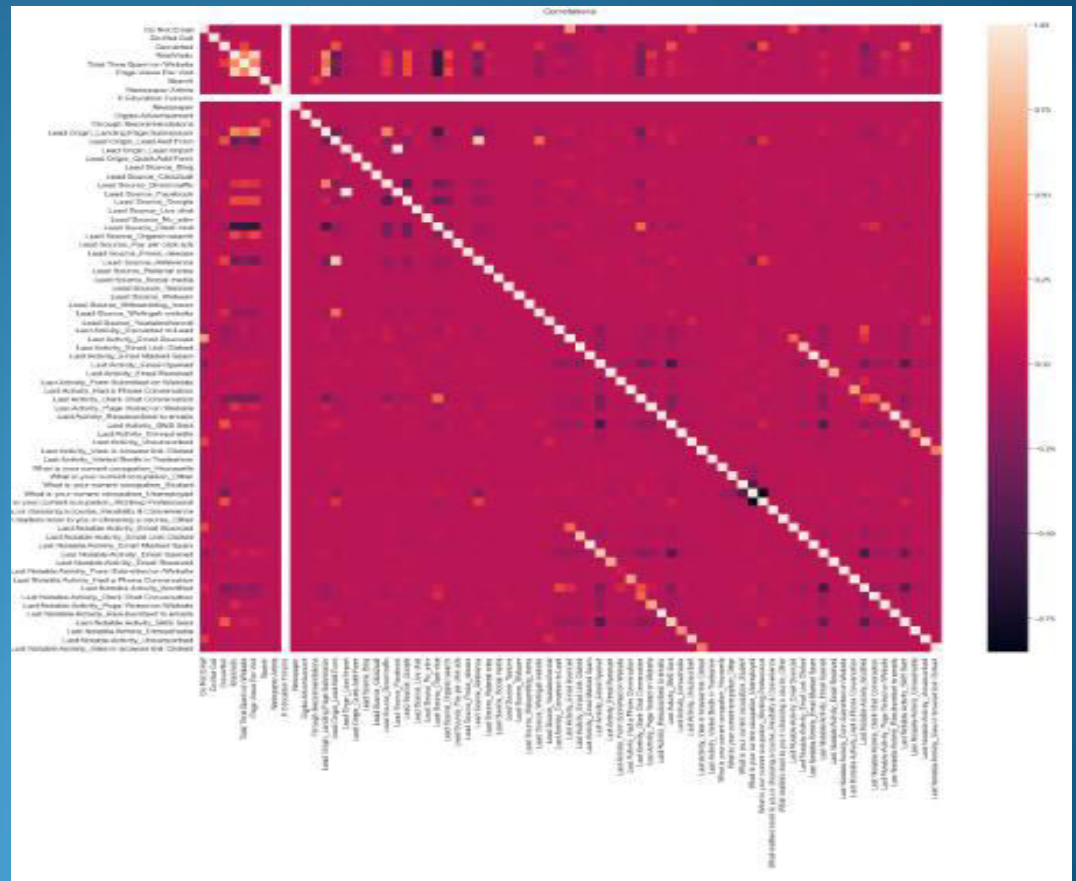
CORRELATION

• 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.
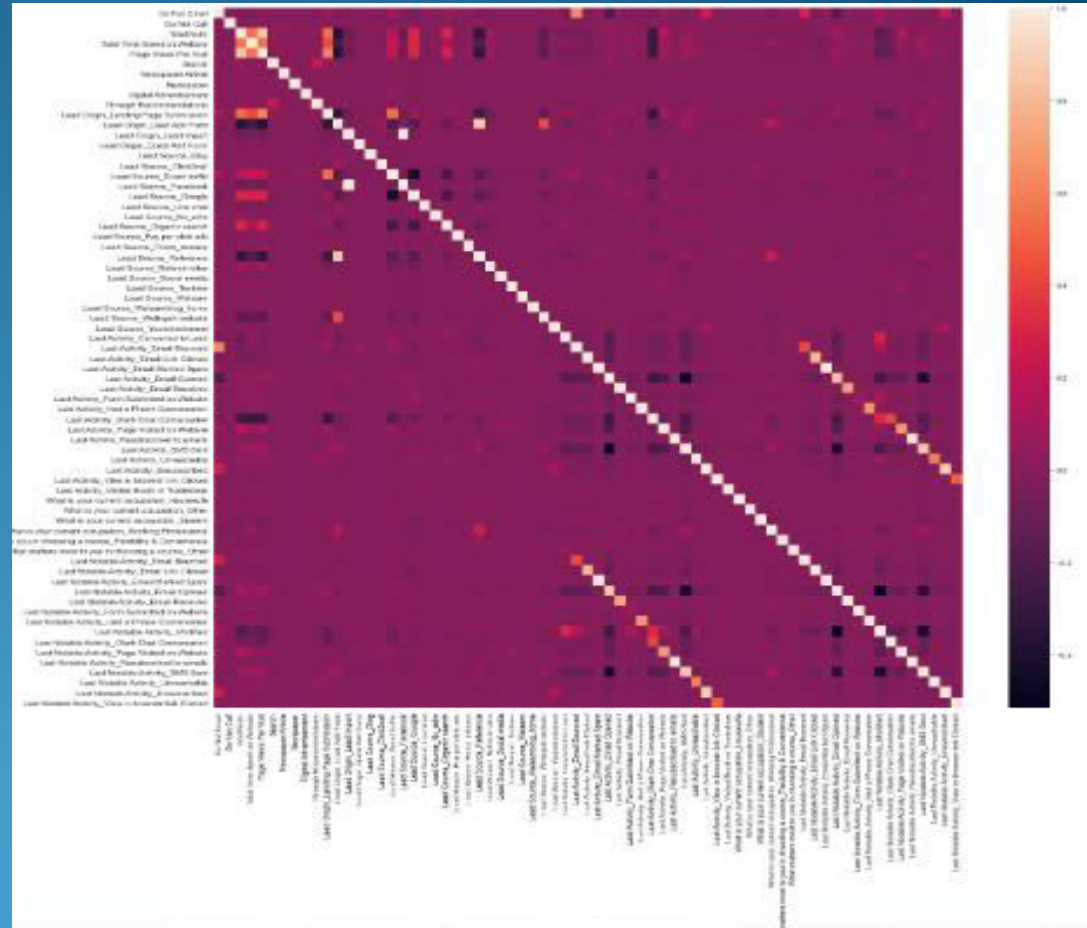
MODEL BUILDING

With the help of RFE, we can identify the insignificant variables present in our model.

	Features	VIF
2	Lead Origin_Lead Add Form	1.46
13	Last Notable Activity_SMS Sent	1.35
8	Lead Source_Welingak website	1.29
3	Lead Source_Direct traffic	1.25
5	Lead Source_Google	1.24
0	Do Not Email	1.19
11	What is your current occupation_Working Profes...	1.18
1	Total Time Spent on Website	1.15
6	Lead Source_Organic search	1.13
9	Last Activity_Converted to Lead	1.10
10	Last Activity_Olark Chat Conversation	1.08
15	Last Notable Activity_Unsubscribed	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:	6363
Model:	GLM	Df Residuals:	6345
Model Family:	Gaussian	Df Model:	17
Link Function:	Identity	Scale:	0.13759
Method:	IRLS	Log-Likelihood:	-2709.2
Date:	Mon, 08 Mar 2021	Deviance:	873.00
Time:	12:27:37	Pearson chi2:	873.
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4041	0.013	30.814	0.000	0.378	0.430
Do Not Email	-0.1824	0.018	-9.966	0.000	-0.218	-0.147
