Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the mostpromising 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 leadscore have a lower conversion chance.

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

Summary:

Step1: **Reading and Understanding Data**. Imported the (.csv) file and did basic analysis

Step2: **Data Cleaning**:
Dropped the variables that had high percentage of NULL values. We also replaced the missing values with median values in case of numerical variables and created new classification variables in case of categorical variables. Outliers were treated.

Step3: **Data Analysis**
We did Exploratory Data Analysis (EDA) of the data set. In this step, we dropped three variables which have only one value in all rows.

Step4: **Creating Dummy Variables**
We created dummy variables for the categorical variables.

Step5: **Test Train Split**:
In the next step we divided the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling
We have used the Min Max Scaling to scale the original numerical variables. Then used the stats model to create our initial model, which would give us a complete statistical view of all the parameters of ourmodel.

Step7: **Feature selection using RFE**:
Using the Recursive Feature Elimination top 20 features were selected. Using the statistics generated, we recursively tried to look for features with high p-values (>=0.05) and dropped those features.

We also check VIF values of the feature and dropped the feature with VIF greater than 5.Finally, we arrived at the most significant variables. The VIF's for these variables were also found tobe good.

We then created the data frame having the converted probability values and we had an initialassumption 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 Accuracyof the model.

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

Step8: Plotting the ROC Curve
We then tried plotting the ROC curve for the features and the curve came out be pretty decent with anarea coverage of 89% which further solidified the of the model.

Step9: 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 cutoffpoint. The cutoff point was found out to be 0.37

Based on the new value we could observe that close to 80% values were rightly predicted by the

model.We could also observe the new values of the 'accuracy=81%, 'sensitivity=71%',

'specificity=88%'.

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

Step10: Computing the Precision and Recall metrics
we also found out the Precision and Recall metrics values came out to be 79% and 71.2% respectively onthe train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

Step11: **Making Predictions on Test Set**
Then we implemented the learning's to the test model and calculated the conversion probability basedon the Sensitivity and Specificity metrics and found out the accuracy value to be 81.4%; Sensitivity=79.4%; Specificity= 82.4%.

Step12: Finalized the case study with conclusion:
Valuable Insights -
The Accuracy, Precision and Recall score we got from test set in acceptable range.
We have high recall score than precision score which we were exactly looking for.
In business terms, this model has an ability to adjust with the company's requirements in coming future.
This concludes that the model is in stable state.
Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- **Total time spent on website**
- **Lead Add Form from Lead Origin and**
- **Had a Phone Conversation from Last Notable Activity**