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 website and browse for courses.

Process→

- First we read the data set in python notebook, then one inspect the data set.
- There are many columns with high % of null values, i.e. null values >35 % ,drop these columns.
- Then with the other columns with lesser % of null values, assign them value 'unknown' nd for those rows which have 'select' as an entry which is as good as null values, replace them with 'unknown.
- After all of this split the dataset into train and test data.
- Then using RFE selected some 25 features out of 115 features.
- After it, started building model and based on VIF and p value dropped some columns.
- Then took out all the values of metrics such as accuracy, sensitivity, specificity, Precision , recall for the train set
- Then plotted ROC curve for the same.
- Then took out probability for conversion
- Calculated threshold probability from accuracy , sensitivity, specificity curve, which came out to be 0.35.
- Calculated all the predictions on test set ie. All the metrics
- Plotted ROC curve for it and took out area under the curve for it, which came out to be 0.9.
- Calculated the lead score for diif lead IDs with the formula lead score= conversion probability *100