models_total_Candice

Candice Yu

2024-03-26

Contents

```r
library(caret)
library(earth)
library(tidyverse)
library(gridExtra)
```

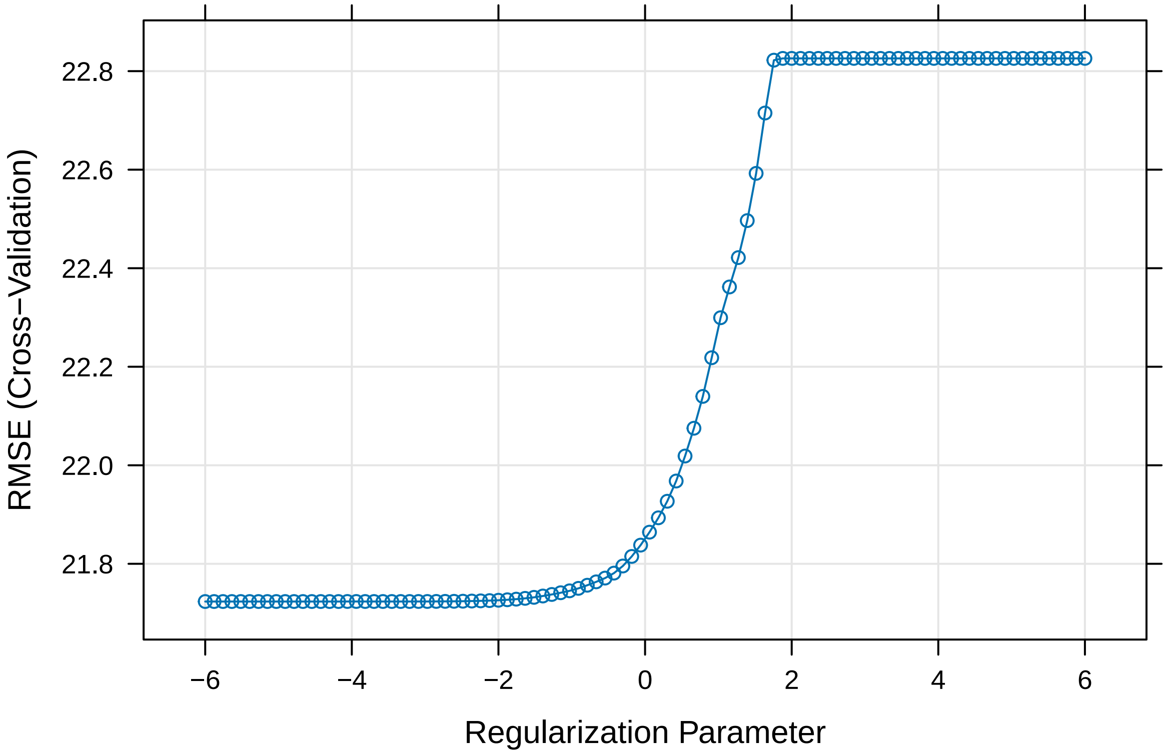
Load the training/test set & control method

```r
set.seed(2716)
# Load the training and test sets
train_data <- read.csv("./Data/train_data.csv")
test_data <- read.csv("./Data/test_data.csv")

# Load the control method
ctrl1 <- readRDS("./Data/train_control.rds")

# change variables to be factors again
train_data <- train_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         vaccine = as_factor(vaccine),
         race = as_factor(race),
         smoking = as_factor(smoking),
         severity = as_factor(severity))

test_data <- test_data %>%
  mutate(gender = as_factor(gender),
         diabetes = as_factor(diabetes),
         hypertension = as_factor(hypertension),
         race = as_factor(race),
         smoking = as_factor(smoking),
         vaccine = as_factor(vaccine),
         severity = as_factor(severity))
# matrix of predictors
x <- train_data %>% select(-recovery_time)
y <- train_data$recovery_time

x_test <- test_data %>% select(-recovery_time)
y_test <- test_data$recovery_time
```

