Lead 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 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:

Step1: Reading and Understanding Data.

First step is to read the data and understand what the data contains.

Step2: Data Cleaning:

We dropped the variables that had high percentage of NULL values in them. There were two columns with more than 45% of NULL values so we dropped them without much thought. There were some columns

which contains ‘Select’ which was as good as NULL and therefore increasing the NULL values of columns hence we removed those columns too except the column Occupation. After that there were some columns which contains very low number of NULL values to deal with them, we removed the rows containing NULL.

Step3: Exploratory Data Analysis (EDA):

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. We analyzed the columns that are contributing in the leads converted.

Step4: Creating Dummy Variables

we went on with creating dummy data for the categorical variables here as the NULL values of the Occupation columns was converted to string we dropped that string while creating dummies and then merged with the main data.

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. Also fixing the same sample using random state.

Step6: 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.

Step7: Feature selection using RFE

Using the Recursive Feature Elimination, we went ahead and selected the 25 top important features. 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 and dropped some columns based on VIF value and were left with 23 columns.

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 89% which further solidified the model.

Step9: Finding the Optimal Cutoff Point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', 'Specificity' and 'Precision' 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 0.4 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=78%’, ’specificity=83%’, ‘precision = 73%’. 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 75% and 78% respectively on the 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 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%; Sensitivity=77%; Specificity= 83% and precision = 72%.