hw3

Johnstone Tcheou

2025-03-24

Contents


```r
library(tidyverse)
library(caret)
library(glmnet)
library(mlbench)
library(tidymodels)
library(pROC)
library(pdp)
library(vip)
library(MASS)
library(earth)
library(plotmo)
library(car)
```

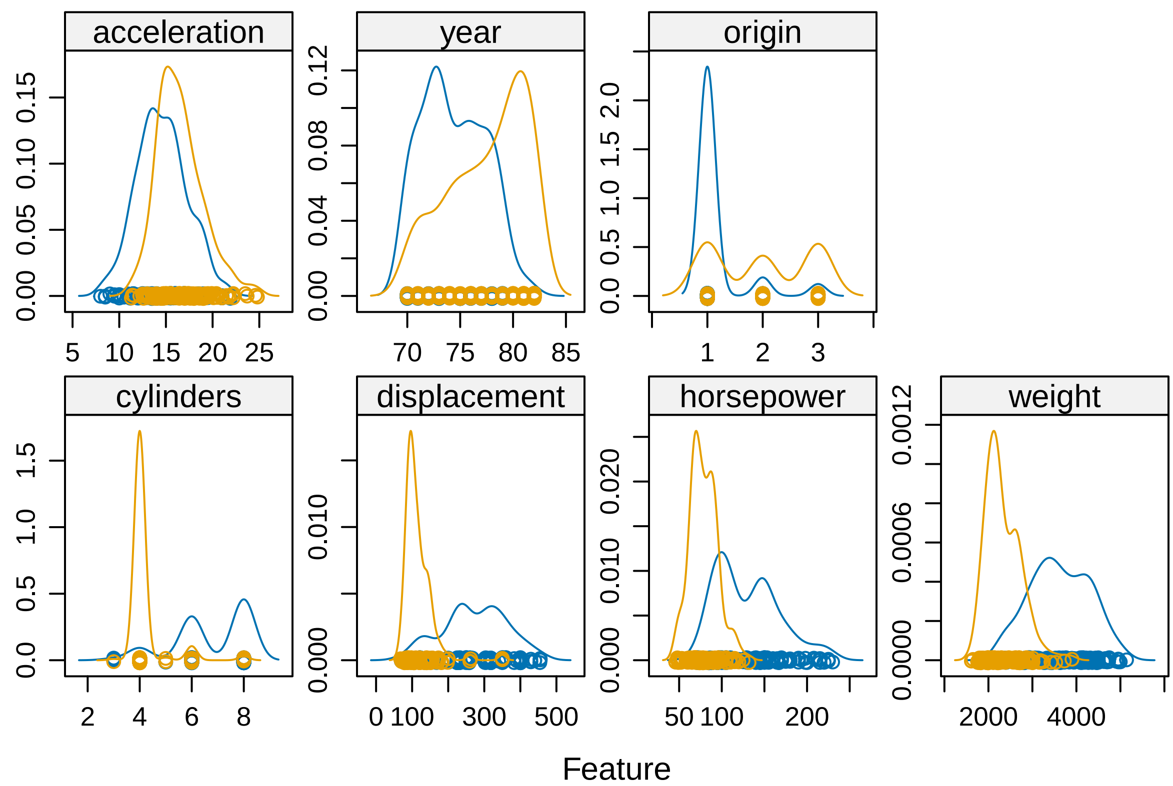
Data import, exploration, and split

Since there are no `NA` observations, we do not need `na.omit`. We do need to coerce the response variable `mpg_cat` to be a factor before visually exploring the data with `featurePlot` prior to any model fitting. Since most of the predictors are continuous, we can use density plots to best visualize their distributions stratified by levels of the response variable with y axes scaled to each predictor.

```r
set.seed(81063)

auto <-
  read.csv("auto.csv") |>
  na.omit() |>
  mutate(
    mpg_cat = factor(mpg_cat, levels = c("low", "high"))
  )

featurePlot(
  x = auto[, 1:7],
  y = auto$mpg_cat,
  scales = list(x = list(relation = "free"),
                        y = list(relation = "free")),
            plot = "density"
  )
```



When stratified by `mpg_cat`, most variables have pretty different distributions, except for `acceleration`. These may indicate potential variable informativeness towards predicting `mpg_cat`.

```r
set.seed(81063)
auto_split <- initial_split(auto, prop = 0.70)

training_data <- training(auto_split)
testing_data <- testing(auto_split)

training_predictors <- training_data[, -ncol(training_data)]
training_response <- training_data$mpg_cat

testing_predictors <- testing_data[, -ncol(testing_data)]
testing_response <- testing_data$mpg_cat
```

