LEADING SCORE CASE STUDY

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.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Objective:-

Objective of this project is increase conversion rate of hot leads by creating Machine learning Model

To help company to make useful calls that will lead to better conversion and efficiency of team

Approach:-

1]Read and Clean data

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data.

2]EDA

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.

3]Data Preparation

Dummy variables were created where for categorical columns

Then we split the data in Test and Train dataset by 70:30 ratio and scaled the variables

4]Model creation

Firstly use RFE to select 15 variable that are most important and then start building Logistic Model creation and step by step we achieved final model with p value less than 0.05 and VIF less than 5 so we conclude that this model is ok.

5] Analyze the Model

After creation Model we analyse the model by calculation accuracy, precision, recall and then assign lead score to each row number so that company can decide to call or not that person

Analyse the Result:-

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6359
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2727.7
Date:	Mon, 23 Jan 2023	**Deviance:**	5455.5
Time:	16:51:48	**Pearson chi2:**	6.56e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3774
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4752	0.135	-3.528	0.000	-0.739	-0.211
Do Not Email	-1.6752	0.184	-9.107	0.000	-2.036	-1.315
Total Time Spent on Website	1.0858	0.039	27.586	0.000	1.009	1.163
Lead Origin_Landing Page Submission	-0.9611	0.128	-7.495	0.000	-1.212	-0.710
Lead Origin_Lead Add Form	3.7027	0.229	16.144	0.000	3.253	4.152
Lead Source_Olark Chat	0.9847	0.117	8.437	0.000	0.756	1.213
Lead Source_Welingak Website	2.3053	1.038	2.220	0.026	0.270	4.340
Last Activity_Email Opened	0.5710	0.083	6.853	0.000	0.408	0.734
What is your current occupation_Other	-1.2577	0.088	-14.332	0.000	-1.430	-1.086
What is your current occupation_Student	-0.1123	0.223	-0.503	0.615	-0.550	0.325
City_Others	-0.9570	0.125	-7.638	0.000	-1.203	-0.711
Last Notable Activity_Others	2.0171	0.276	7.297	0.000	1.475	2.559
Last Notable Activity_SMS Sent	1.9972	0.093	21.487	0.000	1.815	2.179

Above variable contributed most in lead conversion ,but in which below are top 3 variable that improves lead conversion

- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Last Notable Activity_Others

Some Important factor:-

Accuracy of Train Data:-80.39%

Accuracy of Test data:- 80.99

Sensitivity of Test Data:-79%

Specificity of Test Data:-82%