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.

As part of data cleaning & analysis, identified the columns with over 45% missing values and dropped them. Few columns were identified with very high data imbalance. Since these wouldn't influence the analysis dropped these columns. Dropped ID fields as not relevant. Created count plot for visualization of categorical variables.

For categories having low frequency values, merged all these columns as Others. For missing values/"Select" values in categorical attributes, imputed with Mode value as applicable. Out of 4 columns with "Select", 2 had around 50% values as "Select", dropped them. For Specialization imputed with Not Available. For columns with missing values (1.48%), dropped the missing value records.

For Numerical attributes, created Heatmap for correlation & Box plot for outliers. No strong correlation observed with Target Variable. For outlier treatment removed the top and bottom 1% values. 8953 records and 12 attributes retained.

Created dummy variables for categorical columns, concatenated it to lead dataset and dropped the initial category columns. Manually dropped the dummy variable corresponding to "Specialization_Not Available" for "Select" case and dropped dummy variables corresponding to Others. Total features were 56 now.

For train-test split, separated dataset into feature variables (X) and target variable (y). Divided into 70-30 ratio as train and test datasets.

Scaling was done for the numeric variables of both train and test datasets using StandardScaler(). After scaling, evaluated correlations. Due to high feature count couldn't make any major findings. Proceeded with feature selection using RFE.

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

First model had 2 features with high VIF (>5), and 1 feature with high p-value((>0.05). Feature with highest VIF was dropped. Proceeded with building our 2nd model. It has 1 feature with high p-value and no feature with high VIF. Dropped the feature with high p-value. Built our 3rd model and found no feature with high p-value & VIF. This is our final 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, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value and Negative Predictive Value of the model using confusion matrix. 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 0.3.

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 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.