

Wearable fitness devices collect data that can predict proper weight lifting technique with 99.95% accuracy.

Executive Summary:

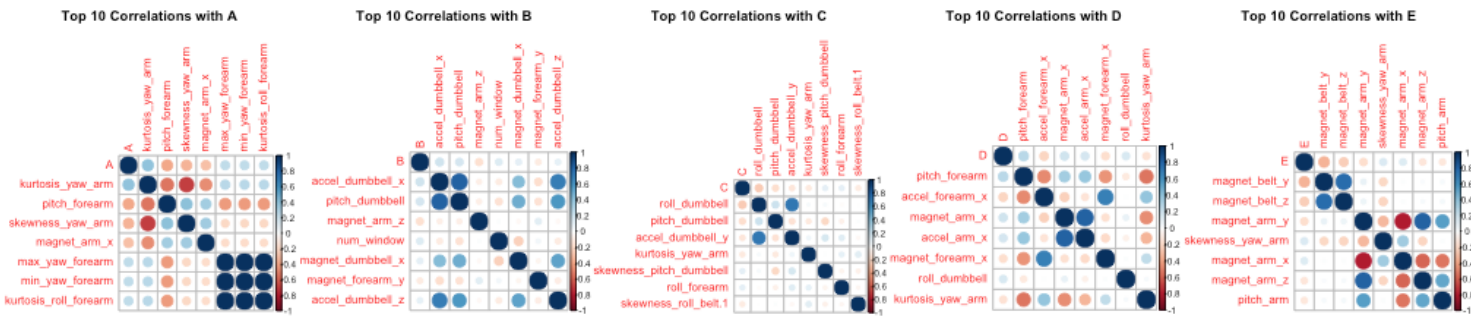
A Gradient Boosting Machine (gbm) is trained using wearable fitness device sensor data to predict whether weight lifting techniques are being performed properly. A trained model is able to achieve 99.95% accuracy predicting the five different outcomes (A = Good Technique; B,C,D,E = Bad Techniques).

Data Preparation Techniques:

- 1. 19,622 observations of 155 variables were loaded from the source
- 2. Row names (X), user_name, and all timestamps were removed as predictors to minimize overfitting
- 3. Training and Testing sets were created using a 70% training random sampling from the overall set
- 4. To improve efficiency of model generation without losing predictive capabilities, 77 variables with near zero variance were removed from the training set
- 5. KNN Imputation was used to impute values that were loaded as NA (in cases primarily where values were #DIV/0! and thus columns were converted to factors)
- 6. Because tree-based methods are a good fit for this multi-class non-linear classification problem, no other monotonic transformations were used

Exploration of Training Data:

Initial data exploration shows many features with significant correlations to each class (A, B, C, D, E):



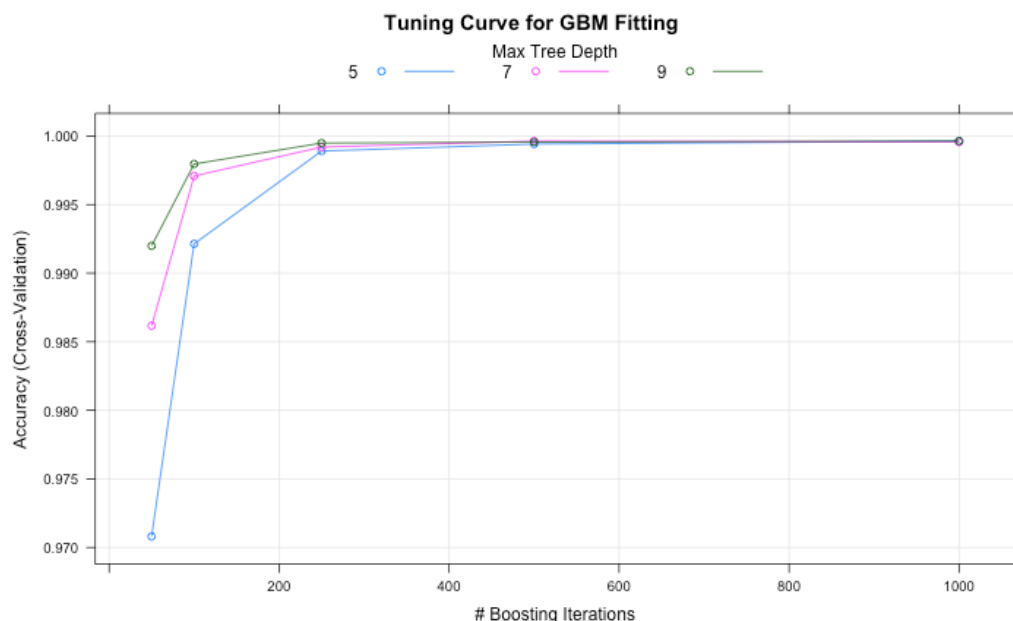
Training the Predictive Algorithms

Three different classification algorithms were tested with varying results: Decision Trees (rpart), Random Forests, and Gradient Boosting Machine (boosted trees). In each trained algorithm, repeated 10-fold cross-validation was performed to minimize overfitting while optimizing tuning parameters for maximum classification accuracy.

