Summary

Current state -

- X Education currently has several ways to get leads but the conversion rate is only around 38% from the provided data.

Goal -

- X Education wants to increase their sale of courses by increasing lead conversion rate around 80%.

Steps followed to achieve the goal -

1. Data Collection

Data was already provided. Used Pandas Library for going through the data

2. Data Cleaning

Some part of the data was in good shape. But there were a lot of null values in many of the columns. Some of these data were not provided - found a lot of columns with value as Select - this indicated that values were not provided while filling the form.
Some of the attributes which had more than 40% missing values were deleted from the dataset. Rest of the data were imputed based on analysis results.

3. EDA

Basic EDA was done to understand the current lead conversion and find out inputs for the next step. It was found that most of the columns were categorical. Also, some of the columns had outliers, as part of data cleaning, they were imputed.

4. Data Preparation

a. **Dummy Variables creation -**
As there were many categorical columns with multiple categories, they were converted to dummy variables for better model building.

b. **Scaling -**
After the dummy variables were created, it was important to scale the variables as there were numerical values with different ranges. Used StandardScaler from sklearn module for scaling those variables.

c. **Feature selection -**
After this based on the correlation matrix, eliminated some of the highly correlated features/variables from the dataset for better model building. Also, RFE technique was applied from better feature selection

d. **Test-Train split -**
Post the feature selection the data was split into train and test with 70% and 30% ratio.

5. Model Building

There were a total 11 models built and during this process few more features were removed based on P-Value and VIF analysis, where VIF is less than 5 and P-Value is less than 0.05.

6. Model Evaluation

Created confusion matrix and ROC curves to evaluate the models performance. Calculated Accuracy, Sensitivity, Specificity which were around 80% and that satisfied the requirement.

7. Prediction

Using the last model, when predictions were made all the confusion matrix values i.e. Accuracy, Sensitivity, Specificity came around 80% - which is considered to be a good model prediction.

8. Final observations

a. Leads who have origin of `**Add Form**` and `**Lead Import**` should be contacted
b. Leads whose sources are `**Olark Chat**` and `**Google**` should be contacted
c. Leads who are currently `working professionals` they must be contacted
d. Leads who have done any activities such as `**SMS Sent**`, `**Olark Chat**`, **Modified** should be contacted
e. Leads who spend more time on the website should be contacted
f. Overall when the leads have engagement with the platform or the sales representative, they should be contacted.
g. Leads who are students should not be contacted as their conversion rate is very low
h. Specialisation mentioned as Others should be be contacted as they are highly likely to not get converted