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

Approach

1. Read the data dictionary
2. Load data using csv file
3. Change the column names to more workable formats (without spacing etc.)
4. Check the given data and look for any missing/null values –
    1. There were 9240 observations in dataset
    2. There were many features with more than 70% of missing values
5. There were many observations with default value 'Select'. Replace this value with nan
6. Drop the features with more than 70% of missing values
7. Drop the features with very less variance in data (e.g. Country, What matters most etc.)
8. Replace the missing values in Occupation with category 'Other' instead of most frequent value
9. Combine few categories (with very less observed values) in Lead Source, Last Activity features into a new category 'Others'
10. Remove the categorical features with NO variation in data(i.e. with only 1 category) – Magazine, ReceiveCourseUpdate, UpdateDMContent etc.
11. For continuous variables check the presence of outliers. If yes then replace the missing values with median. Further check the percentile values for presence of outliers and drop observations in 99_{th} percentile – PagesPerVisit, TotalVisits etc.
12. Do exploratory data analysis to draw some inferences – e.g.
    1. 'Lead Add Form' comparatively has the highest lead conversion rate
    2. 'Welingak Website' comparatively has the highest successful lead conversion rate
    3. Focus should be more on 'Facebook', 'Referral Sites', 'Olark Chat' to improve lead conversion rate etc.
13. Convert categorical columns into dummy variables
14. Do Train-Test split with 70% train and 30% test data on original data set
15. Scale the continuous variables using StandardScalar
16. Create the first logistic regression using stats model by taking in all the features at once and look for log likelihood value
    1. The value is nan – probably due to unimportant features or features with high VIF
17. Use RFE and select top 15 features
18. Create the logistic regression model again using stats model with these 15 features
19. Check for p-value and VIF
    1. p-value and VIF values are within the range
20. Set the cut-off probability initially as 0.5 and check the accuracy, sensitivity and precision scores
21. Draw the plot for accuracy, sensitivity, specificity scores. 0.36 seems to be optimal cut-off point
22. Since the CEO has asked for 80% conversion rate, this means the focus should be on sensitivity and accuracy score
23. Check for various cut-off values like 0.35, 0.36, 0.37

    1. 0.35 is the optimal cut-off point with sensitivity above 80% without compromising accuracy score
24. Use this cut-off on test data as well and predict the lead conversion
25. Generate the lead score on train and test data using converted probabilities based on cut-off set. Combine both the data frames to generate final set of data set with lead scores. These lead scores with value greater than 80% have higher chance of lead conversion
26. Solve business problems of aggressive lead conversion (high sensitivity with cut-off as 0.3) and avoid unnecessary phone calls (high specificity with cut-off as 0.6) by setting the cut-off values appropriately with very less compromising accuracy score
27. List the top three predictors that contribute in lead conversion - Lead Origin, Last Notable Activity and Lead Source
28. List the top three dummy variables that contribute in lead conversion - LeadOrigin_Lead Add Form, LastNotableActivity_Had a Phone Conversation and LeadSource_Welingak Website
29. Conclusion and Recommendations:
    1. The cutoff probability must be set as 0.35 for conversion rate i.e. Sensitivity of model to be 80%
    2. For aggressive lead conversion, the cut-off must be set as 0.3 increasing the Sensitivity of model without compromising much on accuracy of the model
    3. To avoid unnecessary phone calls, the cut-off must be set as 0.6 thereby increasing Specificity of the model without compromising much on accuracy of model
    4. Lead Origin, Last Notable Activity and Lead Source are top predictors in lead conversion model
    5. By marketing more on Welingak Website or approaching more Housewife, Working Professional will increase the chances of lead conversion
    6. Also marketing with Lead Source as Quick Add Form will increase the chances of conversion rate