Data Science II: Homework 3

Name: Jasmin Martinez (JRM2319) Date: 04/01/25

In this problem, you 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 include both continuous and categorical variables:
- **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) - **mpg_cat**: *response variable* indicates whether the miles per gallon of a car is 'high' or 'low'

Import Data

```
auto = read.csv("auto.csv")
head(auto)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         8          307        130   3504         12.0   70      1     low
## 2         8          350        165   3693         11.5   70      1     low
## 3         8          318        150   3436         11.0   70      1     low
## 4         8          304        150   3433         12.0   70      1     low
## 5         8          302        140   3449         10.5   70      1     low
## 6         8          429        198   4341         10.0   70      1     low
```

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

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.3.0 --
```

```
## v broom        1.0.7     v rsample      1.2.1
## v dials        1.4.0     v tibble       3.2.1
## v dplyr        1.1.4     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.3.0
## v modeldata    1.4.0     v workflows    1.2.0
## v parsnip      1.3.0     v workflowsets 1.1.0
## v purrr        1.0.4     v yardstick    1.3.2
## v recipes      1.1.1

## -- Conflicts ---------------------------------------- tidymodels_conflicts() --
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
```

```
datSplit = initial_split(data = auto, prop = 0.7)
trainData = training(datSplit)
testData = testing(datSplit)
head(trainData)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         4          105         63   2215         14.9   81      1    high
## 2         4          105         63   2125         14.7   82      1    high
## 3         6          232        100   3288         15.5   71      1     low
## 4         6          173        115   2595         11.3   79      1    high
## 5         4          135         84   2525         16.0   82      1    high
## 6         4           98         65   2380         20.7   81      1    high
```

```
trainData$mpg_cat = as.factor(trainData$mpg_cat)
testData$mpg_cat = as.factor(testData$mpg_cat)
```

(a) Perform logistic regression analysis. Are there redundant predictors in your model? If so, identify them. If there are none, please provide an explanation. Yes, there are redundant predictors in the model. By using the $\Pr(>|z|)$ in the logistic regression model, the following variables are redundant: cylinders, displacement, horsepower, acceleration, and origin. The predictors stated above have p-values > 0.05 and therefore do not contribute to the model in a statistically significant way.

```
set.seed(2)
glmnGrid = expand.grid(.alpha = seq(0, 1, length = 21),
.lambda = exp(seq(-8, -1, length = 50)))

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

```
glm.fit = train(x = trainData[, c("cylinders", "displacement", "horsepower",
                                  "weight", "acceleration", "year", "origin")],
                y = trainData$mpg_cat,
                   method = "glm",
                   family = "binomial",
                   metric = "ROC",
                   trControl = ctrl)

summary(glm.fit)
```

Perform logistic regression analysis

```
##
## Call:
## NULL
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  25.435399   7.494577   3.394 0.000689 ***
## cylinders     0.577051   0.510022   1.131 0.257877
## displacement -0.010846   0.013826  -0.784 0.432773
## horsepower    0.019437   0.031424   0.619 0.536226
## weight        0.004619   0.001380   3.346 0.000819 ***
## acceleration -0.146270   0.182106  -0.803 0.421853
## year         -0.506200   0.100614  -5.031 4.88e-07 ***
## origin       -0.532410   0.460479  -1.156 0.247596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.61  on 273  degrees of freedom
## Residual deviance: 106.75  on 266  degrees of freedom
## AIC: 122.75
##
## Number of Fisher Scoring iterations: 8
```

```
glm.fit2 = train(x = trainData[c("weight", "year")],
                 y = trainData$mpg_cat,
                 method = "glm",
                 family = "binomial",
                 metric = "ROC",
                 trControl = ctrl)

summary(glm.fit2)
```

Adjusting logistic regression model to include only non-redundant predictors

```
##
```

```
## Call:
## NULL
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 24.3740930  6.0800824   4.009 6.10e-05 ***
## weight       0.0050141  0.0007057   7.106 1.20e-12 ***
## year        -0.5106288  0.0949171  -5.380 7.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.61  on 273  degrees of freedom
## Residual deviance: 115.06  on 271  degrees of freedom
## AIC: 121.06
##
## Number of Fisher Scoring iterations: 7
```

```
mars_grid = expand.grid(degree = 1:2,
                        nprune = 2:4)

ctrl1 = trainControl(method = "cv", number = 10)

trainData$mpg_cat = as.factor(trainData$mpg_cat)

set.seed(2)

mars.fit = train(x = trainData[, c("cylinders", "displacement", "horsepower",
                                   "weight", "acceleration", "year", "origin")],
                 y = trainData$mpg_cat,
                method = "earth",
                tuneGrid = mars_grid,
                trControl = ctrl1)
```

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

```
## Loading required package: earth
```

```
## Loading required package: Formula
```

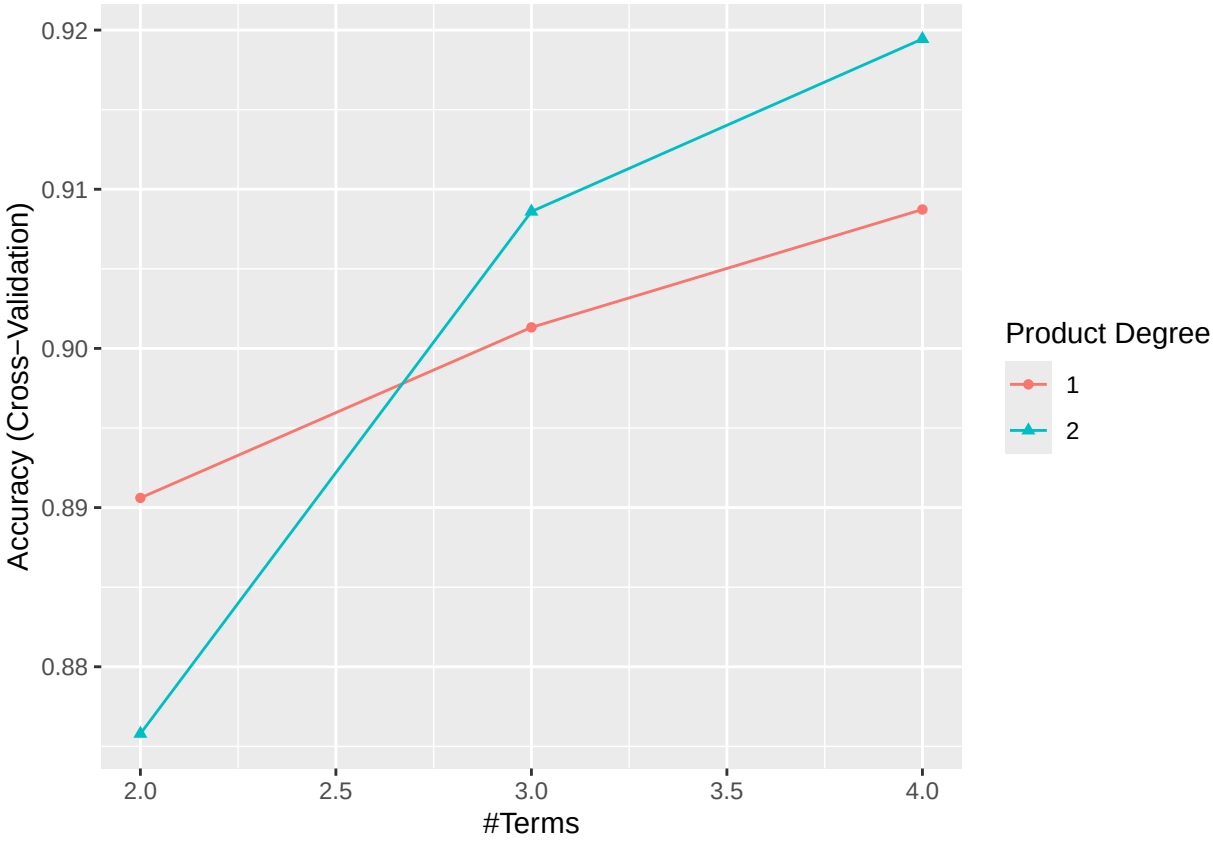
```
## Loading required package: plotmo
```

```
## Loading required package: plotrix
```

```
##
## Attaching package: 'plotrix'
```

```
## The following object is masked from 'package:scales':
##
##     rescale
```

```
ggplot(mars.fit)
```




```
glm.pred = predict(glm.fit2, newdata = testData, type = "raw")
head(glm.pred)
```

Prediction performance of Logistic Regression

```
## [1] low  low  low  low  high high
## Levels: high low
```

```
mars.pred = predict(mars.fit, newdata = testData, type = "raw")
```

Prediction performance of MARS

```r
confusionMatrix(glm.pred, testData$mpg_cat)
```

Model Performance Comparison

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high   53   9
##       low     2  54
##
##                Accuracy : 0.9068
##                  95% CI : (0.8393, 0.9525)
##     No Information Rate : 0.5339
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8142
##
##  Mcnemar's Test P-Value : 0.07044
##
##             Sensitivity : 0.9636
##             Specificity : 0.8571
##          Pos Pred Value : 0.8548
##          Neg Pred Value : 0.9643
##              Prevalence : 0.4661
##          Detection Rate : 0.4492
##    Detection Prevalence : 0.5254
##       Balanced Accuracy : 0.9104
##
##        'Positive' Class : high
##
```

```r
confusionMatrix(mars.pred, testData$mpg_cat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high   54   9
##       low     1  54
##
##                Accuracy : 0.9153
##                  95% CI : (0.8497, 0.9586)
##     No Information Rate : 0.5339
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8313
##
##  Mcnemar's Test P-Value : 0.02686
##
##             Sensitivity : 0.9818
```

```
##             Specificity : 0.8571
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.9818
##              Prevalence : 0.4661
##          Detection Rate : 0.4576
##    Detection Prevalence : 0.5339
##       Balanced Accuracy : 0.9195
##
##        'Positive' Class : high
##
```

The MARS model does not significantly improve prediction performance compared to logistic regression. Both models achieve high accuracy. Logistic regression has a slightly higher overall accuracy and specificity, while MARS has a slightly better sensitivity, meaning it identifies high-mileage cars more effectively.

However, the differences are minimal, and both models perform well. Since there is no substantial improvement in predictive performance, logistic regression may be preferable due to its interpretability and simplicity.

```
library(MASS)
```

(c) Perform linear discriminant analysis using the training data. Plot the linear discriminant(s)

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggplot2)
library(caret)

ctrl3 = trainControl(method = "repeatedcv", repeats = 5,
summaryFunction = twoClassSummary,
classProbs = TRUE)

set.seed(22)

model.lda = train(x = trainData[, c("cylinders", "displacement", "horsepower",
                                    "weight", "acceleration", "year", "origin")],
                  y = trainData$mpg_cat,
                  method = "lda",
                  metric = "ROC",
                  trControl = ctrl3)

print(model.lda)
```

```
## Linear Discriminant Analysis
##
```

```
## 274 samples
##   7 predictor
##   2 classes: 'high', 'low'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 247, 247, 247, 246, 247, 247, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.9653239  0.9601905  0.8371429
```

```
lda.pred2 = predict(model.lda, newdata = testData)
```

The Linear Discriminant Analysis (LDA) model performed well in distinguishing between the two classes ('high' and 'low') of 'mpg_cat'. The ROC (=0.95) suggests excellent overall model discrimination. The sensitivity of 96.35% shows that the model is very effective at identifying instances of the 'high' class, while the specificity of 82.82% indicates that it is reasonably good at classifying the 'low' class.

Overall, the model is performing well, with strong ability to correctly classify both classes.

(d) Which model will you choose to predict the response variable? Plot its ROC curve and report the AUC. Next, select a probability threshold to classify observations and compute the confusion matrix. Briefly interpret what the confusion matrix indicates about your model's performance. The MARS model has the highest accuracy and sensitivity. Therefore, it is the most reliable at predicting both classes ('high' and 'low'). While the LDA model also had a good performance, in terms of ROC and sensitivity, it had a lower specificity. This suggests the LDA model did poorer in correctly classifying the 'low' class compared to the GLM and MARS models.

Therefore, the MARS model is the best to predict the response variable, mpg_cat. The ROC curve was plotted and the AUC is 0.981.

