Data Science II Homework 3

Camille Okonkwo

Contents

```r
library(tidymodels)
library(caret)
library(earth)
library(pROC)
library(vip)
library(MASS)
set.seed(2)
```

Introduction

Background

We will develop a model to predict whether a given car gets high or low gas mileage based on the dataset `auto.csv`. The dataset contains 392 observations.

The response variable is `mpg_cat`, which indicates whether the miles per gallon of a car is high or low.

The predictors are:

- `cylinders`: Number of cylinders between 4 and 8
- `displacement`: Engine displacement (cu. inches)
- `horsepower`: Engine horsepower
- `weight`: Vehicle weight (lbs.)
- `acceleration`: Time to accelerate from 0 to 60 mph (sec.)
- `year`: Model year (modulo 100)
- `origin`: Origin of car (1. American, 2. European, 3. Japanese)

Split the dataset into two parts: training data (70%) and test data (30%)

```
auto = read_csv("data/auto.csv") |>
  drop_na() |>
  mutate(
    mpg_cat = as.factor(mpg_cat),
    mpg_cat = forcats::fct_relevel(mpg_cat, c("low", "high"))
  )

set.seed(2)

# create a random split of 70% training and 30% test data
data_split <- initial_split(data = auto, prop = 0.7)

# partitioned datasets
training_data = training(data_split)
testing_data = testing(data_split)

head(training_data)
```

```
## # A tibble: 6 x 8
##   cylinders displacement horsepower weight acceleration  year origin mpg_cat
##       <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <dbl>  <dbl> <fct>
## 1         4           86         64   1875         16.4    81      1 high
## 2         6          225        100   3651         17.7    76      1 low
## 3         6          231        165   3445         13.4    78      1 low
## 4         5          131        103   2830         15.9    78      2 low
## 5         4           98         65   2380         20.7    81      1 high
## 6         4           97         75   2155         16.4    76      3 high
```

```
head(testing_data)
```

```
## # A tibble: 6 x 8
##   cylinders displacement horsepower weight acceleration  year origin mpg_cat
```

```
##       <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <dbl>  <dbl> <fct>
## 1         8          302        140   3449         10.5    70      1 low
## 2         8          390        190   3850          8.5    70      1 low
## 3         4          113         95   2372         15      70      3 high
## 4         6          200         85   2587         16      70      1 low
## 5         4           97         88   2130         14.5    70      3 high
## 6         4          107         90   2430         14.5    70      2 high
```

1a) Perform a logistic regression analysis using the training data. Are there redundant predictors in your model? If so, identify them. If none is present, please provide an explanation.

```
ctrl <- trainControl(method = "cv",
                     number = 10,
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)
set.seed(2)

# logistic using glm
glm.fit <- glm(mpg_cat ~ .,
               data = training_data,
               family = binomial(link = "logit"))

coef(glm.fit)
```

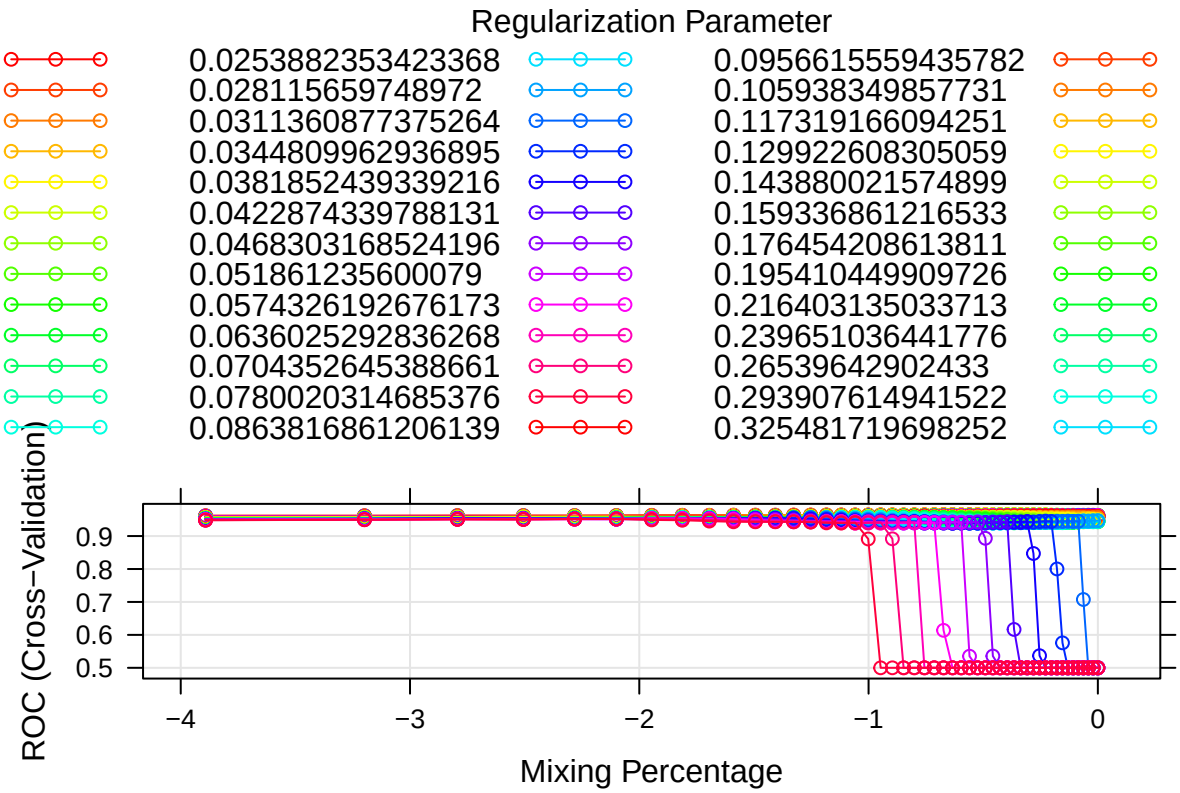
```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -2.054216e+01  2.717063e-02  9.626894e-04 -2.423505e-02 -4.363108e-03
##  acceleration          year        origin
##  1.453842e-01  4.225079e-01  2.459384e-01
```

```
# logistic using caret
set.seed(2)
model.glm <- train(x = training_data[1:7],
                   y = training_data$mpg_cat,
                   method = "glm",
                   metric = "ROC",
                   trControl = ctrl)

model.glm$finalModel
```

```
##
## Call:  NULL
##
## Coefficients:
##  (Intercept)     cylinders  displacement    horsepower        weight
##   -2.054e+01     2.717e-02     9.627e-04    -2.424e-02    -4.363e-03
## acceleration          year        origin
##    1.454e-01     4.225e-01     2.459e-01
##
## Degrees of Freedom: 273 Total (i.e. Null);  266 Residual
## Null Deviance:       379.8
## Residual Deviance: 120.3     AIC: 136.3
```

```
#both models gave same coefficients

# penalized logistic model
glmnGrid <- expand.grid(alpha = seq(0, 1, length = 50),
                        lambda = exp(seq(-5, 0, length = 50)))
set.seed(2)

model.glmn <- train(x = training_data[1:7],
                    y = training_data$mpg_cat,
```

```
                    method = "glmnet",
                    tuneGrid = glmnGrid,
                    metric = "ROC",
                    trControl = ctrl)

# visualizing AUC and regularization parameters
myCol <- rainbow(25)
myPar <- list(superpose.symbol = list(col = myCol),
superpose.line = list(col = myCol))
plot(model.glmn, par.settings = myPar, xTrans = function(x) log(x))
```




```
summary(model.glmn)
```

```
##             Length Class      Mode
## a0             86  -none-     numeric
## beta          602  dgCMatrix  S4
## df             86  -none-     numeric
## dim             2  -none-     numeric
## lambda         86  -none-     numeric
## dev.ratio      86  -none-     numeric
## nulldev         1  -none-     numeric
## npasses         1  -none-     numeric
## jerr            1  -none-     numeric
## offset          1  -none-     logical
## classnames      2  -none-     character
## call            5  -none-     call
## nobs            1  -none-     numeric
## lambdaOpt       1  -none-     numeric
## xNames          7  -none-     character
```

```
## problemType   1    -none-     character
## tuneValue     2    data.frame list
## obsLevels     2    -none-     character
## param         0    -none-     list
```

```
best_model = model.glmn$finalModel


# extracting best lambda for coefficients
best_lambda = model.glmn$bestTune$lambda

# extracting coefficients in the best model with best lambda
coefficients = coef(best_model, s = best_lambda)

# extracting redundant predictors
names(coefficients[coefficients == 0])
```

```
## NULL
```

