Leading Score Summary.

Approach.

From the problem description, the problem is a Classification problem, hence we went with Logistic Regression to calculate the Lead Rate.

Below are the steps followed to solve the problem

1. *Loading the Data and Understanding it.*

Here we loaded the data and inspected it for to understand the data. We learned:

- The dimensions of the data [Number of Rows and Columns].
- Found out whether Duplicates are present in the data.
- Understood the data types of each of the columns.

2. *Data Cleaning and Data Processing.*

The data was examined here for any discrepancies.

- Treated the Null Values present in the data.
- Treated categorical values in a column which had very low frequency by combining them together to form a new categorical value.
- Dropped columns which were dominant be one of the two categorical value (example: Yes and No), as it would not be helpful in our analysis.
- Treated the Outliers present in the numerical values. We used the IQR for this process.
- We renamed few of the columns which were too long to a short and understandable names.

3. *Exploratory Data Analysis.*

In this step, we performed Data visualization.

- We performed Univariate analysis on the Categorical Variables and Numerical Variables, to understand the distribution of the Variables.
- We performed Bivariae analysis on Numerical variables with Converted to understand how the leads are related to these column.
- We plotted a Heat-Map to understand the correlation between the variables.

4. *Data Preparation.*

At this stage, our data was clean and the outliers were also dealt with. Since Logistic Regression needs input parameters as Numerical values, so we done what was necessary.

- Columns with more than two levels were converted to numerical values by creating the Dummy variables (with dropping the first variable).
- The whole data was then Split into two Data-frames, Train and Test Data, which would then be used to build regression model.
- We used StandardScaler to standardize the numerical values.

5. *Model Building.*

- In this step, we first used RFE (Recursive Feature Elimination) to obtain top 20 relevant variables.
- The next step was to make the model more stable, by checking the p-values and VIF (Variance Inflation Factor). Variables were dropped if there p-values were greater than 0.05 and VIF greater than 5.
- Once the model was stable, we predicted probabilities on the Train data and created new Data-Frame, with prediction 1 if 'Converted_Prob' was greater than 0.5 else 0.
- We then calculated the Confusion matrix on this new Data-Frame and calculated the Accuracy (90.38%), Sensitivity (81.6%), Specificity (95.9%), also plotted the ROC curve to find the area under the curve, which came out to be 0.2.
- Found the optimum Cut-Off by comparing the Accuracy, Sensitivity and Specificity for probabilities [0.0 to 0.9], which came out to be 0.3.
- Plotted a graph with Accuracy, Sensitivity and Specificity in the Y-axis and Probability in the X-axis, to find out the intersection point of Accuracy, Sensitivity and Specificity to find the optimum Cut-Off, that came to be at 0.3.
- Also plotted a graph with Precision and Recall in Y-axis and Probability on the X-axis, this also came out to be intersecting at 0.3.
- A new column was added ‘Final_Predicted’, and the values were recorded.

6. *Model Evaluation.*

- In this step, predictions were made on the Test data and the values were recorded.
- We checked the Confusion matrix and calculated Accuracy (91.15%), Sensitivity (86.8%), Specificity (93.74%), Precision (89.17%) and Recall (86.8%) of the final predicted model.
- We created a Lead_Score ['Converted_Prob' * 100] to provide a score between 0 to 100 where higher the value of Lead_Score means the Lead is ‘Hot’ and there is a high possibility that the lead can be converted.

7. *Conclusion.*

It was found that the variables that mattered the most in the potential buyers are (In descending order):

- When 'Tag' was:
    a) Closed by Horizon.
    b) Lost to EINS.
    c) Will revert after reading the email.
- When 'Lead Source' was:
    a) Welingak Website.
    b) Reference.
- When Occupation was:
    a) Working Professional.
- When Specialization was:
    a) Travel and Tourism.