Lead Scoring Case Study Summary:

Objective: The objective of this analysis is to explore strategies for attracting more industry professionals to enrol in courses offered by X Education. The initial dataset provided has furnished us with valuable insights regarding the website's visitor behaviour, including their browsing patterns, duration of stay, referral sources, and the rate of conversion.

The following are the steps used:

1) Data cleaning:

The dataset was mostly clean, aside from a few instances of missing values and the need to replace the option select with a null value due to its limited information. Some of the missing values were substituted with 'not provided' to minimize data loss, although they were subsequently eliminated during the creation of dummy variables. Considering the diverse origin of respondents, with a significant number from India and a few from other countries, the categorical elements were modified to 'India', 'Outside India', and 'not provided'.

2) EDA:

During the preliminary exploratory data analysis (EDA), an assessment of the data was conducted. It was found that certain elements within the categorical variables were deemed irrelevant to the analysis, while the numeric values appeared to be in good condition without any outliers.
In terms of conversion rate, approximately 39% of leads successfully converted. Notably, the primary source of conversions was identified as Google. The analysis also revealed that the "Landing Page Submission" method exhibited high lead conversions. Furthermore, it was observed that most leads preferred not to be informed through email. Additionally, it was found that using fewer copies of interviews and utilizing SMS as a method of communication yielded higher confirmed leads, while email also showed a high conversion rate.

3)Dummy Variables:

The dummy variables were generated and subsequently the dummies containing 'not provided' entries were eliminated. In the case of numerical values, the standard scaler was employed.

4) Train-Test split:

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

5) Model Building:

Initially, a recursive feature elimination (RFE) technique was applied to identify the top 15 relevant variables. Subsequently, the remaining variables were manually eliminated based on their variance inflation factor (VIF) values and p-values. Specifically, P-value is greater than 0.05
and VIF value is greater than 5.

6) Model Evaluation:

The area under the ROC curve was calculated as 0.87, indicating a favorable performance. This suggests that the model is performing well. To further assess its effectiveness, the sensitivity and specificity will be evaluated.

7) Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.5 withaccuracy of 79 %.

8) Precision - Recall:

This method was also used to recheck and a cut off of 0.42 was found with Precision around 80% and recall around 73% on the test data frame.

Conclusion :

Based on the analysis,identifying individuals who spend more time than the average on the website can be a promising approach to target and convert potential leads. Analyzing landing page submissions can provide further insights and help identify additional leads. It has been observed that specializations such as marketing management and human resources management exhibit high conversion rates, indicating that individuals with these backgrounds can be valuable leads. Overall, this model has demonstrated its accuracy in predicting and identifying potential leads for X Education.

Logistic Regression model

Train Dataset :
Accuracy - 78.88%
Sensitivity - 79.47%
Specificity - 78.33%

Test Dataset :
Accuracy - 78.60%

Sensitivity - 78.71%
Specificity : 78.51%