Per the penalized logistic, there are no redundant predictors.Each predictor variable is included in the model and provides valuable information that is duplicated by other predictors. With no redundant predictors, the model tends to be more interpretable because each predictor has a clear and unique role in explaining variation in the response variable. No redundant predictors also tells us the model has a high predictive power and findings can be generalized.

1b) Based on the model in (a), set a probability threshold to determine the class labels and compute the confusion matrix using the test data. Briefly interpret what the confusion matrix reveals about your model's performance.

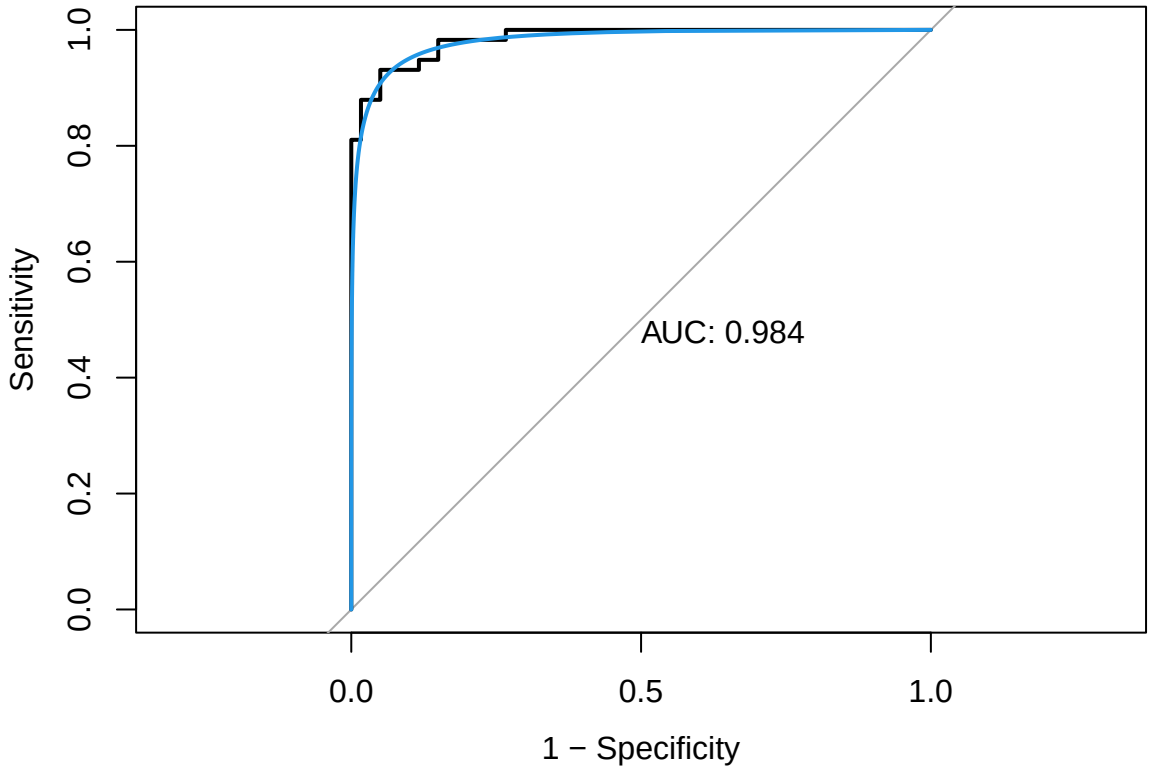
```
# checking coding
contrasts(auto$mpg_cat)
```

