Executive Summary

Following are the process followed:

1. Started with importing all the essential libraries
2. Checked data sanity using head(), shapes, columns, describe(), info()
3. The Info() function helped to identify whether the dataset has null values or not
4. Removed the columns having 40% missing value and columns having missing values greater than 15% were imputed and rows were deleted for columns with less than 15% null values.
5. Then created the dummy variables for columns ['Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Specialization', 'What is your current occupation', 'A free copy of Mastering The Interview', 'Last Notable Activity'].
6. Then split the data into train(70%) and test(30%) datasets.
7. Since few numerical variables had different scales. So next our target was to bring the scales of those columns between 0 & 1.
8. Converted rate comes out to be 37.85
9. Since the date was huge so we checked the correlations using df.corre() as heatmap wouldn't have been clear to understand.
10. **Correlation let us find out that the 'Converted' variable is highly correlated with 'Total Time Spent on Website'.**

Model Building using RFE

11. RFE gave us 15 important feature varibles.
12. Dropped the columns having p-values > 0.05 and VIF's >5.
13. Using Model Evaluation we figured out the relation between Converted, conversion probabilities and Predicted columns.
14. The **Overall accuracy between Converted and Predicted is 81.27%**.
15. Created the confusion matrix to find out:
    Sensitivity = 0.69
    Specificity = 0.88
16. ROC gave, **area under curve = 0.89 or 89%**
17. The probability where accuracy , sensitivity and specificity curve met is 0.38.
18. Then calculated the final_predicted column by mapping conversion_prob column having values>0.38 then 1 else 0.

19. The **Overall accuracy between Converted and finale_Predicted is 80.90%.**

20. Created confusion matrix for Converted and final_predicted

Sensitivity = 0.78

Specificity = 0.82

21. Precision = 0.79

Recall = 0.69

Precision and recall tradeoff was at around 0.18

22. Now checked for **final_predicted the accuracy = 74.79%**

23. While working with the test dataset do the following:

(i) Scaled the features that are called in train dataset

(ii) Then drop the columns which were eliminated in the train dataset while working with p-values and VIFs.

Now make the predictions,

And the Overall accuracy, came out to be 81.12%.

Sensitivity = 0.77

Specificity = 0.83

Observation:

1. Overall accuracy for the train dataset and the test data set were quite close.
2. The important feature variables are:
   'TotalVisits', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website','Do Not Email_Yes', 'Last Activity_Converted to Lead','Last Activity_Had a Phone Conversation','Last Activity_Olark Chat Conversation' etc.
3. We need to separate the best leads from the ones we generated by taking the factors such as, total time, visits, on the website etc. this will help us to identify the leads with the highest probability of the getting converted.