Lead Scoring Case Study Summary

In this case study our primary goal is to build a logistic regression model to assign a lead score of (0 to 100) to each of leads which can be used by the company to target potential leads. A higher score would mean that a lead is hot, i.e. is most likely to convert and a low score means that a lead is cold and will mostly not get converted.

We started this case study with data understanding and data cleaning. We tried to understand the structure of the data-frame after importing data to a data-frame from the csv. In our initial analysis we found out that some columns in the dataset were having null values of more than 30% and there were also a few columns that had null values which were below 30%. The ones with more than 30% null values were dropped and the rest of the columns were imputed with the most appropriate values, so that it doesn't affect our model.

Also figured out the columns which were having significantly high values of either ('Yes' or 'No'). We performed label encoding for all such variables by assigning **1** for '**Yes**' a **0** for '**No**' for that these can be used for our logistic regression model. Then we converted all categorical columns into dummy variable and concatenated the same with our initial data-frame. We have used sklearn and statsmodel library to perform logistic regression. We created our first iteration of LRM. We could see from initial model that there were a lot of columns with higher probability (even close to 1), we repeatedly ran model in successive iterations to eliminate one column at a time with highest probability.

Once we got a stable model it was time to perform feature selection using VIF (variance inflation factor), we found that some columns were having a VIF of greater than 5 and we eliminated these feature one at a time iteratively. Once were done with feature selection, we used predict() to calculate probabilities for the train data set initially and found out that our model's accuracy, sensitivity and specificity is (0.928, 0.881 & 0.957). We created columns with different probability cutoff after plotting ROC, then we went with plotting accuracy, sensitivity and specificity to determine a cutoff point of 0.30.

We again created columns with different probability cutoff set at 0.30, after plotting ROC, we used predict again for our train dataset to calculate probability and found that our model's accuracy, sensitivity, specificity and precision is (0.925, 0.918, 0.929, 0.888). We were at our final step of predicting on test data set, we followed the same procedure of using predict () function on our test set to calculate probabilities and ended up with a similar score on the test data set as well. Our scores for the precision was 0.883 or 88% and it implied that our model is good and we achieved our goal of target lead conversion rate to be around 80%.