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

```
# predict class labels using the logistic regression model and testing data
test.pred.prob <- predict(glm.fit,
                          newdata = testing_data,
                          type = "response")

 # setting a probability threshold of 0.5
test.pred <- rep("low", length(test.pred.prob))
test.pred[test.pred.prob > 0.5] <- "high"


confusionMatrix(data = as.factor(test.pred),
                reference = testing_data$mpg_cat,
                positive = "high")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction low high
##       low   57    5
##       high   3   53
##
##                Accuracy : 0.9322
##                  95% CI : (0.8708, 0.9703)
##     No Information Rate : 0.5085
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8643
##
##  Mcnemar's Test P-Value : 0.7237
##
##             Sensitivity : 0.9138
##             Specificity : 0.9500
##          Pos Pred Value : 0.9464
##          Neg Pred Value : 0.9194
##              Prevalence : 0.4915
##          Detection Rate : 0.4492
##    Detection Prevalence : 0.4746
##       Balanced Accuracy : 0.9319
##
##        'Positive' Class : high
##
```

```r
# plotting test ROC curve
roc.glm <- roc(testing_data$mpg_cat, test.pred.prob)
plot(roc.glm, legacy.axes = TRUE, print.auc = TRUE)
plot(smooth(roc.glm), col = 4, add = TRUE) # smooth ROC curve
```



The matrix reveals the accuracy of the model is 93.22%[87.08%-97.03%], meaning the model correctly predicts the class (low or high mileage car) 93.22% of the time on the testing data and we are 95% confident the accuracy ranges from 87.98% to 93.03%. A p-value less than 0.05 indicates to us there's sufficient evidence that the model's accuracy is better than simply predicting the most frequent class. A Kappa of 0.8643 shows there is high inter-rater agreement between the observed label and predicted label (by chance). A sensitivity of 0.9138 indicates that the model correctly identifies 91.38% of the high mileage cars, and a specificity of 0.95 indicates that the model correctly identifies 95% of the low mileage cars.

In sum, the confusion matrix tells us the model performs well in distinguishing between high and low gas mileage cars. It has high accuracy, sensitivity, specificity, as well as positive and negative predictive values (0.9464 & 0.9194, respectively), suggesting that it is effective in making predictions.

1c) Train a multivariate adaptive regression spline (MARS) model. Does the MARS model improve the prediction performance compared to logistic regression?

```
set.seed(2)

# log stats
coef(model.glm$finalModel)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -2.054216e+01  2.717063e-02  9.626894e-04 -2.423505e-02 -4.363108e-03
##  acceleration          year        origin
##  1.453842e-01  4.225079e-01  2.459384e-01
```

```
summary(model.glm)
```

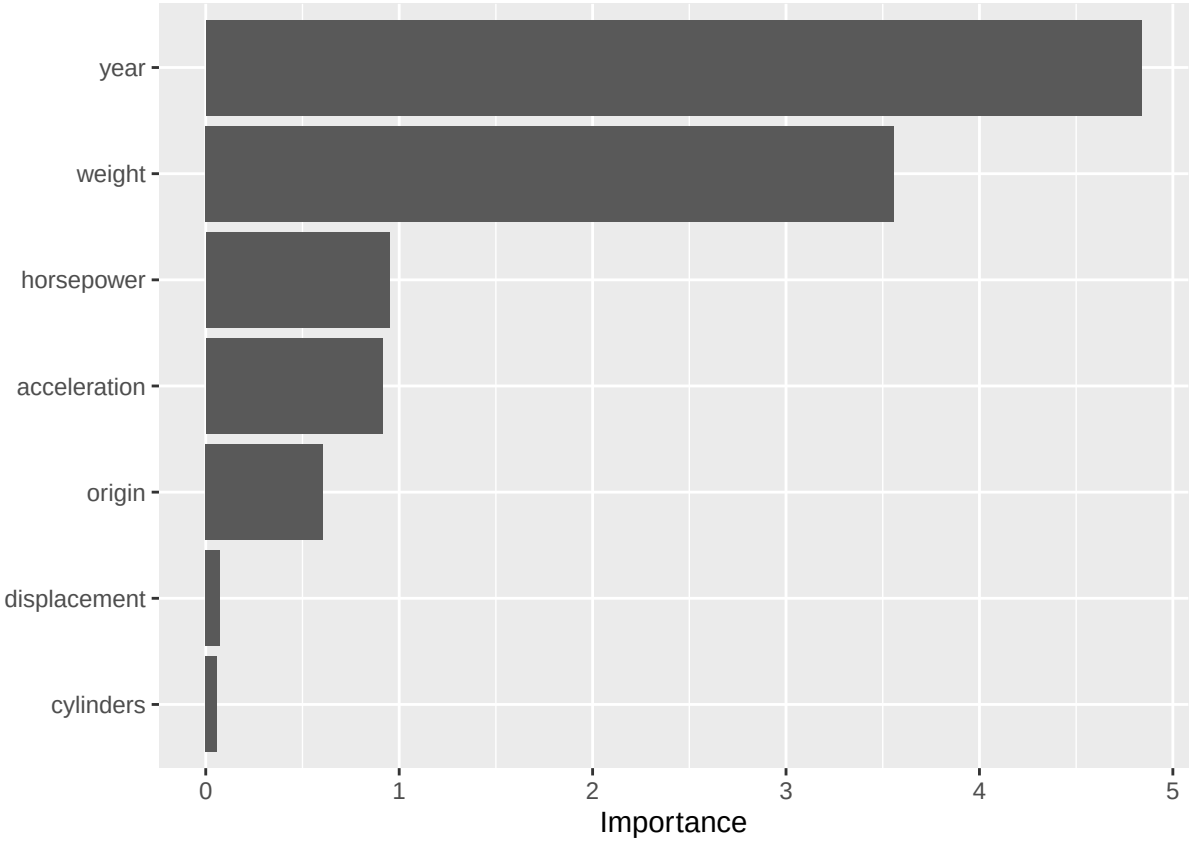
```
##
## Call:
## NULL
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.054e+01  6.889e+00  -2.982 0.002866 **
## cylinders     2.717e-02  4.756e-01   0.057 0.954446
## displacement  9.627e-04  1.305e-02   0.074 0.941212
## horsepower   -2.423e-02  2.554e-02  -0.949 0.342653
## weight       -4.363e-03  1.226e-03  -3.558 0.000373 ***
## acceleration  1.454e-01  1.593e-01   0.913 0.361485
## year          4.225e-01  8.732e-02   4.838 1.31e-06 ***
## origin        2.459e-01  4.057e-01   0.606 0.544401
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.83  on 273  degrees of freedom
## Residual deviance: 120.29  on 266  degrees of freedom
## AIC: 136.29
##
## Number of Fisher Scoring iterations: 7
```

```
model.glm$fin
```

