Executive Summary

This analysis is done for X Education to find ways to get more industry professionals to join their courses. 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**:
    a. For all columns that have values of 'Yes' and 'No', checked the distribution of data
    b. Dropped all columns where there is only one value 'No' as they don't help in buidling the model
    c. Variables with values ‘Yes’ and ‘No’ were mapped to 1 and 0, respectively
    d. Dropped columns if more than 40% of the values were Null
    e. Based on the type of variable, handled Null values using various methods like Mode, Median, Mean and text ‘no data’
2. **Exploratory Data Analysis**:
    a. **For Outlier and Target variables**: ‘TotalVisits’ and ‘Page Views Per Visit’ are the two variables that had outliers. As the number of outlier records were low, we removed the outlier records
    b. **For Other Categorical variables**:
        i. Lead Origin mainly has two values API and landing Page Submission
        ii. Better career prospects is what customers are looking for
        iii. Most common Last Activity of customers are ’Email Opened’ and ‘SMS Sent’
        iv. Most of the customers are Unemployed
        v. Last notable activity are ‘Modified’, ‘Email Opened’ and ‘SMS Sent’
    c. **For Numerical variables**:
        i. ‘TotalVisits’ and ‘Page Views Per Visit’ are two variables that are correlated but as the correlation co-efficient is 0.68 and not very strong, we will retain both the features
        ii. Similarly, ‘X Education Forums’ and ‘Newspaper Article’ are correlated, but as the correlation co-efficient is 0.71, we will retain both the features
3. **Dummy Variables**: Created dummy variables for some of the categorical variables and dropping the first one
4. **Scaling**: StandardScaler method was used to scale the variables
5. **Train-Test split**: The split was done at 70% and 30% for train and test data respectively
6. **Model Building**: Firstly, RFE was done to attain the top 20 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)
7. **Model Evaluation**: A confusion matrix was prepared. Later on the optimum cut off value (using ROC curve) was used to find the Accuracy Score, Sensitivity and Specificity which came to be around 90% each

8. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.29 with Accuracy Score, Sensitivity and Specificity of 90%
9. **Key Takeaways**:
    a. The top three variables contributing the most towards probability of a lead getting converted are Tags, Lead Source and Last Activity
    b. The top three dummy variables that can help in increasing the probability of lead conversion are Tags_Closed by Horizzon, Tags_Lost to EINS and Tags_Will revert after reading the email

The Sales team should target those customers who have the Tags assigned with values of [“Closed by Horizzon”, “Lost to EINS” and “Will revert after reading the email”]. Also, they should consider the top 3 variables [“Tags”, “Lead Source”, “Last Activity”] to increase the Lead conversion percentage