MACHINE LEARNING COMPETITION

ORANGE TEAM 12

YUTING (CRYSTAL) CHENG

WILL ELMORE

THOMAS GOW

HIRSH GUPTA

SARAH WOTUS

NOVEMBER 25, 2019

EXECUTIVE SUMMARY

Researchers are interested in knowing which attributes influence the occurrence of events Target 1 and Target 2. Machine learning algorithms, including Random Forests, Extreme Gradient Boosting (XGBoost), Ridge and LASSO Regressions, and Generalized Additive Models (GAMs) were developed to predict each target individually. Based on our analysis, XGBoost had the best predictive power for Target 1, and Ridge Regression had the best predictive power for Target 2. The optimal XGBoost used a sample of 85% of the variables for each tree, a learning rate of 0.15, and 15 trees with a max depth of three for each tree. The optimal lambda for the Ridge Regression was 0.028.

DATA USED

The data provided was split into three groups: training, validation, and test. We used the training dataset (or a subset of depending on the algorithm and computational time) to build the models. The models were then evaluated using the validation dataset. For model comparison, we calculated the area under the ROC curve (AUC) and chose the model with the highest value. After choosing the best model for each target, we supplied the researchers with the predicted probabilities of Target 1 and Target 2 for the observations in the test dataset. Target 1 and Target 2 were not mutually exclusive events, therefore supplying the researchers with predicted probabilities would enable them to choose which target had a higher probability of occurring for each observation.

MODELING

RANDOM FOREST

The team's first model attempt was a Random Forest as it is both computationally fast and capable of handling thousands of input variables. To improve speed, a random sample of 100,000 observations was obtained from the training dataset to build the model and tune parameters. The following parameters were autotuned using SAS Viya to find the highest AUC on the validation dataset: number of trees, maximum nodes in each tree, the number of variables randomly sampled as candidates for each tree split, and maximum/minimum size of terminal nodes for each tree. A final AUC of 0.765 was obtained for Target 1 and 0.753 for Target 2.

XGBOOST

When properly trained and tuned, the XGBoost model can outperform Random Forests. Therefore, the team once again used a random subset of 100,000 observations from the training set to build and tune an XGBoost model for Target 1 and Target 2. The following parameters were tuned to find the highest AUC: the number of variables used in the model, the learning rate, the number of trees used, and the max depth for each tree. After fitting the model on each variable, AUC was calculated using the validation dataset. This model resulted in an AUC of 0.796 for Target 1 and 0.779 for Target 2. This was the highest AUC for Target 1 out of all the models we tested.

RIDGE AND LASSO REGRESSION

Both Random Forests and XGBoosts are prone to overfitting the training data, therefore the team pursued regularized regression to retain all variables in the model while avoiding overfitting the training data. Ridge and LASSO models were built for both targets. To serve the computational abilities of R software, 10% of the training data was randomly sampled to build the models. To determine the most optimal model, lambda was chosen based on the value that provided the minimum average error on K-fold cross validation. For Target 1, Ridge Regression resulted in an AUC of 0.788 while LASSO Regression resulted in an AUC of 0.786 on the validation dataset. For Target 2, Ridge Regression also outperformed Lasso with an AUC of 0.781 versus 0.690 on the validation dataset. This was the highest AUC for Target 2 out of all the other models.

GENERALIZED ADDITIVE MODEL

GAMs were advantageous since the computational time for the size of the data was much quicker. Therefore, the training data (omitting all missing values) did not have to be sampled in order to build the model. Using piecewise linear functions, the GAM for Target 1 reported an AUC on the validation set of 0.769 and an AUC of 0.766 for Target 2. If the researchers desire quick model building with comparable AUCs, the team would recommend pursuing GAMs. However, for the purposes of reporting the models with the highest AUC, XGBoost outperformed GAM for Target 1 and Ridge Regression outperformed GAM for Target 2.

CONCLUSION

Out of all the models tested by the team, the XGBoost model proved to be the most accurate predictor of Target 1 and Ridge Regression was most accurate for Target 2 based off AUC. Random Forests, GAMs, and regularized regression were also all tuned to optimize their predictive abilities. We recommend the researchers go forward with XGBoost and Ridge Regression. Predicted probabilities for the test dataset have been provided using these recommended models.