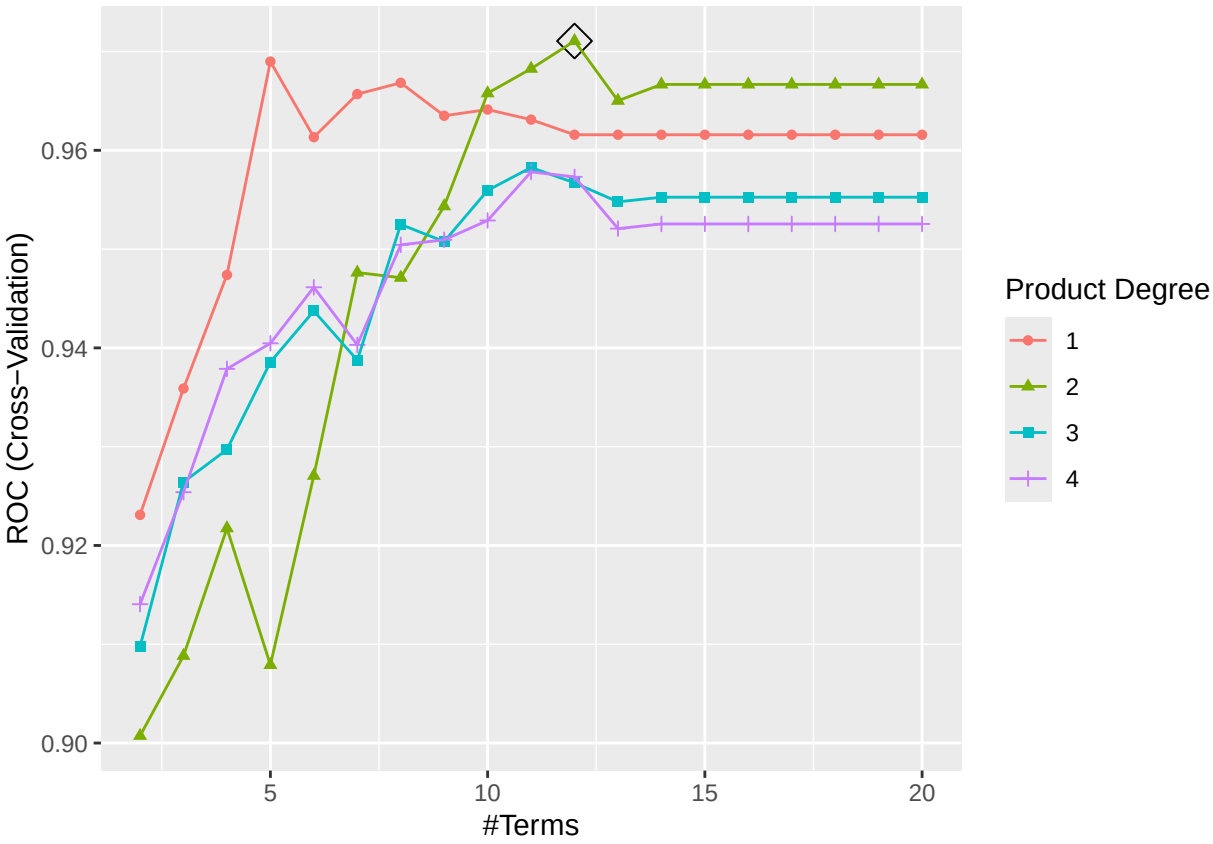
```
##
## Call:  NULL
##
## Coefficients:
##  (Intercept)     cylinders  displacement    horsepower        weight
##   -2.054e+01     2.717e-02     9.627e-04    -2.424e-02    -4.363e-03
## acceleration          year        origin
##    1.454e-01     4.225e-01     2.459e-01
##
## Degrees of Freedom: 273 Total (i.e. Null);  266 Residual
## Null Deviance:       379.8
```

```
## Residual Deviance: 120.3     AIC: 136.3
```

```
vip(model.glm$finalModel, type = "stat")
```




```
# MARS
set.seed(2)

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

ggplot(model.mars, highlight = TRUE)
```



