Lead Scoring Case Study Summary

In this assignment we have to build a logistic regression model which predicts whether the given lead will convert or not.

First we read the file and see its details to get an intuition about file. Then we have to treat the missing value. First we drop the column having more than 70% missing value. Since most of the missing values are from categorical variable, we looked each column separately by plotting them and looking ow to impute them. After missing value treatment we did the Univariate Analysis with the target variable. And for numerical variable we did the outlier treatment. By doing univariate analysis we get to know few variables which would not help in the model building because most of them having very little variance in them. Then after that we converted the binary variable into 1 and 0 and created the dummy variables for the categorical variable. Then we separate the target variable into y and other variables in X. Then we separate the training and test data set. Then we scaled the numerical variables by Standard Scaler. Then we build the model using stats model. To reduce the number of variable we use the RFE and reduce the number of variables to 15. And again build the model by stats model and displayed the summary. From there we saw the significance of each variable by seeing the p-value and it should be less than 0.05 if any variable which has the p-value greater than 0.05 we dropped it one by one and build the model after dropping each variable. Then we checked the VIF for each variable to know whether each variable are free or have very less multicollinearity. If the VIF would have been greater than 5 then we would have dropped it. Then we did the prediction of whether the particular lead will convert or not in the form of probability. And initially we set the cut-off of probability for deciding whether any lead will convert or not at 0.5. Then we derived the confusion matric and overall accuracy. By the help of confusion matric we get True Positive, True Negative, False Positive and False Negative. Which help us in deriving the sensitivity, specificity, False Positive Rate, Positive Predictive Value and Negative Predictive Value. Then we plot the ROC curve and to calculate optimum cut off we plot the accuracy, sensitivity and specificity for various probabilities and we get the optimum value at which all three graph meet. Then after getting optimum cut off value we predict for each lead whether they will convert or not, and calculate all the matrices that we have calculated above again related to confusion matric. And we have to see that Specificity and Sensitivity are above 80% to reach our goal. Then we tested our model on the test data set. And checked its Specificity and Sensitivity are above 80% to reach our goal.