Summary

From the problem statement, we understood the company wants to target the high potential customers and increase their conversion rate.

Data Understanding- We came across many types of independent variables which includes input variables, score variables etc. We removed the score variables like asymetrique scores/index, Tags since they were tagged by the salesperson and won't be available to us while actual analysis. The target variable was identified as 'Converted'

Data Modification- We came across default values like ('Select') which we converted into a null value and calculated as percentages. We removed the columns with missing values more than 40%. Most of the columns with high missing values were removed because they had very skewed data. For example- Variables like Search, Magazine, Newspaper etc had data of 90-95% of 1 category and rest for others. Outlier treatment was done by removing those rows.

EDA - At first we split the dataset into converted and not converted dataset to analyze better. We observed leads with core specialization in their masters have a higher chance of getting converted. Apart from email, SMS was identified to be an efficient option.Leads originated from add forms tend to convert and not prioritizing import. Apart from Olark chat, other sources showed high conversion rates.

Data Transformation- While dummy variables were made for all the categorical variables (dropping the first column to reduce redundancy), all the continuous variables were scaled using min max scaler and then split the data into training and test data

Modeling- Using RFE, we found top 20 variables which contributed most to the model ie. variables which were better able to explain the conversion. Further using manual selection, we arrived at 15 variables. This was done by checking the P values and VIF values at each step.

Model Validation and Prediction - To predict, a proper cut off needed to be estimated. So we analyzed sensitivity, specificity and accuracy for different cut off points and arrived at a point where there was convergence of all the 3 score metrics specially with sensitivity higher than 80%. Finally the conversion probabilities were predicted for the test data and the lead scores were calculated by multiplying 100 with the conversion probabilities. The ROC curve was generated for the predicted conversion rates.The curve inclined towards the left top corner confirming the accuracy of our test and generating maximum area under the curve.

Inference- Further we segregated our leads into 3 categories- cold lead (lead score <30%), warm lead (lead score 30%-70% lead score) and hot lead (70% above lead score). Finally an aggressive sales team can be made who can directly target these hot leads(predicted as converted by the machine) for faster conversion.