SUMMARY

Problem Statement:

An X Education company 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 where every lead is to be given a lead score, such that a customer with higher lead score will have a higher chance of conversion similarly a customer with lower lead score will have a lower chance of conversion.

We're given to reach a target percentage of 80% from 30%.

Solution Summary:

1.Reading and Understanding the Data

We start by importing necessary libraries and using pandas we read the csv file and started to understand the data.

2. Cleaning the Data

We started by checking the null values of each column and dropped the columns with very high number of null values. We individually checked some columns and decided to either drop them as they were of no use in the analysis or to replace the null values.

3. Preparing the Data

We start by doing the univariate analysis and then bivariate analysis with respect to target variable 'Converted' on categorical and numerical variables.

4. Creating Dummy Variables

We created dummy data for categorical variables, and after creation dropped the variables for which dummy variables had been created.

5. Train-Test Split

We next divide the data set into two parts train and test sections with a proportion of 70 – 30.

6. 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.

7. Feature selection using RFE

Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features. Using the statistics generated, we recursively tried looking at the p-values and VIF in order to select the most significant values that should be present and dropped the insignificant values. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

8. Model Evaluation

A confusion matrix was made, and later using the ROC curve an optimal cut off point (0.34) was used to find the Accuracy, Sensitivity and Specificity which came around to be 80% for each them.

9. Prediction

Prediction was done on the test data frame with the optimal cut off point of 0.34, Accuracy was found to be 80%, Sensitivity was around 80% and Specificity was 80%.

10. Precision and Recall

This method was used to recheck and with this on the test data frame Precision was 71% and Recall was around 80%