# Problem Statement:

X Education is selling online courses and Leads which are generated from various sources are captured. We needed to find the potential leads and convert it to opportunity.

# Solution:

The following steps are involved to find out the potential leads:

1. **Understanding the data**

   To understand the data, we have imported the Leads.csv file to our python notebook and viewed the number of rows, columns, number of missing values, describe the data frame was done.

2. **Clean the data**

   The data was cleaned partially where we have the null values. The column which has unique entries were also removed. After converting the data to small case, we also observed that there were some entries where only Select was mentioned. These data were converted to NA and later was removed while making dummy entries. Mostly all the data was from India, Null values and very least number of data were from outside India. So it was categorized as 'India', 'Outside India' and 'NA'.

3. **Model Building Preparation**

   A quick Explanatory Data Analysis was performed where we have also observed that some of our categorical values were not relevant. There were some outliers present, but those were not removed as it would hampered our analysis.

4. **Model Building**

   - To build the model we first created dummy variables for all the categorical variables using get_dummies method and all the NA variables were removed. For numerical variables we have used MinMaxScaler.
   - Then we have split the data into Train and Test where 70% of the data was for Train and 30% of the data was for Test.
   - Then we have performed RFE to attain top 15 relevant variables. Depending on the VIF and the P values i.e., VIF < 5 and P value < 0.05, we have also dropped some columns manually using drop.

5. **Model Evaluation**

   A confusion matrix was created and with an optimum cut off value using ROC curve we found the overall accuracy, sensitivity and specificity which turns around 80% on an average.

6. **Prediction on Test set**

   A prediction was also performed on the Test data with cut value as 0.35 and found the overall accuracy, sensitivity and specificity which turns around 80% on an average. This

method was also used to recheck with an optimum cut off value of 0.4 and the same turns around 75% on the test data frame.

## 7. Conclusion

Below are the list of variables which impact the most:

a) Total number of visits
b) When their current occupation is working professional
c) The total time spend on the Website
d) When the Lead Source is :

    i. Direct Traffic
    ii. Google
    iii. Organic Search
    iv. Wellingak Website
    v. Olark Chat
    vi. Reference

X Education can keep the above points to get more conversion rate for their courses.