Question a

Logistic regression

We can use the `contrasts` function to ensure we are using the correct predictor labels. Afterwards, we can fit a logistic regression model to the training data and get predicted probabilities with the testing data to evaluate the model.

```r
set.seed(81063)
contrasts(auto$mpg_cat)
```

```
##      high
## low     0
## high    1
```

```r
ctrl <- trainControl(
  method = "cv",
  number = 10,
  summaryFunction = twoClassSummary,
  classProbs = TRUE
)

logit <- train(
  x = training_data[, -ncol(training_data)],
  y = training_response,
  method = "glm",
  metric = "ROC",
  trControl = ctrl
)

summary(logit)
```

```
##
## Call:
## NULL
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -13.883316   7.133612  -1.946  0.05163 .
## cylinders     -0.092058   0.525410  -0.175  0.86091
## displacement   0.005974   0.014922   0.400  0.68891
## horsepower    -0.061502   0.031976  -1.923  0.05443 .
## weight        -0.004089   0.001411  -2.898  0.00376 **
## acceleration  -0.125432   0.188102  -0.667  0.50488
## year           0.416420   0.090016   4.626 3.73e-06 ***
## origin         0.695208   0.467909   1.486  0.13734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.48  on 273  degrees of freedom
## Residual deviance: 109.21  on 266  degrees of freedom
## AIC: 125.21
##
## Number of Fisher Scoring iterations: 8
```

```r
coef(logit$finalModel)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -13.883315553  -0.092057931   0.005973898  -0.061502176  -0.004089208
##  acceleration          year        origin
##  -0.125432046   0.416420203   0.695208015
```

```r
logit_pred_prob <- predict(
  logit,
  newdata = testing_data,
  type = "raw"
)

(logit_confusion_matrix <- confusionMatrix(
  data = logit_pred_prob,
  reference = testing_data$mpg_cat,
  positive = "high"
))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction low high
##       low   46    3
##       high   8   61
##
##                Accuracy : 0.9068
##                  95% CI : (0.8393, 0.9525)
##     No Information Rate : 0.5424
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8108
##
##  Mcnemar's Test P-Value : 0.2278
```

```
##
##             Sensitivity : 0.9531
##             Specificity : 0.8519
##          Pos Pred Value : 0.8841
##          Neg Pred Value : 0.9388
##              Prevalence : 0.5424
##          Detection Rate : 0.5169
##    Detection Prevalence : 0.5847
##       Balanced Accuracy : 0.9025
##
##        'Positive' Class : high
##
```

```
# logit <- glm(
#   mpg_cat ~ .,
#   data = training_data,
#   family = binomial(link = "logit")
# )
#
# logit_pred_prob <- predict(logit, newdata = testing_data, type = "response")
# logit_pred <- rep("low", length(logit_pred_prob))
# logit_pred[logit_pred_prob > 0.5] <- "high"
#
#
# confusionMatrix(data = factor(logit_pred, levels = c("low", "high")),
#                 reference = testing_data$mpg_cat,
#                 positive = "high")
#
# roc_logit <- roc(training$mpg_cat, logit_pred_prob)
#
# plot(roc_logit, legacy.axes = TRUE, print.auc = TRUE)
# #plot(smooth(roc_logit), col = 4, add = TRUE)
#
# vip(logit)
```

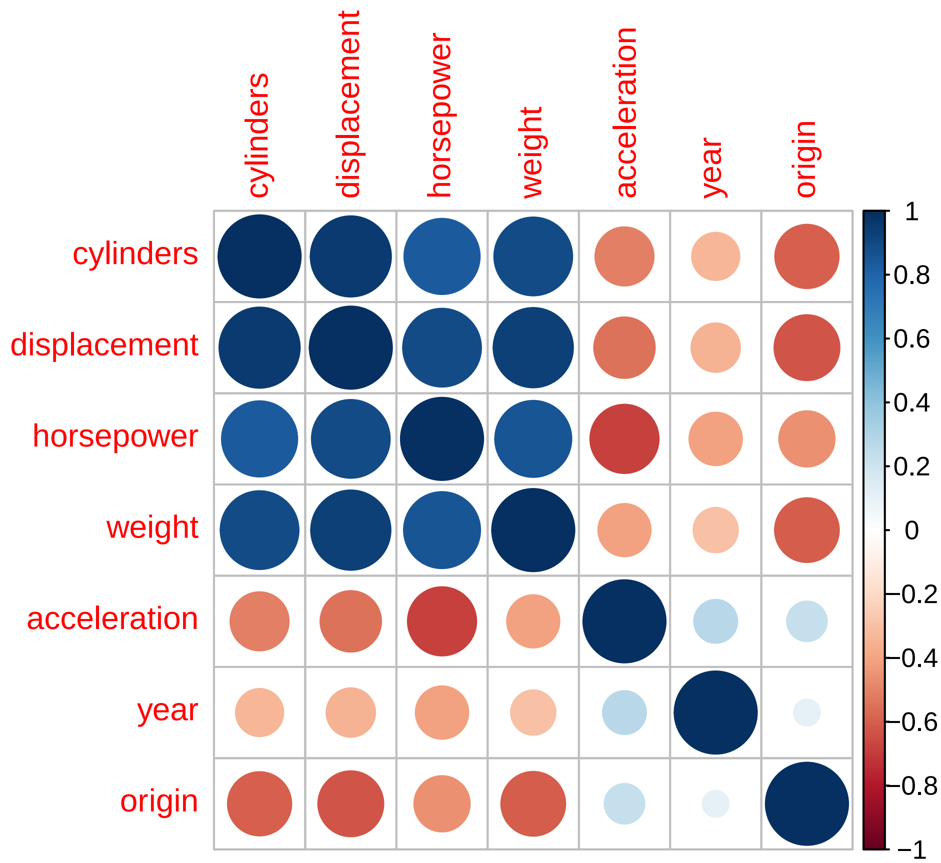
The fitted logistic regression model has 7 predictors, for `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, and `origin`. When estimating predictions against the training dataset, we can get the confusion matrix to assess the robustness of the model's classification.

With an accuracy of 0.9067797, it is greater than the no information rate, which means that this classifier is meaningful. Additionally, the kappa is 0.8108423. Being greater than 0.6, it indicates good agreement.

Are there redundant predictors in your model?

