FML Assignment 2

Pooja Babu

2025-02-04

Loading the required packages

```
library(caret)
```

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

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

```
library(ISLR)
```

View the Carseats dataset

```
head(Carseats,n=10)
```

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1   9.50       138     73          11        276   120       Bad  42        17
## 2  11.22       111     48          16        260    83      Good  65        10
## 3  10.06       113     35          10        269    80    Medium  59        12
## 4   7.40       117    100           4        466    97    Medium  55        14
## 5   4.15       141     64           3        340   128       Bad  38        13
## 6  10.81       124    113          13        501    72       Bad  78        16
## 7   6.63       115    105           0         45   108    Medium  71        15
## 8  11.85       136     81          15        425   120      Good  67        10
## 9   6.54       132    110           0        108   124    Medium  76        10
## 10  4.69       132    113           0        131   124    Medium  76        17
##    Urban  US
## 1    Yes Yes
## 2    Yes Yes
## 3    Yes Yes
## 4    Yes Yes
## 5    Yes  No
## 6     No Yes
## 7    Yes  No
## 8    Yes Yes
## 9     No  No
## 10    No Yes
```

Checking the dimensions of the dataset

```
dim(Carseats)
```

```
## [1] 400  11
```

Summary of the dataset

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

Splitting the dataset into 3 sets (training_data, validation_data, test_data)

Step 1: Partitioning carseats datasets into temp (300 records) and test (100 records)

```
temp = createDataPartition(y=1:nrow(Carseats),p=300/nrow(Carseats),list = FALSE)
```

Checking the dimensions of temp dataset

```
dim(temp)
```

```
## [1] 300   1
```

Subsetting the dataset using temp so that the values in train_index (300 records) does not repeat in test_data (100 records)

```
train_index=Carseats[temp,]
test_data=Carseats[-temp,]
```

Checking the dimensions of train_index and test_data

```
dim(train_index)
```

```
## [1] 300  11
```

```
dim(test_data)
```

```
## [1] 100  11
```

Step 2: Now we divide the train_index into 2 sets in such a way that 200 records are put in training_data and 100 records in validation_data.

We create a temp_train_index to store the 200 records that we select randomly from the train_index

```
temp_train_index<-createDataPartition(y=1:nrow(train_index),p=200/nrow(train_index),list = FALS
E)
```

Checking the dimensions of temp_train_index

```
dim(temp_train_index)
```

```
## [1] 200   1
```

Subsetting the data into training_data (200 records) and validation_data(100 records)

```
training_data=train_index[temp_train_index,]
validation_data=train_index[-temp_train_index,]
```

Checking the dimensions of training_data and validation_data

```
dim(training_data)
```

```
## [1] 200  11
```

```
dim(validation_data)
```

```
## [1] 100  11
```

Checking the first 6 rows of all the datasets

```
head(validation_data)
```

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 16  8.71       149     95           5        400   144    Medium  76        18
## 17  7.58       118     32           0        284   110      Good  63        13
## 21  6.41       125     90           2        367   131    Medium  35        18
## 22 12.13       134     29          12        239   109      Good  62        18
## 27  8.33       107    115          11        496   131      Good  50        11
## 28  5.27        98    118           0         19   107    Medium  64        17
##    Urban  US
## 16    No  No
## 17   Yes  No
## 21   Yes Yes
## 22    No Yes
## 27    No Yes
## 28   Yes  No
```

```
head(training_data)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 6 10.81       124    113          13        501    72       Bad  78        16
## 8 11.85       136     81          15        425   120      Good  67        10
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 6    No Yes
## 8   Yes Yes
```

```
head(test_data)
```

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 5   4.15       141     64           3        340   128       Bad  38        13
## 7   6.63       115    105           0         45   108    Medium  71        15
## 19 13.91       110    110           0        408    68      Good  46        17
## 20  8.73       129     76          16         58   121    Medium  69        12
## 23  5.08       128     46           6        497   138    Medium  42        13
## 26 14.90       139     32           0        176    82      Good  54        11
##    Urban  US
## 5    Yes  No
## 7    Yes  No
## 19    No Yes
## 20   Yes Yes
## 23   Yes  No
## 26    No  No
```

Summary of training_data , validation_data and test_data

