Summary for the lead scoring case study

This analysis is done for X Education to optimize the lead conversion ratio. Following steps are performed during the overall model building:

1) Inspect Data: To check what all data elements are available for the case study, overall volume of the data. Also check statistics to see how data is distributed or there could be potential outliers.
   There are some columns like 'Total Vists', 'Total visits per site' , 'Page views per visit' seem to have some outliers.

2) Data cleansing: Along with null values, many columns have 'Select' as their values, which ideally means that values are not available. These are as good as missing values so checks were performed by talking a look at the total % of null and 'select' together.
   Some of the columns like City and Specialization has 35%+ such values so they are not ideal while making the predictions.

3) EDA: Count plots for some of the columns to check data distribution helped to find if there are any dominating values. For example, count of leads by country has 'India' as the far dominating value compared to all the other countries. These attributes will not be helpful to make the prediction so can be removed.

   Other EDA were heatmap to check the correlation of the attributes with each other. For example,
   Lead Source_Olark Chat and Lead Origin_Landing Page Submission have correlation of -51% . This means they are inversely corelated so if one of them goes up, another would go down.
   Also last Notable Activity_Modified and Last Notable Activity_SMS Sent has high correlation value.

4) Data preparation: This includes converting 'Yes' and 'No' values to 1 and zero. 'Do not call' and 'Do not email' are examples for binary value.
   Outliers detection is also part of it, some values for total visits and other columns are trimmed.
   Scaling of the data were done for TotalVisits, Total Time Spent on Website etc.

RFE is also used find corelated columns that can be removed to select more important 15 columns. For example, Newspaper Article has 54th rank whereas Magazine has 61st so both are least important columns.

5) Model building: Logical Regression is used for model building. Training data is taken as 70% and 30% is used to test.
   RFE is used to select 15 important attributes first and then VIF is used to gradually remove columns having >5 VIF value.

6) Model Evaluation: ROC is used to find balanced point w.r.t sensitivity, accuracy, and specificity. 0.4 seems to be a reasonable cut off.

7) Precision and Accuracy: Accuracy is 81% and precision is around 79 which are good matrix for the model.

8) Making Predictions: Using the test data, the conversion probability is the column which assigns predicted probability of the leads to be converted. Even here the accuracy is 81% which means our model is behaving as expected.

   CEO's expectation to target for 80% lead conversion seems possible though this model.