Per section 3.3.3, potential problems, in the Introduction to Statistical Learning textbook, **redundancy is when there is multicollinearity between predictors, so the information that one predictor provides is redundant when the other collinear predictor already provides that information**. Multicollinearity can be visualized in a correlation plot, and it can be quantified using the variable inflation factor, which is the ratio of the variance of the predictor in the saturated model to the variance of that same predictor in a univariate model. A VIF of 5 or 10 and above indicates multicollinearity, and can be calculated with `vif()` from the `car` package.

```
corrplot::corrplot(
  cor(model.matrix(mpg_cat ~ ., training_data)[,-1]),
  type = "full"
)
```




```
vif(logit$finalModel)
```

```
##     cylinders displacement   horsepower       weight acceleration         year
##      5.546834    10.662354     3.662072     5.468094     3.465300     1.713242
##        origin
##      2.133171
```

As illustrated in the above correlation plot, the predictors are pretty heavily correlated with each other. When looking at their VIFs, `cylinders` and `weight` have VIFs > 5 and `displacement` has a VIF > 10. These would be predictors are redundant in our final logistic regression model.

question b

MARS model

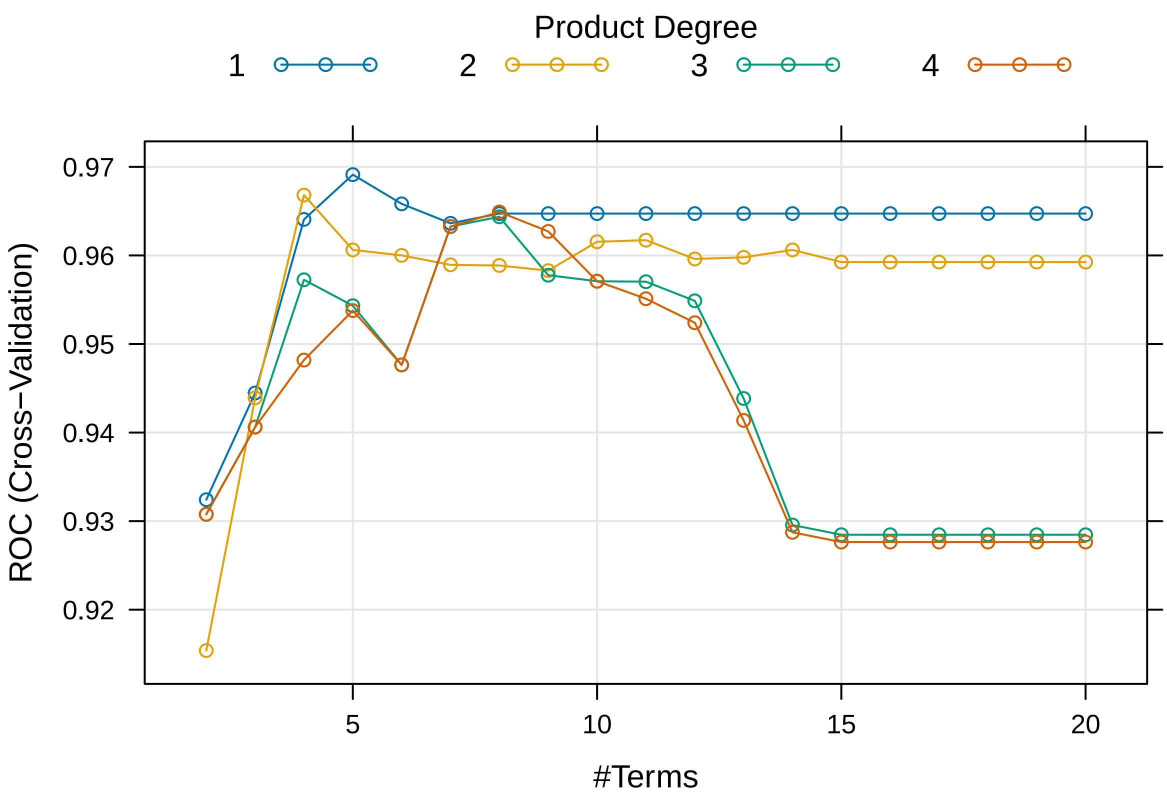
Next, we can fit a MARS model to the training data, passing the `preProcess` argument `"scale"` to scale the data.

Worth noting, **fitting MARS model gets Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred**. This should be okay in our case, since our predicted outcome is binary anyways, with levels of `low` or `high`.

```
set.seed(81063)

mars <- train(
  x = training_data[1:7],
  y = training_data$mpg_cat,
  method = "earth",
  tuneGrid = expand.grid(
    degree = 1:4,
    nprune = 2:20
  ),
  preProcess = "scale",
  metric = "ROC",
  trControl = ctrl
)

plot(mars)
```




```
coef(mars$finalModel)
```

```
##           (Intercept)      h(year-19.4294)      h(4.015-weight)
##             -4.561637             1.949816             4.060890
```

```
## h(displacement-1.40162) h(displacement-2.15348)
##               -4.389948               6.293828
```

```
mars_pred <- predict(
  mars,
  newdata = training_data
)

(mars_confusion_matrix <- confusionMatrix(
  data = mars_pred,
  reference = training_data$mpg_cat,
  positive = "high"
))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction low high
##       low  127    8
##       high  15  124
##
##                Accuracy : 0.9161
##                  95% CI : (0.8767, 0.946)
##     No Information Rate : 0.5182
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8322
##
##  Mcnemar's Test P-Value : 0.2109
##
##             Sensitivity : 0.9394
##             Specificity : 0.8944
##          Pos Pred Value : 0.8921
##          Neg Pred Value : 0.9407
##              Prevalence : 0.4818
##          Detection Rate : 0.4526
##    Detection Prevalence : 0.5073
##       Balanced Accuracy : 0.9169
##
##        'Positive' Class : high
##
```

Does the MARS model improve prediction performance compared to logistic regression?