Model Training: Linear models

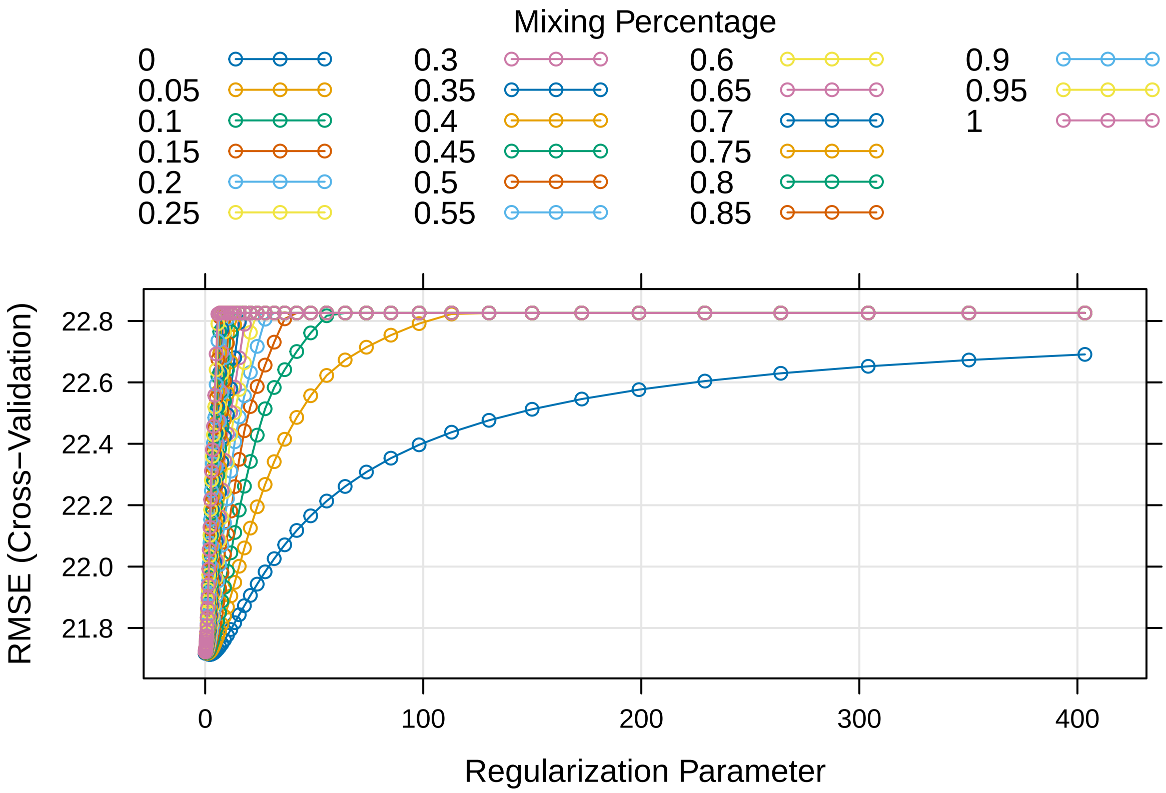
Lasso Regression Model

```r
set.seed(2716)
lasso_grid <- expand.grid(
  alpha = 1,
  lambda = exp(seq(-6, 6, length.out = 100))
)

lasso_fit <- train(x, y,
  method = "glmnet",
  tuneGrid = lasso_grid,
  trControl = ctrl1
)

plot(lasso_fit, xTrans = log)
```



