Predicting Basketball Home Team Wins

Hanna Grossman and Citlally Reynoso

12/8/2019

Introduction

We began the project with NCAA data from Kaggle, consisting of basketball statistics. This data consisted of 218 variables and 9520 observations. Our goal was to create a model that could predict whether the home team had won the game or not. From there we applied this model to our testing data to see how well our model performed.

Methods

- Step 1: First, we read in the training and testing data into R.
- Step 2: We then cleaned the data by deleting the columns that were repeated.
- Step 3: From there we began to fit models to the train data and explore what worked best on our particular data set.
  - Logistic regression, Random forest, XGBoost, Adaboost, LDA, QDA
- Step 4: We observed that the random forest and XGBoost models performed best, so from there we worked to tune these models.
  - 4a: We tuned the following parameters using tuneRF: mtry and ntree
  - 4b: We tuned the following parameters using xgb.train: nrounds, eta, max_depth, verbose, min_child_weight, gamma, subsample, colsample_bytree, early_stopping_rounds, eval_metric
- Step 5: In addition, we created two new variables to add to our dataset for our XGBoost model.
  - We created two variables, one looking at the proportion of games that each team won at home, and the other being the proportion of games that each team won away from home.

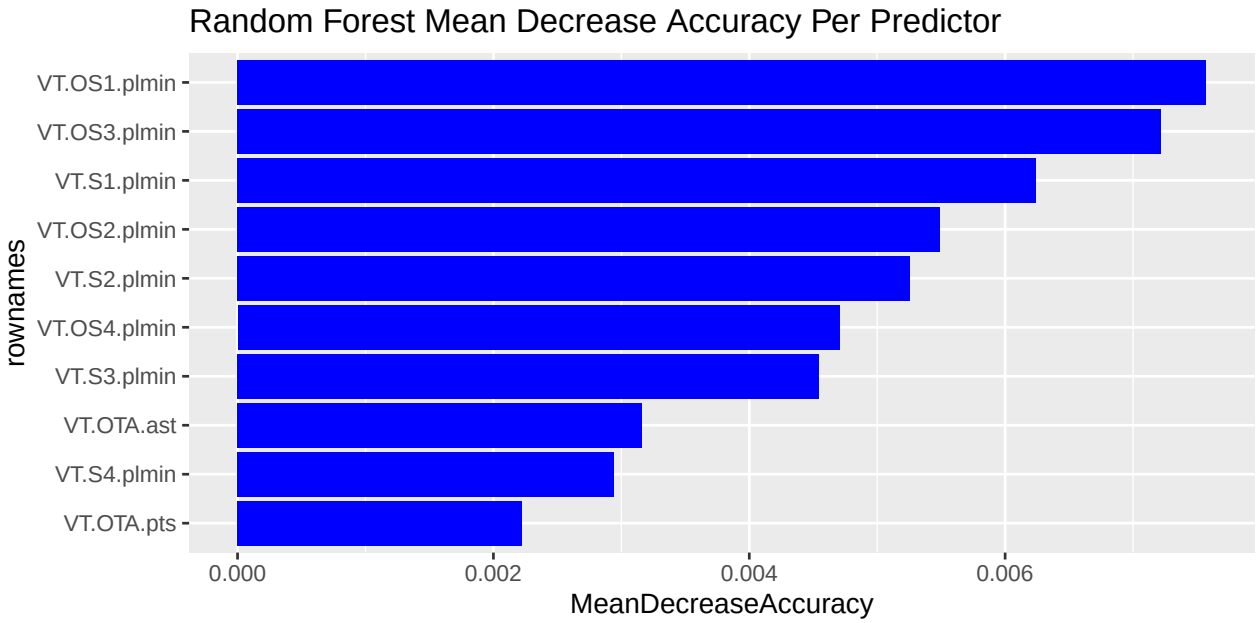
Results

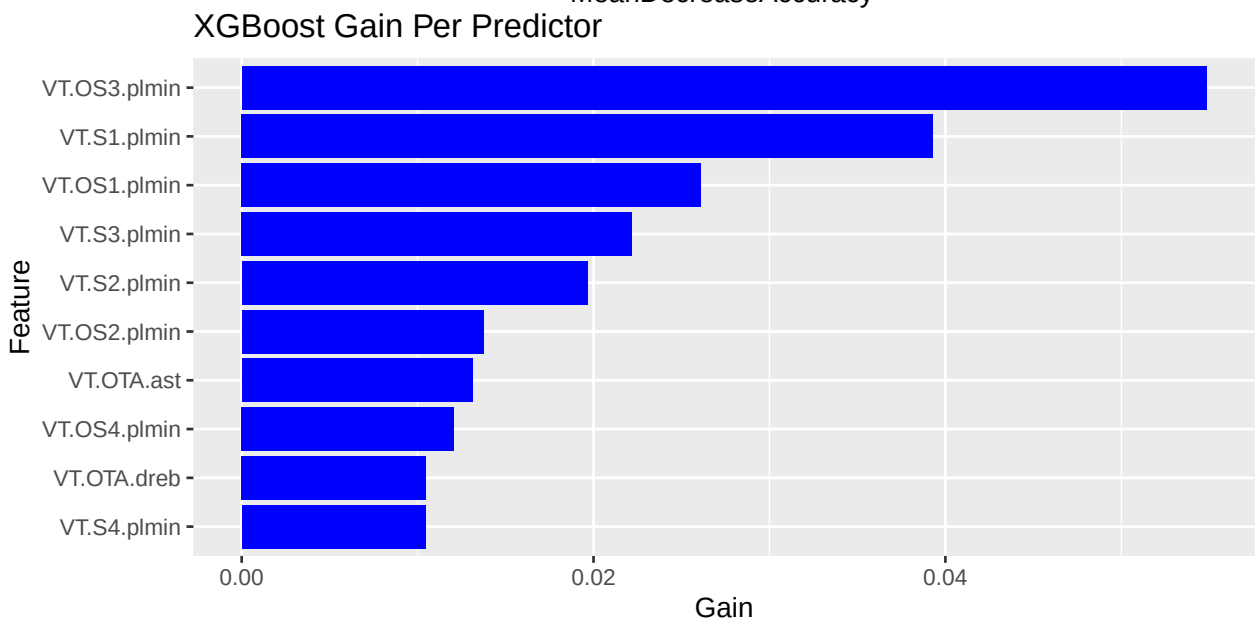
Random Forest Model:

model_forest <- randomForest(HTWins~., data=data.train, mtry=ceiling(sqrt(ncol(data.train)-1)), importance=TRUE)

XGBoost Model:

model_boost <- xgboost(data=data.train.x, label=data.train.y, objective="binary:logistic", nrounds=250, eta=0.09, max_depth=6, verbose=0, min_child_weight=1, gamma=0, subsample=1, colsample_bytree=1, early_stopping_rounds = 10, eval_metric="error", silent=1)





Conclusion and Next Steps

In conclusion, we did see that our random forest model performed best, both on the subsetted and full test data in Kaggle. However, we do believe with more time we would be able to further tune our XGBoost model to further improve our accuracy.

In the future, we would bring in external data that would provide us with the rankings for the basketball teams for each year. This would improve the model because team rankings would be strongly correlated with which team wins the game. We could look into creating new predictors from the variables available in our current data set. This would allow us to create a more concise and possibly more accurate model.