```
summary(training_data)
```

```
##      Sales          CompPrice         Income        Advertising
##  Min.   : 0.000   Min.   : 77.0   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.360   1st Qu.:115.0   1st Qu.: 40.75   1st Qu.: 0.000
##  Median : 7.555   Median :125.0   Median : 69.50   Median : 6.000
##  Mean   : 7.412   Mean   :124.9   Mean   : 69.23   Mean   : 6.765
##  3rd Qu.: 9.322   3rd Qu.:135.0   3rd Qu.: 92.25   3rd Qu.:11.250
##  Max.   :15.630   Max.   :175.0   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 12.0   Min.   : 24.0   Bad   : 55   Min.   :25.00   Min.   :10.00
##  1st Qu.:133.8   1st Qu.:101.0   Good  : 43   1st Qu.:39.75   1st Qu.:12.00
##  Median :274.0   Median :119.5   Medium:102   Median :54.00   Median :14.00
##  Mean   :266.0   Mean   :116.8                Mean   :53.09   Mean   :13.87
##  3rd Qu.:388.8   3rd Qu.:132.0                3rd Qu.:64.25   3rd Qu.:16.00
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.00
##  Urban       US
##  No : 61   No : 68
##  Yes:139   Yes:132
##
##
##
##
```

```
summary(validation_data)
```

```
##      Sales          CompPrice         Income        Advertising
##  Min.   : 0.530   Min.   : 89.0   Min.   : 21.00   Min.   : 0.00
##  1st Qu.: 5.577   1st Qu.:114.8   1st Qu.: 45.00   1st Qu.: 0.00
##  Median : 6.900   Median :124.0   Median : 69.00   Median : 5.00
##  Mean   : 7.313   Mean   :124.8   Mean   : 67.74   Mean   : 6.36
##  3rd Qu.: 8.815   3rd Qu.:133.0   3rd Qu.: 88.00   3rd Qu.:11.00
##  Max.   :13.550   Max.   :162.0   Max.   :120.00   Max.   :25.00
##    Population        Price        ShelveLoc       Age          Education
##  Min.   : 10.0   Min.   : 49.0   Bad   :15   Min.   :25.00   Min.   :10.00
##  1st Qu.:143.5   1st Qu.:103.0   Good  :16   1st Qu.:42.00   1st Qu.:11.00
##  Median :271.5   Median :115.0   Medium:69   Median :54.50   Median :14.00
##  Mean   :266.3   Mean   :116.1               Mean   :53.48   Mean   :14.25
##  3rd Qu.:412.0   3rd Qu.:131.0               3rd Qu.:65.00   3rd Qu.:17.00
##  Max.   :508.0   Max.   :171.0               Max.   :80.00   Max.   :18.00
##  Urban      US
##  No :27   No :36
##  Yes:73   Yes:64
##
##
##
##
```

```
summary(test_data)
```

```
##      Sales          CompPrice         Income        Advertising
##  Min.   : 0.160   Min.   : 88.0   Min.   : 21.00   Min.   : 0.00
##  1st Qu.: 5.140   1st Qu.:115.0   1st Qu.: 42.00   1st Qu.: 0.00
##  Median : 7.580   Median :125.0   Median : 66.50   Median : 5.00
##  Mean   : 7.847   Mean   :125.4   Mean   : 68.42   Mean   : 6.65
##  3rd Qu.:10.065   3rd Qu.:137.0   3rd Qu.: 89.25   3rd Qu.:12.25
##  Max.   :16.270   Max.   :161.0   Max.   :120.00   Max.   :23.00
##    Population      Price        ShelveLoc       Age          Education
##  Min.   : 16   Min.   : 53.0   Bad   :26   Min.   :25.00   Min.   :10.00
##  1st Qu.:143   1st Qu.: 95.5   Good  :26   1st Qu.:37.75   1st Qu.:11.00
##  Median :270   Median :116.5   Medium:48   Median :55.50   Median :14.00
##  Mean   :261   Mean   :113.4               Mean   :53.63   Mean   :13.62
##  3rd Qu.:381   3rd Qu.:130.0               3rd Qu.:71.00   3rd Qu.:16.00
##  Max.   :504   Max.   :173.0               Max.   :80.00   Max.   :18.00
##  Urban      US
##  No :30   No :38
##  Yes:70   Yes:62
##
##
##
##
```