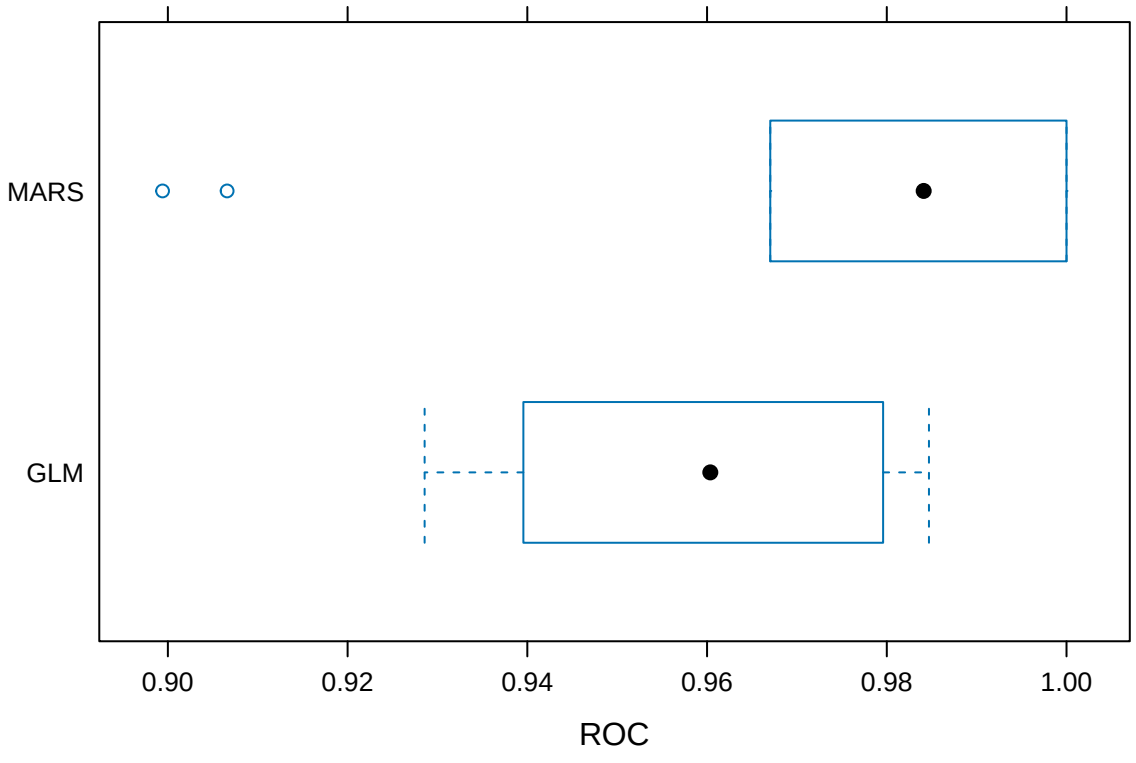

```
coef(model.mars$finalModel)
```

```
##                                 (Intercept)
##                                -7.7330069243
##                          h(displacement-232)
##                                -0.9356130852
##                          h(232-displacement)
##                                 0.0847147717
##                                   h(year-72)
##                                 0.7579418640
##       h(4-cylinders) * h(232-displacement)
##                                -0.6891088932
##           h(displacement-173) * h(year-72)
##                                -0.0257570538
##           h(173-displacement) * h(year-72)
##                                -0.0115097096
##       h(232-displacement) * h(weight-2648)
##                                -0.0001223507
##           h(displacement-232) * h(year-75)
##                                 0.2998569916
##          h(acceleration-14.5) * h(year-72)
##                                 0.2342242317
##                h(4-cylinders) * h(year-72)
##                                16.3097304420
## h(232-displacement) * h(acceleration-14.1)
##                                -0.0037792287
```

```r
set.seed(2)


res <- resamples(list(GLM = model.glm,
                      MARS = model.mars))
summary(res)
```

```
##
## Call:
## summary.resamples(object = res)
##
## Models: GLM, MARS
## Number of resamples: 10
##
## ROC
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM  0.9285714 0.9395604 0.9603611 0.9593316 0.9795918 0.9846939    0
## MARS 0.8994083 0.9676217 0.9841052 0.9710633 1.0000000 1.0000000    0
##
## Sens
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM  0.6923077 0.7857143 0.8846154 0.8527473 0.9271978 0.9285714    0
## MARS 0.7857143 0.8008242 0.8846154 0.8829670 0.9285714 1.0000000    0
##
## Spec
##           Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## GLM  0.7692308 0.8736264 0.9285714 0.9049451 0.9285714    1    0
## MARS 0.6923077 0.9285714 0.9285714 0.9329670 1.0000000    1    0
```

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



```r
# predicted probabilities with testing data
glm.pred <- predict(model.glm, newdata = testing_data, type = "prob")[,2]
mars.pred <- predict(model.mars, newdata = testing_data, type = "prob")[,2]

# ROC curves
roc.glm <- roc(testing_data$mpg_cat, glm.pred)
roc.mars <- roc(testing_data$mpg_cat, mars.pred)
```

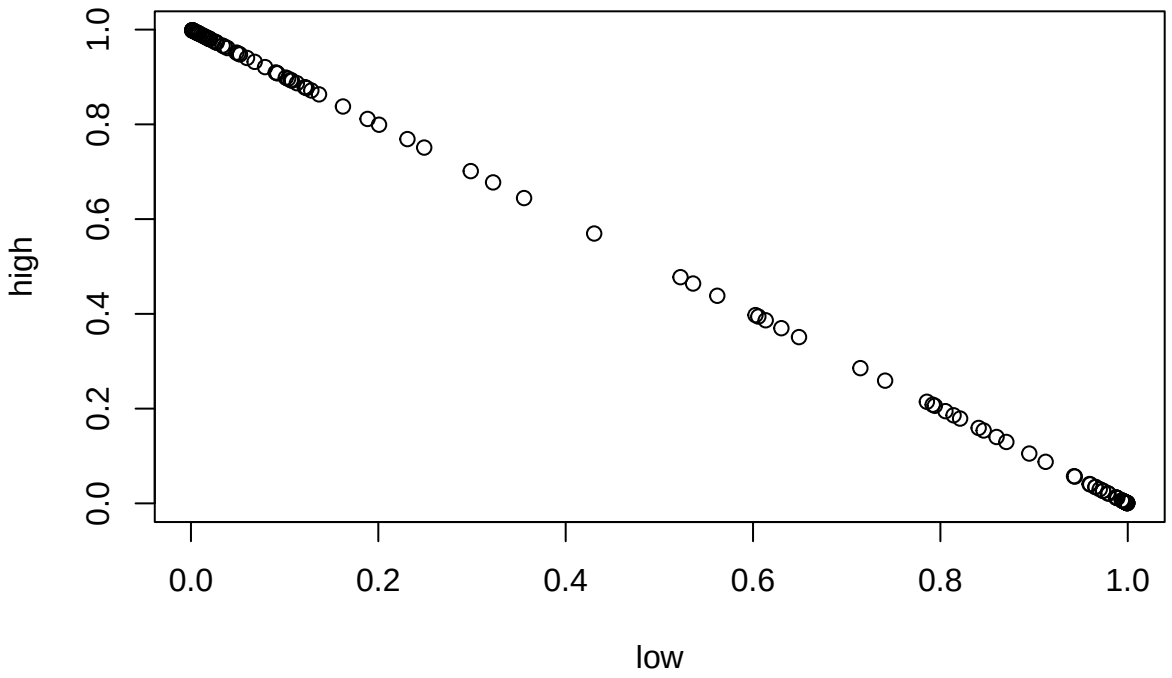
MARS has the highest mean and median ROC values, already pointing us that this model will have a better discriminatory power compared to the logistic model. However, in order to properly determine if MARS has better **predictive power**, we must consider the predicted probabilites using the testing data and compare the area under the curve values for each respective ROC curve for the log model and the MARS model. Based on the AUC values, the MARS model does marginally improve predictive performance compared to logistic regression.

