Summary

The Following steps are done on the data :

1. **Cleaning Data :** Initially there are a lot of columns which have high number of missing values. Clearly, these columns are not useful. Since, there are 9000 data points in our data frame, let's eliminate the columns having greater than 3000 missing values as they are of no use to us. After that there are missing values which are dropped . There are some columns which are imbalanced , for all those necessary changes have been taken.
2. **EDA :** For all categorical and numerical columns we have drawn necessary plots to showcase the linearity. For finding out outliers also we can use some kind of plots.
3. **Dummy variables** : Dummy variables are created and later original columns are dropped.
4. **Train test split** : split is done according normal mode only like 70% train data and 30% test data. For scaling we used MinMaxScaler.
5. **Model Building** : Model building is done by taking RFE variables . Here we considered 15 RFE variables. Later rest of the variables are removed by considering VIF values as well as P-values. For VIF the values should be below 5 . For p-values it should be below0.05.
6. **Model evaluation and plotting Roc curve** : An ROC curve demonstrates several things:
   It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. A confusion matrix is made and using the optimal cutoff probability we get balanced sensitivity and specificity.
7. **Prediction**: This is done on test data frame with optimal cut off of 0.41 to calculate accuracy, sensitivity, specificity.
8. **Precision and Recall**: Normally this is used to recheck the values of test data frame with same cut off value.
9. **Conclusion**: We got the results like Accuracy: 79.1% , Sensitivity: 79.9% , Specificity: 78.4%, Precision: 77.5%, Recall: 79.9% . These results are on train data. The results on test data are Accuracy: 78.39% , Sensitivity: 78.82%, Specificity: 78.01%, Precision: 76.72%, Recall: 78.82%.
   By looking at the Evaluation we have built a decent model with accuracy of 80%.

Business Strategy Report:

1. The top variables that contributed towards leads being converted are:
   - Lead Origin_Lead Add Form
   - What is your current occupation_Working Professional
   - Total Time Spent on Website

2. The top 3 categorical/dummy variables which should be focused the most in order to increase the probability of lead conversion are:
   - Lead Origin_Lead Add Form
   - Lead Source_Olark Chat
   - Do Not Email.

3. Order of importance of the variabels in the data set are:
   - Lead Origin_Lead Add Form
   - What is your current occupation_Working Professional
   - Total Time Spent on Website
   - Do Not Email
   - Lead Source_Olark Chat
   - Last Activity_SMS Sent
   - Last Activity_Had a Phone Conversation
   - Lead Source_Reference
   - Lead Source_Welingak Website
   - What is your current occupation_Housewife
   - What is your current occupation_Unemployed
   - Last Notable Activity_Had a Phone Conversation
   - Last Notable Activity_Unreachable