Lead Scoring Case Study Summary

Problem Statement:

X Education 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%.

Solution Summary:

Reading and Understanding Data: Initial data exploration and analysis.

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.

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, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Creating Dummy Variables: We went on with creating dummy data for the categorical variables.

Test-Train Split: Divided the dataset into 70% training and 30% testing data.

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

__Feature Selection Using RFE__: Using the Recursive Feature Elimination we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 15 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.

Refined model using P-values and retained 15 significant variables with acceptable VIF values.

__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 89% which further solidified the of the model.

__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 0.37.

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=80.4%', 'sensitivity=80.7%', 'specificity=80.2%'. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

__Precision and Recall Metrics__: We also found out the Precision and Recall metrics values came out to be 79.2% and 70.7% respectively on the train data set.

Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.41.

__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 81.3%; Sensitivity=74.3%; Specificity= 77.4%.

,