Lead Scoring Case Study Summary

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps followed for building the model:

Step 1: Importing Libraries and Data:

Here we import the necessary libraries and import the data from the csv file.

Step 2: Reading and Understanding the Data:

Here we used functions like head (), info () to understand the data. few categorical columns we need to create dummy variables and need to take care of null values present in some columns.

Step 3: EDA (Exploratory Data Analysis):

There are few columns with level called 'Select'. These values are as good as missing values and hence we will convert 'Select' values to 'Nan'.

We have handled missing values, checked for highly biased/imbalanced categorical variables. We have also analysed the categorical and numeric variables using count and box plot. Based on this analysis we have identified the variables that need to be dropped. After EDA we have a dataset in good shape that can be considered for further model building.

Step 4: Data Preparation:

We have converted some columns from yes, no to numeric and created dummy variables for categorical columns.

Step 5: Test-Train Split:

We have considered train dataset of 70% and test dataset 30%.

Step 6: Feature Scaling:

Normalisation has been used for rescaling of numeric variables.

Step 7: Model building using Stats Model and RFE:

RFE method has been used to select 15 features for model building. We are checking for P-values and VIF in our logistic model and removing variables which are insignificant.

Step 8: Prediction on Train set:

Accuracy: 74%, Sensitivity :84%, Specificity: 67%

Step 9: Plotting the ROC Curve:

The ROC curve value for our model is sufficiently good.

Step 10: Finding Optimal Cutoff Point:

The cutoff point was determined by the intersection of accuracy, sensitivity, and specificity.

Step 11: Prediction on Test Set:

Accuracy: 74.55%, Sensitivity :85.22%, Specificity: 67.97%.

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction. Accuracy, Sensitivity and Specificity values of test set are around 74%, 85% and 67% which are approximately closer to the respective values calculated using trained set.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Last Activity_SMS Sent