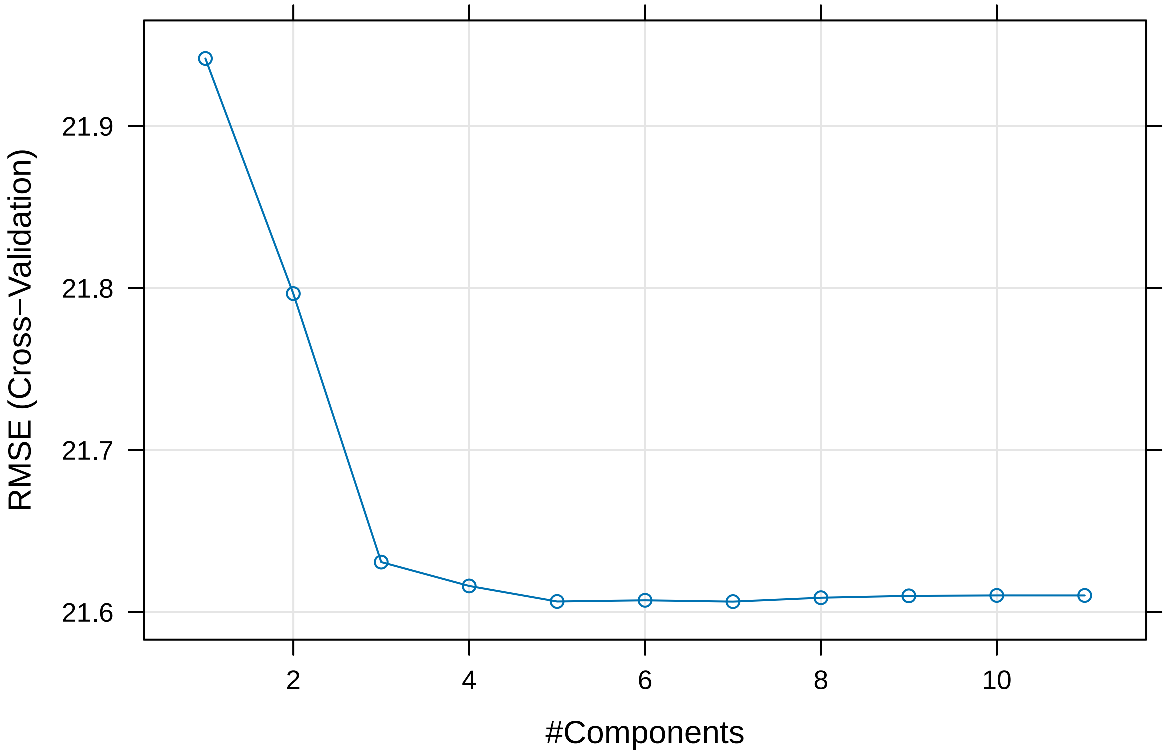
Elastic Net Model

```
set.seed(2716)
enet_grid <- expand.grid(
  alpha = seq(0, 1, length.out = 21),
  lambda = exp(seq(-8, 6, length.out = 100))
)

enet_fit <- train(x, y,
                  method = "glmnet",
                  tuneGrid = enet_grid,
                  trControl = ctrl1,
                  preProcess = c("center", "scale")
)
plot(enet_fit)
```



Partial Least Squares

```
set.seed(2716)
pls_fit <- train(x, y,
                 method = "pls",
                 tuneLength = 20,
                 trControl = ctrl1,
                 preProcess = c("center", "scale")
                 )
plot(pls_fit)
```



Evaluate the performance of linear models

```
lasso_pred <- predict(lasso_fit, newdata = x_test)
enet_pred <- predict(enet_fit, newdata = x_test)
pls_pred <- predict(pls_fit, newdata = x_test)

lasso_performance <- postResample(pred = lasso_pred, obs = test_data$recovery_time)
lasso_performance
```

```
##       RMSE   Rsquared        MAE
## 21.6711297  0.1592601 12.5830955
```

```
enet_performance <- postResample(pred = enet_pred, obs = test_data$recovery_time)
enet_performance
```

```
##       RMSE   Rsquared        MAE
## 21.7323312  0.1589219 12.5660278
```

```
pls_performance <- postResample(pred = pls_pred, obs = test_data$recovery_time)
pls_performance
```

```
##       RMSE   Rsquared        MAE
## 21.5903221  0.1631082 12.8407993
```

Model Training: Nonlinear Methods

The EDA plots show that the relationship between predictors and recovery time is likely non-linear, and there may be interactions between variables, especially considering the difference between study groups A and B.

Given the results from the EDA plots and the nature of the data, both generalized additive models (GAM) and multivariate adaptive regression splines (MARS) could be suitable choices for modeling. They both are capable of modeling complex, non-linear relationships in the data.

Multivariate Adaptive Regression Spline (MARS)

Build the MARS model

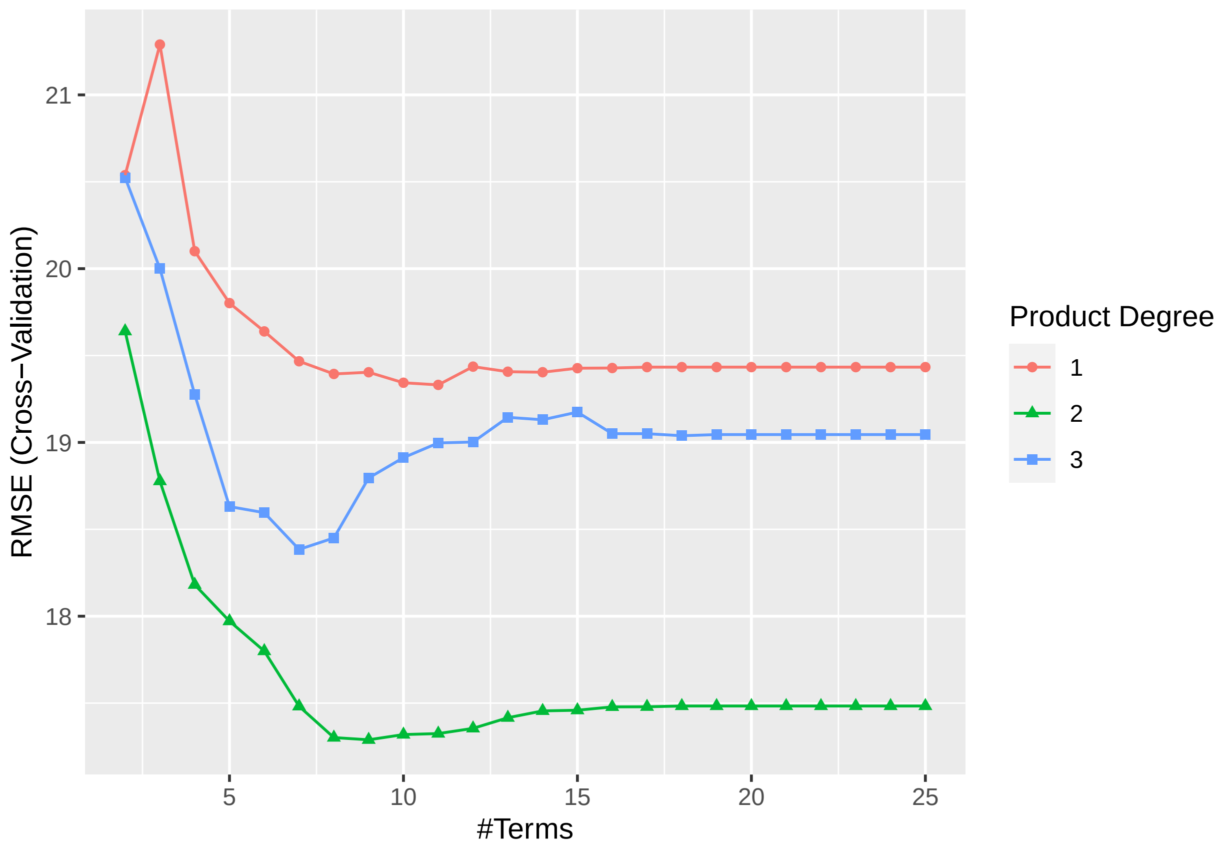
```
# train the MARS model
mars_grid <- expand.grid(degree = 1:3, nprune = 2:25)

set.seed(2716) # set the same seed
mars_fit <- train(x, y,
                    method = "earth",
                    tuneGrid = mars_grid,
                    trControl = ctrl1)
```

