

Modeling Report

Data Preparation

First we conducted basic statistical analysis on the modeling dataset, and we were particularly focused on the missingness of all predictors. It turned out that the data quality is not very bad. Only couple of predictors have severe missing problem. We created MVIs (Missing Value Indicators) for all the predictors that have missing values. We also made imputations for the missing values, using CART from Salford Systems. Other new derived variables were also created to be tested.

Variable Selection

We used TreeNet from Salford Systems to evaluate the importance of predictors. TreeNet gives importance scores for all the predictors used in the models so that we could obtain the ranking of all the predictors. Then we compared the performance of models of different combinations of variables with the consideration of their ranking. Finally we determined 17 variables would be contained in our final model. We also tried logistic regression to select variables. Interestingly, by using chi-square test on individual predictor, similar variables with imputation and MIVs stood out.

Building Model

With these 17 predictors, we built TreeNet models. We also built logistic regression models using these 17 predictors (imputed variables are used for the predictors having missing values). We compared the area under ROC curve of different models, both of which are around 0.7. TreeNet performs slightly better than logistic regression, although the correlations of their predictions are not very high (about 0.75). This means that TreeNet and logistic regression don't agree on some of the predictions while their overall performances are close. Hence, we decided to combine the results of the two different models in hopes of making the models complement each other. Two ranks from the two models were assigned to each prediction record. The transformed average rank for each record was the final score for that record.

Model Performance and Results

It is seen that the ensemble model has a more stable performance than individual TreeNet, logistic regression model. However, due to the complexity of the model, not much understandable insights could be obtained from the model except some kind of variable importance ranking. For the further understanding of the business insights, it is better to use CART tree to give rule-based interpretation.