1d) Perform linear discriminant analysis using the training data. Plot the linear discriminant variable(s).

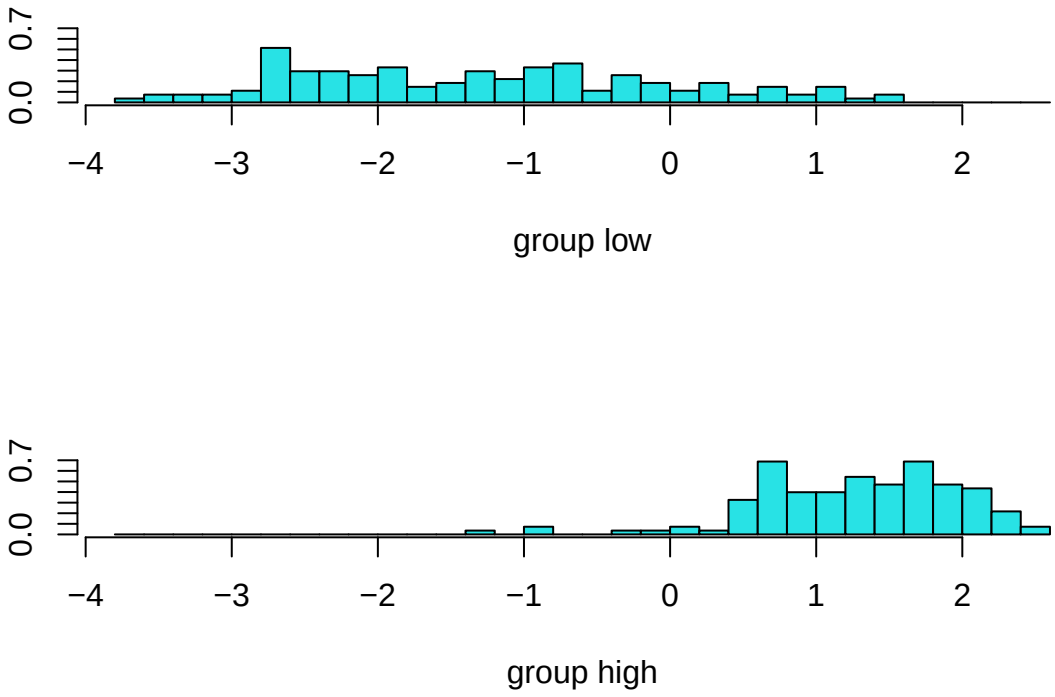
```
set.seed(2)

# LDA using caret
model.lda <- train(x = training_data[, 1:7],
                   y = training_data$mpg_cat,
                   method = "lda",
                   metric = "ROC",
                   trControl = ctrl)


# prediction
lda.pred2 <- predict(model.lda, newdata = testing_data, type = "prob")
head(lda.pred2)
```

```
##         low        high
## 1 0.9955980 0.004401999
## 2 0.9963409 0.003659114
## 3 0.2007719 0.799228125
## 4 0.8408375 0.159162536
## 5 0.1235133 0.876486667
## 6 0.3555758 0.644424200
```

```
plot(lda.pred2)
```




```
# LDA using MASS
lda.fit <- lda(mpg_cat~.,
               data = training_data)

# linear discriminant variables
plot(lda.fit)
```

