Summary

This analysis pertains to X Education and its efforts to attract more industry professionals to enrol in their courses. The initial dataset provided valuable insights into how prospective customers engage with the website, the duration of their visits, their referral sources, and the conversion rates.

To address the business requirements, a lead scoring case study was conducted utilizing a logistic regression model. The following steps were followed:

1. **Data Cleaning: **

The dataset required some cleaning, including handling null values and converting certain options to null since they did not provide significant information. Some null values were replaced with top value(mode zero) to retain data

2. **Exploratory Data Analysis (EDA): **

An initial EDA was performed to assess the data's quality. Many elements were removed from the dataset. Numeric values were scaled using MinMaxScaler.

3. **Dummy Variables:**

Dummy variables were created, and dummies with "not provided" elements were subsequently removed.

4. **Train-Test Split:**

The dataset was divided into a 70% training set and a 30% test set.

5. **Model Building:**

Initially, Recursive Feature Elimination (RFE) was employed to select the top 25 relevant variables. Subsequently, the remaining variables were manually removed based on Variance Inflation Factor (VIF) values and p-values. Variables with VIF < 5 and p-value < 0.05 were retained.

6. **Model Evaluation:**

A confusion matrix was constructed. An optimal cutoff value (determined using the ROC curve) was utilized to calculate accuracy, sensitivity, and specificity, all of which were approximately between 75-80%.

7. **Prediction:**

Predictions were made on the test dataset using an optimal cutoff of 0.4, resulting in accuracy, sensitivity, and specificity of approximately 75-80%.

8. **Precision-Recall Analysis:**

Precision-recall analysis was performed, yielding a cutoff of 0.44 with precision around 42.31% and recall around 89.24% on the test dataset.

The analysis revealed that the following variables were most influential in identifying potential buyers:

1. Total Time Spent on Website
2. Total Number of Visits
3. Lead Source (specifically, Olark Chat and Referral Sites)
4. Last Activity (particularly, SMS)
5. Lead Origin (Lead Add Form)
6. Current Occupation (including Working Professionals, Students, Unemployed, and Others)

Considering these key findings, X Education can implement strategies to attract and convert potential buyers, ultimately boosting course enrolments.