Lead Scoring Case Study Summary

Problem Statement:

This case study is done for the Education X company which sells online courses to industry professionals. Education X markets their courses on several websites. Candidates land on their website, they brows website or watch some videos or fill the form on website. Once they fill the form and providing their email address or mobile number they are considered as lead. They also get leads through referrals. Then Employees from sales team starts making calls to these leads, out of which some converts to join any course and most do not. There current lead conversion rate is 30%.

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

Solution Summary:

Step1: **Reading and Understanding Data**.
Read and analyze the data.

Step2: **Data Cleaning**:
We dropped the variable columns that had high percentage of NULL values in them. Also, in this step we imputed missing values in numerical variables. The outliers were identified and they were capped using IQR.

Step3: **Data Analysis**
Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step we found that many categorical columns had only single type of category such and yes or no so we dropped those columns. Also many categories were too less in number like 1 to 10 so we clubbed all such categories in one category naming 'Other'

Step4: **Creating Dummy Variables**
we went on with creating dummy variable for the categorical variables.

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.

Step7: Model Creation
Using the stats model library wecreated our initial model, which would give us a complete statistical view of all the parameters of ourmodel.

Step8: **Feature selection using RFE**:
Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features. Using the statistics generated, we recursively tried looking at the P-values and VIF score in order to drop the insignificant variables from the model

Finally, we arrived at the 12 most significant variables. 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.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

Step9: 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 93% which further solidified the of the model.

Step10: Finding the Optimal Cutoff 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 cutoff point. The cutoff point was found out to be 0.35

Based on the new value we could observe that close to 85% values were rightly predicted by the model.

We could also observe the new values of the 'accuracy=85%, 'sensitivity=87%', 'specificity=84%'.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step10: Computing the Precision and Recall metrics
we also found out the Precision and Recall metrics values came out to be 77% and 86% respectively on the train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.38

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 84%; Sensitivity=87%; Specificity= 83%.

Conclusion:

Major indicators that lead will get converted to hot lead:
Following are the top variables which are responsible for converting leads successfully. Sales team should focus on these variables while making calls to make efficient calling.
Also, company should avoid wasting resources on these factors as they are performing good.

a) Tags_Closed by horizon: with positive coefficient of 8.81 these leads are mostly likely to convert.
b) Total Time Spent on Website: with positive coefficient of 4.96, leads who spend more time on website are more likely to convert.
c) Tags_Will revert after reading the email: with positive coefficient of 3.68 these leads are also likely to convert. Company should contact leads who have reverted to email.
d) Lead Origin_Lead add form: with coefficient 2.83 these are likely to convert

Major indicators that lead will get converted to cold lead:

Following are the top variables which are responsible for not converting leads successfully. Sales team should not focus on these variables that much while making calls .

Also company can focus on these variables to improve the reaction of candidates.

a) Last Activity_Olark chat conversation: with coefficient of -1.39, we can say that students having questions goes to olark chat to seek answers and they are not getting it so company can focus on this point.
b) Tags_Ringing: with coefficient -1.12, to make efficient calling team should avoid calling to these leads
c) Total visits_1-4: with coefficient -0.97
d) Last Activity_other activity: With coefficient -0.73