Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. The company acquire leads from various sources. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goals of the case study

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

Solution Summary:

Step1: **Reading and Understanding Data**.

Load, read and inspect the data.

Step2: **Data Cleaning**:

Firstly, we replaced 'select' values with NaN and dropped the variables having >40% of NULL values. For others, we imputed the missing values with median in case of numerical variables or created new category in case of categorical variables. Also, dropped categorical columns with highly skewed data; identified and removed outliers.

Step3: **Data Analysis**

Visualised categorical variables against Converted variable for insights.

Step4: Data Preparation (Creating Dummy Variables)

Created dummy data for the categorical variables.

Step5: **Test-Train Split**:

We split the dataset into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

We used the Min-Max Scaling to scale the original numerical variables

Step7: **Model Building using Stats Model & RFE**

Initially, using RFE we selected the 15 features. Then using the stats model we created our first model with those 15 features. We recursively tried looking at the P-values and VIF in order to select the most significant features and dropped the insignificant features.

Finally, we finalized our model with 9 features.

Created a dataframe with the actual converted flag and the predicted the leads assuming the initial probability cutoff 0.5.

Step8: Plotting the ROC Curve

We then plotted the ROC curve for the features which came out be pretty decent with an area coverage of 87%.

Step9: 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 0.31

Based on the new value we could observe that close to 83% values were rightly predicted by the model.

We could also observe the new values of the ‘accuracy=78%, ‘sensitivity=84%’, ’specificity=75%’.

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

Step10: Making predictions on the test set

We applied scaling and made predictions on test set.

Step11: **Evaluating Overall Metrics**

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 =77%; Sensitivity=82%; Specificity= 73%.