Summary

This analysis for X Education aims to identify strategies to attract more industry professionals to their courses. The data provided insights into customer behavior, including site visits, time spent, referral sources, and conversion rates.

Steps Taken:

1.Data Cleaning: The data was mostly clean, with a few null values. 'Option select' was replaced with null, and some null values were changed to 'Others'. These were later removed when creating dummy variables.

2. Exploratory Data Analysis (EDA): A quick EDA revealed that many categorical variable elements were irrelevant. Numeric values were good with no outliers. Outliers were present only for 'Total Visits' variable. Some spelling errors were corrected in 'Lead Source' variable. Tried to understand each variable with both classifications 'Converted = 0 and Converted =1'. Also some of the variables had only 'No' in each rows that is for Indicating whether the customer had seen the ad in any of the listed items(Magazine, Search etc).

3. Dummy Variables: Created dummy variables and removed those that did not add meaningful information to the analysis.

4. Train-Test Split: Data was split into 80% training and 20% on testing sets. Used StandardScaler for numeric values. Fit and transform was done for training data and only transform for testing data.

5. Model Building: First model was built after doing scaling, which should too many variable with high p-value. Hence used Recursive Feature Elimination (RFE) technique to identify the top 20 relevant variables. Dropped all other variables were RFE did not support. Checked for Variance Inflation Factors (VIF< 5) and none of the independent variables had high VIF that is greater than 5. Built 4 more models by manually checking and dropping variables which had higher p-value that is where p-value was > 0.05.

6. Model Evaluation: Created a confusion matrix and used the ROC curve to find the optimal cutoff value, achieving around 92.7% accuracy, sensitivity, and specificity.

7. Prediction: Predictions on the test data with an optimal cutoff of 0.35 resulted in 92.0% accuracy, sensitivity, and specificity.

8. Precision-Recall: Rechecked using Precision-Recall trade off, finding a cutoff of 0.41 with approximately 89% precision and 91% recall on the test data.

9. Assigned Lead Score for both training and testing data.

Key Variables Influencing Potential Buyers (in descending order):

1. Total time spent on the website
2. Total number of visits
3. Lead source: Google Direct traffic Organic search Welingak website
4. Last activity: SMS Olark chat conversation
5. Lead origin as Lead add format
6. Current occupation as a working professional
7. Do Not Email

By focusing on these factors, X Education can significantly increase their chances of converting potential buyers into customers, thereby boosting their course sales.