Yes, it does. For one thing, the accuracy is higher - 0.9160584 compared to 0.9067797. Secondly, the kappa is also much higher - 0.8322062 vs 0.8108423, indicating great agreement. However, the ROC AUC is slightly lower, with the best fit MARS with the highest ROC had an ROC (`nprune` = 12, `degree` = 2) of 0.9691209 compared to 0.9703715.

```
mars
```

```
## Multivariate Adaptive Regression Spline
```

```
##
## 274 samples
##   7 predictor
##   2 classes: 'low', 'high'
##
## Pre-processing: scaled (7)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 247, 247, 245, 247, 247, 247, ...
## Resampling results across tuning parameters:
##
##   degree  nprune  ROC        Sens       Spec
##   1        2      0.9324176  0.8519048  0.9104396
##   1        3      0.9444689  0.8519048  0.9395604
##   1        4      0.9640659  0.8947619  0.9472527
##   1        5      0.9691209  0.9076190  0.9318681
##   1        6      0.9658242  0.9076190  0.9318681
##   1        7      0.9636264  0.9076190  0.9318681
##   1        8      0.9647253  0.9147619  0.9318681
##   1        9      0.9647253  0.9147619  0.9241758
##   1       10      0.9647253  0.9147619  0.9241758
##   1       11      0.9647253  0.9147619  0.9241758
##   1       12      0.9647253  0.9147619  0.9241758
##   1       13      0.9647253  0.9147619  0.9241758
##   1       14      0.9647253  0.9147619  0.9241758
##   1       15      0.9647253  0.9147619  0.9241758
##   1       16      0.9647253  0.9147619  0.9241758
##   1       17      0.9647253  0.9147619  0.9241758
##   1       18      0.9647253  0.9147619  0.9241758
##   1       19      0.9647253  0.9147619  0.9241758
##   1       20      0.9647253  0.9147619  0.9241758
##   2        2      0.9153846  0.8161905  0.8873626
##   2        3      0.9439194  0.8519048  0.9472527
##   2        4      0.9668132  0.8876190  0.9620879
##   2        5      0.9606227  0.9080952  0.9401099
##   2        6      0.9600026  0.9076190  0.9324176
##   2        7      0.9589377  0.8938095  0.9543956
##   2        8      0.9588645  0.9290476  0.9467033
##   2        9      0.9582810  0.9290476  0.9472527
##   2       10      0.9615411  0.9295238  0.9395604
##   2       11      0.9617216  0.9223810  0.9472527
##   2       12      0.9595997  0.9157143  0.9472527
##   2       13      0.9597828  0.9157143  0.9395604
##   2       14      0.9606253  0.9157143  0.9395604
##   2       15      0.9592517  0.9157143  0.9395604
##   2       16      0.9592517  0.9157143  0.9395604
##   2       17      0.9592517  0.9157143  0.9395604
##   2       18      0.9592517  0.9157143  0.9395604
##   2       19      0.9592517  0.9157143  0.9395604
##   2       20      0.9592517  0.9157143  0.9395604
##   3        2      0.9307692  0.8519048  0.9181319
##   3        3      0.9406227  0.8519048  0.9472527
##   3        4      0.9572527  0.8800000  0.9406593
##   3        5      0.9543223  0.8871429  0.9549451
##   3        6      0.9476374  0.8938095  0.9395604
```

```
##   3         7      0.9632601  0.8938095  0.9318681
##   3         8      0.9643590  0.9009524  0.9164835
##   3         9      0.9577656  0.9080952  0.9164835
##   3        10      0.9570879  0.9080952  0.9010989
##   3        11      0.9570330  0.9080952  0.9087912
##   3        12      0.9548718  0.9080952  0.9087912
##   3        13      0.9438462  0.8938095  0.9087912
##   3        14      0.9295604  0.9009524  0.8934066
##   3        15      0.9284615  0.9009524  0.8934066
##   3        16      0.9284615  0.9009524  0.8934066
##   3        17      0.9284615  0.9009524  0.8934066
##   3        18      0.9284615  0.9009524  0.8934066
##   3        19      0.9284615  0.9009524  0.8934066
##   3        20      0.9284615  0.9009524  0.8934066
##   4         2      0.9307692  0.8519048  0.9181319
##   4         3      0.9406227  0.8519048  0.9472527
##   4         4      0.9481868  0.8800000  0.9406593
##   4         5      0.9537729  0.8871429  0.9549451
##   4         6      0.9476374  0.8938095  0.9395604
##   4         7      0.9632601  0.8938095  0.9318681
##   4         8      0.9649084  0.9009524  0.9164835
##   4         9      0.9627106  0.9080952  0.9164835
##   4        10      0.9570879  0.9080952  0.9010989
##   4        11      0.9551099  0.9080952  0.9087912
##   4        12      0.9523993  0.9080952  0.9087912
##   4        13      0.9413736  0.8938095  0.9087912
##   4        14      0.9287363  0.9009524  0.8934066
##   4        15      0.9276374  0.9009524  0.8934066
##   4        16      0.9276374  0.9009524  0.8934066
##   4        17      0.9276374  0.9009524  0.8934066
##   4        18      0.9276374  0.9009524  0.8934066
##   4        19      0.9276374  0.9009524  0.8934066
##   4        20      0.9276374  0.9009524  0.8934066
## 
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were nprune = 5 and degree = 1.
```

```r
mars$bestTune
```

```
##   nprune degree
## 4      5      1
```

```r
max(mars$results$ROC)
```

```
## [1] 0.9691209
```

Question c

Linear discriminant analysis

We can also fit the data with linear discriminant analysis.

```r
set.seed(81063)

lda <- train(
  x = training_predictors,
  y = training_response,
  method = "lda",
  metric = "ROC",
  trControl = ctrl
)

lda$results$ROC
```

```
## [1] 0.9496337
```

```r
lda_pred <- predict(
  lda,
  newdata = training_data,
  type = "raw"
)

(lda_confusion_matrix <- confusionMatrix(
  data = lda_pred,
  reference = training_data$mpg_cat,
  positive = "high"
))
```

