Finn Tomasula Martin

COSC-4557

Warmup Exercise

Report

Set up

I chose to use the winequality-red.csv dataset for both the models created in this exercise.

Regression

The code for the regression model is in LinearRegression.R. For this model, we are trying to predict wine quality using the rest of the variables as features. The results of the linear regression model are the following:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As you can see the most significant features are volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol, with higher amounts of volatile acidity, chlorides and total sulfur dioxide reducing the quality of the wine, while higher levels of sulphates and alcohol increasing the quality of the wine. We can use adjusted R-squared to evaluate the fit of our model. Adjusted R-squared is 0.3561 which means that 35.61% of the variation in quality can be explained by the rest of the features. So, the fit isn’t great but it isn’t terrible either.

Classification

The code for the classification model is in LinearClassification.R. For this model, we are trying to predict classes of wine quality using the rest of the variables as features. The two classes for quality are good and bad, where good is anything that has a quality score greater than or equal to the mean of all quality scores and bad is anything less than the mean. The classification model chosen for this task is logistic regression. The results of the classification model are the following:

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          42.949948  79.473979   0.540  0.58890
fixed.acidity         0.135980   0.098483   1.381  0.16736
volatile.acidity     -3.281694   0.488214  -6.722 1.79e-11 ***
citric.acid          -1.274347   0.562730  -2.265  0.02354 *
residual.sugar        0.055326   0.053770   1.029  0.30351
chlorides            -3.915713   1.569298  -2.495  0.01259 *
free.sulfur.dioxide   0.022220   0.008236   2.698  0.00698 **
total.sulfur.dioxide -0.016394   0.002882  -5.688 1.29e-08 ***
density              -50.932385  81.148745  -0.628  0.53024
pH                    -0.380608   0.720203  -0.528  0.59717
sulphates             2.795107   0.452184   6.181 6.36e-10 ***
alcohol               0.866822   0.104190   8.320  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As you can see the results seem to mirror our linear regression model as the most significant features are the same and the sign of those features are the same too. To evaluate the performance of this model we can make a confusion table to see how accurate it is, The confusion table is given here:

```
pred   bad good
  bad  549  214
  good 195  641
```

The percentage that our model evaluated the quality correctly is ((549 + 641) / 1599) * 100 = 74.24%. This percentage is pretty large so our model does a pretty good job predicting the quality of wine.

Sources

Used for help building classification model:

https://rstudio-pubs-static.s3.amazonaws.com/349547_0f2e13952bda439da6faee386190bc2c.html