p8106_hw3_qz2266

Qing Zhou

2023-03-23

Data import

In this problem, 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.

```
auto = read.csv("data/auto.csv") %>%
  mutate(
    mpg_cat = as.factor(mpg_cat),
    mpg_cat = fct_relevel(mpg_cat, c("low", "high")),
    year = factor(year),
    origin = as.factor(origin))
```

Data spliting

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

```
set.seed(1)

# training data
trainRows <- createDataPartition(y = auto$mpg_cat, p = 0.7,list = FALSE)
train_data = auto[trainRows, ]
test_data = auto[-trainRows, ]

x = model.matrix(mpg_cat ~ ., train_data)[,-1]
y = train_data$mpg_cat
x_test = model.matrix(mpg_cat ~ ., test_data)[,-1]
y_test = test_data$mpg_cat
```