Summary

The analysis conducted for X Education aims to attract more industry professionals to enroll in their courses. The initial dataset offers insights into customer behavior: website visits, duration spent, referral sources, and conversion rates, essential for devising strategies to enhance enrolment.

Following Strategy is deployed for the same:

1. **Data Cleaning:**
   The data underwent initial cleaning, addressing most null values. The placeholder 'Select' was replaced with null, lacking substantial information. A few remaining nulls, deemed insignificant, were removed. Additionally, columns with over 70% null values were excluded from the analysis for relevance.
2. **Exploratory Data Analysis:**
   An initial Exploratory Data Analysis (EDA) revealed several insights. Categorical variables contained numerous irrelevant elements, while numerical variables appeared satisfactory. Outliers were identified within the dataset during this assessment.
3. **Dummy Variables:**
   Dummy Variables were created for categorical features then original features were dropped after concatenating with the newly created dummy variables.
4. **Train- Test Split:**
   Split was done at 70-30 % ratio i.e 70% data was used for training the model and 30% was used for testing the model
5. **Feature Scaling:**
   Numerical features were scaled using standard Scaler.
6. **Model Building:**
   Initially, Recursive Feature Elimination (RFE) was utilized to select the top 15 relevant variables. Subsequently, the remaining variables underwent manual selection based on VIF and P-values. Variables meeting the criteria of VIF<5 and p-values <0.05 were retained, while others were removed.

7. **Model Evaluation:**

   A confusion matrix was generated, followed by the determination of the optimal cutoff value using the ROC curve. This method yielded accuracy, sensitivity, and specificity scores of 82%, 93%, and 75%, respectively.

8. **Prediction:**

   Prediction was done on the test data frame with optimum cutoff as 0.27.

9. **Precision-Recall:**

   Additionally, this method was employed for a recheck, resulting in a cutoff value of 0.3. On the test data frame, this yielded a precision score of 71%.