Lead Scoring Model

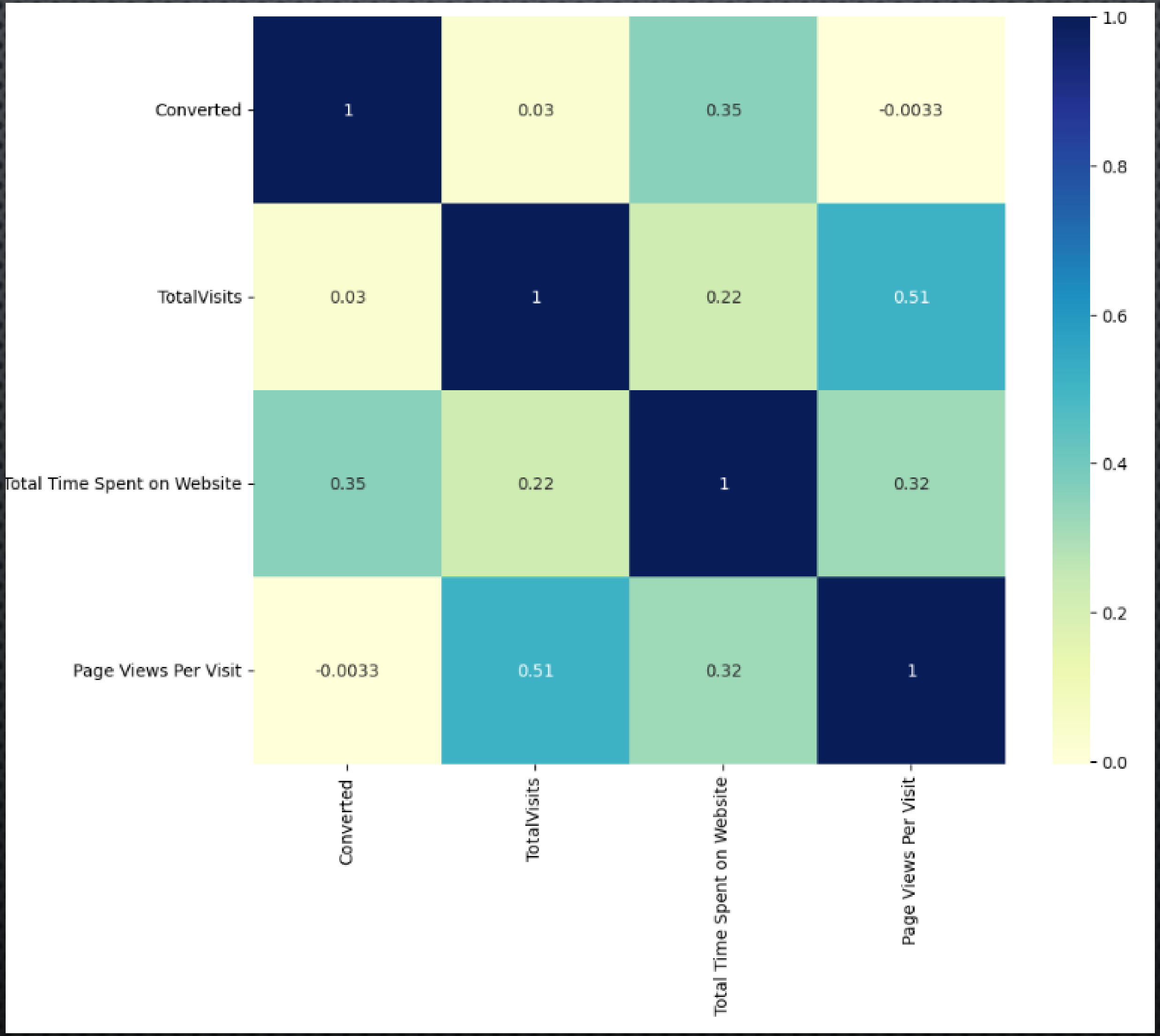
Your Company Name

X
EDUCATION

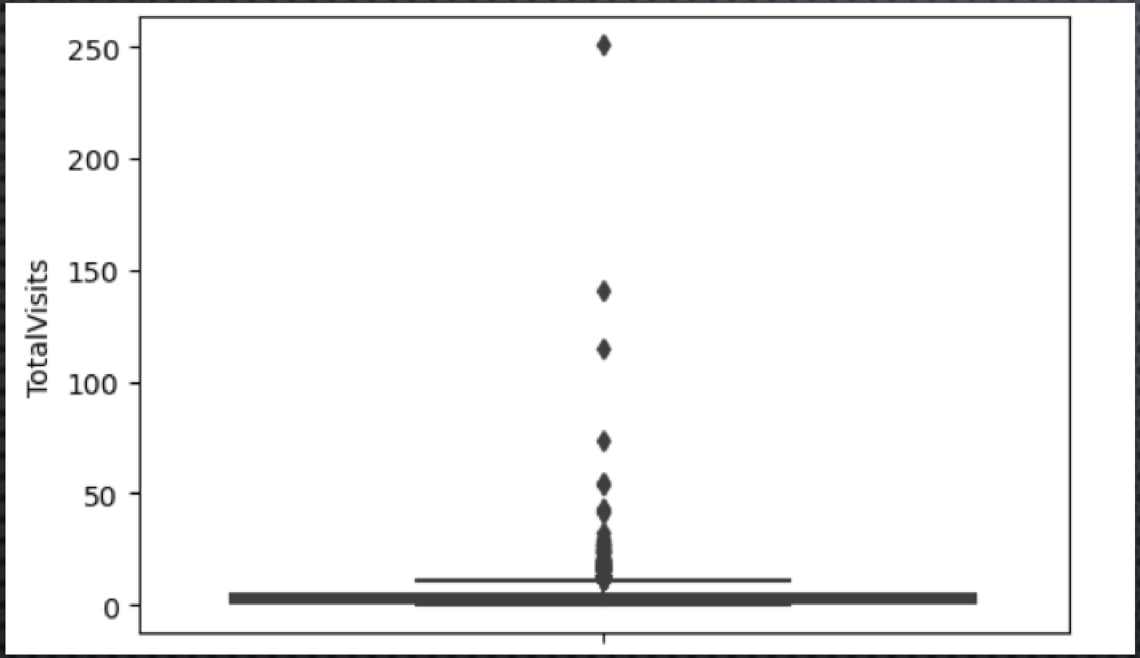
BUSINESS OBJECTIVES

