Lead Scoring Case Study Summary

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%

The company requires you to build a model wherein you need to assign a lead score 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%.

Summary:

Below are the steps with which we approached the problem

1.) Data reading and understanding
    Through python commands, we read the csv data into dataframes and analyse the shape , info and understand the basic structure of data.
2.) Identify the missing values present in the data. Drop columns/rows with huge missing value percentage.
3.) Impute missing values with mean, median or mode values wherever appropriate, or with any other suitable values for categorical variables.
4.) Analyse outliers present if any in numerical variables.
5.) Perform EDA on the remaining variables with univariate and bivariate analysis to understand the data better and correlations among other variables and proportion related to target variable

6.) Data preparation for model building- Create dummy variables for all categorical variables , divide them as train and test data. Perform scaling.
7.) Using RFE method, we have selected features for model building
8.) By checking p value and vif value, dropped variables with high p and vif values
9.) Reached optimized model with 13 variables .
10.) Model is then evaluated by calculating parameters Accuracy,Sensitivity, Specificity for train and test data.
Below are the resultant values:
Test:
Accuracy : 88.5
sensitivity :86.9
specificity : 89.4

Train:
Accuracy : 89.8
sensitivity :89.5
specificity : 90.06

11.) ROC Curve is plotted for which the area under curve s 0.96 which suggests the derived model is optimal.
12.) Lead Score value between 0 to 100 is calculated and predictions are made.
13.) Optimal cutoffs points are calculated
14.) Precision and Recall values are calculated which are 92.2 and 84.3 respectively.
15.) With the derived cut offs model is evaluated on test data.