MARS Model Summary

```
# Model summary
summary(mars_fit)
```

```
## Call: earth(x=data.frame[2402,12], y=c(44,29,40,31,5...), keepxy=TRUE,
##             degree=2, nprune=9)
##
##                       coefficients
## (Intercept)               9.358604
## gender1                  -3.292896
## hypertension1             3.841507
## vaccine1                 -5.668675
## h(bmi-25.1)               6.146988
## h(30.5-bmi)               5.345431
## severity1 * studyB       16.973744
## h(bmi-30.5) * studyB     13.128169
## h(bmi-35.2) * studyB    198.572004
##
## Selected 9 of 21 terms, and 6 of 15 predictors (nprune=9)
## Termination condition: Reached nk 31
## Importance: bmi, studyB, severity1, vaccine1, hypertension1, gender1, ...
## Number of terms at each degree of interaction: 1 5 3
## GCV 290.93    RSS 686648.1    GRSq 0.4529171    RSq 0.4619934
```

```
ggplot(mars_fit)
```



```
mars_fit$bestTune
```

```
##    nprune degree
## 32      9      2
```

```
coef(mars_fit$finalModel)
```

```
##          (Intercept)          h(30.5-bmi) h(bmi-30.5) * studyB
##             9.358604             5.345431            13.128169
##         h(bmi-25.1) h(bmi-35.2) * studyB             vaccine1
##             6.146988           198.572004            -5.668675
##        hypertension1              gender1   severity1 * studyB
##             3.841507            -3.292896            16.973744
```

MARS Model Description:

The MARS model is a flexible regression method capable of uncovering complex nonlinear relationships between the dependent variable (recovery_time) and a set of independent variables. It does this by fitting piecewise linear regressions, which can adapt to various data shapes. This is particularly useful for modeling the recovery time from COVID-19 since the relationship between predictors and recovery time could be highly nonlinear and interaction-heavy.

Assumptions:

- The relationships between predictors and the response can be captured using piecewise linear functions.
- Interactions between variables can be important and are modeled by products of basis functions.
- There is no assumption of a parametric form of the relationship between predictors and the response.

Final Model Selection:

- The optimal hyperparameters were degree (degree of interaction) = 3 and nprune (number of terms)

= 16.

- The selected model terms involve interactions between patient characteristics, their biometrics, the specific study group they belong to, and some non-linear transformations of these variables.

Evaluate performance on the test set

```
# Evaluate its performance on the test set:
predictions <- predict(mars_fit, newdata = test_data)
postResample(pred = predictions, obs = test_data$recovery_time)
```

```
##       RMSE   Rsquared        MAE
## 34.8585974  0.3524154 12.9107465
```

The results from evaluating the MARS model on the test set provide three key metrics:

1. **Root Mean Squared Error (RMSE):** RMSE measures the average magnitude of the prediction error. It represents the square root of the average squared differences between the predicted and actual values. An RMSE of 19.629 suggests that, on average, the model's predictions of the recovery time are about 19.629 days off from the actual recovery times.
2. **R-squared (R^2):** R^2 is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. In your case, the R^2 value is 0.2177, which means approximately 21.77% of the variance in the recovery time is explained by the model. This is a relatively low value, indicating that there is a lot of variability in the recovery time that is not captured by the model.
3. **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values, providing a linear score that reflects the average error magnitude without considering its direction. An MAE of 12.409 suggests that the model's predictions are, on average, 12.409 days different from the actual recovery time.

Interpretation

- The **RMSE** of 19.629 days is relatively high, depending on the context of the recovery times' range. If the typical recovery time is on the order of a few days, this is a substantial error. However, if recovery times are generally several weeks, the error may be more acceptable.
- The **R-squared** value of 0.2177 is not very high, suggesting that there might be other factors not included in the model that affect the recovery time. It also indicates that the relationship between the predictors and the recovery time has a significant amount of unexplained variability.
- The **MAE** gives us an indication that, despite the direction of the errors, the model's predictions are off by about two weeks on average. MAE is less sensitive to outliers than RMSE, so this value suggests that the model has a consistent average error across the test dataset.