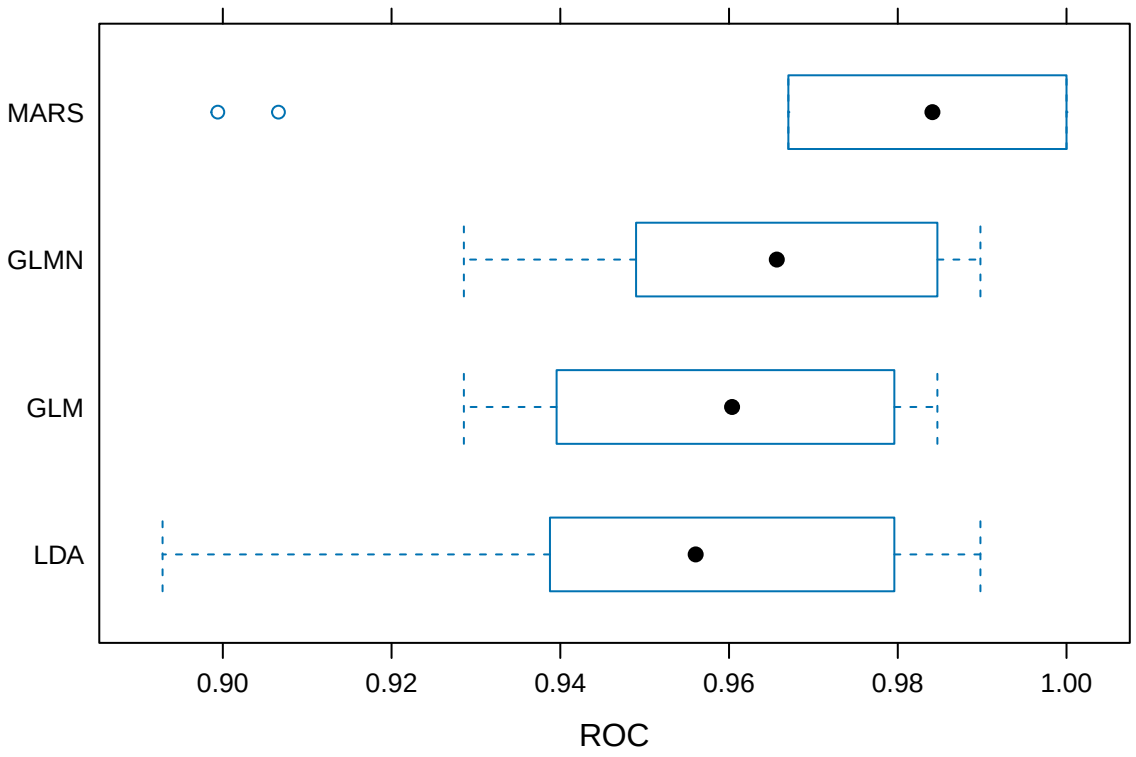

0.0
0.7
−4
−3
−2
−1
0
1
2
group low
0.0
0.7
−4
−3
−2
−1
0
1
2
group high

1e) Which model will you use to predict the response variable? Plot its ROC curve using the test data. Report the AUC and the misclassification error rate.

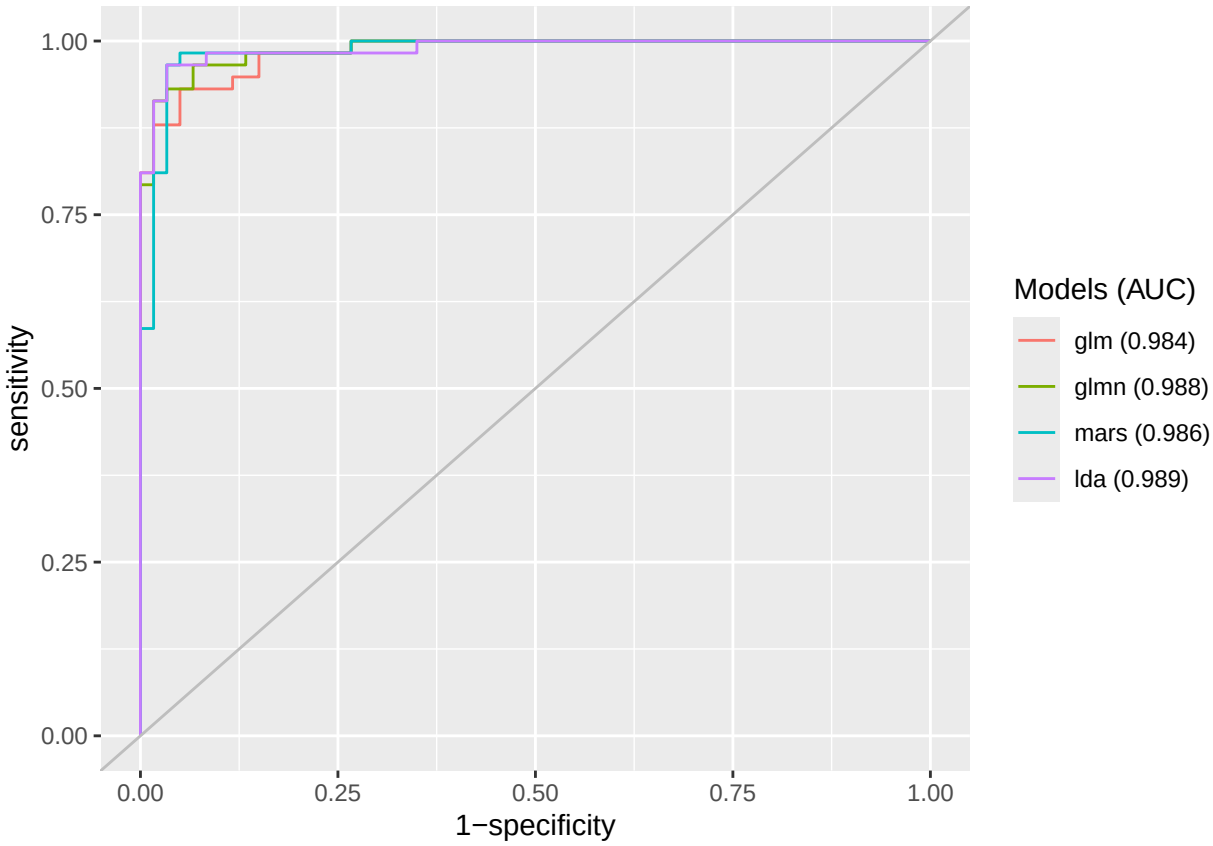
```
# resampling comparison
res = resamples(list(
                     GLM = model.glm,
                     GLMN = model.glmn,
                     MARS = model.mars,
                     LDA = model.lda))
summary(res)
```

```
##
## Call:
## summary.resamples(object = res)
##
## Models: GLM, GLMN, MARS, LDA
## Number of resamples: 10
##
## ROC
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM  0.9285714 0.9395604 0.9603611 0.9593316 0.9795918 0.9846939    0
## GLMN 0.9285714 0.9507457 0.9656593 0.9651823 0.9846939 0.9897959    0
## MARS 0.8994083 0.9676217 0.9841052 0.9710633 1.0000000 1.0000000    0
## LDA  0.8928571 0.9389717 0.9560440 0.9516000 0.9783163 0.9897959    0
##
## Sens
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM  0.6923077 0.7857143 0.8846154 0.8527473 0.9271978 0.9285714    0
## GLMN 0.6923077 0.7857143 0.8846154 0.8527473 0.9271978 0.9285714    0
## MARS 0.7857143 0.8008242 0.8846154 0.8829670 0.9285714 1.0000000    0
## LDA  0.6923077 0.7733516 0.8516484 0.8230769 0.8571429 0.9285714    0
##
## Spec
##           Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## GLM  0.7692308 0.8736264 0.9285714 0.9049451 0.9285714    1    0
## GLMN 0.8461538 0.8736264 0.9285714 0.9126374 0.9285714    1    0
## MARS 0.6923077 0.9285714 0.9285714 0.9329670 1.0000000    1    0
## LDA  0.9230769 0.9285714 0.9285714 0.9560440 1.0000000    1    0
```

