Lead Scoring Case Study -- Summary

This analysis was conducted for X Education to explore strategies for increasing enrolment of industry professionals in their online courses. The data provided offered insights into how potential customers interact with the website, including their time spent on the site, how they arrived, and the overall conversion rate.

The following steps are taken:

1. **Importing modules:**
    - The basic modules like numpy, pandas, matplotlib and seaborn.
    - Import the statsmodel and sklearn libraries and their modules like VIF, RFE, test-train split, Logistic Regression, metrics and MinMaxScalar.
2. **Reading and understanding the data:**
    - Read data from CSV File and derive its information, statistical data, shape etc.
3. **Data Cleaning:**
    - Some columns have minute value counts; we can club them to a single value to retain data.
    - The "Select" option was replaced with null since it has no meaningful information.
    - Removing highly skewed columns.
    - The data has a few null values which needs imputation where mode is filled for categorical columns.
4. **Data Visualization:**
    - Done with Univariate Analysis -- Bar Graph for Categorical and Histograms for Numerical features.
    - Done with Bivariate Analysis – For all features with respect to Target variable.
    - Done with Multivariate Analysis – Heat Map is plotted for correlation between the numerical features.

5. **Data Preparation:**
    - Dummy variables were created for non-binary class features. The binary valued columns (Yes/No) are mapped to 1/0.
    - The data was split into 70% for training and 30% for testing.
    - For Rescaling, MinMaxScaler () was applied to scale numeric values.
6. **Model Building:**
    - Create X, y sets for train and test sets.
    - Create a logistic regression model using RFE (was used to identify the top 20 relevant variables) and fit it using training dataset.
    - Variables were further refined manually based on Variance Inflation Factor (VIF) and p-value criteria, retaining those with VIF < 5 and p-value < 0.05.
7. **Model Evaluation:**
    - A confusion matrix was constructed, and the optimal cutoff value was determined using the **ROC curve**, resulting in accuracy, sensitivity, and specificity of around 80% and for **Precision-Recall Trade-off**, resulting in Precision and Recall of around 75-76%.
8. **Predictions on Test set:**
    - Predictions were made on the test dataset, with an **optimal cutoff of 0.35**, yielding accuracy, sensitivity, and specificity of 80% and also with an **optimal cutoff of 0.41**, yielding Precision-Recall of 75-76%.

Key Findings:

The most significant factors influencing the likelihood of a lead converting are:

1. **Total Time Spent on Website**
2. **Total Number of Visits**
3. **Lead Source:**
    - Google, Direct Traffic, Organic Search
4. **Last Activity:**

    - SMS sent
5. **Lead Origin:**
    - API, Lead Ad Format
6. **Current Occupation:**
    - Working Professional

By focusing on these factors, X Education can significantly enhance its ability to convert potential leads into paying customers, thereby increasing overall course enrolment.