GAM model

Build the GAM model

```
set.seed(2716) # set the same seed
gam_fit <- train(x = x, y = y,
                 method = "gam",
                 trControl = ctrl1)
```

Display the summary of the final model

```r
gam_model_final <- gam_fit$finalModel
summary(gam_model_final)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender + hypertension + diabetes + vaccine + severity +
##     study + smoking + race + s(age) + s(SBP) + s(LDL) + s(bmi)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.45538    0.96899  43.814  < 2e-16 ***
## gender1       -3.32709    0.77911  -4.270 2.03e-05 ***
## hypertension1  3.73230    0.78134   4.777 1.89e-06 ***
## diabetes1     -2.12159    1.09065  -1.945   0.0519 .
## vaccine1      -6.15159    0.79351  -7.752 1.33e-14 ***
## severity1      8.47963    1.24622   6.804 1.28e-11 ***
## studyB         4.90825    0.82831   5.926 3.56e-09 ***
## smoking1       2.23808    0.88018   2.543   0.0111 *
## smoking2       3.02769    1.30043   2.328   0.0200 *
## race2          1.11665    1.74207   0.641   0.5216
## race3          0.06618    0.99249   0.067   0.9468
## race4         -1.12042    1.42608  -0.786   0.4321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df       F p-value
## s(age) 8.829e-07      9   0.000   0.736
## s(SBP) 4.201e-07      9   0.000   0.420
## s(LDL) 1.296e-01      9   0.016   0.285
## s(bmi) 8.924e+00      9 104.072  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.321   Deviance explained = 32.7%
## GCV = 364.15  Scale est. = 360.96    n = 2402
```

Evaluate the GAM model's performance

```r
test_predictions <- predict(gam_model_final, x_test)
postResample(pred = test_predictions, obs = test_data$recovery_time)
```

```
##       RMSE   Rsquared        MAE
## 21.5949142  0.3554629 13.4168111
```

Random Forest

Build the rf model

```r
set.seed(2716)
# Parameters for Random Forest training
tunegrid <- expand.grid(mtry = 1:5)

# build the rf model
rf_fit <- train(
   x = x, y = y,
   method = "rf",
   trControl = ctrl1,
   tuneGrid = tunegrid
 )
rf_model_final <- rf_fit$finalModel
```

Evaluate the rf model's performance

```r
# Calculate and print the RMSE for training and test datasets
rf_predictions <- predict(rf_model_final, x_test)
postResample(pred = rf_predictions, obs = test_data$recovery_time)
```

```
##       RMSE   Rsquared        MAE
## 18.1800886  0.4208812 12.1826118
```

