Summary Report:

Step1: Reading, Understanding Data and inspecting the data.

Step2: Data Cleaning:

a) Replaced the "select" values with nan

b) Dropped the variables with 40 % missing values

c) Dropped the following categorical variables:

City: as this column has 39.71 % missing values and imputing them with mode value i.e "Mumbai", will make the data skewed.

What matters most to you in choosing a course: This variable has 29.32 % missing values. 99.95% customers have selected 'better career prospects' which is massively skewed and will not provide any insight

Country: This column has 26.63 % missing values and 95.77 % of the customers are from India. Since X Education sells online courses, imputing this with India does not make business sense.

d) Imputed the missing values of following columns:

1. Specialization with "Others"
2. Tags with "Others"
3. Lead Source with "Google"
4. What is your current occupation with "unemployed"
5. Last Activity with "Email Opened"

e) Handling Invalid values & Standardising Data

- Google and google were standardised.

Step 3: Data Transformation: Changed the binary variables into '0' and '1'

Step 4: Categorical Analysis (Univariate/bivariate)
Step 5: Numerical Analysis

Step 4: Dummy Variables Created for

“Lead Origin", "Lead Source", "Last Activity", "Specialization", "What is your current occupation", "Tags", "Last Notable Activity

Step 5: Test Train Split: The data was divided into train and test set with a proportion of 70- 30% values.

Step 6: Feature Rescaling

Step 7: Model Building:

- Using the Recursive Feature Elimination, we selected the 20 top important features.

-Using the statistics generated, we recursively tried looking at the P-values along with VIF in order to select the most significant values that should be present and dropped the insignificant values.

-Finally, we arrived at the 14 most significant variables.

- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.

-We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 97% which further solidified the of the model.

- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.

- Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.29.

-Then we implemented the learnings to the test data and calculated conversion probability based on the Sensitivity and Specificity metrics

Step 8: Conclusions:

Train Data Set:

- Accuracy: 92.32%

- Sensitivity: 92.11%

- Specificity: 92.44%

Test Data Set:

- Accuracy92.42%

- Sensitivity:92.48%

- Specificity: 92.39%

The evaluation matrics are pretty close to each other so it indicates that model is performing consistently across different evaluation metrics in both test and train dataset.

- The model achieved a sensitivity of 92.11% in the train set and 92.48% in the test set, using a cut-off value of 0.29.

- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting