<https://topepo.github.io/caret/measuring-performance.html>

## Part I: Implementing a Simple Prediction Pipeline

#Question 1

```{r}

#Q1: Fit two prediction models using different subsets of the features in the training data. Features can overlap in the two models, but the feature sets should not be exactly the same across models. Clearly state which features were used in the two models.

#NTS: Now I'm tuning my hyperparameter using cross validation (through the train control function.) and doing a 3-fold cross-validation.

control.settings <- trainControl(method = "cv", number = 3)

#Now I'm running my train function

set.seed(1000)

model1\_again <- train(healthydays ~ gpaq8totmin + gpaq11days + bmi, data = class4, method = "glm", family = "gaussian", trControl = control.settings)

model1\_again

model2\_again <- train(healthydays ~ gpaq8totmin + gpaq11days + bmi + povertygroup + agegroup, data = class4, method = "glm", family = "gaussian", trControl = control.settings)

model2\_again

```

My first model includes 3 variables: BMI, gpaq8totmin(minutes of total physical activity on home chores on an average day) and gpaq11days (during the last 7 days, on how many days did you walk to get to and from places?). The root mean square error was 7.654.

My second model includes these three variables plus poverty group (is your household’s annual income from all source: 1 = <100%, 2 = 100-199%, 3 = 200-399%, 4 = 400-599%, 5 = 600%, 6 = Don’t know) and agegroup (group 1 = 18-24, group 2 = 25-44, group 3 = 45-64, and group 4 = 65+). The root mean square error was 7.444.