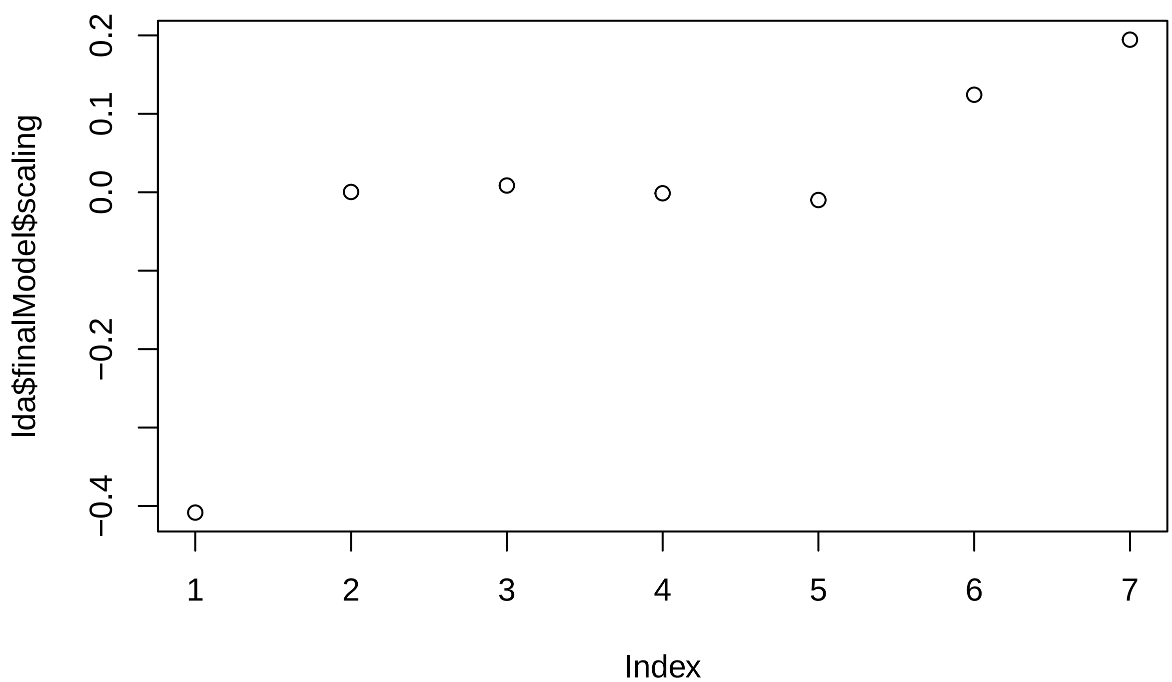
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction low high
##       low  122    5
##       high  20  127
##
##                Accuracy : 0.9088
##                  95% CI : (0.8683, 0.9401)
##     No Information Rate : 0.5182
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.818
##
##  Mcnemar's Test P-Value : 0.00511
##
##             Sensitivity : 0.9621
##             Specificity : 0.8592
##          Pos Pred Value : 0.8639
##          Neg Pred Value : 0.9606
##              Prevalence : 0.4818
##          Detection Rate : 0.4635
##    Detection Prevalence : 0.5365
##       Balanced Accuracy : 0.9106
##
##        'Positive' Class : high
##
```

Getting predictions against the training dataset again, we see that the LDA model has an accuracy of 0.9087591. It also has a good kappa of 0.8180031. Though a good model, it still does not have as high agreement or accuracy as MARS, and also has a lower ROC with 0.9496337.

Plot the linear discriminants

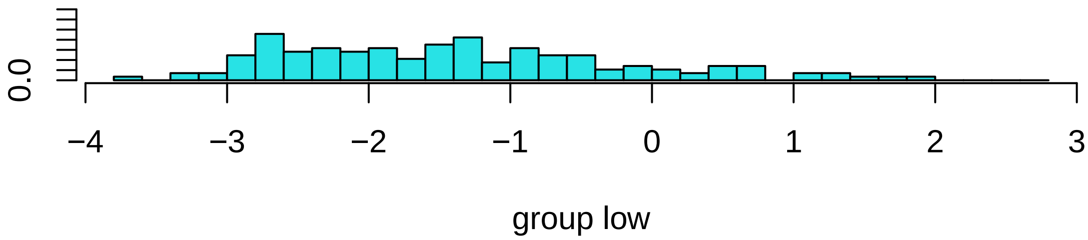
Below are the discriminant coordinates for the LDA model and a histogram of the discriminant variables for each class, `mpg_cat=low` and `mpg_cat=high` in this case.

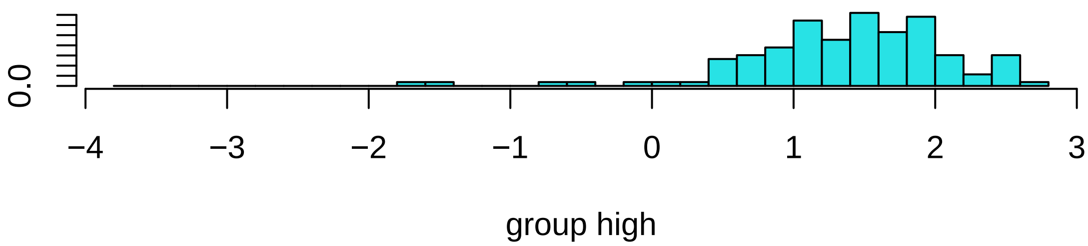
```
plot(lda$finalModel$scaling)
```




