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 where you need to assign a lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

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

Goals of the Case Study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Solution Summary:

Reading and Understanding Data

Read and analyze the data.

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.

Visualization of Data

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

Creating Dummy Variables: We went on with creating dummy data for the categorical variables. We created dummy variables and dropped the columns for those which dummy variables created.

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

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 15 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.

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.

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 86% 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 0.41 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=79.1%, 'sensitivity=80%', 'specificity=78.5%'.

Computing the Precision and Recall

We also found out the Precision and Recall metrics values came out to be 78% and 79% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

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 78.9%; Sensitivity=78%; Specificity= 80%.