SUMMARY

THE ANALYSIS PERFORMED IS FOR 'X EDUCATION' AND TO FIND WAYS TO GET MORE INDUSTRY PROFESSIONALS TO JOIN THEIR COURSES. THE BASIC DATA PROVIDED GAVE US LOT OF INFORMATION ABOUT HOW THE POTENTIAL CUSTOMERS VISIT THE SITE, THE TIME THEY SPEND THERE, HOW THEY REACHED THE SITE AND THE CONVERSION RATE.

THE STEPS PERFORMED ARE LISTED AS FOLLOWING:

1. BUSINESS UNDERSTANDING:
   We spent some time in analyzing the business problem and data dictionary to get a clearer understanding of the statement and the attributes.

2. CLEANING DATA:
   - We checked the duplicated data but there were not any.
   - We removed the redundant variables/features from the data frame.
   - Dropped the high percentages of null values in the dataset and also dealt with the null values in the rows.
   - Did the outliers treatment and data imputation techniques.

3. EDA:
   - Performed Univariate Analysis for both Continuous and Categorical variables
   - Performed Bivariate Analysis with respect to target variable.

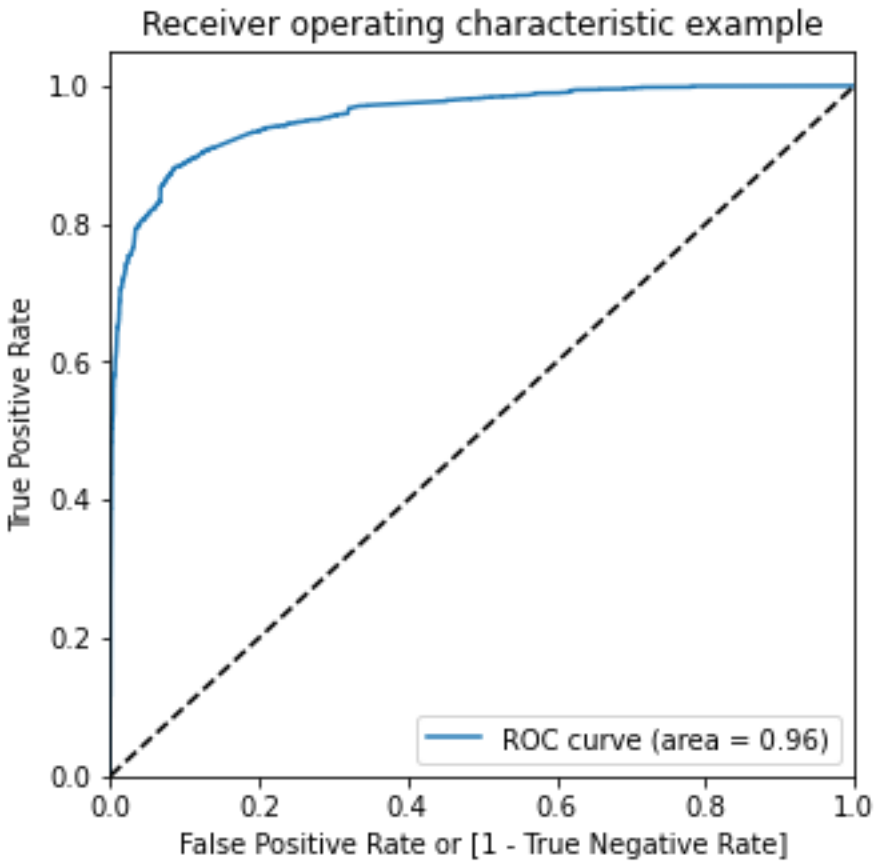
4. DUMMY VARIABLES:
   - The dummy variables are created for all categorical columns for others categorical columns with "Yes", "No" variables we simply did mapping with 1 and 0.

5. SCALING:
   - Used standard scalar to scale the data for continuous variables

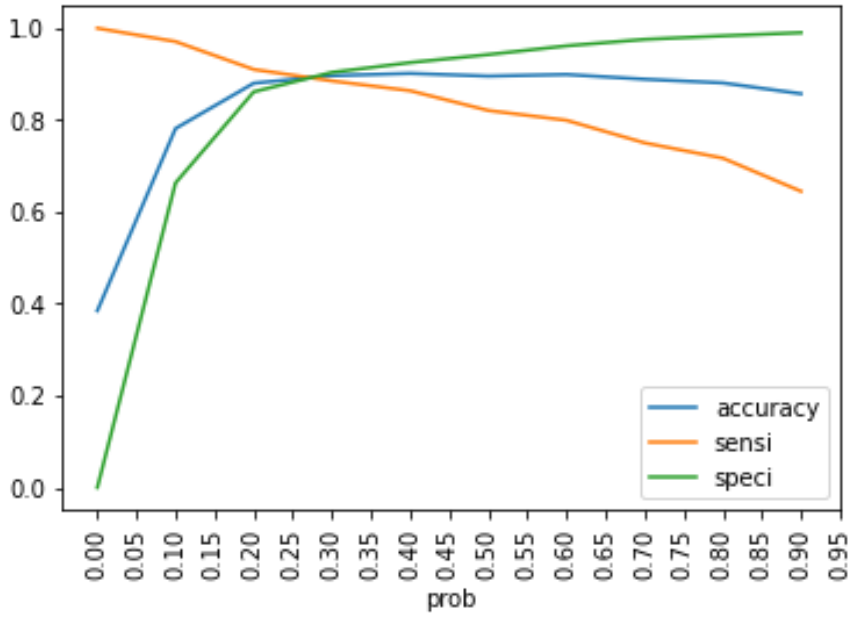
6. TRAIN-TEST SPLIT:
   - The split was done at 70% and 30% for train and test data respectively.

7. MODEL BUILDING:
   - We did RFE by providing 20 variables. It gave top 20 relevant variables. Later the irrelevant features were removed manually taking help of VIF values and the p-values reference.

8. MODEL EVALUATION:
   - A confusion matrix was made. And the model accuracy was found to be over 90 % . The tradeoff between sensitivity and specificity is well and good with area under ROC to be 96 %



We found the Optimal Cutoff Point using the plot for accuracy sensitivity and specificity for various probabilities and found that 0.28 is the optimum point to take it as a cutoff probability.



9. PREDICTION:
    - Was done on the test data frame an optimum cut-off as 0.28 . The overall accuracy of our model comes out to be nearly 90 %
    - The sensitivity of our model is nearly 89%
    - The specificity of our model is coming to be about 90 %

10. CONCLUSION:

1. Leads generated through past referrals are hot leads.
2. Leads with Tags : Email Opened, Will revert after reading the mail and SMS Sent seem to resulting in a conversion.
3. Leads closed by horizzon have a very strong chance of actually becoming a customer.
4. The company should focus more on Working Professionals specially at manager levels as there is a high chance of them getting converted as they could be looking for a boost in their current role.
5. Leads generated by Add Forms are also hot leads.