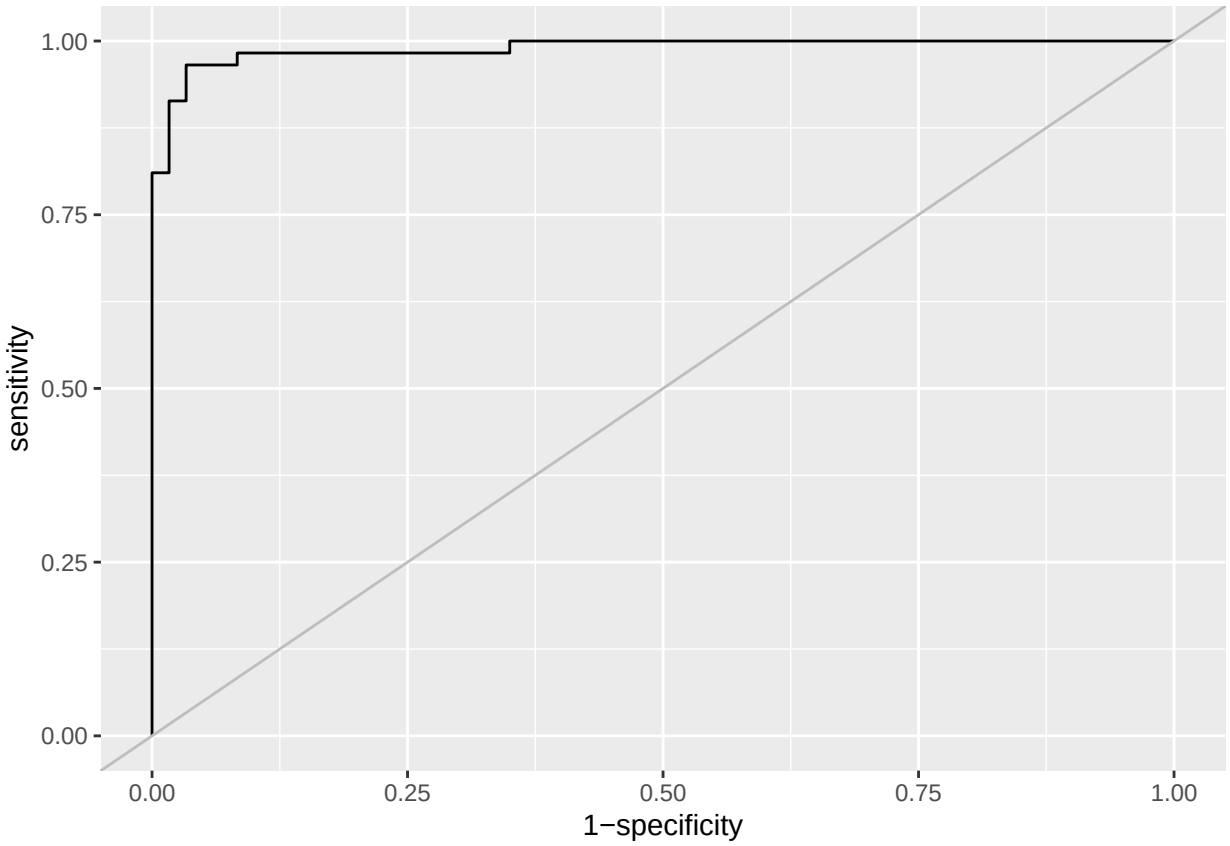
```
bwplot(res, metric = "ROC")
```




```
# predicition
glm.pred <- predict(model.glm, newdata = testing_data, type = "prob")[,2]
glmn.pred <- predict(model.glmn, newdata = testing_data, type = "prob")[,2]
mars.pred <- predict(model.mars, newdata = testing_data, type = "prob")[,2]
lda.pred <- predict(model.lda, newdata = testing_data, type = "prob")[,2]

# ROC curves (for AUC)
roc.glm <- roc(testing_data$mpg_cat, glm.pred)
roc.glmn <- roc(testing_data$mpg_cat, glmn.pred)
roc.mars <- roc(testing_data$mpg_cat, mars.pred)
roc.lda <- roc(testing_data$mpg_cat, lda.pred)

# AUC values
auc <- c(roc.glm$auc[1], roc.glmn$auc[1], roc.mars$auc[1], roc.lda$auc[1])

modelNames <- c("glm", "glmn", "mars","lda")

# combined ROC curves
ggroc(list(roc.glm, roc.glmn, roc.mars, roc.lda),
      legacy.axes = TRUE) +
  scale_color_discrete(labels = paste0(modelNames, " (", round(auc,3),")"),
                       name = "Models (AUC)") + geom_abline(intercept = 0, slope = 1, color = "grey")
```




```
# ROC for best model only (LDA)
ggroc(roc.lda, legacy.axes = T) +
  geom_abline(intercept = 0, slope = 1, color = "grey")
```




```
# AUC for LDA
roc.lda$auc
```

```
## Area under the curve: 0.9891
```

```
# create a logical vector where we compare the predictions from the MARS model to the actual test data
misclass = (roc.lda$response != testing_data$mpg_cat)
# take the mean of the logical vector to calculate the proportion of misclassifications
mean(misclass)
```

```
## [1] 0
```

From the re-sampling summary, we can see the MARS model has the highest mean and median ROC so I would use the MARS model to predict miles per gallon. The LDA model, however, has a greater AUC indicating better predictive performance in comparison to the MARS model. Considering the context of the data, while having high sensitivity and specificity (and therefore higher ROC values) are important for understanding and interpreting model performance, I will prefer to choose a model with higher discriminatory power and a better overall performance in correctly classifying high and low mileage cars. Therefore, I will choose the LDA model over the MARS. The difference in their ROC values is marginal. The AUC of the LDA model is 0.9890805. The mis-classification error rate is 0, which essentially means that the LDA model's predictions matched every actual value in our response variable `mpg_cat`.