Predicting Wine Quality Through Linear Regression

Stephen Bernardo

2023-12-10

Installing Packages

```
#install.packages('tidyverse')
```

Loading Libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(lubridate)
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Reading the data set

```
df <- read.csv('winequality-red.csv')
```

Checking the structure and summary to see if data types are correct

Make sure that all columns to be used in the analysis are set to the correct data types.

```
str(df)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
summary(df)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

The concept of training and testing sets is often associated with predictive modeling, it still makes sense to use a similar approach in explanatory modeling

Linear Regression Model

```
#install.packages("caTools")
library(caTools)

set.seed(123)

split <- sample.split(df$quality, SplitRatio = 0.8)
dftrain <- subset(df, split == TRUE)
```

```r
dftest <- subset(df, split == FALSE)

# Linear Regression Model
model <- lm(quality ~ volatile.acidity + chlorides + total.sulfur.dioxide + density + pH + sulphates + a
summary(model)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol, data = dftrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59109 -0.36172 -0.06145  0.47410  1.91217
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.977456  12.200015   0.162 0.871264
## volatile.acidity      -0.926640   0.112899  -8.208 5.48e-16 ***
## chlorides             -2.041464   0.435139  -4.692 3.01e-06 ***
## total.sulfur.dioxide -0.002373   0.000565  -4.200 2.86e-05 ***
## density                2.367245  12.021204   0.197 0.843920
## pH                    -0.496746   0.135527  -3.665 0.000257 ***
## sulphates              0.880198   0.128261   6.863 1.05e-11 ***
## alcohol                0.302022   0.021651  13.949  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6518 on 1270 degrees of freedom
## Multiple R-squared:   0.35,  Adjusted R-squared:  0.3464
## F-statistic:  97.7 on 7 and 1270 DF,  p-value: < 2.2e-16
```

```r
predictions <- predict(model, newdata = dftest)
```

Reporting model performance

The business problem calls for an explanatory model, where we need to investigate the relationship of different chemical metrics (volatile.acidity, chlorides, total.sulfur.dioxide, density, pH, sulphates, alcohol) on quality.

Linear regression was chosen as the preferred machine learning because it provides an analysis to study the relationships between variables. In this setting, linear regression is in line with the objective of trying to understand how different chemical metrics affect quality of wine.

The output model presents one statistically insignificant variable while the rest of the variables are statistically significant. volatile.acidity, chlorides, total.sulfur.dioxide, pH, sulphates, and alcohol has an impact on the quality of the wine, where we can observe from the output above that all the p-values are less than 0.05. On the other hand, density is not statistically significant because the p-value of 0.843920 is above the 0.05 threshold.

Based from this model, we can make recommendations to improve wine quality. We can tell the winery to consider adjusting their wine-making process to improve volatile.acidity, chlorides, total.sulfur.dioxide, pH, sulphates, and alcohol. The density of the wine may not be a strong indication of quality. The winery should investigate the quality of wine and determine which step of the wine-making process will need improvement (ex. aging and storage).

The linear regression model provides insights into the relationship between chemical metrics (volatile.acidity,

chlorides, total.sulfur.dioxide, density, pH, sulphates, alcohol) and quality of the wine produced. These findings could help refine the wine production process, allocate budgets (e.g., the need for new wine equipment and quality raw material). Looking at the Adjusted R-squared of 0.3464, another recommendation is to consider additional variables outside of chemical composition in order to further determine other factors that affect quality (ex: room temperature, cleanliness of the barrels measured by microbial levels).