Analysis

2025-09-22

```
library(ISLR2)
library(MASS)
```

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

```
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data1 <- read.csv('AmazonWithScores.csv')
data1 <- na.omit(data1)
```

```
train_idx <- sample(1:1264,1000)
data1_train <- data1[train_idx, ]
data1_test  <- data1[-train_idx, ]
table(data1$review.score)
```

```
##
##   1   2   3   4   5
## 173 103 103 182 703
```

```
lda_fit <- lda(factor(review.score) ~ compound + Num_Characters + Num_Words+but_count+Num_Exclamations,
lda_fit
```

```
## Call:
## lda(factor(review.score) ~ compound + Num_Characters + Num_Words +
##     but_count + Num_Exclamations, data = data1, subset = train_idx)
##
## Prior probabilities of groups:
##     1     2     3     4     5
## 0.138 0.090 0.082 0.136 0.554
##
## Group means:
##     compound Num_Characters Num_Words but_count Num_Exclamations
## 1 -0.1631014       331.2826  62.93478 0.4130435        0.6304348
## 2  0.3030000       430.1556  82.17778 0.7444444        0.4888889
## 3  0.4137561       406.9024  76.18293 0.9512195        0.2317073
## 4  0.5112426       432.0294  82.72794 0.8529412        0.5735294
## 5  0.6557004       405.5072  76.37004 0.5018051        0.9981949
##
## Coefficients of linear discriminants:
##                            LD1          LD2         LD3          LD4
## compound          2.140326065 -0.233546898  0.06148116 -0.037210232
## Num_Characters    0.003161898  0.003063572 -0.03538717  0.008475826
## Num_Words        -0.017910514 -0.013768221  0.19339318 -0.029795481
## but_count        -0.154525206 -1.082166240 -0.29279503 -0.643765023
## Num_Exclamations  0.073912195  0.310380690  0.02756471 -0.344479587
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.8312 0.1429 0.0244 0.0015
```

```
lda_pred <- predict(lda_fit, data1_test)
names(lda_pred)
```

```
## [1] "class"     "posterior" "x"
```

```
head(lda_pred$posterior)
```

```
##             1          2          3         4          5
## 7  0.01386084 0.09440896 0.12401295 0.2539192 0.51379807
## 9  0.22245999 0.15197629 0.05346032 0.1470291 0.42507434
## 11 0.01288470 0.06275294 0.08237688 0.1548114 0.68717411
## 17 0.02347852 0.08930222 0.08276893 0.2132974 0.59115294
## 24 0.01633055 0.04163332 0.04118413 0.0880945 0.81275751
## 28 0.29441187 0.14112286 0.37566788 0.1462833 0.04251408
```

```r
head(lda_pred$class)
```

```
## [1] 5 5 5 5 5 3
## Levels: 1 2 3 4 5
```

```r
real_val <- factor(data1_test$review.score)

tab_lda <- table(Predicted = lda_pred$class, Actual = real_val)
tab_lda
```

```
##          Actual
## Predicted   1   2   3   4   5
##         1  18   3   7  12  10
##         2   0   0   0   0   0
##         3   0   1   2   1   1
##         4   0   0   1   1   0
##         5  17   9  11  32 138
```

```r
acc_lda <- mean(lda_pred$class == real_val)
acc_lda
```

```
## [1] 0.6022727
```

```r
caret::confusionMatrix(lda_pred$class, real_val)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2   3   4   5
##          1  18   3   7  12  10
##          2   0   0   0   0   0
##          3   0   1   2   1   1
##          4   0   0   1   1   0
##          5  17   9  11  32 138
##
## Overall Statistics
##
##                Accuracy : 0.6023
##                  95% CI : (0.5405, 0.6618)
##     No Information Rate : 0.5644
##     P-Value [Acc > NIR] : 0.1189
##
##                   Kappa : 0.2489
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity           0.51429  0.00000 0.095238 0.021739   0.9262
## Specificity           0.86026  1.00000 0.987654 0.995413   0.4000
```

```
## Pos Pred Value         0.36000      NaN 0.400000 0.500000   0.6667
## Neg Pred Value         0.92056  0.95076 0.926641 0.828244   0.8070
## Prevalence             0.13258  0.04924 0.079545 0.174242   0.5644
## Detection Rate         0.06818  0.00000 0.007576 0.003788   0.5227
## Detection Prevalence   0.18939  0.00000 0.018939 0.007576   0.7841
## Balanced Accuracy      0.68727  0.50000 0.541446 0.508576   0.6631
```