```
lda_for_plot <-
  lda(mpg_cat ~ ., training_data)

plot(lda_for_plot)
```





Question d

Which model will you choose to predict the response variable?

To select our best model, we should **evaluate it based on CV and not on test data -** hence, a boxplot of the CV-ROC is shown below.

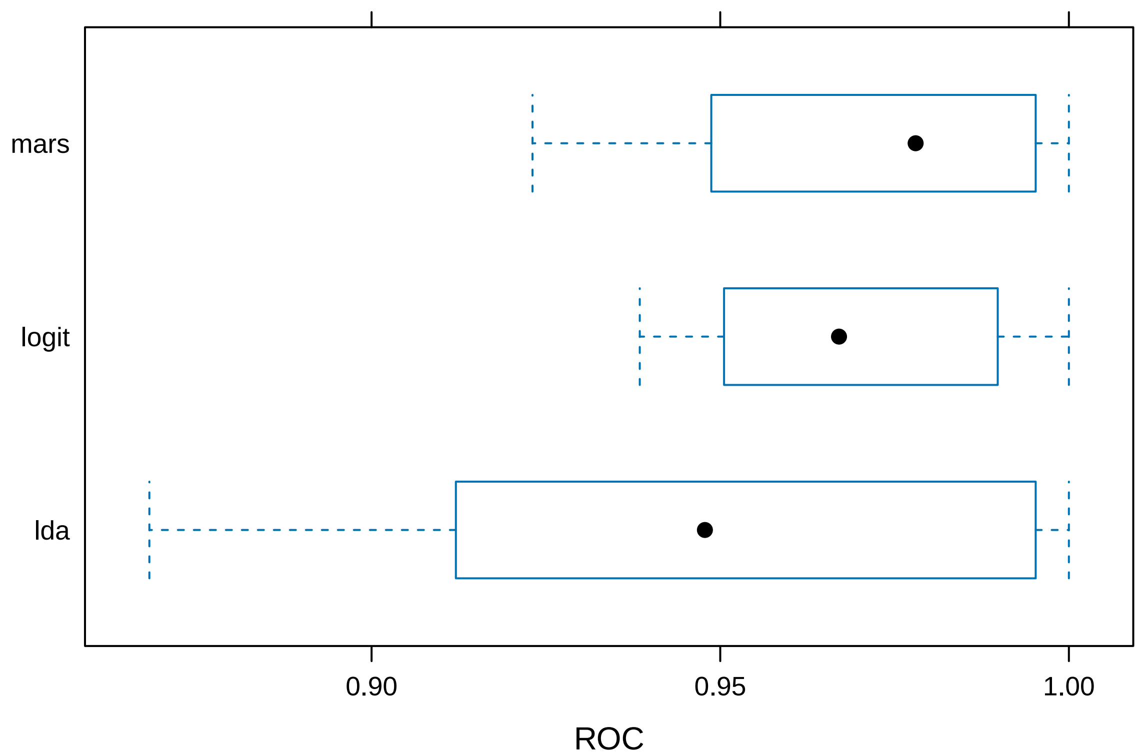
```r
res <- resamples(list(logit = logit,
                      mars = mars,
                      lda = lda))

summary(res)
```

```
##
## Call:
## summary.resamples(object = res)
##
## Models: logit, mars, lda
## Number of resamples: 10
##
## ROC
##            Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## logit 0.9384615 0.9546703 0.9670330 0.9703715 0.9882261    1    0
```

```
## mars   0.9230769 0.9505495 0.9780220 0.9691209 0.9923077    1    0
## lda    0.8681319 0.9175824 0.9478022 0.9496337 0.9923077    1    0
##
## Sens
##              Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## logit 0.7857143 0.7857143 0.8976190 0.8938095 1.0000000    1    0
## mars  0.7857143 0.8571429 0.9285714 0.9076190 0.9833333    1    0
## lda   0.6428571 0.7321429 0.8642857 0.8447619 0.9321429    1    0
##
## Spec
##              Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## logit 0.7692308 0.8489011 0.9230769 0.9016484 0.9271978    1    0
## mars  0.8461538 0.8736264 0.9230769 0.9318681 1.0000000    1    0
## lda   0.9230769 0.9230769 0.9285714 0.9549451 1.0000000    1    0
```

```
bwplot(res, metric = "ROC")
```




```
median_logit_roc <- median(res$values$`logit~ROC`)

median_lda_roc <- median(res$values$`lda~ROC`)

median_mars_roc <- median(res$values$`mars~ROC`)
```

This illustrates that between LDA, logistic regression, and MARS, MARS has the highest median ROC with CV, with 0.978022 compared to 0.967033 for logistic regression and 0.9478022 for LDA.

Plot its ROC curve

We can compare the ROC curves generated when fitting each model's predicted values against the training dataset.

```
logit_pred <- predict(logit, newdata = training_data, type = "prob")[,2]

mars_pred <- predict(mars, newdata = training_data, type = "prob")[,2]

lda_pred <- predict(lda, newdata = training_data, type = "prob")[,2]

