Assignment: Data Analysis for X Education

Objective

The primary objective of this analysis is to identify effective strategies for X Education to attract more industry professionals to enroll in their courses. By analyzing customer behavior data, we aim to uncover trends, patterns, and key variables that influence decision-making and conversions.

Methodology

The following steps were employed during the analysis:

1. Data Cleaning

The initial dataset was mostly clean, but several null values required attention. Fields labeled as 'option select' were replaced with null values, as they did not provide meaningful insights. Some null values were re-categorized as 'not provided' to preserve as much data as possible, though they were later removed when creating dummy variables. Since many users originated from India, geographic data was reclassified into three categories: 'India', 'Outside India', and 'Not provided'.

2. Exploratory Data Analysis (EDA)

A preliminary exploratory data analysis (EDA) was conducted to assess the condition of the dataset. It was noted that many categorical variables contained irrelevant elements, while numeric values appeared well-distributed without the presence of significant outliers.

3. Dummy Variables Creation

Dummy variables were generated, with any dummies containing 'not provided' values removed. The numeric data was standardized using the MinMaxScaler technique, ensuring consistent scaling of numerical features.

4. Train-Test Split

The dataset was split into training and testing sets, with a 70% allocation for training data and 30% for testing.

5. Model Building

Recursive Feature Elimination (RFE) was performed to identify the top 15 most relevant variables. The remaining variables were then manually removed based on their Variance Inflation Factor (VIF) values and p-values. Only variables with $VIF < 5$ and $p\text{-value} < 0.05$ were retained for further analysis.

6. Model Evaluation

A confusion matrix was constructed to evaluate the model's performance. The optimal cutoff value was identified using the Receiver Operating Characteristic (ROC) curve, yielding accuracy, sensitivity, and specificity metrics of approximately ranging between 63%-90% each.

7. Prediction

Predictions were generated on the test dataset, with an optimal cutoff value of 0.35. This resulted in an accuracy, sensitivity, and specificity of around 78%.

8. Precision-Recall Analysis

A precision-recall analysis was conducted to validate the results. A cutoff value of 0.41 was identified, with precision at 77% and recall at 63% on the test dataset.

Conclusion

Based on these findings, X Education has a significant opportunity to convert potential buyers into customers. By focusing on the key variables identified, X Education can tailor their marketing efforts to target industry professionals more effectively, thereby increasing course enrollments.