Lead Scoring Case Study Summary

Problem Statement:

This analysis is done for X Education which sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned 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%. The main objective of this logistic regression model building is to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Solution Summary:

Step1: Reading and Understanding Data.

Read and analyse the data.

Step2: Data Cleaning:

We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed. 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. And variables having missing values more than 40% were dropped. Few of categorical variables had vivid categories of data we have combined the less % of categories into 'Others' so as to not lose much data. Although later some were removed for making a better model.

Step3: Exploratory Data Analysis:

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. 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 seems good and less outliers were found

Step4: Creating Dummy Variables:

We went on with creating dummy data for the categorical variables. Later on the dummies with 'others' elements were removed. For numeric values we used the Standard Scalar.

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

We used the Min Max Scaling to scale the original numerical variables. Then using standard scalar from stats model, we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: Model Building and Feature selection using RFE:

After building First model, to focus on few Potential Predictor variables we have considered balanced Approach for feature elimination i.e Combination of Automated (Recursive Feature Elimination /Coarse Tuning) and Manual (Fine Tuning) Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values manually. The VIF's for these variables were also found to be good. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Step8: Plotting the ROC Curve.

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 88% which further solidified the of the model.

Step9: Finding the Optimal Cut-off Point:

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cut-off point. The cut-off point was found out to be 0.34 Based on the new value we could observe that close to 80.83% values were rightly predicted by the model. We could also observe the new values of the 'accuracy=80.83%, 'sensitivity=81.07%', 'specificity=80.69%'.

Step10: Computing the Precision and Recall metrics.

This method was used to check Precision around 80.03% and recall around 69.50% on the train data frame.

Step11: Making Predictions on Test Set:

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.38%; Sensitivity=79.47%; Specificity= 80.91%.

The Education company needs to focus on the following factors to improve conversion rate of leads:

- The total time spent on the website impacts the conversion rate.
- Leads from lead origin 'Lead Add Form' have a high conversion rate.
- Leads both sourced from 'Olark Chat' also have decent conversion rate.
- Leads with lead source 'Welingak website' have a high conversion rate.
- Leads with last notable activity as 'SMS sent' have a high conversion rate.
- Leads who are working professionals have a pretty high conversion rate.
- Can also focus on leads with other last notable activities as: modified, page visited on website.

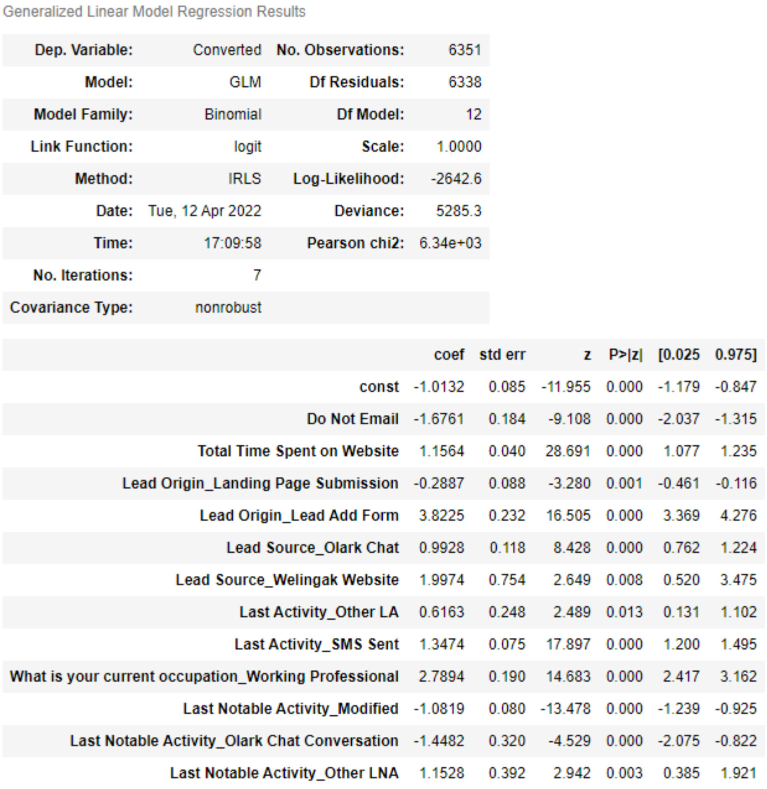
In order to elevate business, X Education has good opportunity to seize credible Customers to buy their courses.

• At about 0.34 is our final probability threshold where the three metrics (Accuracy, Sensitivity, and Specificity) seem to be almost equal with decent values to predict if a lead will convert or not.

• Lead Score is obtained by multiplying 100 to probability of getting converted.

• Lead score is directly and positively proportional to probability of a lead getting converted.

<u>Following are The Final Selected Features (Significant Predictors):</u>



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2642.6
Date:	Tue, 12 Apr 2022	**Deviance:**	5285.3
Time:	17:09:58	**Pearson chi2:**	6.34e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0132	0.085	-11.955	0.000	-1.179	-0.847
Do Not Email	-1.6761	0.184	-9.108	0.000	-2.037	-1.315
Total Time Spent on Website	1.1564	0.040	28.691	0.000	1.077	1.235
Lead Origin_Landing Page Submission	-0.2887	0.088	-3.280	0.001	-0.461	-0.116
Lead Origin_Lead Add Form	3.8225	0.232	16.505	0.000	3.369	4.276
Lead Source_Olark Chat	0.9928	0.118	8.428	0.000	0.762	1.224
Lead Source_Welingak Website	1.9974	0.754	2.649	0.008	0.520	3.475
Last Activity_Other LA	0.6163	0.248	2.489	0.013	0.131	1.102
Last Activity_SMS Sent	1.3474	0.075	17.897	0.000	1.200	1.495
What is your current occupation_Working Professional	2.7894	0.190	14.683	0.000	2.417	3.162
Last Notable Activity_Modified	-1.0819	0.080	-13.478	0.000	-1.239	-0.925
Last Notable Activity_Olark Chat Conversation	-1.4482	0.320	-4.529	0.000	-2.075	-0.822
Last Notable Activity_Other LNA	1.1528	0.392	2.942	0.003	0.385	1.921