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

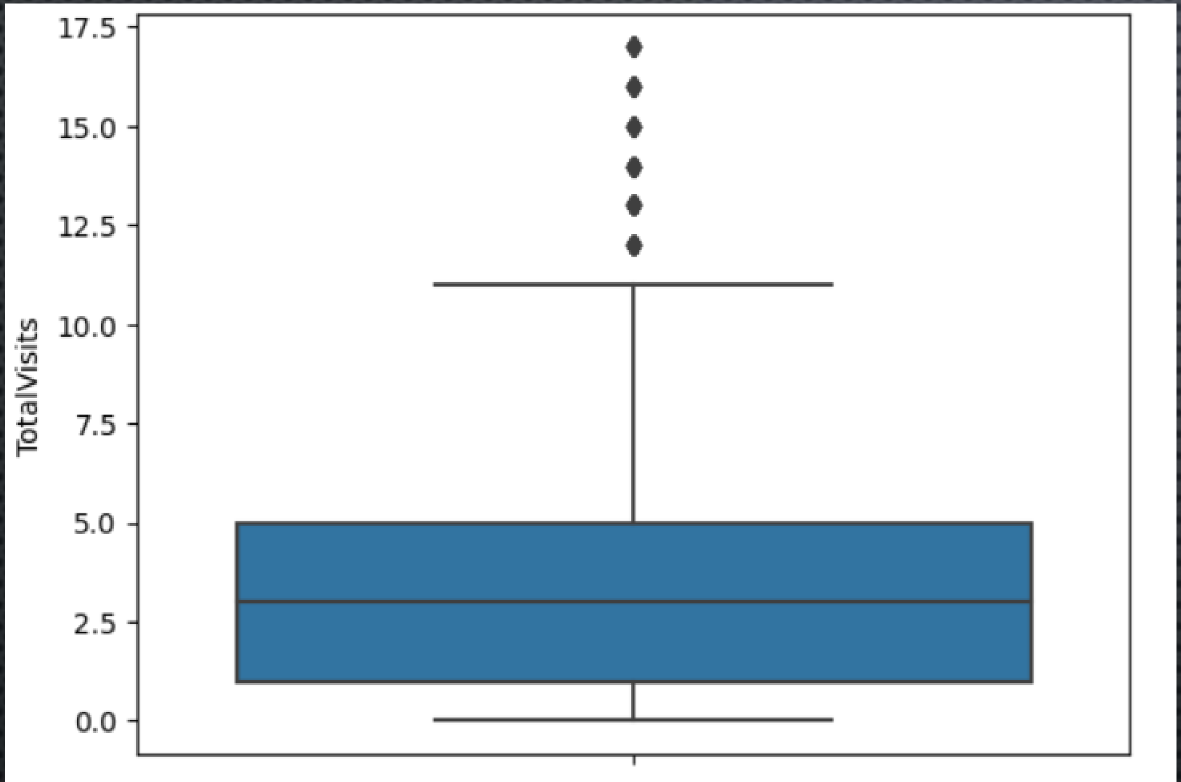
Heat MAP



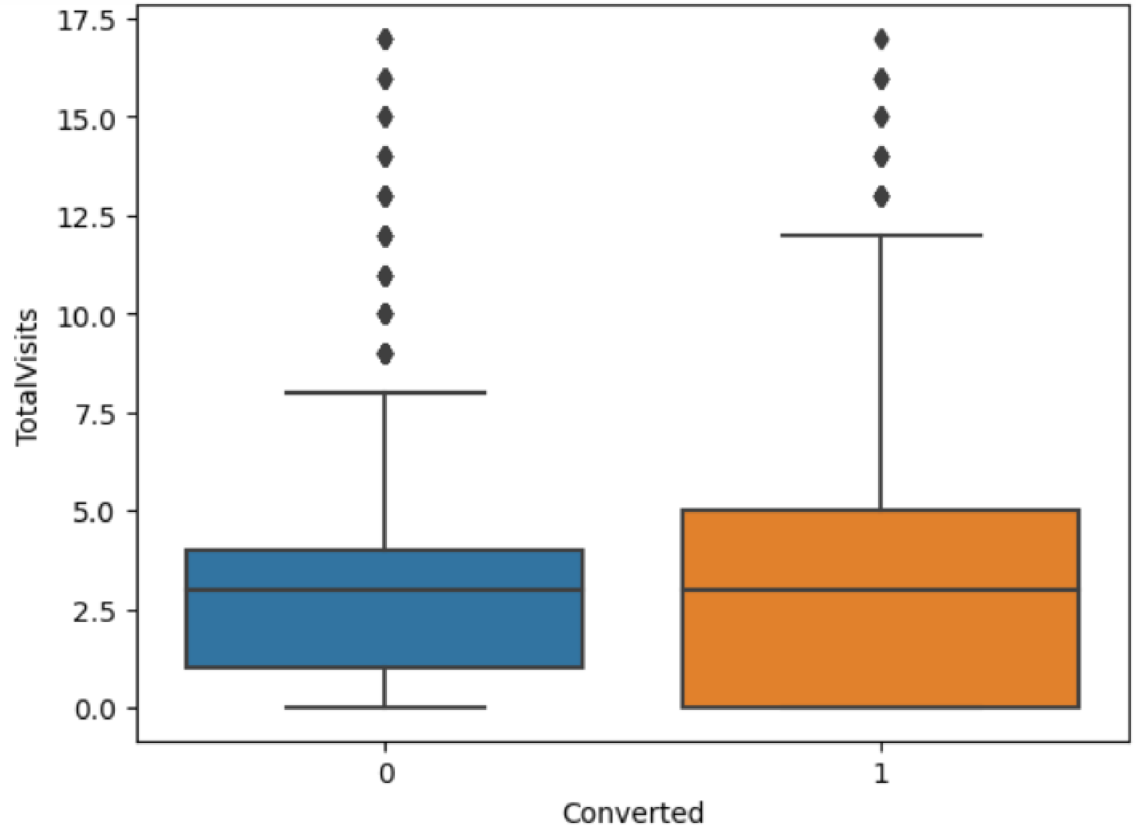
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.03
0.35
-0.0033
0.03
1
0.22
0.51
0.35
0.22
1
0.32
-0.0033
0.51
0.32
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0



Total Visit have some outlier which needs to be removed.



