

## **Summary**

This analysis was conducted for X Education to explore strategies for attracting more industry professionals to enroll in their courses. The data provided offered insights into how potential customers interact with the site, including visit patterns, time spent, traffic sources, and conversion rates.

### **Steps Followed:**

#### **1. Data Cleaning**

- The data was mostly clean but had some null values.
- The "option select" entries were replaced with null values as they were uninformative.
- Some null values were labeled as "not provided" to retain more data but were later excluded during dummy variable creation.
- Geographic data was simplified into three categories: "India," "Outside India," and "Not Provided."

#### **2. Exploratory Data Analysis (EDA)**

- Categorical variables included irrelevant entries, which were identified during EDA.
- Numeric data was clean, with no significant outliers detected.

#### **3. Dummy Variables**

- Dummy variables were created for categorical data, excluding "not provided" values.
- Numeric data was normalized using the MinMaxScaler.

#### **4. Train-Test Split**

- The dataset was split into 70% training data and 30% testing data.

#### **5. Model Building**

- Recursive Feature Elimination (RFE) was used to identify the top 15 variables.

- Variables with **VIF < 5** and **p-value < 0.05** were retained to avoid multicollinearity and ensure statistical significance.

## 6. Model Evaluation

- A confusion matrix and ROC curve were used to determine an optimal cutoff value of 0.35, achieving ~80% accuracy, sensitivity, and specificity.

## 7. Prediction

- Predictions were made on the test data with similar metrics (~80% accuracy, sensitivity, and specificity) at the cutoff of 0.35.

## 8. Precision-Recall

- This method was also applied, leading to a cutoff of 0.41 with precision ~73% and recall ~75%.

---

## Key Insights:

The most important factors influencing potential buyers (in order of importance) are:

1. **Total time spent on the website.**
2. **Total number of visits.**
3. **Lead source**, specifically:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
4. **Last activity**, specifically:
  - SMS
  - Olark chat conversations
5. **Lead origin**, particularly when it is "Lead Add Format."
6. **Current occupation**, specifically if the lead is a working professional.

**Conclusion:** By focusing on these key factors, **X Education** has a high potential to convert nearly all potential buyers into paying customers, significantly boosting enrollment in their courses.