roc_logit <- roc(training_data$mpg_cat, logit_pred)
```

```
## Setting levels: control = low, case = high

## Setting direction: controls < cases
```

```
roc_mars <- roc(training_data$mpg_cat, mars_pred)
```

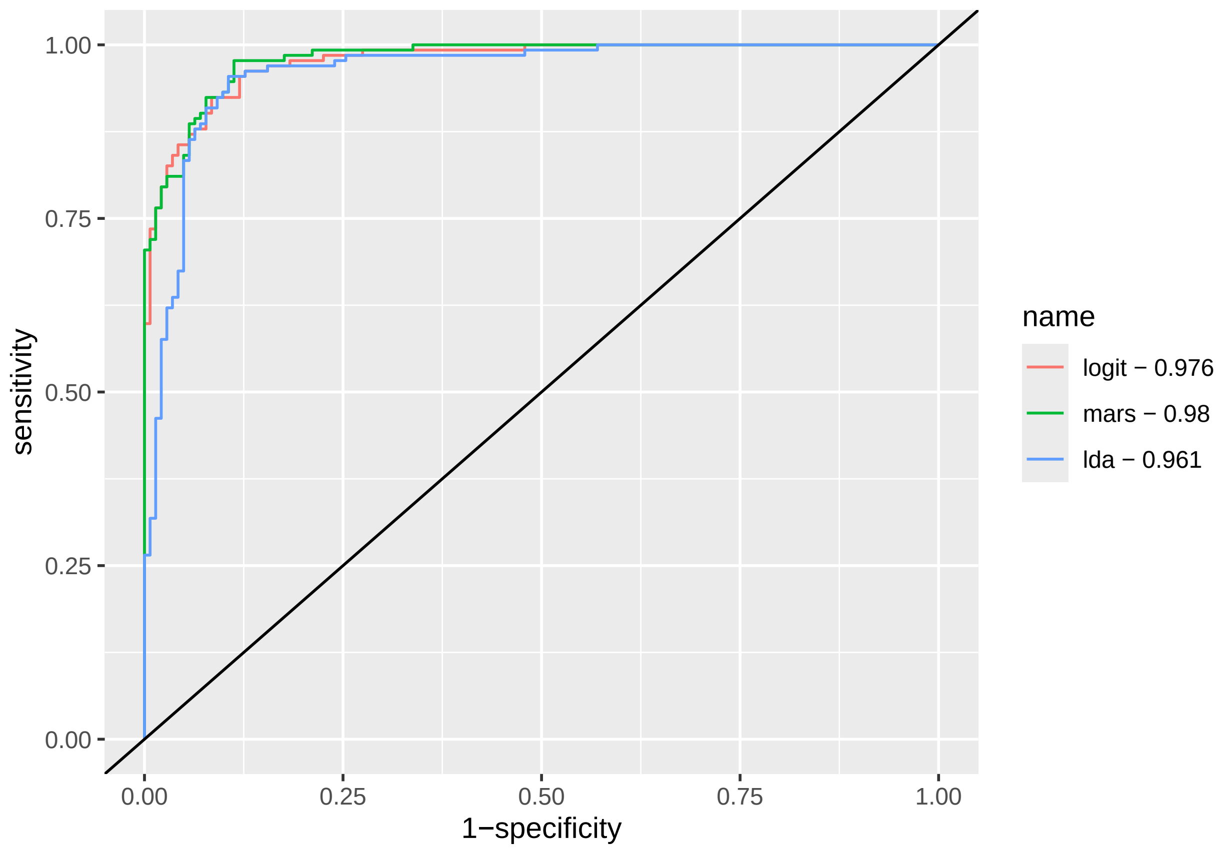
```
## Setting levels: control = low, case = high
## Setting direction: controls < cases
```

```
roc_lda <- roc(training_data$mpg_cat, lda_pred)
```

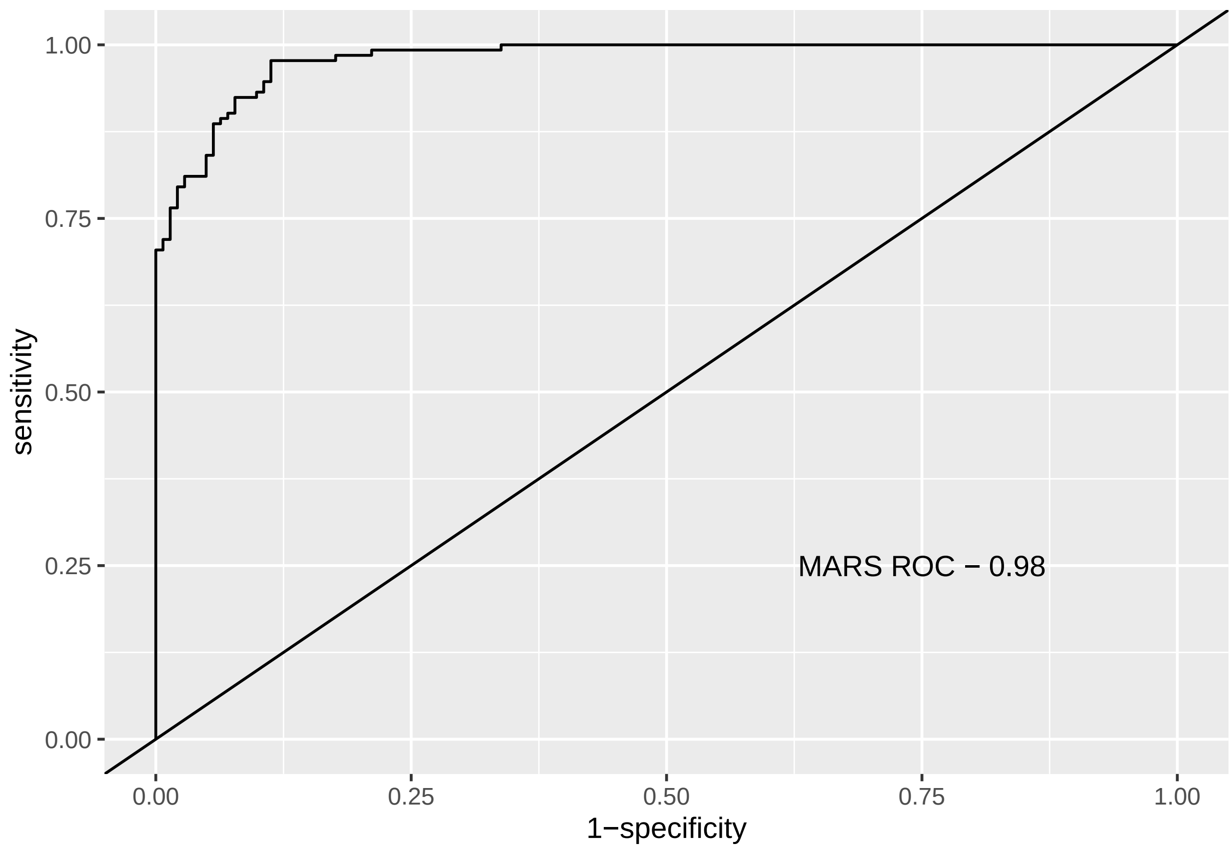
```
## Setting levels: control = low, case = high
## Setting direction: controls < cases
```

```
auc <- c(roc_logit$auc[1], roc_mars$auc[1], roc_lda$auc[1])
model_names <- c("logit", "mars", "lda")

ggroc(list(roc_logit, roc_mars, roc_lda), legacy.axes = TRUE) +
  scale_color_discrete(labels = paste(model_names, "-", round(auc,3), sep = " ")) +
  geom_abline(intercept = 0, slope = 1, color = "black")
```




```
ggroc(roc_mars, legacy.axes = TRUE) +
  geom_abline(intercept = 0, slope = 1, color = "black") +
  annotate("text", x = 0.75, y = 0.25, label = paste("MARS ROC -", round(auc[2], 3), sep = " "))
```



The first graph shows the ROC AUC for all 3 models against the training dataset, which gives that the MARS actually has the highest training ROC, in addition to the highest CV ROC. The MARS ROC from predicting on the training dataset is also visualized.