Total Visit after removing top 1% and bottom 1% outlier values.



Inference

Nothng conclusive can be said on the basis of Total Visits

Median for converted and not converted leads are the close.

MODEL -1

As p-value of variable Lead Source, referral Sites is high, so we can drop it.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1899	0.088	-13.480	0.000	-1.363	-1.017
Total Time Spent on Website	0.8970	0.053	16.999	0.000	0.794	1.000
Lead Origin_Lead Add Form	1.6712	0.450	3.714	0.000	0.789	2.553
Lead Source_Direct Traffic	-0.8320	0.129	-6.471	0.000	-1.084	-0.580
Lead Source_Referral Sites	-0.5284	0.465	-1.138	0.255	-1.439	0.382
Lead Source_Welingak Website	3.9043	1.110	3.518	0.000	1.729	6.079
Last Activity_SMS Sent	1.2373	0.223	5.555	0.000	0.801	1.674
Last Notable Activity_Modified	-1.2839	0.150	-8.532	0.000	-1.579	-0.989
Last Notable Activity_Olark Chat Conversation	-1.7123	0.490	-3.496	0.000	-2.672	-0.752
Last Notable Activity_SMS Sent	1.0151	0.257	3.943	0.000	0.511	1.520
Tags_Closed by Horizzon	6.9834	1.019	6.853	0.000	4.986	8.981
Tags_Interested in other courses	-2.1641	0.407	-5.321	0.000	-2.961	-1.367
Tags_Lost to EINS	5.7302	0.608	9.419	0.000	4.538	6.923
Tags_Other_Tags	-2.4417	0.210	-11.633	0.000	-2.853	-2.030
Tags_Ringing	-3.5858	0.243	-14.752	0.000	-4.062	-3.109
Tags_Will revert after reading the email	4.4263	0.185	23.989	0.000	4.065	4.788

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6251
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1254.7
Date:	Mon, 17 Jul 2023	**Deviance:**	2509.3
Time:	01:03:15	**Pearson chi2:**	8.34e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6048
Covariance Type:	nonrobust		

MODEL – 2

Since All the p-values are less we can check the Variance Inflation Factor to see of there is any correlation between the variables.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2029	0.088	-13.729	0.000	-1.375	-1.031
Total Time Spent on Website	0.8963	0.053	16.979	0.000	0.793	1.000
Lead Origin_Lead Add Form	1.6795	0.450	3.735	0.000	0.798	2.561
Lead Source_Direct Traffic	-0.8224	0.128	-6.409	0.000	-1.074	-0.571
Lead Source_Welingak Website	3.9060	1.110	3.520	0.000	1.731	6.081
Last Activity_SMS Sent	1.2437	0.223	5.584	0.000	0.807	1.680
Last Notable Activity_Modified	-1.2791	0.150	-8.501	0.000	-1.574	-0.984
Last Notable Activity_Olark Chat Conversation	-1.7079	0.489	-3.491	0.000	-2.667	-0.749
Last Notable Activity_SMS Sent	1.0150	0.257	3.943	0.000	0.510	1.520
Tags_Closed by Horizzon	6.9868	1.019	6.857	0.000	4.990	8.984
Tags_Interested in other courses	-2.2028	0.409	-5.391	0.000	-3.004	-1.402
Tags_Lost to EINS	5.7337	0.608	9.426	0.000	4.541	6.926
Tags_Other_Tags	-2.4401	0.210	-11.625	0.000	-2.852	-2.029
Tags_Ringing	-3.5818	0.243	-14.740	0.000	-4.058	-3.106
Tags_Will revert after reading the email	4.4234	0.184	23.993	0.000	4.062	4.785

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6267
Model:	GLM	Df Residuals:	6252
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1255.3
Date:	Mon, 17 Jul 2023	Deviance:	2510.7
Time:	01:03:15	Pearson chi2:	8.34e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6047
Covariance Type:	nonrobust		

Data-frame that will contain the names of all the feature variables and their respective VIFs.

