Lead Scoring Assignment Summary

Problem Statement:

• Online courses are offered by X Education to business professionals. X Education wants assistance in identifying the leads that have the best chance of becoming paying clients.
• The business needs a model where each lead is given a lead score, and leads with higher lead scores have a better chance of converting, while leads with lower lead scores have a lower chance of converting.
• The CEO in particular has stated that an approximate 80% lead conversion rate is the Aim.

The following are the steps used:

1. Importing data:

All the required libraries have been imported to perform the lead score case study. Moreover, validated the info, shape, and other dataset details.

2. Cleaning data:

The data has been cleaned by the following methods. Finding the Pearson correlation for continuous variables and the low correlation with the target variable has been dropped. Dropped the missing values which don't have relevance with the target Variable.

3. EDA:

Performed EDA on both Numerical and Categorical data and attained imperative insights. Moreover, plotted heat map was to check the collinearity between variables.

4. Data preparation for model building:

a. The variable which has Yes/No values is converted to 1/0 and dummies are created. Furthermore, original variables are removed from the dataset.

b. Analyzed the correlation between variables.

c. Dropped the variables which have high multi-collinearity (greater than 0.9).

d. Performed train/test split of (70% - 30%) using sklearn model.

e. To perform feature scaling used standard scaler.

5. Model Building:

Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. 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.

7. Prediction on Test dataset:

The prediction was done on the test data frame and with an optimum cut-off of 0.5 with accuracy, sensitivity, and specificity of more than 80%.

According to research, the factors that affected potential customers the most were

1. The total time spends on the Website.

2. When the lead source was:

a. Facebook
b. Olark chat
c. Welingak website

3. When the last activity was:

a. SMS

4. The leads which are tagged as ‘closed by horizon’, ‘Tags_Busy’, ‘interested in next batch’, ‘lost to EINS’, and ‘will revert after reading the email’ can also prove hot leads. Prompt actions from the sales team on such tagged leads can improve the flow of customers.

With these in mind, X Education can succeed since they have a very good probability of persuading nearly all of the possible customers to purchase their courses.