Summary Report

Introduction:

This summary report presents the findings of a lead scoring case study conducted for X Education, with the objective of attracting more industry professionals to enroll in their courses. The analysis utilized provided data, which offered insights into potential customers' website visits, time spent on the site, referral sources, and conversion rates. The following steps were undertaken to achieve the study's goals:

1. Data Cleaning: The initial dataset was relatively clean, with only a few null values. The "option select" category was replaced with a null value since it didn't provide significant information. Some null values were transformed into "not given" to retain data. However, these "not given" values were subsequently removed when creating dummy variables. Additionally, the geographical data was categorized into "India," "Outside India," and "not given" to differentiate between locations.

2. Exploratory Data Analysis (EDA): A preliminary EDA was conducted to assess the quality of the data. It was observed that certain elements in the categorical variables were irrelevant. On the other hand, the numeric values appeared to be sound, with no outliers detected.

3. Dummy Variables: Dummy variables were created for the categorical variables, and the dummy variables containing "not given" elements were eliminated. The MinMaxScaler technique was applied to normalize the numeric values.

4. Train-Test Split: The dataset was divided into 70% training data and 30% testing data.

5. Model Building: The Recursive Feature Elimination (RFE) technique was employed to identify the top 15 relevant variables. The remaining variables were manually removed based on their Variance Inflation Factor (VIF) and p-values. Variables with VIF < 5 and p-values < 0.05 were retained.

Summary Report

6. Model Evaluation: A confusion matrix was created to assess the model's performance. The Receiver Operating Characteristic (ROC) curve was used to determine the optimal cutoff value, resulting in an accuracy, sensitivity, and specificity of approximately 80% each.

7. Prediction: Predictions were made on the test dataset using an optimal cutoff of 0.35, yielding an accuracy, sensitivity, and specificity of 81%.

8. Precision-Recall: The Precision-Recall method was also employed to validate the results. A cutoff value of 0.41 was identified, resulting in a precision of approximately 73% and a recall of approximately 77% on the test dataset.

Key Findings: The analysis revealed the variables that hold the most significance in identifying potential buyers. In descending order of importance, they are as follows:

1. Total time spent on the website.
2. Total number of visits.
3. Lead source:
   a. Google
   b. Direct traffic
   c. Organic search
4. Last activity:
   a. SMS
   b. Olark chat conversation
5. Lead origin as a Lead add format.
6. Current occupation as a working professional.

Conclusion: Based on the findings, X Education can greatly benefit from focusing on the identified variables to attract potential buyers, particularly industry professionals. By leveraging these insights, X Education has a high probability of successfully influencing potential buyers and encouraging them to enroll in their courses.