Summary Report

I started this case study problem solving by importing required libraries of Python into python notebook. Then I applied these steps-

1. Data Preparation-

Data Loading- Importing Leads.csv file data and display it.
Checking Duplicates- Checked for duplicate values but no duplicates found in result.

2. Data Inspection-

Data.shape- checked shape of data in row by column that is 9240 rows and 37 columns

Data.Info- Info about data types of columns
Data.Describe- Data count,mean, std etc.

3. Data Cleaning-

(a) **Select Value converted to NaN**
(b) **Null Value Count & Percent checking**
(c) **Lead Quality Column Imputation- Not Sure to NaN**
(d) **index and score columns having 45% Null values**
(e) **City column Imputation**- 60% Mumbai data imputed to NaN.
(f) **Tags column** - "Will revert after reading the email" imputed to NaN.
(g) **'What matters most to you in choosing a course'** imputed to NaN.
(h) **'What is your current occupation'** imputed to NaN.
(i) Country is India for around 96% of data so India imputed in missing values.
(j) Rest missing values are under 2% so we dropped these rows.

4. Exploratory Data Analytics specificity

Univariate Analysis

(a) We got rate of conversion as 37.85%.
(b) Most of the leads are converted from Lead originated by **API** and **Landing Page submission**. Very less amount of leads are originated from **lead add form**.
(c) **Direct Traffic** and **Google** are the major lead sources. To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
(d) Capped the outliers to 95% value for **Page Views Per Visit**.
(e) The rate of conversion of the customers whose last activity is Email_opened or SMS_Sent is higher.
(f) Number of retained rows 98%

5. Data Preparation-

- 'Do Not Email', 'Do Not Call' converted to Binary variables 'Yes' & 'No'.
- Dummy variable creation for 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization' ,'What is your current occupation', 'Tags', 'Lead Quality', 'City' and 'Last Notable Activity'.
- Dropping Original columns from dataframe of dummy variables.

6. Performing Train-Test Split On Data-

Splitting the data into Train- Test in 70-30 ratio.

7. Feature Scaling

Standard Scaling applied. Rate of conversion is 38%.

8. Model Building-

Logistic regression model used.

9. Feature Selection Using RFE

running RFE with 15 variables as output. Assessing the model with StatsModels. Confusion matrix created.

10. Checking VIFs

11. Metrics beyond simply accuracy

12. Plotting the ROC Curve

13. Finding Optimal Cutoff Point

0.2 is the optimum point to take it as a cutoff probability.

14. Assigning Lead Score

Sensitivity 85% , Specificity 94%,

15. Precision and Recall

confusion matrix again

[3756,149],

[363, 2083]

Precision 93%

Recall 85%

Using sklearn utilities for the same.

```
precision_score 93%

Recall_score  85%
```

16. Precision and recall tradeoff

17. Making predictions on the test set

18. Checking the overall accuracy that is 90%

sensitivity of our logistic regression model 84%

specificity is 94%