Brief summary report of the lead-scoring Case study

X-Education is a company which wants to sell their course to industry professionals, has approximately 30% conversion rates with the available current data. We need to solve this problem by building a model with 80% conversion rates.

We started inspecting the data, checking for missing values, unnecessary columns and values like 'Select' is same as missing value. Thus, imputing these values and removing them by:

- Dropping the columns with missing values >40%
- Dropping columns with unique values throughout
- Dropping the columns with same value throughout the column
- Checking for skewed values the columns with yes/no by checking its value counts and drop those columns
- Checking the value counts of other columns for skewed results and drop them if it is significantly skewed.
- Drop the rows with ~ (1-2) % missing values

After checking the value counts we can impute the columns wherever necessary with a mode or others accordingly. Performed univariate, bivariate and multivariate analysis to check the significance of each of the variable and check for outliers in the numerical variables. Create dummy variables for easy analysis and drop the unwanted columns after dummy variable creation. Re-check the data by inspecting the missing values, shape.

The data is ready for model building after the above steps:

- Create train (70%) - test (30%) split
- Scale the data using MinMax scaling
- Create logistic regression with rfe = 20
- Check for p-value and VIF
- Drop the columns with p-value > 0.05 or VIF > 5, one by one and keep checking the p-value & VIF
- Once all the values satisfy p-value < 0.05 or VIF < 5

Finally, the model is ready for model evaluation. Now we use the x-train to predict y-train and create a confusion matrix with a random cut-off 0.5 to check how the model is working. Plot the ROC-curve to check the performance of a classification model at all classification thresholds. Followed by plot for accuracy, sensitivity and specificity against all the probability ranges. The intersection of this plot acts as a cut-off for better results (we got a cut-off 0.35). Thus, we got the following results on train-set:

- Accuracy = 80.77%
- Sensitivity = 80.74%
- Specificity = 80.79%

The results were good enough to proceed further with the test set. Analysed the same for the test set and created the confusion matrix with 0.35 cut-off, the results were as follows:

- Accuracy = 81.31%
- Sensitivity = 80.38% = Recall
- Specificity = 81.83%
- Precision = 71.62%

The results were good enough and thus we plotted the precision recall graph to evaluate the results. Now, we calculate the conversion probability for our target variable (Converted) and converted it to percentage format which is our Lead-score. Finally, we had a model with good accuracy, sensitivity and specificity. From the above analysis we will get top variables to look carefully to work on them for better conversion rate and lead generations.

Report by
Sanmitha S J and Anjali Bhalla