This analysis is done for X Education and 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. **Understanding the data and checking data quality:**

After importing the libraries and data, first step followed is to read and understand the data using commands like shape and info. The data was also check for duplicate and missing values to understand the type of data cleaning that need to be performed.

2. Data Cleaning and Treatment:

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 available' so as to not lose much data.

3. Exploratory Data Analysis:

A quick EDA was done to check the condition of our data in multiple columns and done following steps.

- Dropped all the columns that have more than 45% missing values.
- In categorical variable analysis found lot of irrelevant variables which are not considered towards final analysis.
- Additionally, there are some variables with very skewed data which are also dropped from final analysis.
- Also clubbed some variables in some columns to make data comparable which helps in accurate analysis.

After numeric values cleaned and no outliers were trimmed, did plot EDA plots between all relevant variables and their conversion ratios which helped in find lot of good suggestions to focus on.

4. Identifying categorical variable columns and creating Dummy Variables:

The dummy variables were created and later dropped the columns for which dummy variables were created. For numeric values we used the MinMaxScaler.

5. **Train-Test split:**

The split was done at 70% and 30% for train and test data respectively. Used sklearn package and import train_test_split methods.

6. Rescaling the numerical variables

7. **Model Building:**

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were dropped manually depending on the VIF values and p-value (The variables with VIF > 5 and p-value > 0.05 were removed).

8. **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 90% each.

9. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.2 with accuracy, sensitivity and specificity of 90%.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. Direct traffic
2. Welingak website
3. Last Activity-Email Bounced
4. Last Activity-Olark Chat Conversation
5. Tags-Busy
6. Tags-Closed by Horizon
7. Tags-Lost to EINS
8. Tags-Not Specified
9. Tags-Ringing
10. Tags-Will revert after reading the mail.
11. Last Notable Activity_SMS Sent

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.