X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The basic data provided gave us a lot of information about the potential customers visiting the website, the time they spend there, how they reached the website and the conversion rate. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The following are the steps used:

1. **Data Cleaning:** We dropped the variables that had more than 30% 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 in the Current occupation & Lead source columns were changed to ‘unemployed’ & ‘Google’ respectively so as to not lose much data. Although they were later removed while making dummies. As 96% of leads are from India we feel that the country variables don't show any impact on the final model so we dropped it.

2. **EDA:** After Data Cleaning we performed EDA(Univariate & Bi-variate analysis) to check the condition of the data. It was found that a lot of elements in the categorical variables were irrelevant so we dropped a few in this step. The numeric values seem good and no outliers were found.

3. **Data Preparation:** The dummy variables were created for categorical variables. For numeric values we used the StandardScaler.

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. **Prediction:** Prediction was done on the test data frame and with an optimum cut off as 0.345 with accuracy, sensitivity and specificity of around 80%.

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

- Lead source -
  - welingak website
  - reference
- Current occupation
  - Working Professional
- Total time spent on website
- Lead origin
  - Landing page submission
- Last Activity
  - Had a Phone conversation
  - 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.