PREDICTION OF RED WINE QUALITIES

Sercan Öncü

3/25/23

```r
options(repos = list(CRAN="http://cran.rstudio.com/"))
install.packages("readr")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("gridExtra")
install.packages("grid")
library(readr)
library(ggplot2)
library(dplyr)
library(gridExtra)
library(grid)
```

```r
winequality_red <- read_csv("C:/Users/serca/Downloads/archive/winequality-red.csv")
```

In this project we are going to use red wine qualities dataset. With 11 independent value such as (fixed acidity, volatil acidity, citric acid, residual sugar, etc.) we try to predict quality.

1. Our dataset has 1599 row and 12 column(variable).

```r
dim(winequality_red)
```

```
[1] 1599   12
```

2. Datasets have 11 column and all columns are double type.

```r
glimpse(winequality_red)
```

```
Rows: 1,599
Columns: 12
$ `fixed acidity`        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7~
$ `volatile acidity`     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600~
$ `citric acid`          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00,~
$ `residual sugar`       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.~
$ chlorides              <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069~
$ `free sulfur dioxide`  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, ~
$ `total sulfur dioxide` <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 10~
$ density                <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978,~
$ pH                     <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39,~
$ sulphates              <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47,~
$ alcohol                <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 1~
$ quality                <dbl> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5,~
```

3. This part we are going to split our dataset as a train and test. Our aim is train the regression model and predict dependent value.

```r
set.seed(1)
index <- sample(1:nrow(winequality_red), round(nrow(winequality_red)*0.80))
train <- winequality_red[index, ]
test <- winequality_red[-index, ]
x <- lm(quality ~ . , data =train)
summary(x)
```

```
Call:
lm(formula = quality ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.52213 -0.35627 -0.04738  0.44215  2.00517 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.342e+01  2.323e+01   1.008 0.313441    
`fixed acidity`        2.725e-02  2.862e-02   0.952 0.341210    
`volatile acidity`    -1.032e+00  1.353e-01  -7.625 4.77e-14 ***
`citric acid`         -1.857e-01  1.626e-01  -1.142 0.253523    
`residual sugar`       2.223e-02  1.623e-02   1.370 0.170964    
chlorides             -1.456e+00  4.849e-01  -3.002 0.002731 ** 
`free sulfur dioxide`  4.483e-03  2.452e-03   1.828 0.067725 .  
```

```
`total sulfur dioxide` -3.086e-03  8.191e-04  -3.767 0.000173 ***
density                -1.976e+01  2.372e+01  -0.833 0.404986
pH                     -3.457e-01  2.165e-01  -1.597 0.110515
sulphates               9.192e-01  1.257e-01   7.311 4.69e-13 ***
alcohol                 2.834e-01  2.909e-02   9.741  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6474 on 1267 degrees of freedom
Multiple R-squared:  0.3536,    Adjusted R-squared:  0.348
F-statistic: 63.02 on 11 and 1267 DF,  p-value: < 2.2e-16
```

4. Adjusted R-squared is 0.3536 which is not good at explain of red wine quality. That mean is that independent variables do not explain the target variable %35.36.

```
y <- predict(x, test[,-12])
head(y)
```

```
       1        2        3        4        5        6
5.146827 5.057541 5.047935 5.674195 5.379635 5.343364
```

```
e <- test$quality - y
head(e)
```

```
          1           2           3           4           5           6
-0.14682721 -0.05754095 -0.04793474 -0.67419529 -0.37963524  0.65663602
```

```
rmse_model <- sqrt(mean(y ^ 2))
rmse_model
```

```
[1] 5.642643
```

```
rmse_train <- sqrt(mean((x$residuals) ^ 2))
rmse_test <- rmse_model
rmse_train - rmse_test
```

```
[1] -4.998299
```

5. Difference is negative. It means that the performance of the model is better on test set than train set. It's meaning is maybe overfitting problem.