Problem Statement:

An education company named X Education which sells online courses to industry professionals, wants to come with an efficient process to identify the most potential leads, also known as 'Hot Leads'.

The company tracks various parameters when individuals visit their website and wishes to use this data in identifying patterns.

Steps Taken:

1. **Data Understanding**

   We followed the Data dictionary and looked at the data set to get an initial understanding of the data.

2. **Data Cleaning:**

   We took the following steps to clean the data.

   - We dropped the columns with more than 30% missing values.
   - We observed that the remaining missing values were in categorical variables. To handle those, we imputed missing values with mode of the respective column.
   - Columns having same value for each row, and those having unique value for each row were dropped.
   - For numerical features, some of the feature had outliers. To handle this, we capped outliers to Q3 + 1.5 * IQR value.

3. **Exploratory Data Analysis**

   We performed Univariate as well as bivariate analysis.

   - For Univariate analysis,
     - Bar plots were plotted for Categorical variables
     - Box plots were plotted for Numerical variables

     Inferences:
     - We observed that Google, Direct Traffic, Olark Chat dominate in "Lead Source"
     - We observed that Landing Page Submission, API, Lead Add Form dominate in "Lead Origin"
     - Unemployed and Working Professional dominated among the profession of Visitors.
   - Bivariate analysis
     - Bar plots were plotted for Categorical variables against target variable

       Inferences:
       - Last Notable Activity / Last Activity = SMS Sent, has higher conversion rate
       - Lead Source = 'Reference' and 'Welingak Wesite' has higher conversion rate
       - Lead Origin = 'Lead Add Form' has higher conversion rate
       - Occupation = 'Working Professional' has higher conversion rate
     - Numerical variables were plotted against each other

Inference:

- TotalVisits had linear relationship with Page Views per Visit, Same was evident from correlation matrix as well.

4. **Data Preparation and Splitting into Train and Test set**

- Certain columns had value in Yes/No, we mapped them to 1/0.
- For categorical variables having more than 2 distinct values, we created dummy variables.
- We scaled the numerical features in train data set, using MinMax Scaling technique, to bring their values in 0 to 1 range.
- We split the data into training and test data set with **70:30** ratio

5. **Building Models**

We followed the hybrid of automatic and manual approach

- We selected top 15 features using RFE technique, then recursively dropped the features one-by-one based on their p-value and VIF

6. **Model evaluation on Train and Test data set**

- To evaluate our model, we plotted an ROC curve**. The area under the curve, AUC, came out to be 0.88.**
- To find the optimal threshold, we plot line chart of Accuracy, Sensitivity and Specificity for different thresholds. **The optimal threshold came out to be 0.35**

Other metrics were as follows:

	Overall model Accuracy	**Sensitivity**	**Specificity**	**Positive Rate**	**Negative Rate**
Train data	80.19%	0.8	0.8	0.72	0.87
Test data	79.82%	0.8	0.8	0.71	0.87

Conclusion:

- The resulting model can predict at 80% accuracy. This was observed in both train and test data set. Values of other metrics are also very much close to each other for training and test data set.
- Top 3 contributing Factors came out to be:
  - 'Total Time Spent on Website'
  - 'Lead Origin_Lead Add Form'
  - Occupation_Working Professional