Total Time Spent on Website	0.1806	0.005	34.615	0.000	0.170	0.191
Lead Origin_Lead Add Form	0.3821	0.022	17.002	0.000	0.338	0.426
Lead Source_Direct traffic	-0.1843	0.016	-11.651	0.000	-0.215	-0.153
Lead Source_Facebook	-0.1739	0.062	-2.793	0.005	-0.296	-0.052
Lead Source_Google	-0.1211	0.015	-8.030	0.000	-0.151	-0.092
Lead Source_Organic search	-0.1639	0.019	-8.805	0.000	-0.200	-0.127
Lead Source_Referral sites	-0.1517	0.044	-3.482	0.000	-0.237	-0.066
Lead Source_Welingak website	0.2118	0.041	5.125	0.000	0.131	0.293
Last Activity_Converted to Lead	-0.1343	0.023	-5.894	0.000	-0.179	-0.090
Last Activity_Olark Chat Conversation	-0.1753	0.017	-10.418	0.000	-0.208	-0.142
What is your current occupation_Other	0.2088	0.118	1.777	0.076	-0.021	0.439
What is your current occupation_Working Professional	0.3430	0.018	18.770	0.000	0.307	0.379
Last Notable Activity_Had a Phone Conversation	0.5719	0.131	4.353	0.000	0.314	0.829
Last Notable Activity_SMS Sent	0.2786	0.011	24.272	0.000	0.256	0.301
Last Notable Activity_Unreachable	0.3308	0.081	4.071	0.000	0.172	0.490
Last Notable Activity_Unsubscribed	0.1942	0.068	2.858	0.004	0.061	0.327

