Summary Lead Score Case Study

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%.

Step1: Reading and Understanding Data:-

Read and analyse the data.

Step2: 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.

Step3: 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.

Step4: Creating Dummy Variables:-

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

Step5: Test Train Split:-

The next step was to divide the data set 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. 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.

Step7: Model Building:-

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Step8: Model Evaluation:-

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 79% each.

Step9: Prediction:-

Prediction was done on the test data frame and with an optimum cut off is good with accuracy, sensitivity and specificity of 79%.

Step10: Precision – Recall:-

This method was also used to recheck and a cut off of 0.44 was found with Precision around 77% and recall around 78% on the test data frame.