Lead Scoring Summary

After a rough inspection of the file ‘leads.csv’, we checked the data imbalance in the target variable-‘Converted’. It was in the ratio 62:38 which is normal and similar to what was given in the problem statement.

Next we looked at all the columns and their data types. There were 9240 columns in total, 29 Categorical and 5 Continuous variables. We excluded ‘Converted’, ‘Prospect ID’ and ‘Lead Number’ from the analysis.

Exploratory Data Analysis

1. Many columns like ‘Do Not Email’, ‘Do Not Call’, ‘Magazine’ etc had extreme data imbalance or one unique value. So we dropped these columns
2. Most of the leads were from India. We segmented this column’s values as ‘India’, “Outside India’ and ‘NA’
3. Select’ is a recurring value in many columns and per the problem statement it is to be considered as blank. The reason is that ‘Select’ is the default option in forms. So we replaced ‘Select’ with np.nan
4. Most of the leads had the origin ‘Landing Page Submission’ and ‘API’
5. Most of the leads source was Google followed by Direct Traffic and Olark Chat
6. The last activity of most leads was ‘Email opened’ followed by ‘SMS Sent’
7. Leads originated from ‘Lead Add Form’ have a higher conversion rate
8. Leads having Source ‘Reference’ and ‘Welingak Website’ have a higher conversion rate
9. High conversion is seen in leads whose last activity was ‘SMS Sent’
10. Working professionals show high chance of conversion
11. Leads tagged as ‘Will revert after reading the email’ and ‘Closed by Horizon’ show higher conversion.
12. From the Continuous variables, we inferred that the converted leads spent on average more time on the website than the others

Missing Values

- We deleted columns that have >30% missing values,
- deleted the rows of columns that have <1.5% missing values
- imputed missing values with ‘NA’ for columns that had missing value between 1.5% and 30%

Lead Scoring Summary

Data Preparation

- Dummy Variables
    a. Binary variables (Yes/No) were converted to 1/0
    b. Categorical variables were converted to ‘n-1’ dummy variables (n is the unique values of each variable)
- Outliers
    a. Outliers in the numeircal column were identified and replaced with 99%-ile value
- Train-Test split was done in the ratio 7:3
- Lastly, we scaled the numerical variables using standard scaler

Model Building

1. First we did RFE and filtered the top 15 variables.
2. Then we removed the variables with p-value>0.05 and VIF<5
3. The Country_NA does not make business sense, so we dropped this variable

The p-values<0.05 and VIF<5, we finalized this model

Prediction and Model Evaluation

The Area under ROC curve was 0.88 which is good and using the Accuracy, Sensitivity and Specificity curve, we found the optimal cut-off to be 0.35. While, using the precision-recall curve, the optimal cut-off was 0.45. But to keep a higher recall value, we took cut-off=0.35.
Using this cut-off, we predicted on the train and test data.
The metrics were as follows -

Metric	**Train data**	**Test Data**
Accuracy	80.35%	79.55%
Recall	81.77%	81.24%

The drop in accuracy is only 1%, and the precision value is >70% for both. Hence we can confirm this model.

Based on the conversion probability, we multiply the probability values by 100 to get the lead scores