Case Study Summary

The approach in solving this problem is to first clean the given data. We will drop the columns having more than 70% NA values.Four columns have a 45.65% missing value. Data has too much variation in these parameters so it is not reliable to apply any value imputation in it. Therefore we drop these columns. For imputation of missing values we use mode value. After that we perform replacing operation on some of the attributes and we also clubbing the low frequency value better model building.

We start EDA by checking the imbalance percentage in target variable. Performing a univarite analysis to the attributes Based on the analysis we have seen that many columns are not going to add any information to the model, and for further analysis we drop this attributes.

We also remove attributes which are generated or updated by the sales team once the member completed the call with students. This attributes are not relevant for model making because this is not a system generated attributes.

Those attributes are not system generated is also significant for conversion to a positive lead.

In the data preparation stage we first convert categorical binary variables (Yes/No) to 1/0. Performing binary mapping for few of the categorical variables, the next step would be to create dummy variables. Then we split the data into train and test, 70% & 30% respectively. In the final step of data preparation we perform feature scaling.

We start building the model on the train set. In model-I summary we analyze that there are many variables whose p-values are very high, implying that that variable is statistically insignificant. So we eliminate some of the variables in order to build a better model.

We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manually eliminating features based on observing the p-values and VIFs.

Once the model is ready with acceptable p-value and VIFs. We proceed towards making predictions using this model with arbitrary value and after that we checking the confusion metrics and accuracy. For optimal cut-off we draw a ROC curve and check the area under the curve and we find out that its capturing 84% of the area. After that we create columns with different probability cutoffs curve. From the above curve, 0.32 is the optimum point to take it as a cutoff probability. Once this process is over we check model Precision and Recall.

Once the model is ready we start making predictions on the test set. We also validate the test set by comparing over all accuracy, Sensitivity and Specificity. We analyse that the model seems to predict the conversion very well. It looks like we have created a decent model for the converted dataset as the metrics are decent for both the training and test datasets.