Lead Scoring Case Study Summary

Problem Statement:

X Education specializes in offering online courses to professionals within various industries. The company is seeking assistance in identifying the most favorable leads – those with the highest potential to become paying customers. They require a system that assigns a lead score to each potential customer, reflecting the likelihood of their conversion. The goal is to prioritize leads with higher scores, as they are more likely to convert into paying customers. Conversely, leads with lower scores are expected to have a lower chance of conversion. The CEO has indicated a rough target lead conversion rate of approximately 80%.

Solution:

Step 1: Reading and Understanding Data.

Read and analyse

Step 2: 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.

Step 3: 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.

Step 4: Creating Dummy Variables

we went on with creating dummy data for the categorical variables.

Step 5: Test Train Split:

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

Step 6: 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.

Step 7: Feature selection using RFE:

Using the Recursive Feature Elimination, we went ahead and selected the 35 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.

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.

Step 8: 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 90% which further solidified the of the model.

Step 9: 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=81.5%, 'sensitivity=80.2%', 'specificity=82.29%'.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step 10: Computing the Precision and Recall metrics

we also found out the Precision and Recall metrics values came out to be 80% and 71.4% respectively on the train data set.

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

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 81.39%; Sensitivity=79.4%; Specificity= 82.58%.