Summary Report

Problem Statement:

An Education company named X education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their websitre and browse.

Methodology:

Data cleaning: It is the first part of the methodology. After reading and understanding the data and its columns, we thought of perfoming the Data Cleaning as it plays vital role in the model building. This part started with the missing value treatment, We have purged the columns where missing values are more than 30%. All the data set values are converted to lower case to maintain the uniformity of data. After the missing values trreatment, we found that there are few columns contains “select” as the column value and this is changed to “other” and “unknown” as per the data avaialble in the different columns. We did missing value treatment for the columns where missing values are less than 30% and treated the outliers with the help of percentiles, paiplot and boxplots as part of data cleaning. Further, binary enoding is added for the categorical variables and checked for the correaltion matrix using heatmap and purged the highly corrrelated variables.

Model Building:

After the Data cleaning is completed, we have divided the dataset into train-test with 70-30 ratio and defined X and Y variables. Feature Scaling is performed using StandardScaler and fit transform for numerical variables.

We proceeded to build the model using the Logistic Regression, and variables are selected using RFE Method and removed the features using manually where the feature with P-value is more than 0.05 adn VIF is more than 5.

We checked the metrics such as confusion matrix, and Accuracy Score for the model. We also checked the Sensitivity, Specificity, false positive rate, positive predictive value and Negative predictive value.

We drawn the ROC curve with the Optimal Cut-off point, Accuracy, Sensitivity and Specificity plot. We also checked the Accuracy, precision and recall for Test and train data sets.

Finally we merged the lead score data with the original dataset given, and we added the Probability cut-off and projected leads for the model in the future.

Darpan Shah
VenkataNagarjuna T H