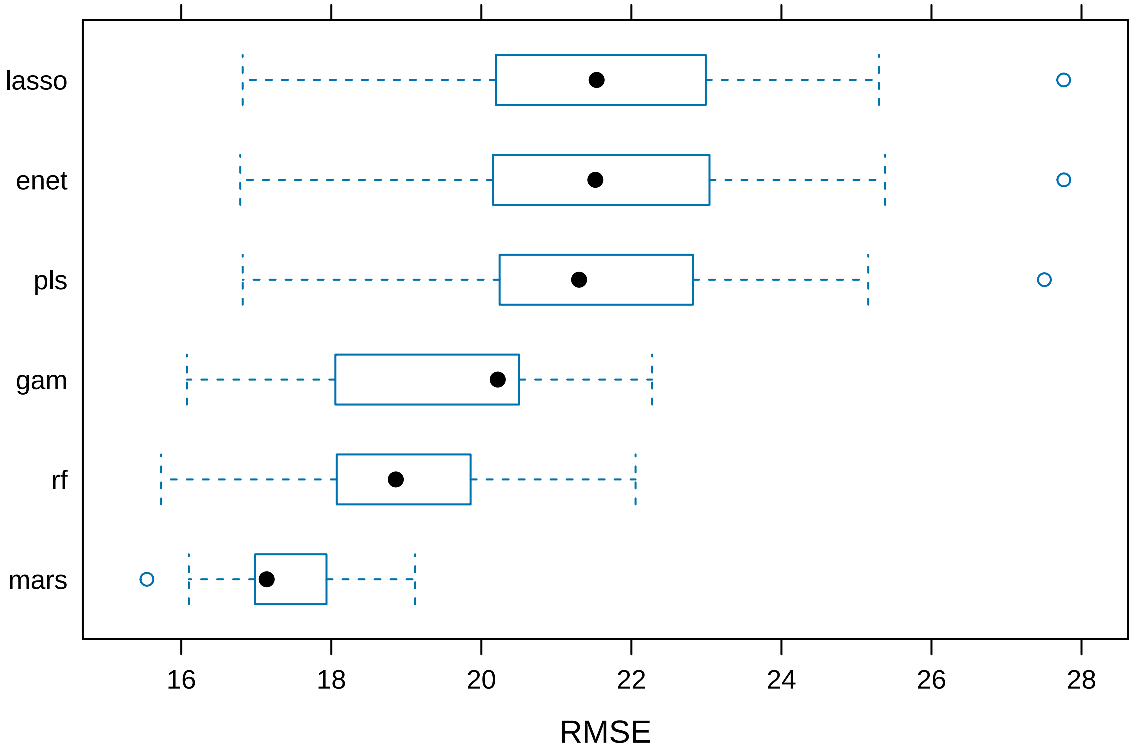
Model Comparison

```r
set.seed(2716)
resamp =
  resamples(list(lasso = lasso_fit,
                 gam = gam_fit,
                 enet = enet_fit,
                 pls = pls_fit,
                 mars = mars_fit,
                 rf = rf_fit))
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: lasso, gam, enet, pls, mars, rf
## Number of resamples: 10
##
## MAE
##           Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lasso 11.61557 13.07002 13.39915 13.38478 13.90405 14.52028    0
## gam   11.42544 12.50332 12.78923 12.69601 12.95004 13.66259    0
## enet  11.53360 13.00630 13.31441 13.30771 13.83171 14.44532    0
## pls   11.78212 13.14454 13.49967 13.47720 13.88119 14.76913    0
## mars  10.99323 11.68725 11.85075 11.79140 12.04602 12.23483    0
## rf    11.06627 11.66679 12.22921 12.08414 12.37878 12.80023    0
##
## RMSE
##           Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lasso 16.81778 20.27198 21.53662 21.72343 22.69250 27.76260    0
## gam   16.07328 18.19815 20.21817 19.45737 20.49510 22.27783    0
## enet  16.78687 20.24242 21.51955 21.71395 22.73373 27.76476    0
## pls   16.81815 20.27157 21.30276 21.60646 22.53514 27.50506    0
## mars  15.54181 17.01221 17.13794 17.28887 17.81367 19.11708    0
## rf    15.73258 18.20358 18.85918 18.79321 19.63882 22.05531    0
##
## Rsquared
##             Min.    1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.04406941 0.07134251 0.1143569 0.1023988 0.1262758 0.1725520    0
## gam   0.16842839 0.21972879 0.2804230 0.3024507 0.3985251 0.4775290    0
## enet  0.04422923 0.07102989 0.1137048 0.1026903 0.1267357 0.1741316    0
## pls   0.04558123 0.07561878 0.1255099 0.1125534 0.1369174 0.1791608    0
## mars  0.18856074 0.28716270 0.3861960 0.4071805 0.5113954 0.6510990    0
## rf    0.16870684 0.25535301 0.3215659 0.3405483 0.4240297 0.5263780    0
```

Using bw-plot to compare their RMSE

```r
bwplot(resamp, metric = "RMSE")
```


lasso
enet
pls
gam
rf
mars
RMSE