Lead Scoring Case Study Assignment

Summary Report

Team

Krishna Murthy, Ankit Pandey, Ajinkya Chinchwadkar

Plan

We have formed a team and discussed problem statement. We understood the requirements and decided approach. As mentioned in the problem statement we have to use logistic regression model. We discussed about timeline and planned tasks. We have decided to meet everyday in the evening for half an hour for progress review. So we have worked for 5 days like mentioned above and prepared logistic regression model and supporting documentation.

Act

We followed typical data analysis steps as learned in the program. We have decided to use multivariate logistic regression model. We have studied input files. We have started with data preparation as follows key steps:

1. Duplicate values removed
2. 'Select' values converted to 'NaN'
3. Missing values treatment – Dropping columns having more than 40% missing values
4. Missing value treatment – Filling with 'NaN' for retaining important columns
5. Grouping of low count values in a columns – e.g. country
6. Outlier treatment with box plot
7. Dropping categorical columns having only one type input
8. Dropping high correlation variables using heat map
9. Dummy variables for categorical type columns

Then we have proceeded with model building steps using GLM logistic regression.

We have used RFE technique (Recursive feature elimination) for automatically eliminating variables. RFE technique identifies lower important variables with rankings. Out dataset has more than 40 variables, hence we have randomly selected 20 variable criteria.

Then we proceeded with model building first iteration. We have observed less important variables having P value > 0.05. We have eliminated such variables.

We have use VIF technique to identify highly correlated variables. We have decided VIF criteria as <5 according to industry best practice.

After dropping less important variables as identified in P value, we have re-run GLM model to confirm P-value and VIF are within limits mentioned above.

After confirmation we proceeded with prediction on train and test data. We have started with train data. We have predicted probabilities on train data.

We have randomly selected 0.5 as cut off on training data and derived prediction ranking as 0 or 1.

Then we put train model for evaluation. We prepared confusion matrix for deriving evaluation metrics of accuracy, sensitivity and specificity. We have confirmed these metrics using ROC curve. We have verified area under ROC curve to be maximum.

Then we created columns with different probability cut offs. We have identified evaluation metrics for each of cutoff. We have run train model with each cut off. We have plotted accuracy, specificity and sensitivity curves on single graph. We have derived probability value as cut off at which point all 3 curves meet. We have fine tuned final derived prediction values based on this cut off of 0.35.

We again confirmed model training set using evaluation metrics.

We have checked derived model on test set using same cut off of 0.35. We observed precision and recall values are satisfactory. However we have decided to use tradeoff method to revise cut off value. We have revised probability cut off value to 0.41 based on p-r curve where precision and recall curve meets.

Learnings

We have learnt how to work together on data science project. We are using Git and Github first time. It gave us a fill how actually a team of data scientist would be working together.

It has improved our understanding of data analysis flow and requirement. We have understood that iterations are required for feature selection. Business understanding helps in feature selection.

We have done multiple discussions on improving VIF. After iterations we could achieve VIF of 3. Balancing technique required for model metric selection. In this case we have decided to use balance of accuracy, precision and recall as all are important in identifying potential leads.