Leads Scoring Case Study Summary Report

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

For the given problem statement, we have followed the following steps elaborated below:

- **Data Cleaning and Manipulation:**

  - Removing redundant columns.
  - Imputing the Select option in few of the variables with null values.
  - Removed columns having more than 30% null values.
  - For the remaining columns, we imputed values with maximum number of occurrences for each column.
  - One column had two identical label names in different format (Upper and Lower Case). We fixed it by changing labels names to one format.

- **Data Transformation:**

  - Assigned numerical values 1 and 0 to categories, Yes and No.
  - Dropped unique valued columns.
  - Converted some variables to numerical as they were imported as object.

- **EDA:**

  - Generalized the Country column.
  - Performed Univariate Analysis.
  - Performed Bivariate Analysis.
  - Performed Outlier Analysis and handled them by creating bins for those variables.

- **Data Preparation:**

  - Created Dummy variables for multiple levels of categories.
  - Removed the original columns as they would be redundant.
  - Created the Train-Test Split.
  - Feature Standardization.

❖ **Model Building:**

- Ran RFE for 15 variables.
- Created multiple models consequently dropping off insignificant variables.
- Finalized a model on the basis of variables having low p-values, i.e. being most significant.

❖ **Model Validation:**

- Performed predictions on train set.
- Plotted the ROC curve to find out that the model has a good accuracy with area under curve equal to 88%.
- Found the optimal probability cut-off point to be 0.4
- Ran predictions with a cut-off of 0.4
- Calculated precision and recall to be approximately 73% and 79% respectively.

❖ **Model Testing:**

- Performed predictions on the test set.
- Evaluated the model finally, and calculated the accuracy, precision and recall to be approximately 82%, 76% and 80% respectively.
- Assigned the lead scores.

❖ **Conclusion and Recommendations:**

- The Accuracy, Precision and Recall score we got from test set are in an acceptable range.
- We have higher recall score than the precision score which is what we were looking for.
- The model is in a stable state which means that this model has an ability to adjust with the company's requirements in the coming future.
- Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are:
    - **Lead Origin_Lead Add Form**
    - **What is your current occupation_Working Professional**
    - **Last Notable Activity_Had a Phone Conversation**