SUMMARY REPORT

By

Nitin Koundilya & Priyadarshan
PGDDS, UPGRAD

TASK REQUIRED FROM DATA ANALYST

Given a set of data for X Education Company, we were asked to build a logistic regression model, which will boost up the current conversion rate of company from 30% to more than 75%

APPROACH USED AS A DATA ANALYST

1. Imported and inspect the data. While inspection noticed there were many columns with missing values in them, two continuous columns having outliers and some unnecessary columns from business point of view as they were assigned variables after sales executive contacted potential customer. Dropped such columns and columns having missing values more than 50%, as those columns do not give any significance information.

   For other columns which had less % of missing values, inspected them thoroughly and for some of them categorized missing values with other insignificant values of that column and stored it under Other_Column_name of that column with intent to drop those values while the process of creating dummies.

2. Created dummies for the categorical columns, and for continuous columns scaled them with StandardScaler, so that it would not distort the significance of other features.
3. While EDA and preparation found that some of the columns have quite high correlation with others. So dropped them accordingly.
4. Now that data was cleansed and analyzed thoroughly, split the data into train and test set with ratio of 70%-30% of original data.

5. Created Logistic regression model and checked for its statistical summary for the p value, many features were having high P-value, but number of features was also too high. Hence reduced the number of features using mixed approach of Recursive Feature Elimination and manual method. Finally, finalized 16 features which helps in predicting the probability of conversion from customer to student.
6. After finalizing all the features, checked for other matrices such as Accuracy, Specificity, Sensitivity and ROC curve @0.5. All seems to be good, but to be sure, checked for the optimum cut-off for all the three matrices and it resulted out to be 0.43. Checked revised Accuracy, Sensitivity and Sensitivity, all were between .77-.80. ROC curve is also 0.86.
7. Validated our model on the Test data set and it turns out that accuracy is .79, Sensitivity is .77 and Specificity turns out to be .80 along with ROC curve area .87.
8. Hence, all seems to be okay and model seems to be robust. Assigned Lead Score to every lead number of the test data set, so that X Education can formulate their strategies based on this model.