Summary Report

Lead Scoring Case Study

This research is being conducted for X Education in order to discover new ways to attract more industry professionals to their courses. The basic information provided us with a lot of information about how potential customers visit the site, the time they spend there, how they arrived at the site, and the products they purchase and the rate of conversion.

The steps we took to finish our assignments are as follows:

1. Understanding data:
   We imported the required libraries. We the read and analysed the data.
2. Data Cleaning:
   The variables with a high percentage of NULL values were removed. In the case of numerical variables, this step also included imputing missing values as needed with median values, and in the case of categorical variables, creating new classification variables. The outliers were discovered and it was taken care of.
3. Exploratory Data Analysis(EDA):
   To check the condition of our data, we ran a quick EDA. Many elements in the categorical variables were found to be irrelevant. The numerical values appear to be accurate, and no outliers have been discovered.
4. Dummy Variables:
   The dummy variables were created, and the ones that had 'not provided' elements were later removed. We used the MinMaxScaler for numeric values.
5. Train-Test split:
   The following step was to divide the data set into test and train sections with a 70-30 percent split.
6. Model Building:
   First, RFE was used to identify the top ten relevant variables. Later, the remaining variables were manually removed based on their VIF values and p-values (the variables with VIF 5 and p-value 0.05 were kept).
7. Plotting the ROC Curve:
   We then plotted the ROC curve for the features, and the curve came out pretty good with an area coverage of 87 percent, further solidifying the model.
8. Model Evaluation:
   We decided to create a confusion matrix. Later, the optimum cut off value (using the ROC curve) was determined to be around 80% for accuracy, sensitivity, and specificity.
9. Prediction:
   Prediction was performed on the test data frame with an optimum cut off of 0.4 and an accuracy, sensitivity, and specificity of 80%.

It was found that the variables that mattered the most in the potential buyers are:

- When their current occupation is as a working professional.
- When the lead origin is Lead add format.
- When the lead source was Welingak website
- Total number of visits
- The total time spend on the Website.
- Last Notable Activity_Had a Phone Conversation

Keeping these in mind, X Education can thrive because they have a very high chance of convincing almost all potential buyers to change their minds and purchase their courses.