SUMMARY REPORT

Lead score case study is about an education company name X Education which sells online courses to industry professionals but having very poor lead conversion rate. The expectation of the case study is to build a model where in a lead score can be generated to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
The provide data set has leads from the past with 9240 data points. This dataset consists of 37 attributes. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
So, to work on model building, the first action is Exploratory Data Analysis (EDA). In EDA columns were dropped based on the below criteria:
1. Columns having missing values more than 40%. (5 columns dropped)
2. Tags column as dropped as value of missing percentage.
3. Lead profile column also dropped as it does not have to many details.
4. Data imputed for columns where missing values were observed.

Dummy variables were created and data was split in train set and test set with 70-30 ratio respectively.

RFE is used to select 15 most suitable variables for creating logistic regression model in such a way that P-values of the model is less than 0.05 and VIF is less than 5.

The final model is evaluated by calculating accuracy, sensitivity, specificity, ROC, precision and recall.

The Probability cut off for Sensitivity-Specificity view is 0.40 and for Precision-Recall view is 0.42.

The equation for final model is P = 1/(1+e^(-A))

> ln[P/(1-P)] = -2.79 + 3.8924*(Total Time Spent on Website) + 3.2760*(Lead Origin_Lead Add Form) – 0.5035*(Lead Source_Direct Traffic) + 2.4205*(Lead Source_Welingak Website) - 1.5070*(Do Not Email_Yes) + 1.3212*(Last Activity_SMS Sent) + 1.5297*(Last Activity_SMS Sent) + 2.4078*(What is your current occupation_Other) + 1.2552*(What is your current occupation_Student) + 1.1513*(What is your current occupation_Unemployed) + 3.6658*(What is your current occupation_Working Professional) – 0.8341*(Last Notable Activity_Olark Chat Conversation) + 1.8430*(Last Notable Activity_Unreachable).

Variables with positive coefficient contribute towards the probability of a lead getting converted and with negative coefficient degrades the probability of a lead getting converted.
Higher the Log odd of the model, higher is the probability of a lead getting converted into paying customers.