Lead Scoring Case Study Report

As the first step, imported the dataset and performed few checks to understand the data. The leads dataset had 9240 records and 37 attributes. Referred the data dictionary for better understanding of the data. During the analysis, identified multiple columns with missing values in the dataset. Few columns had value “Select” which needed to be handled just like missing values.

As part of data cleaning, identified the columns with over 35% missing values and decided to drop such columns. Few columns were identified with very high data imbalance. Since these wouldn’t influence the analysis dropped these columns as well. Also dropped ID fields which wouldn’t contribute to the analysis.

Out of 4 columns with “Select”, 2 had around 50% values as “Select”, hence discarded those columns. Also, dropped City and Country columns which doesn’t seem to impact Conversion probability. At this point only “Specialization” column had “Select” value.

For columns with small percentage of missing values, dropped the missing value records. Retained 69% of the records and 12 attributes at the end of the cleansing activity.

As next step, created dummy variables for category columns, concatenated it to leads dataset and dropped the category columns. Special care was taken for “Specialization” column to manually drop the dummy variable corresponding to “Specialization_Select” to handle the “Select” value case. Total feature count was 74 at the end of this process.

For train-test split, initially separated dataset into feature variables dataset (X) and target variable dataset(y). Then divided it in the ratio 70-30 as train and test datasets.

Scaling was done for the numeric variables of both train and test datasets using MinMaxScaler (). After scaling, looked at the correlations. Due to high count of features couldn’t make any major findings. Hence proceed with feature selection using RFE

Selected 15 features using RFE, which was then used to build our first model using statsmodel.

First model had few features with high p values, so checked for vif values. Feature with highest p-value and vif was then deleted.

Proceeded with building our 2nd model. Performed the same process until optimal p-values and vif were achieved. After multiple iterations, finalised the 5th model.

Next, generated predictions for train set using the final model with an arbitrary cut-off of 0.5 and performed model evaluation.

Evaluated accuracy, confusion matrix, sensitivity and specificity of the model. The value ranges were optimal. To further evaluate, plotted the ROC curve, which looked good. From the trade-off plot between accuracy, sensitivity and specificity found the optimal value for cut-off as 0.42.

Using the new cut-off evaluated the model again. The matrix values were very comparable at this stage. Calculated the Precision and Recall metrices as well from confusion matrix. The trade-off curve between precision and recall also gave the optimal cut-off value as 0.42

As the last step we made predictions on our test set. Evaluated the metrices for the test predictions. The matrix values were all in an optimum range and were very comparable to that of train set.