Lead Scoring Case Study Summary

Problem Statement:

An online education company wants to know its potential lead (unpaid customers) which can convert into paid customers. The company wants a model which assign a lead score to each lead based on chances of its converting to paid customer. Higher the lead score, higher the chance of conversion.

Solution Summary:

Step-1: Reading the data

We read the data and analyzed various points like shape, dimension and statistical part of data.

Step-2: Data cleaning

In this step, we first replace the select value with null and then check for null values in all columns. Columns having more than 30% null values are dropped and remaining null values were replaced by the mode value in their respective column.

Step-3: Data transformation

In this step, we converted the categorical values (Yes/No) in to 1 and 0 respectively. After this we created the dummy variables and then dropped the duplicate and redundant variables.

Step-4: Train-Test split and feature scaling

In this step, we split the data in train and test part in 70-30 ratio. After that we used Min Max scaling to scale the numerical variable.

Step-5: Model building and feature selection using RFE

We first build the model with all dummy variables but after that we used RFE to select 20 variables and calculated p value. We then eliminate variables one by one which has p value more that 0.05. After 6 iterations we got our final model where every variable has p value less than 0.05 and VIF value less than 3.

Step-6: Calculating the model evaluation metrics

In this step we calculated the evaluation metrics such as accuracy, sensitivity and specificity which came out to be 81%,70.4% and 88.7%.

Step-7: Plotting ROC and finding optimal cutoff point

In this step, we plotted the ROC curve for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.
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 0.36. 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=79.6%', 'specificity=81.8%'.

Step-8: Calculating the precision and recall metrics

In this step, we calculated the precision and recall which comes out as 79.4% and 70.4% respectively.

Step-9: Prediction on test set

In this step, 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 81.9%, Sensitivity=79.6% and Specificity= 83.4%