Lead Scoring Case Study Assignment Summary

Problem Statement:

X Education 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 wherein you a lead score is assigned 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.

Solution Summary

This analysis is done for X Education and to find ways to get more leads to join their courses. The basic data provided gave us a lot of information about how the potential customers and their conversion rate.
Effective way of working on the leads is to start with potential leads i.e. leads that have higher probability of getting converted. This will not only result in higher conversion ratio but also effective use of time. Time spent on nurturing hot leads can be increased whereas time spent on leads with low score can be minimized.
Determining potential and not so potential leads can be done by using a logistic regression model using the data provided for each lead, and assign score to each lead.

Step1: Reading and Understanding Data.

Read and analyse the data.

Step2: Cleaning data:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Outside India’ and ‘not provided’.

Step3: Exploratory Data Analysis

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step 4: Dummy Variables:

The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMaxScaler.

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value .

Step7: Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

Step8: Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

There are many learning we gathered from this assignment. These are as below:

1. How to handle missing value and outliers in a data set.
2. How to create dummy variables/labels on categorical columns.
3. How to use python libraries to perform logistic regression on selected features.
4. How to choose best model based on balanced sensitivity and specificity.
5. Finally, how to solve a problem with team effort.