Summary:

This EDA analysis and Lead classifier are done for X Education and to find ways to get more industry professionals to join their courses and also predict the customers who are likely to join the course. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Cleaning data: The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Outside India’ and ‘not provided’.
2. EDA: A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and outliers were found in a few numeric data which we have imputed with the 99th percentile value using the method of capping.
3. Label encoder: The inbuilt label encoder function is used to convert the categical columns into the numerical values and later using StandardScaler to scale the values within the standard deviation values
4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.
5. Model Building:Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
6. Model Evaluation: A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
7. The model demonstrates a balanced performance across various evaluation metrics. With a **precision of 0.72**, the model is effective in predicting true positives among the positive predictions. The **specificity of 0.83** indicates a strong ability to identify true negatives. Furthermore, the **sensitivity (recall) of 0.74** highlights the model's competence in detecting true positives among all actual positives. Finally, the **accuracy score of 0.80** reflects the model's overall correctness across all predictions.

8. It was found that the variables that mattered the most in the potential buyers are (In descending order):
    1. The total time spent on the Website.
    2. Total number of visits.
    3. When the lead source was:
        a. Google
        b. Direct traffic
        c. Organic search
    4. When the last activity was:
        a. SMS
        b. Olark chat conversation
    5. When the lead origin is Lead add format.
    6. When their current occupation is as a working professional.
    7. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.