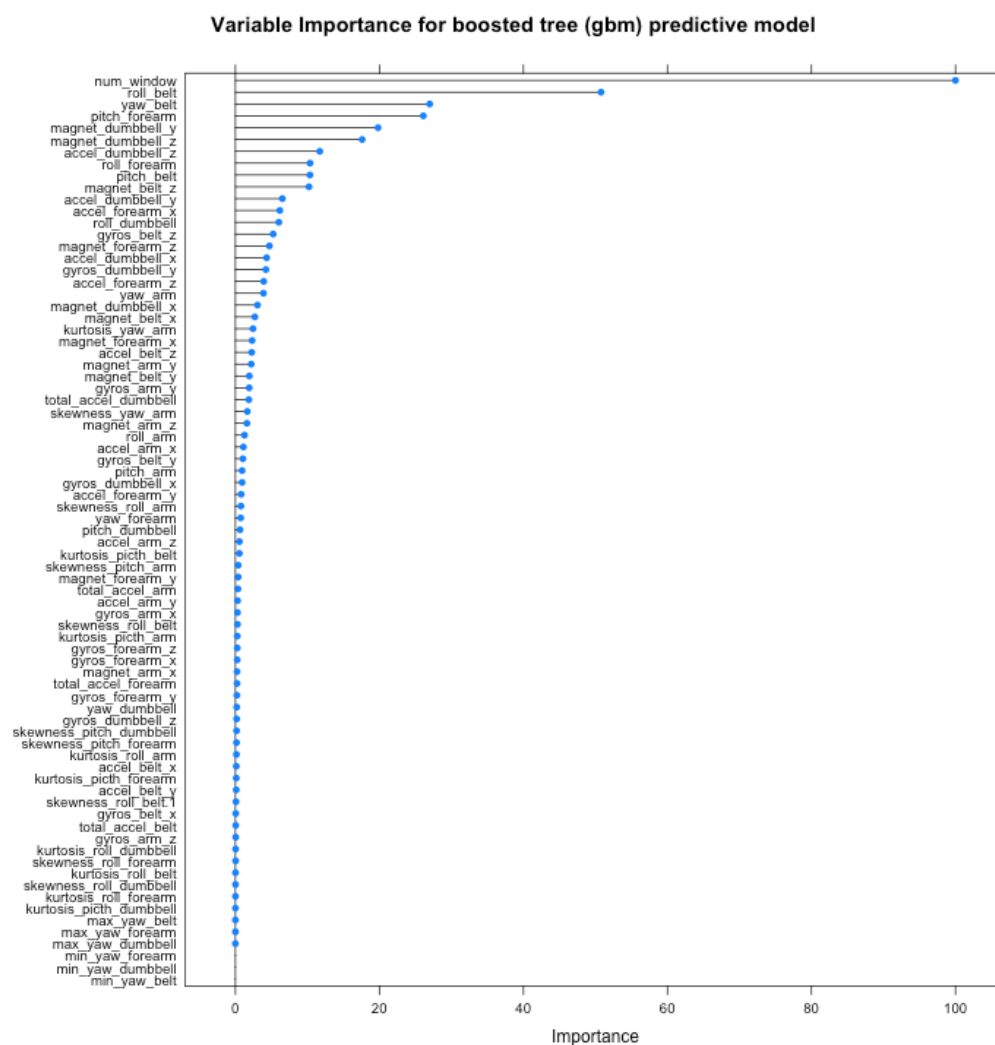
The best performing algorithm was created using **Gradient Boosting Machine (gbm)** package which combines boosting and trees to achieve an outstanding **99.95% overall class prediction accuracy** on test data.

Tuning the Algorithm for Maximum Accuracy

Through experimentation, excellent tuning parameters were found using 10-fold cross-validation with 10 repeats. These parameters include 500 trees with an interaction depth of 7 levels, shrinkage of 0.1, and 50 minimum observations in each leaf node.



Variables with non-zero effects in the final predictive model are plotted below (from most influential to least).



Results from the Testing Dataset: Confusion Matrix and Summary of Accuracy

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   A     B     C     D     E
```

```

##          A 1674      0      0      0      0
##          B      0 1139      0      0      1
##          C      0      0 1026      0      0
##          D      0      0      0 964      2
##          E      0      0      0      0 1079
##
## Overall Statistics
##
##          Accuracy : 0.9995
##          95% CI : (0.9985, 0.9999)
##          No Information Rate : 0.2845
##          P-Value [Acc > NIR] : < 2.2e-16
##
##          Kappa : 0.9994
##          McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##          Class: A Class: B Class: C Class: D Class: E
## Sensitivity      1.0000      1.0000      1.0000      1.0000      0.9972
## Specificity      1.0000      0.9998      1.0000      0.9996      1.0000
## Pos Pred Value    1.0000      0.9991      1.0000      0.9979      1.0000
## Neg Pred Value    1.0000      1.0000      1.0000      1.0000      0.9994
## Prevalence        0.2845      0.1935      0.1743      0.1638      0.1839
## Detection Rate    0.2845      0.1935      0.1743      0.1638      0.1833
## Detection Prevalence 0.2845      0.1937      0.1743      0.1641      0.1833
## Balanced Accuracy  1.0000      0.9999      1.0000      0.9998      0.9986

```