Summary Report

First, we needed to import all the libraries. then we downloaded /read the dataset, view and understand the dataset. From viewing the data set, we understood that the majority of the country in dataset is “India”. from the dataset we can see that 96% of the leads are from India only, we combined the rest 4% as others. In the dataset "Select" means not available, so considered it as null value. then we checked if any of the columns having null values. Dropped the columns with null values>40%. Replaced the yes/no with 1 and 0. Categorical variables were converted to dummy variables and then dropped all the original variables. Splitted the target variable "Converted" into two datasets for better visualization - converted & not converted (n_converted).

From Plotting different variables, we found that:

- The lead origin from Quick add forum and Lead add forum is very high likely to get converted
- Lead origin from Landing page submission, API and Lead import is unlikely to get converted. Unemployed is the highest number in leads.
- Working professionals are the second highest and Working professionals are very highly converted.
- Last activity of both converted and not converted are majority in ‘Email opened and SMS sent’.
- If the last activity is in OLark chat or page visited, they are unlikely to get converted.
- Lead source through ‘Reference’ is very high likely to get converted.
- Tags, ‘Closed by horizon’ and ‘will revert after mail’ are most likely to get converted compared to other tags.

From the above analysis it was clear that a Logistic regression model will let us further analyse the model better, since the target variables are binary (1/0).

As a starting we kept the feature variables in to ‘X’ and response variable to ‘y’. Then we splitted the data into test and train data. Converted rate of total data was found to be close to 38.5. We also MinMax scaled some variables for the consistency throughout other variables.

Since there was a lot of feature variable it was difficult to select the individual variables manually, so we used RFE for feature selection. From the resulting model after setting the cutoff to 0.5 we got an accuracy of 0.936 which was pretty high. So, we checked the VIF of the model and found that one variable was having very high VIF value (>9) and all other variables was within limit (<2). Even after dropping the variable with high VIF, the Model accuracy remained unchanged.

The next step was to see the sensitivity, Specificity which was 0.919 and 0.947 respectively. Further we needed to know the optimal cutoffs, so we calculated accuracy sensitivity and specificity for various probability cutoffs and graphed it. From the graph the optimum cutoff was 0.5 itself. Since all the evaluation was good enough, we proceeded to do prediction in our test set. After making the prediction on test set, model showed very good Accuracy (0.928), Specificity (0.936) and Sensitivity (0.916) in Test and train set consistently.