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

Receiver Operating Characteristic

ROC Curve (area= 0.88)
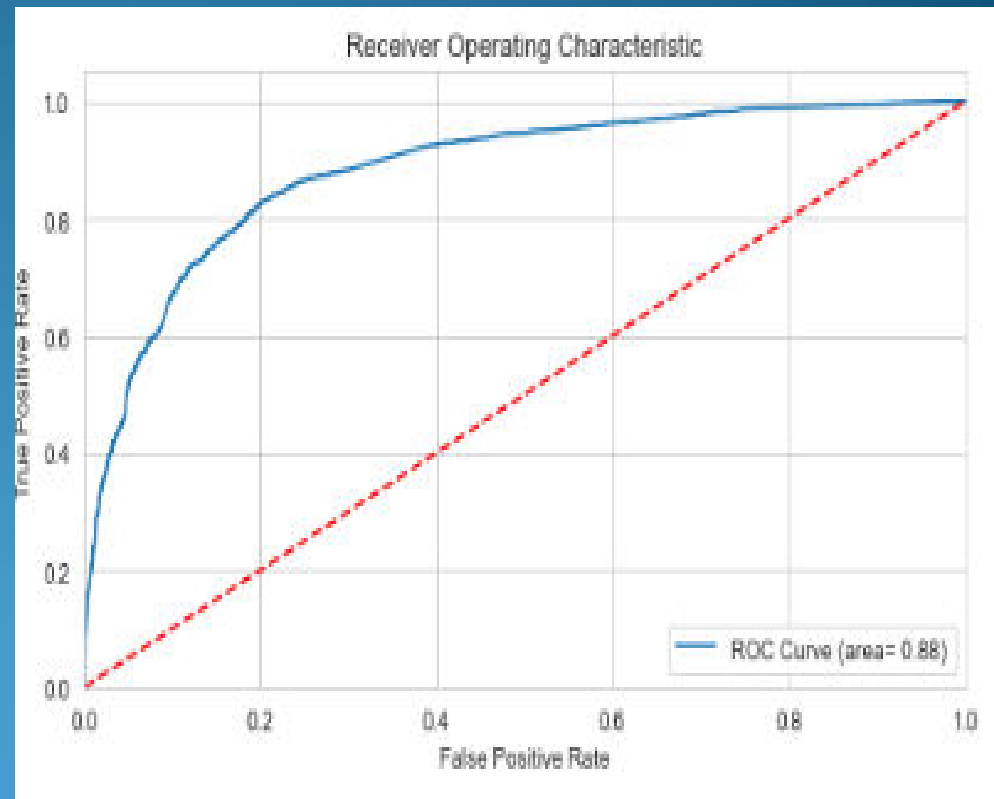
FINDING THE OPTIMAL CUT OFF POINT
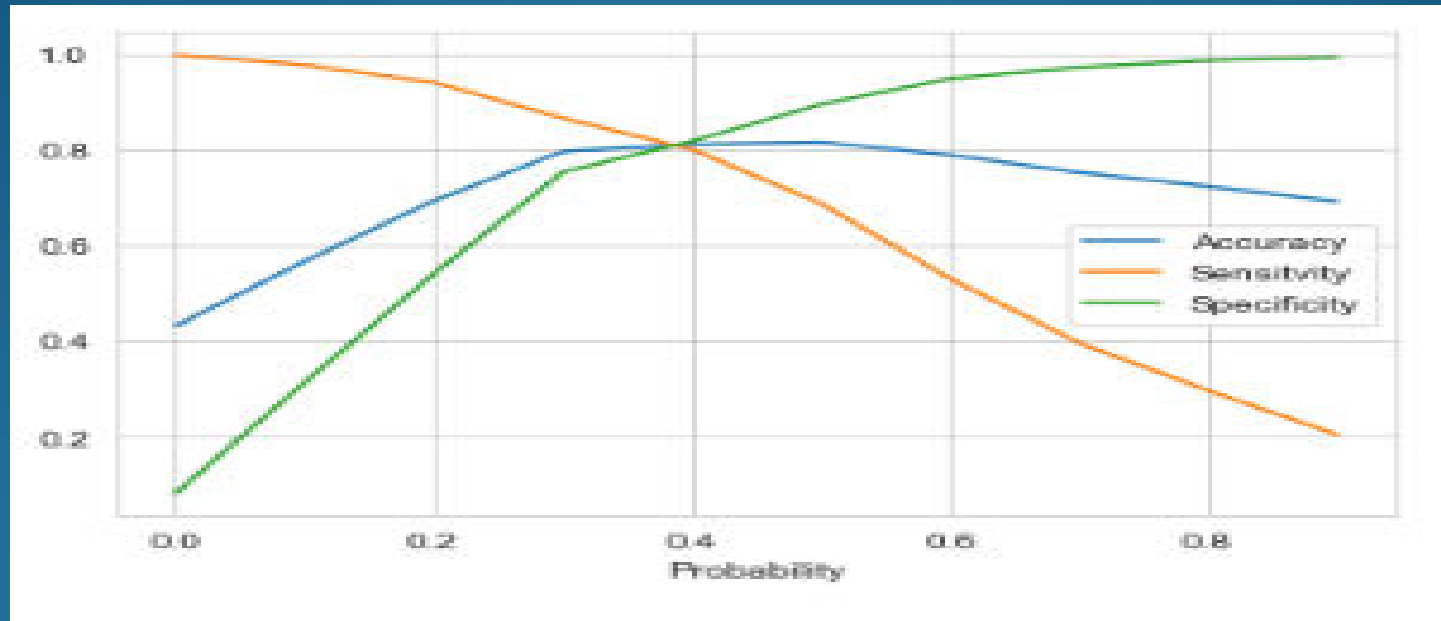
We found that on 0.4 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.

