Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Task - Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Solution Methodology:

I divided my assignment in the following steps: -

1. Data Inspection and Cleaning
2. Outlier Treatment
3. Exploratory Data Analysis
4. Data Preparation
5. Modeling

Data Inspection and Cleaning:

- Data Inspection is done for getting an overall view of the data and checking for any abnormalities in the data before feeding it to the model
- One the data is inspected, the necessary changes to the data needs to be made which leads to Data Cleaning

- Data Cleaning methods:
  - Prospect ID & Lead Number are two variables that are just indicative of the ID number of the contacted people & can be dropped
  - Checked and removed duplicates
  - The response from the Google form recorded as “Select” was replaced by Null values for the below mentioned columns:
    - Specialization
    - How did you hear about X education
    - Lead Profile
    - City
  - Dropped the columns that have Null values of >45%
  - Missing value handling:
    - Dropped missing values in numeric variables
    - Replaced missing values of categorical variables as a different category
    - Impute missing values of categorical variables as Mode

Outlier Treatment:

- Having visits greater than 100 is a very extreme case and very rare. Since the values are too extreme and there are only 3 records, we can drop these records
- We can also cap the values to 30

Exploratory Data Analysis:

- Standardization of values
  - Reduced number of categories for some features

- Data Imbalance
  - We will also check which columns can be dropped due to the imbalanced nature of the columns
    - Combined the variables with less than 1% values as Others
    - Do not email, Do not call - Columns are imbalanced. Though these feature might be very important but the resulting model will always favor the higher category. Since we can't sample data again, and other techniques might be an overkill, we will drop this column
    - Dropped many columns that has imbalanced values

After removing all the unnecessary columns, the final columns to be kept for further analysis are:

```
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9074 non-null   object
 1   Lead Source                             9074 non-null   object
 2   Converted                               9074 non-null   int64
 3   TotalVisits                             9074 non-null   float64
 4   Total Time Spent on Website             9074 non-null   int64
 5   Page Views Per Visit                    9074 non-null   float64
 6   Last Activity                           9074 non-null   object
 7   Specialization                          9074 non-null   object
 8   Current Occupation                      9074 non-null   object
 9   Tags                                    9074 non-null   object
 10  City                                    9074 non-null   object
 11  A free copy of Mastering The Interview  9074 non-null   object
 12  Last Notable Activity                   9074 non-null   object
```

- Univariate Analysis - Plotted histogram for all the variables and noted down observations
- Bivariate Analysis - Plotted pair plot between all variables and identified few relationships amongst them
- Multivariate Analysis - Plotted a correlation heatmap and observed some significant correlations between some variables

Data Preparation:

- Converting some binary variables (Yes/No) to 1/0
- Creating Dummy variables
- Train-Test split (70-30)
- Feature Scaling

Modeling:

- The total number of columns in the data is 64, I reduced it to 15 using RFE (Recursive Feature Elimination)
- I built the model using these 15 variables
- Out of these 15, 2 of them had high p-value (>0.05), so they were dropped and now we have 13 variables

- The final model was built with 13 variables and the predicted values were found on the Train dataset
- Then I took a random threshold value of 0.5 and calculated accuracy
  - Accuracy - 92.4%
- Calculation of VIF - VIF value gives the dependency of one variable on the other. As all the features had low VIF values, no features will be dropped; they do not depend highly on each other
- **Sensitivity - Specificity Framework -** ROC curve was drawn for optimal cut-off for Sensitivity and Specificity which turned out to be 0.3
- In this use case, the Recall should be at least 80%, we first find the optimum cut-off for Precision-Recall Framework and then the Precision and Recall are calculated
- **We prepared the final dataset by we appending the Lead Score in the range 0-100**