Summary

The analysis centered around X Education, primarily identifying strategies to attract more industry professionals to enroll in their courses. The initial dataset has provided valuable insights into the behaviors of potential customers, including details about their website visits, duration of engagement, sources of arrival, and the observed conversion rate. This comprehensive understanding serves as a foundation for devising targeted approaches to enhance the enrolment of industry professionals in X Education courses.

The following are the steps used:

1. **Cleaning data:**

The initial data cleaning process revealed partial cleanliness, focusing on addressing null values. The 'Select' option was replaced with a null value due to its limited informational value. Additionally, some null values were temporarily replaced with 'not provided' to retain data integrity, although they were eventually removed during the creation of dummy variables. Considering the diverse geographical distribution, elements were standardized to 'India,' 'Outside India,' and 'not provided' to enhance clarity and consistency in the dataset.

2. **EDA:**

A brief Exploratory Data Analysis (EDA) was conducted to assess the state of our data. Notably, it was observed that numerous elements in the categorical variables were deemed irrelevant. On the other hand, the numeric values appeared satisfactory, with no identified outliers. This preliminary examination provides insights into the distribution and characteristics of the data, setting the stage for more in-depth analyses and targeted interventions if needed.

3. **Dummy Variables:**

The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMaxScaler.

4. **Train-Test split:**

The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**

The feature selection process began with Recursive Feature Elimination (RFE), which identified the top 15 relevant variables. Subsequently, the remaining variables underwent manual removal based on Variance Inflation Factor (VIF) values and p-values. Variables meeting the criteria of VIF < 5 and p-value < 0.05 were retained, ensuring a refined set of predictors for further analysis. This step aims to enhance model efficiency and interpretability by focusing on the most influential variables while mitigating multicollinearity issues.

6. **Model Evaluation:**

Following the creation of a confusion matrix, the optimal cut-off value was determined using the ROC curve. Subsequently, this cut-off value was utilized to calculate accuracy, sensitivity, and specificity, with each metric converging around 80%. These metrics serve as valuable indicators of the model's performance, with a balanced and satisfactory level of accuracy, sensitivity, and specificity achieved through the optimized cut-off value.

7. **Prediction:**

The prediction was executed on the test data frame using an optimal cut-off value set at 0.35. The resulting model demonstrated an accuracy, sensitivity, and specificity of 80%. This implies that the model's predictions on the test dataset, guided by the 0.35 cut-off, yielded a balanced performance across these key metrics.

8. **Precision – Recall:**

The methodology was rechecked, revealing an alternative cut-off value of 0.41. Under this threshold, the model demonstrated a precision of approximately 73% and a recall of around 75% when applied to the test data frame. This re-evaluation provides additional insights into the model's performance under different cut-off scenarios, emphasizing the trade-offs between precision and recall based on the chosen threshold.

9. **Recommendations:**

Targeted Calling Strategy:

Prioritize calls to leads from the following sources:

- Welingak Websites
- Reference
- Olark Chat

Focus on leads who:

- Are working professionals
- Have spent more time on the websites
- Showed the last activity as SMS Sent

Avoid calling leads with the following characteristics:

- Last activity recorded as Olark Chat Conversation
- Lead origin being Landing Page Submission
- Specialization listed as Others
- Selected "Do not Email" as "yes"

This strategy aims to concentrate efforts on leads with a higher likelihood of conversion while excluding those with attributes that suggest a lower likelihood of success.