Summary

Background:

X Education, an education company sells online courses to industry professionals. They have a website to browse courses. People can either land here directly or through other websites and fill up form.

Though there are lot of leads coming in and filling up form, the lead conversion rate is very low.

Solution:

As this is binary classification problem, one of the possible solutions could be to create Logistic Regression model that outputs lead conversion rate with good metrics.

The metrics in focus could be accuracy and specificity/Recall as we need to be mindful about the possible leads to be converted to hot leads.

Technical Approach:

1. **Importing data and Understanding Data**:
    a. Importing Lead.csv and understanding the shape, null values etc.

2. **Data Preparation/Cleaning**
    a. Verified that there are no duplicate records in the dataset.
    b. Replace 'Select' with nan.
    c. Drop columns with missing or null values >40%
    d. Handle nan /null values by replacing with
        i. Mean, Median or Mode appropriately,
        ii. most occurring value if the number of missing values is less in that column
        iii. "Not available" if the number of missing values is more in that column
    e. Grouped least occurring categorical values to "Others" category. Grouped related values for example 'Finance Management', 'Human Resource Management' into Management specialization.
    f. Dropped columns that are very inclined towards single column like country or highly skewed.

3. **Exploratory Data Analysis: Univariate and Bivariate analysis**
    a. Visualized Numeric Variables using boxplot with respect to "converted" field.
    b. Visualized Categorical Variables using countplot with respect to "converted" field.
    c. Handling outliers for numeric variables.

4. **Data modification post EDA**:

a. Converted yes/no fields to 1/0.
b. Dummification of Categorical Variables

5. **Splitting the Data into Training and Testing Sets**

a. Splitting the Data into Training and Testing Sets: 70% for training and 30% for testing.
b. Feature Scaling/Scaling variables: used standardscalar for continuous variables.

6. **Building a Logistic model:**

a. Identify Correlation using VIF and heat maps
b. Build model:
   i. Top ‘n’ features: Recursive Feature elimination (RFE): taken top 20 features
   ii. Manual recursive elimination process based on pvalue (<0.05), VIF (<3)

7. **Model Evaluation:**

a. Plotting ROC Curve: Area under curve is 0.89.
b. Finding Optimal Cutoff Point for different probabilities using Specificity, accuracy, sensitivity: Optimal cut off point is 0.35.
c. Identify metrics based on the cutoff:
   i. Confusion matrix

	Predicted Negatives	Predicted Positives
Actual Negatives	3260	742
Actual Positives	481	1985

   ii. Specificity: 81.5%
   iii. Accuracy: 81.1%
   iv. Sensitivity: 80.5%
   v. Precision: 72.8%

d. Finding Optimal Cutoff Point for different probabilities using precision, recall curve: Optimal cut off point is 0.4

8. **Making predictions on the test set using final model**:

a. Make predictions and calculate below metrics
   i. Confusion matrix

	Predicted Negatives	Predicted Positives
Actual Negatives	1380	297
Actual Positives	213	882

   ii. Specificity: 82.3%

iii. Accuracy: 81.6%

iv. Sensitivity: 80.5%

v. Precision: 74.81%

b. Validate if the test and the train set metrics align

9. **Calculate lead score of both the train and the test data set for all the leads**.

Learnings:

1. There could be values such as “Select” in the columns. Such values occur when the column is not mandatory on the user interface but the default selection in the dropdown is “Select” which the user leaves as is.
2. Grouping column values in case there is not much data. For ex: 'Businessman', 'Housewife' occupations are replaced with ‘Other’.
3. Grouping column values to a parent category. For ex: 'Finance Management', 'Human Resource Management' specializations are replaced with 'Management Specialization'.