```
logit_pred_testing <- predict(logit, newdata = testing_data, type = "prob")[,2]

mars_pred_testing <- predict(mars, newdata = testing_data, type = "prob")[,2]

lda_pred_testing <- predict(lda, newdata = testing_data, type = "prob")[,2]

roc_logit_testing <- roc(testing_data$mpg_cat, logit_pred_testing)
```

```
## Setting levels: control = low, case = high

## Setting direction: controls < cases
```

```
roc_mars_testing <- roc(testing_data$mpg_cat, mars_pred_testing)
```

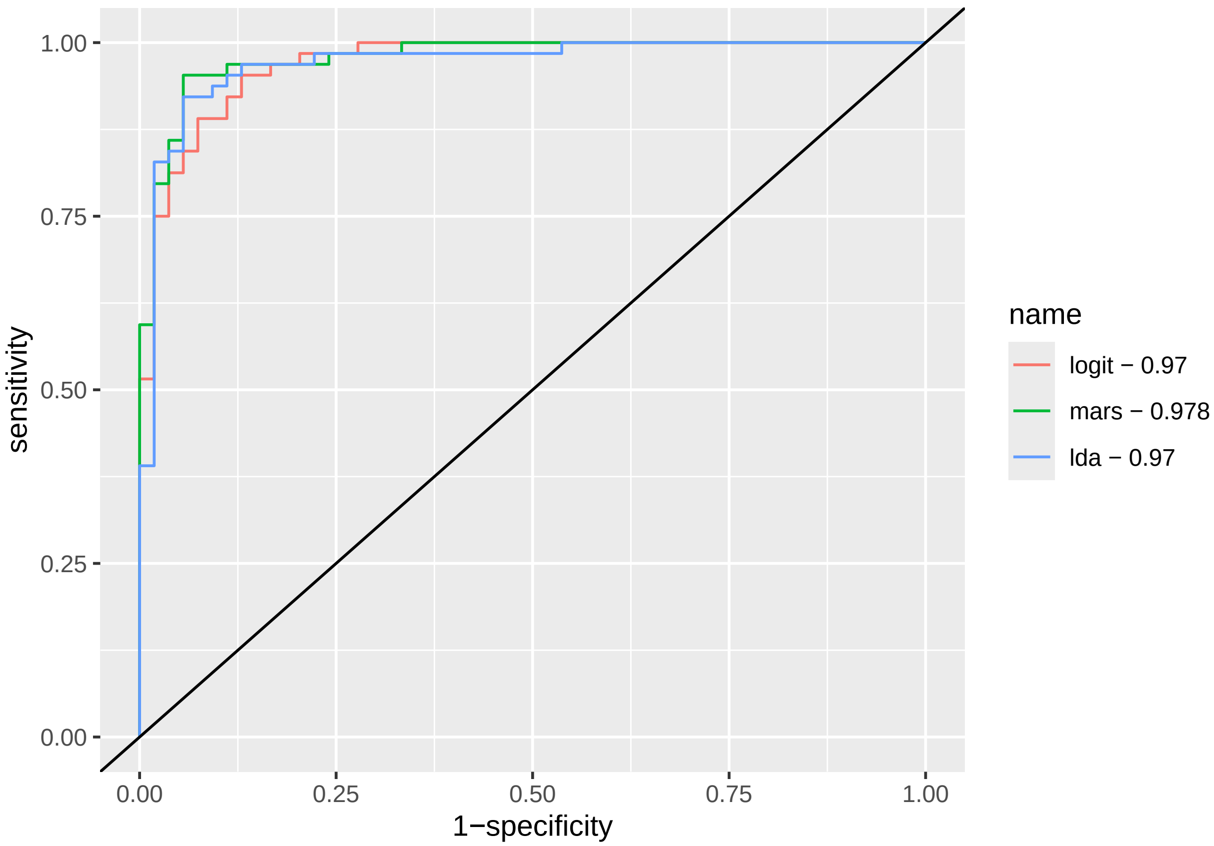
```
## Setting levels: control = low, case = high
## Setting direction: controls < cases
```

```
roc_lda_testing <- roc(testing_data$mpg_cat, lda_pred_testing)
```

```
## Setting levels: control = low, case = high
## Setting direction: controls < cases
```

```r
auc_testing <- c(roc_logit_testing$auc[1], roc_mars_testing$auc[1], roc_lda_testing$auc[1])

ggroc(list(roc_logit_testing, roc_mars_testing, roc_lda_testing), legacy.axes = TRUE) +
  scale_color_discrete(labels = paste(model_names, "-", round(auc_testing,3), sep = " ")) +
  geom_abline(intercept = 0, slope = 1, color = "black")
```



When the models are fit to the testing dataset, the MARS model still has the highest ROC with 0.9780093, compared to LDA and logistic regression with 0.9699074 and 0.9704861, respectively.

Select a probability threshold to classify observations and compute the confusion matrix.

lda_pred_prob generates 274 obs for each of the 2 classes, low and high, for a total of 548 obs while the training data is only 274 obs