PRECISION AND RECALL TRADE OFF POINT

•We created a graph which will show us the trade off 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.

Precision vs Recall tradeoff
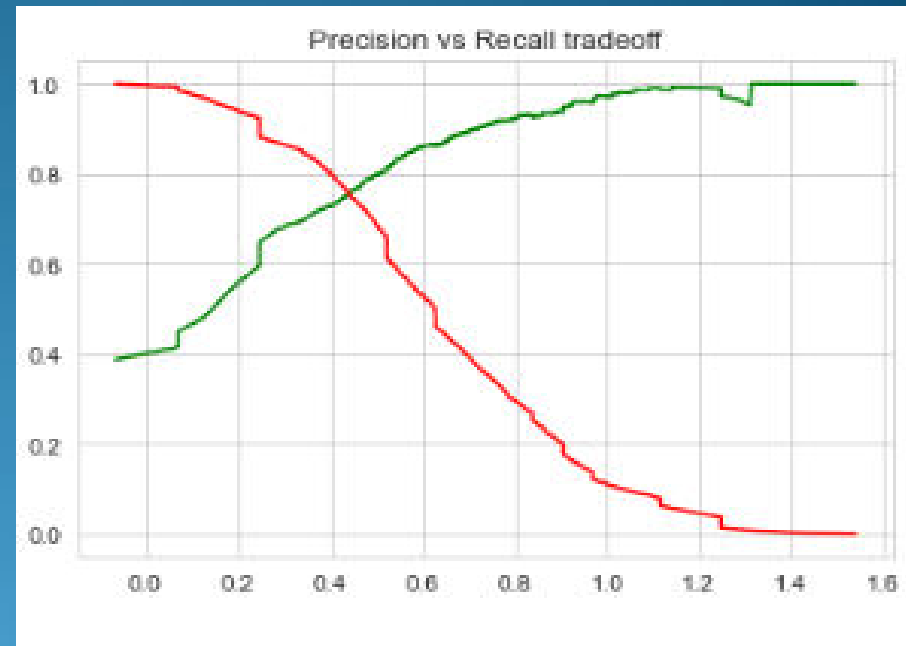
CONCLUSION

- The Accuracy, Precision and Recall score we got from the test data are in the acceptable region.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:
  - **Last Notable Activity_Had a Phone Conversation**
  - **Lead Origin_Lead Add Form**
  - **What is your current occupation_Working Professional.**