Summary Report – Lead Score Analysis Case Study

- Initially, import all the required libraries such as pandas, numpy, matplotlib, seaborn, sklearn and statsmodels. Moreover, inspect the data and make it ready for analysis by handling null records, dropped the columns which has high correlation with others, transforming data points as required and drop the irrelevant columns which are not necessary for analysis by performing EDA. Although dataset begins with 9240 records and 37 columns, after cleaning up data due to above mentioned reasons as well as data imbalance reason, record and column count reaches to 9029 records with 18 columns. During the clean-up of data, Transformation such as Yes to 1 and No to 0 and Select as null has taken place.

- Furthermore, transform categorical variables into numeric by creating multiple level of dummy variables with values as 1s and 0s. Hence, more number of columns are added and column count, however, due to high correlation, some of the columns are dropped. On top of that, split the dataset into train and test with 70% (6320 records with 32 columns) and 30% respectively plus where train and test dataset had performed min max scaling on numeric variables.

- Begin building the logistic regression model (with target variable as Converted) within 32 variables and reduce it to 15 by using RFE and further reduce variables manually until it satisfies the criteria of p-value<0.05 and VIF<5. Where model 1 starts with 31% pseudo R-square value this remain around the same percentage throughout the Final model (Model 5). The reason we keep on going rebuilding model is because of not satisfying the criteria of p value should be less than 0.05 and Variance Inflation Factor should be less than 5.

- Finally, evaluating the model by employing methods such accuracy, specificity and sensitivity are around 77% for train model where as it is 80% in test model which clearly shows that model evaluation is successful. Apart from we also plot ROC curve to figure out about optimal point which is just under 0.4, Nevertheless, area of ROC curve is about 80% which means this model is healthy one.