Problem Statement:

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 Approach:

Step wise solution for performing Logistic regression =>

1. **Data cleaning and preparation (EDA):**
   a. **Missing value treatment:**
      i. Identify number of missing values in all columns.
      ii. Also replace 'Select' value with null as this value is as good as missing value.
      iii. Drop columns which has null values more than 70%
      iv. Explore those columns to remove rows which has low percentage null values.

   b. **Impute Null values:**
      i. Now for remaining columns we'll impute null values as 'Others'

   c. **Create Dummy Variable (One-Hot Encoding):**
      i. Identify those columns which are having only one value majorly present for all data points and drop them.

      Column examples: `Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement.`

      ii. For Binary variables, we convert them to 0/1 from Yes/no
      Column examples: `'Do Not Email', 'A free copy of Mastering The Interview'`

      iii. For multivariable we create dummy variables and drop original variable.
      Column examples: `'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'Last Notable Activity'`

2. **Outlier check and capping:**
   a. We will box plot to check Outliers in numerical dataset points.
   b. We will cap those outliers as we do not want to lose any data.
   c. We will use capping 1% and 99% and again plot a boxplot.

3. **Train Test Split:**
   a. Split the data into train and test set as 70% and 30%.

4. **Feature Scaling:** Now there are a few numeric variables present in the dataset has different scales. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the

regression model might be very large or very small as compared to the other coefficients. So Standard Scaling is used to scale ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'] these columns

5. **Feature Selection using RFE:**
    a. We will perform RFE for feature selection (We will take top 15 features for our model)
    b. After that we will apply logistic regression model using these 15 features.
    c. Then we will perform VIF and p value for checking those features.
    d. Remove those high p values and then HIGH VIF values one by one for each iteration then perform modelling it until we get all features with low p and low VIF.
6. **Model Building (on Train set):** Logistic regression using RFE features.
7. **Model Evaluation and it is metrics:**
    a. Train model will used for predicting the values of output variables in set.
    b. We will plot RoC Curve and Precision recall trade off
    c. We will find optimal cut-off point (from curve we see cutoff as 0.45)
    d. Now we choose 0.45 as cut off to convert predicated values 0 and 1
    e. Check Accuracy, sensitivity, specificity, precision, and confusion matrix.
    f. Precision should be achieved as 80% and our model achieved 92.14%
8. **Prediction on Test Data:** Apply model on test data and check the confusion matrices.
9. **Observation:**
    a. We have good model with precision around 92.14% on Train Data set and 91% on test data set.
    b. Column finally_predicted, Converted_prob and Converted with corresponding cut off used as 0.45. probability score less than 0.45 is cold lead meaning low chance of getting converted. Greater than is hot lead meaning higher chances of getting converted.