

Summary

This analysis for X Education aims to identify strategies to attract more industry professionals to enroll in their courses. The provided data offered insights into customer behavior, including website visits, time spent on the site, sources of traffic, and conversion rates. The process involved the following steps:

The following are the steps used:

1. **Data Cleaning:**

The dataset was mostly clean, with **a few null values** and **irrelevant entries**. The option **"select"** values were replaced with nulls due to their lack of informational value. Some null values were converted to "Not Given" to retain data during processing, although these were later excluded while creating dummy variables. Additionally, **location data** was simplified into three categories: "India," "Outside India," and "Not Provided."

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

A quick EDA revealed that many **categorical variable** entries were irrelevant. **Numeric data** appeared well-structured with **no outliers**.

3. **Feature Scaling and transformation:**

Dummy variables were created for categorical fields, and entries with "Not Given" were removed. For **numerical variables**, the **MinMaxScaler** was applied for normalization.

4. **Train-Test Split:**

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

5. **Model Building:**

Recursive Feature Elimination (RFE) was applied to identify the **top 15** relevant variables. Other variables were manually removed based on **Variance Inflation Factor** (VIF) and **p-value** criteria, retaining those with **VIF < 5** and **p-value < 0.05**.

6. **Model Evaluation:**

A **confusion matrix** was created, and the **optimal cut-off** value was determined using the **Accuracy, Sensitivity, Specificity vs Threshold** which is **0.36**. **Accuracy, sensitivity, and specificity** were all approximately **80%**.

7. **Prediction:**

Predictions on the **test data** achieved accuracy, sensitivity, and specificity around **80%** with an optimal cut-off of **0.36**.

8. **Precision-Recall Analysis:**

A cut-off of **0.42** was identified, yielding a **precision and recall** of about **76%** on the train data. Also, the Sensitivity dropped to **75%** while Specificity rose to **85%**.

Key findings highlighted the factors most influencing potential buyers:

1. **Total time spent on the website (most critical).**
2. **Total number of website visits.**
3. **Lead sources**, particularly:
 - a. **Welingak website**
 - b. **Olark Chat**
4. **Last activity**, specifically:
 - a. **SMS**
 - b. **Olark chat conversation**
5. **Lead origin from lead add forms.**
6. Current occupation as a **working professional**.

By focusing on these factors, X Education can significantly improve its ability to convert potential customers into buyers, particularly industry professionals. Strengthening engagement through effective lead sources, optimizing website user experience, and leveraging targeted communication strategies can help maximize conversions.

X----X---X----X