SUMMARY

This analysis is conducted for X Education with the objective of attracting more industry professionals to enroll in their courses. The initial dataset has provided valuable insights into customer behavior, including their site visitation patterns, duration of stay, referral sources, and conversion rates.

1. Data Cleaning

   The dataset was predominantly clean, although several columns contained a significant number of null values, which were subsequently dropped as they were deemed unhelpful for the model. Additionally, unnecessary columns such as 'city' and 'country' were removed to streamline the dataset.

2. EDA

   A brief exploratory data analysis (EDA) was conducted to assess the quality of our dataset. It was observed that many elements within the categorical variables were irrelevant, while the numeric values appeared satisfactory, with no outliers detected.

3. Categorical Variables

   Categorical variable analysis was undertaken, resulting in the creation of dummy variables. Subsequently, dummies containing 'not provided' elements were eliminated. Numeric values were standardized using MinMaxScaler.

4. Data Distribution for Train and Test

   Data was distributed in 70:30 ration for training and testing data sets.

5. Model Building

   Initially, Recursive Feature Elimination (RFE) was employed to identify the top 15 relevant variables. Subsequently, the variables were manually removed based on their Variance Inflation Factor (VIF) values and p-values, with variables having VIF < 5 and p-value < 0.05 being retained.

6. Model Evaluation

   A confusion matrix was generated, followed by the determination of the optimal cutoff value using the ROC curve. This process yielded accuracy, sensitivity, and specificity values, each approximately reaching 80%.

7. Prediction

   Predictions were made on the test data frame using an optimal cutoff value of 0.42, resulting in an accuracy, sensitivity, and specificity of approximately 80%.

8. Precision & Recall

   This method was also employed for reevaluation, revealing a cutoff of 0.42. This yielded precision of approximately 78% and recall of approximately 77% on the test data frame.