1. **Business Understanding**
    a. The CEO wants the lead conversion rate to be over 80%.
    b. The requirement is to correctly identify True Positives, meaning the Recall should be over 80%.
2. **Read and Understand the Data** for missing values, data types of columns, and the spread of numerical variables.
3. **Fix Data Quality Issues**
    a. Columns with only one unique value were dropped.
    b. Columns with the value 'Select' were treated as missing values.
4. **Checking and Handling for Missing Values**
    a. Columns > 40% missing values were removed.
    b. Null values were replaced by 'not provided' in the categorical columns with high percentage of missing values.
    c. Null values were dropped for categorical columns with less than 2% missing values.
    d. Columns with values with very low percentages were clubbed together under a Single Category.
5. **Visualizing the Data**
    a. The data has a 62:38 ratio which seems to be balanced.
    b. Dataset split based on Conversion for Univariate (UA) and Bivariate Analysis (BA).
    c. *UA - Numerical Variables* using *distplots*. Noticed that time spent on Website sharply increased for Converted Leads.
    d. *UA - Categorical Variables* using *countplots*. Noticed *Working professionals and leads originating from add forms* showed a high conversion rate.
    e. *BA - Numerical Variables* using *scatterplots.* No discernible relationships observed.
    f. *BA - Categorical Variables* using *countplots.* Noticed that working professionals and leads sourced through referrals have a high conversion rate.
    g. *Multivariate Analysis* using *heatmaps*. We didn't observe any strong correlations here.
6. **Outlier Detection and Treatment**
    a. Outlier Detection was done using *boxplots*.
    b. The statistical summary of the columns with outliers were studied and outliers were replaced with median values.
7. **Prepare the Data for Modelling**
    a. We will drop the columns generated by Sales Team.
    b. Dummy Variables created for Categorical columns.
    c. Dataset split into training and test sets in a 70:30 ratio.
    d. Standardization of numerical variables.
8. **Modelling**
    a. Feature Selection using mix of RFE and manual method.
    b. Multiple iterations of the model till our final list of features were both significant (p-value < 0.05) and had very low multicollinearity (VIF < 5).
9. **Evaluation**
    a. Prediction was made on the training set and ROC curve was created. Area under the ROC curve was 0.87 which is a good value.
    b. Calculated the accuracy, sensitivity, and specificity for various probability cut-offs and arrived at using 0.3 as the optimum value.
    c. Assigned each lead to a converted value and found the sensitivity - 83%. And the specificity - 73%. This is a good model.
10. **Prediction and Lead Score**

a. We made predictions on the test set and calculated the Lead Score.
b. The Sensitivity - 82.9% and Specificity - 73.41%. Thus, the model is performing well.

11. Insights

a. The company should focus on working professionals, leads generated from add forms, and on people who spend more time on website.
b. People who have requested to not be contacted via email should not be pursued.