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:

Step 1: **Reading and Understanding Data**.
Read and analyze the data.

Step 2: **Data Cleaning**:
The variables that had high percentage of NULL values in them have been dropped. 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.

Step 3: **Data Analysis**
Exploratory Data Analysis of the data set has been done to get a feel of how the data is oriented. The variables with high correlation has been identified.

Step 4: **Creating Dummy Variables**
Dummy data for the categorical variables has been created.

Step 5: **Test Train Split**:
The next step is to divide the data set into train and test sections with a proportion of 70-30% values.

Step 6: **Feature selection using RFE**:
Using the Recursive Feature Elimination the 20 top important features were selected. Using the statistics generated, the most significant variables were selected and least significant variables were dropped based on p-value.

Based on the above assumption, the Confusion Metrics has been created and parameters such as Accuracy, Sensitivity and Specificity of the model has been calculated for both train ad test data set.

Step 7: Plotting the ROC Curve

The ROC curve for the features has been generated and the curve came out be pretty decent with an area coverage of 90% which further solidified the of the model.

Step 8: Finding the Optimal Cutoff Point

The probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' has been generated for different probability values. The intersecting point of the graphs has been considered as the optimal probability cutoff point. The cutoff point is found to be 0.37

Based on the new value it has been found that 80% values has been rightly predicted by the model. The model gives an accuracy, sensitivity and specificity of 81.64%, 70.44%, 88.66% respectively for the train data set.

Step 9: Computing the Precision and Recall metrics.

The Precision and Recall metrics values came out to be 79.55% and 70.44% respectively on the train data set.

Based on the Precision and Recall tradeoff, a cut off value of approximately 0.42 has been obtained.

Step 10: **Making Predictions on Test Set**

The learnings have been implemented to the test model and model parameters have been calculated. The conversion probability based on the Sensitivity and Specificity metrics has been found out the accuracy value to be 80.5%; Sensitivity=80.5%; Specificity= 80.5% has been recorded.

Submitted By:
Aayushi Singh
Abhijeet Singh