Lead Scoring Case Study – Brief Summary

Started the assignment by reading business requirement and revising the material on logistic regression.

Habitually, the input data is reviewed and initiated the data cleaning procedures. Executed the foremost step of data preparation, missing values are handled as follows in the data elements.

- above 40% missing are removed due to lack of information.
- 2% of missing values are removed as its minute and required undue time in analysis.
- In-between range of 2 – 40% are analysed for data imputation and the decisions are made as required.

EDA simplifies the data understanding through univariate, bivariate and multivariate analysis using plots, also outlier, relationship of each variable to the target variables, data skewness and correlation are identified comprehensively.

As a first step of modelling, Logistic Regression requires binary classification of categorical attributes. Dummy variables are created for each category of categorical features. Class-imbalance is measured as ~37% which is good rate. This shows whether the data has imbalance on target ratio.

Train (70%) and test (30%) data are split. Continuous numerical variables are scaled to avoid disproportionate effect on model results. Built the very first multivariate logistic regression model using GLM (Generalized Linear Models with Binomial stats models). Most of the features had high p-values which are identified as statistically insignificant.

To build a better model, balanced feature-modelling combination of auto-tuning (using Recursive Feature Elimination (RFE)) and fine-tuning (manual selection/elimination) are performed. Elements which are having High p-value and VIF (variance inflation factor)>=5 are removed from model as it signifies poor modelling. The feature elimination is handled one at a time since ones removed make considerable changes in other feature aspects.

Probability of lead conversion is measured using “predict” function. At first step, Probability measure is set to 0.5 for predicting if the customer is “Converted” or not. Based on this, the chances of identifying “Converted” as “Not-Converted” lead and “Not-Converted” as “Converted” lead errors are common.

Confusion matrix is created for analysing these error-measures. The optimal threshold metrics are calculated the cut-off point as ~0.36 which is almost equal to precision and recall trade-off 0.4. By using optimal cut-off, train-data accuracy is calculated as 0.810, Sensitivity= 0.806, Specificity= 0.813, False Positive Value= 0.18, Precision= 0.73, Recall= 0.80, F1-Score= 0.767, Threshold cut-off=0.4, ROC Curve Area=0.88

Prediction on test data is Accuracy= 0.813, Sensitivity= 0.779, Specificity= 0.832, False Positive Value= 0.16, Precision= 0.72, Recall= 0.77, F1-Score= 0.75, Roc Curve Area = 0.87.

Logistic Regression modelling is the exciting procedure as it has involved with sequence of steps and formulae execution. Data understanding and business perspective are the key elements for achieving the best data model. The modelling required handful of experiment to obtain the expected result. It is incredibly important to understand each element and their categories, relations, correlation, as all the facts determine the end-result of model and its standards. This exercise helped in gathering knowledge on how data science is implemented in validated online courses and forums also the how this model will be benefitting the marketing team on Hot Lead customer prediction.