The confusion matrix shows that the model performs well in predicting both "high" and "low" gas mileage cars, with an overall accuracy of 92.37%. The strong Kappa value (0.8455) indicates a high level of agreement between the predicted and actual outcomes. Overall, the model demonstrates strong performance, especially in identifying high-mileage cars.

```
library(pROC)
```

ROC Curve

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

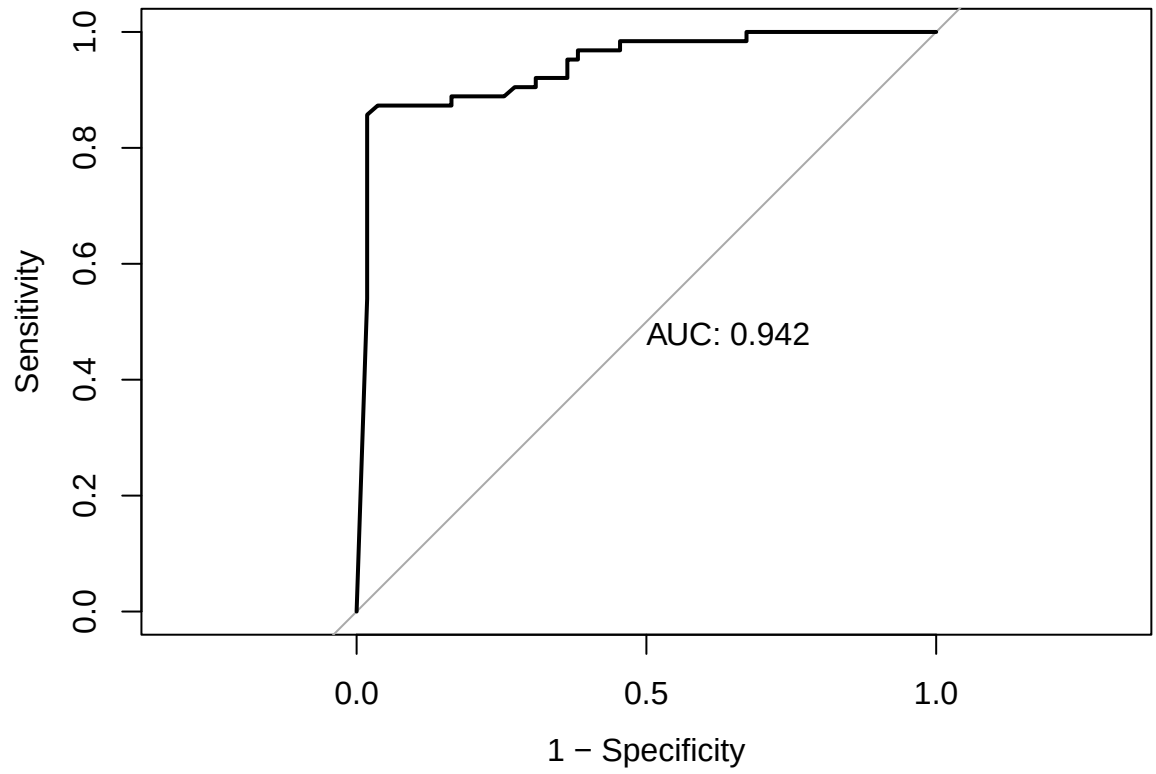
```r
mars.prob = predict(mars.fit, newdata = testData, type = "prob")[, 2]

roc.mars = roc(testData$mpg_cat, mars.prob)
```

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

## Setting direction: controls < cases
```

```r
plot(roc.mars, legacy.axes = TRUE, print.auc = TRUE)
```



Confusion Matrix

```r
test.pred.prob <- predict(mars.fit, newdata = testData, type = "prob")

test.pred.prob_pos <- test.pred.prob[, "high"]  # Assuming "high" is the positive class

test.pred <- ifelse(test.pred.prob_pos > 0.5, "high", "low")  # Binary classification: "high" or "low"

test.pred <- factor(test.pred, levels = levels(testData$mpg_cat))

confusionMatrix(data = as.factor(test.pred),
                reference = testData$mpg_cat,
                positive = "high")  # Change to "low" if "low" is the positive class
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction high low
##       high   54   9
##       low     1  54
##
##                Accuracy : 0.9153
##                  95% CI : (0.8497, 0.9586)
##     No Information Rate : 0.5339
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8313
##
##  Mcnemar's Test P-Value : 0.02686
##
##             Sensitivity : 0.9818
##             Specificity : 0.8571
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.9818
##              Prevalence : 0.4661
##          Detection Rate : 0.4576
##    Detection Prevalence : 0.5339
##       Balanced Accuracy : 0.9195
##
##        'Positive' Class : high
##
```