1/18/2022

Danielle sherman

CTO, Blackwell Electronics

Overview:

As per your request, I analyzed the CompleteResponses data set along with the corrupted data set, SurveryIncomplete. I utilized the following classification methods: C5.0, RandomForest, Stochastic Gradient Boosting, and Partial Least Squares. Stochastic Gradient Boosting yielded the best results for this dataset, followed by C5.0, RandomForest, and Partial Least Squares. Below, you will find the results and predictions.

**C5.0**:

C5.0 classification was chosen because it was a specific requirement. Out of the two models built in C5.0 using automatic tuning with a tuneLength of 5 and 10, the first yielded higher results: Model 1 Accuracy: 0.9195635 Kappa: 0.8307364; Model 2 Accuracy: 0.9179466 Kappa: 0.8271194.

**RandomForest**:

RandomForest was another classification model that was a specific requirement. I built two models using manual tuning with an mtry of 5 and 20. The second model produced better results: Model 1 Accuracy: 0.8799515 Kappa: 0.7451652; Model 2 Accuracy: 0.9143088 Kappa: 0.8184389

**Stochastic Gradient Boost**:

Stochastic Gradient Boost for classification modeling was also a requirement and works on the principle that many shallow trees together can make a more accurate predictor, and it produced the highest results. I built two models using automatic tuning with a tuneLength of 5 and 10. The second model produced slightly higher results: Model 1 Accuracy: 0.9240097 Kappa: 0.8394613; Model 2 Accuracy: 0.9244139 Kappa: 0.8402157

**Partial Least Squares**:

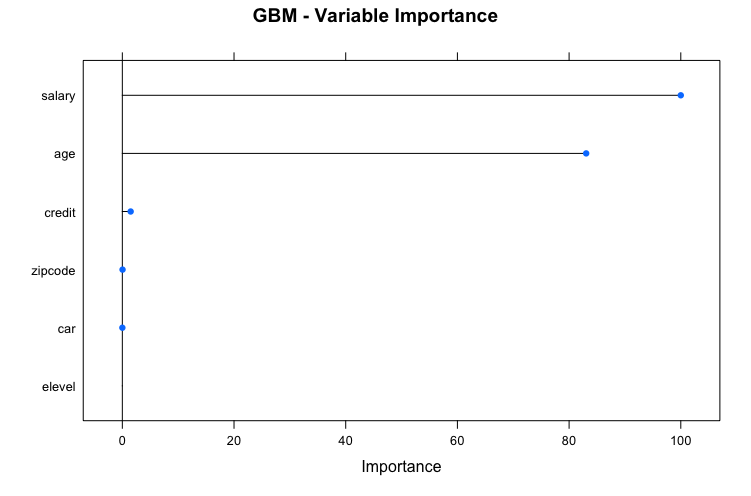
Partial Least Squares was chosen because it is used for classification modeling and was available in the R library. I built two models with automatic tuning with tuneLengths of 5 and 10. It was not a good method for this dataset and produced very low results: Model 1&2 Accuracy: 0.58286176 Kappa: -0.03578476

**Predictions**:

Making predictions with the corrupted brand data (SurveyIncomplete.csv) using the Stochastic Gradient Boost model yielded much lower results: Accuracy: 0.39580000 Kappa: 0.01191735.

Whereas the “Ground Truth” in CompleResponses yielded an accuracy score above 90%, introducing corrupt data into the model shows that we cannot predict with a high degree of accuracy.

Regarding the importance of the features according to Stochastic Gradient Boost, “Salary” and “Age” are the most important. Almost all other features are irrelevant to this problem.



According to the Stochastic Gradient Boost model and the CompleteResponses dataset, more customers prefer Sony to Acer, at 9924 to 5674, respectively.

GRANT JOHNSON