Summary

We have analysed the data give provided by X Education to convert more leads to customers. Step by step process in arriving at the final conclusions are listed below.

1. **Reading the Data**:

We were given two data files.

a. Leads.csv – The actual data file with the leads data, this will be used for all the analysis and modelling purpose.
b. Leads Data Dictionary – This is to get the preliminary understanding of the data provided. Will not be needed for analysis.

2. **Cleaning the Data:**

Preliminary Data cleaning data steps.

a. Removed entire columns with null values more than 40% of the entire data. Columns Removed:
    i. How did you hear about X Education - 78.46 %
    ii. Lead Profile - 74.19 %
    iii. Lead Quality - 51.59 %
    iv. Asymmetrique Profile Score - 45.65 %
    v. Asymmetrique Activity Score - 45.65 %
    vi. Asymmetrique Activity Index - 45.65 %
    vii. Asymmetrique Profile Index - 45.65 %
b. Rest of the columns with null values are imputed with mean, median and mode accordingly. Columns imputed:
    i. City – Imputed with Mumbai since it is the mode.
    ii. Specialization – Replaced with '**others'** since the lead may have not selected any option for this.
    iii. Tags – Imputed with 'Will revert after reading the email' since it is the mode.
    iv. What is your occupation – Imputed with 'Unemployed' since it is the mode.
    v. Country – Imputed with 'India' since it is the mode.
c. Rows are deleted for the lesser null value percentage columns. Columns whose rows are deleted:
    i. TotalVisits – 1.48 %
    ii. Page Views Per Visit – 1.48 %
    iii. Last Activity – 1.11 %
    iv. Lead Source – 0.39 %

d. Column ‘What matters to you most in choosing a course’ is completely imbalanced so we removed the column.

Note: Categorical columns with value as ‘Select’ as value are replaced with **NA** as they might not have been chosen by the lead.

3. **Exploratory Data Analysis:**
    a. Outlier Analysis:
        i. TotalVisits – Limited the data to consider only values upto 95% of the data.
        ii. Page Views Per Visit – Limited the data to consider only values upto 95% of the data.
    b. Data imbalance:
        i. Converted rate – 37.9%
        ii. Un-converted rate – 62.1%

4. **Creating Dummy Variables:**

    Dummy variables are created for the following columns:

    Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation, City, Last Notable Activity.

5. **Split the Data:**
    We split the train and test data into 70 % and 30 % respectively with a random shuffling of 100.

6. **Model Building:**
    a. **Data Scaling:** We scaled the numerical columns using StandardScaler.
    b. **Model 1:** Dropped column ‘What is your current occupation_Housewife’ because p value (**0.999**) of this column is too high.
    c. **Model 2:** Dropped column ‘Last Notable Activity_Had a Phone Conversation’ because p value (**0.247**) of this column is too high.
    d. **Model 3:** Dropped column ‘What is your current occupation_Student’ because p value (**0.098**) of this column is too high.
    e. **Model 4:** Dropped column ‘Lead Origin_Lead Add Form’ because p value (**0.081**) is too high.
    f. **Model 5:** Dropped column ‘What is your current occupation_Unemployed’ because VIF (**9.72**) is too high for this column.

g. **Model 6:** Dropped column 'Lead Origin_Lead Import' because p value (**0.072**) is too high.
h. **Model 7:** Dropped column 'Last Activity_Unsubscribed' because VIF is too high for this column.
i. **Model 8:** Dropped column 'Last Notable Activity_Unreachable' because VIF is too high for this column.
j. **Final Model:** All the remaining variables' p values are ~0 and VIF is also low indicating low multi-colinearity. Final model has 12 independent variables.

7. **Model Evaluation Metrics:**
   a. **Accuracy on train set: 81 %**
   b. **Sensitivity on train set: 81.7 %**
   c. **Specificity on train set: 80.6 %**

8. **ROC Curve:**
   a. **Area under curve: 0.89**
   b. **Optimal cut off: 0.35**

9. **Prediction on test set:**
   a. **Accuracy on test set: 80.4 %**
   b. **Sensitivity on test set: 80.4 %**
   c. **Specificity on test set: 80.5 %**

Conclusion:

We were able to stabilize the model with a cut off of **0.35**. We considered a lead score of 85 and above as hot leads and the rest of them as cold leads.

Most Important features that converts the lead into customers:

1. When the lead source is:
   a. Welingak Website
   b. Reference
   c. Olark Chat
2. Lead origin is:
   a. Landing Page Submission
3. Current occupation is:
   a. Working Professional
4. Last Activity is:

    a. Other_Activity
    b. SMS Sent
    c. Olark Chat Conversation
5. Last Notable Activity is:
    a. Modified
6. Total Time Spent on Website.
7. The customers don't have a specialization already.

X Education should consider the above variables carefully while trying to approach a lead to have higher conversion rate.