Problem Statement: 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.

To Achieve the model following methods were followed:

1. **Cleaning The Data**:
   The data is cleaned where null values and columns with single values and duplicate values were handled. The columns with more than 30% of null values were dropped.
   Many variables are having dataframes with single value which is leading to imbalance so we will drop the following column
   Do Not Call,Search,Newspaper Article,X Education Forums, Newspaper,Digital Advertising,Through Recommendations
   Some variables were affecting the model accuracy rate due to which these variables were removed from the data set. The variables are namely Asymmetrique Activity, Asymmetrique Activity Score, Asymmetrique Profile Score, Country, Lead Profile, Tags, Lead Quality, 'How did you hear about X Education, City, etc.

2. **EDA**:
   Exploratory Data Analysis was done to check categorical variables .
   We can find that some numerical variables consisted of very high values as compared to their respective means. That's why we have created charts using boxplot to understand the patterns. We have observed that the outliers are very high and we need to treat it. That was the reason we have retained 99% quantile of the data and removed the max value from it.

3. **Dummy Variables:**
   Dummy variables were created to identify the categorical variables to convert
   The variables were not numerical so we needed it to be converted into numeric values to build the logistic regression model.
   Now the data is numeric so we have split the data set into train and test data frames 70 % and 30 % Respectively.. MinMaxScaler is used for further scaling.

4. **Model Building**:
   Now all the data is numeric we can create a logistic regression model.
   As it was difficult to understand the correlation between the variables because of this we have performed Recursive Feature Elimination method to take the top 15 variables,from the train dataset this will help in training the model.

We have iterated through 3 models and evaluated each model based on p- values and VIF. Finally we have ended up with a model with accuracy of 74%.

5. **Model Evaluation:**

   To understand the model and optimise the cut off, we have plotted the ROC curve. In the ROC curve we have seen that the model has generated 87% of AUC which means that the model can be successful and effective.

6. **Prediction**:

   In this, the model gives the sensitivity, specificity and accuracy of the model 76%, 81%, 80% respectively.

7. **Precision – Recall**:
   Here we have rechecked the model and found the cutoff as .4 and the Precision is around 73aand Recall around 74%.

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

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
   E. Referral sites
4. When the last activity was:
   a. SMS Sent
   b. Had a Phone conversation

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional. 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.