	Features	VIF
7	Last Notable Activity_SMS Sent	6.22
4	Last Activity_SMS Sent	6.12
1	Lead Origin_Lead Add Form	1.82
5	Last Notable Activity_Modified	1.69
13	Tags_Will revert after reading the email	1.61
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
11	Tags_Other_Tags	1.26
0	Total Time Spent on Website	1.22
8	Tags_Closed by Horizzon	1.21
12	Tags_Ringing	1.18
9	Tags_Interested in other courses	1.13
10	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

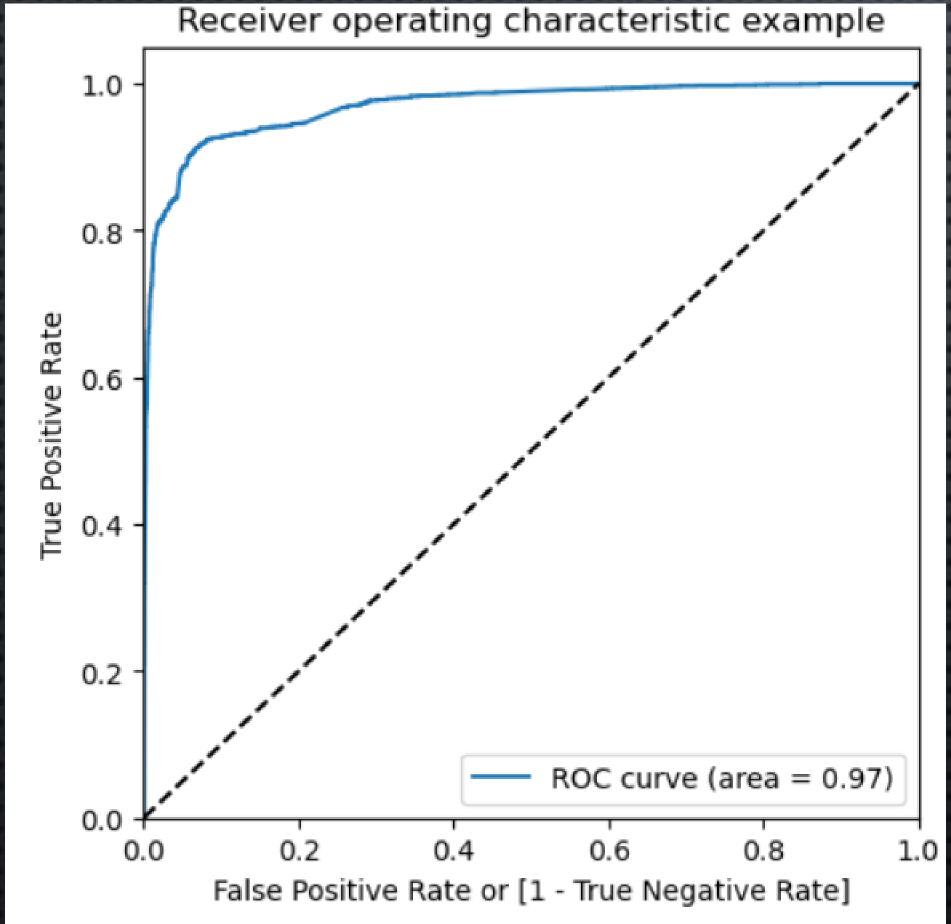
As there is a high correlation between two variables so we can drop the variable with the high valued VIF value.

After removing high VIFs values

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

The Values all seem to be in order so now, Moving on to derive the Probabilities, Lead Score, Predictions on Train Data.

PLOTTING ROC CURVE



The ROC Curve should be a value close to We are getting a good value of 0.97 indicating a good predictive model.

After checking the overall accuracy the observations are –

1. Accuracy : 92.78%
2. Sensitivity : 91.98%
3. Specificity : 93.26%

Final Observation:
Let us compare the values obtained for Train & Test:

Train Data:
1. Accuracy : 92.29%
2. Sensitivity : 91.70%
3. Specificity : 92.66%

Test Data:
1. Accuracy : 92.78%
2. Sensitivity : 91.98%
3. Specificity : 93.26%

SUMMARY

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

First, sort out the best prospects from the leads you have generated. 'Total_Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted. Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.

Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.