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 30%. (6 columns dropped)
2. Columns not much significant of analysis. (4 columns dropped)
3. Columns having value as "Select" more than 30%. (2 columns dropped)
4. Columns having very high data imbalance. (13 columns dropped)

Still there were 6 columns left with missing value which were treat by deleting rows as the count of missing rows were not much. Post deleting rows 69% of original rows were retained.

Dummy variables were created for 11 categorical variables.

Now the data set was ready for model building, so target variable("Converted") is separated as Y data frame and remining variable are under X data frame and later divided into train and test set as 70:30 ratio. Further Min-Max scaling is used for 3 continuous variables.

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 0.42 and for Precision-Recall view is 0.44.

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

where,

A = 0.204 + 11.1489* TotalVisits + 4.4223*(Total Time Spent on Website) + 4.2051*(Lead Origin_Lead Add Form) + 2.7846*(Last Notable Activity_Unreachable) + 2.7552*(Last Activity_Had a Phone Conversation) + 2.1526*(Lead Source_Welingak Website) + 1.4526*(Lead Source_Olark Chat) + 1.1856*(Last Activity_SMS Sent) - 1.50307*(Do Not Email_Yes) – 2.5445*(What is your current occupation_Unemployed) – 2.3578*(What is your current occupation_Student)

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.