Steps:

1)Got started with adding various libraries like Matplotlib, Seaborn, NumPy, Pandas, Stats model and Sklearn. Further after adding the data made the checks regarding whether there are any duplicates or not, what’s the shape of the data? whether there are any null values in the variables or not and statistical view of the data

2) Further to check the null values among the columns started plotting graphs for each and every column and replaced null values of important columns with a variable weighing majorly among the column. Also, in some of the columns like Tags, Last Activity, there were a greater number of variables than needed so were replaced

3) Some of the columns like 'Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Index','Asymmetrique Profile Score’ were not important for any kind of analysis so dropped them .

4) Now performed Exploratory Data Analysis on the data where:

I. In Lead Origin API and Landing Page are having more than 30 % conversion rate and Lead Add form is having more than 90% conversion rate

II. For Lead Source Google and Direct Traffic generates maximum number of leads

III. For Calls and Email, most of the individuals asked not to mail and call them

IV. In Total number of visits, majority of individuals have visited nearly 50 times, converting it into percentile and comparing it with converted data its found that both are having same median at nearly 3.25 in the data

V. In Last Acitivity majority of lead conversion was after SmS sent

VI. Majority of the individuals were Unemployed

VII. Majorly the individuals were from India

VIII. Most of the individuals chose the program for better career prospects

IX. Most number of Individuals were from Mumbai

5) Further we did data preparation and started by converted ‘Do not Email’ and ‘Do not Call’ into 0 and 1 and created dummy varibales for 'Lead Origin', 'Lead

Source', 'Last Activity', 'Specialization','What is your current occupation', 'Tags','Lead Quality','City','Last Notable Activity' columns

6) Then made a split of data into train-test keeping train size= 0.7 and test size = 0.3 and then using standard scaler scaler fit is applied.

7) In Model building using Generalized Linear Model Regression stated the p-value,coefficients and standard variables for all the variables

8) Further done feature selection using Recurssive Feature Elimination where the best 15 variables got churned out, there using statsmodels the columns with high p-value like Tags_invalid number, Tags_wrong number given got removed and VIF for all the columns was checked and found it under 7.50 which looked good

9) Now Sensitivity, Specificity, False Positive Rate and Positive value was at 0.85, 0.96,0.38 and 0.91 then found the optimal cuttoff points for Sensitivity, Accuracy, Sensitivity and Specificity, then precision and recall were found at 0.93 and 0.85.

10) In the last step applied it on test set and Sensitivity and Specificity came at 0.86 and 0.93 which were similar to train set