Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X education needs help in selecting the most hot leads, i.e. the leads that are most likely to convert into paying customers.

The company requires a model where a Lead score is assigned to the Leads such that the customers with higher lead score have a higher conversion change and the customers with lower lead score have a lower conversion change.

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

Solution Summary:

Data reading and Understanding.

Read and analyse the data.

Data Cleaning

The variables with the high percentage of null values was dropped. We also imputed the missing values as and where required.

Data Analysis

We perform Exploratory Data Analysis on the data set to get a better understanding of the orientation of the data. Here, we found that multiple variables were identified to have only one value in all rows and these variables were dropped.

Creating Dummy Variables

Here we created dummy variables for the categorical variables.

Test Train Split

Then we divided the data set into test and train sections with a proportion of 80-20% values.

Feature Scaling

We used the Standard Scaling to sale the original numerical variables. 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

We used the Recursive feature Elimination and selected the 27 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 retained whereas the insignificant variables are to be dropped.

We reiterate this process and arrive at the 18 most significant variables. The VIF's for these variables were also found to be good.

Based on the assumption that a probability value of more than 0.5 means 1 else 0, we derived the confusion matrix and calculated the overall Accuracy of the Model.

The 'Sensitivity' and 'Specificity' was also calculated to understand how reliable the model is.

Plotting the ROC Curve

Then we plotted the ROC curve for the features and the curve turned out to be quite decent with an area coverage of 88%.

Finding the Optimal Cut-off Point

Optimal cut-off point is the intersection point between the accuracy, sensitivity, specificity when we plot the probability graph for accuracy, sensitivity, specificity for different probability points. The optimal cut-off pint was found to be 0.345.

We also observe that close to 72% values were rightly predicted by the model.

We also calculated the new values of the 'accuracy = 81%','sensitivity = 80%', 'specificity = 81%'.

Computing the Precision and Recall Metrics

We also calculated the Precision and Recall on the Train set and that turned out to be 80% and 66 % respectively on the train.

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

Making Prediction on Test Set

We implemented the learning to the test set and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.4%, Sensitivity = 80.1%, Specificity = 80.7%.