Insights from Data

Extra Credit Report - Predicting Wine Quality Using Linear Regression

Maxwell Fundi Njiru

0. **Project code**

All project code for this task may be found on my Github repository

1. **Explain the concept of linear regression and its functionality.**

   Linear regression is a type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables by fitting a linear equation to the data. It is the most common type of regression approach. It generally uses the mathematical equation of **y = ax+b.** There are majorly two types of linear regression namely
   - Simple linear regression - This is the most basic form of linear regression which uses one independent variable and one dependent variable.
   - Multiple linear regression – This is used where there is more than one predictor variables

2. **Utilize linear regression to predict wine quality in the dataset**
   In this task I did the following activities
   - **Data Preprocessing**
     - I imported the data using the read_excel function from the readxl package
     - As required by this task, I did binarizing of the type variable with white = 1 and red=0.
     - I split the dataset into 80% training and 20% test data

   - **Model Building**
     - I constructed using the training data to predict the Quality of wine. All predictors were used including the binarized type variable. The regression model was fitted using the lm() function
     - Interpreting the coefficients of linear regression – from the summary function I found that
       - Volatile Acidity, Residual Sugar, Free Sulfur Dioxide, Alcohol, and Sulphates were the most significant predictors. They had low p-values and large coefficients.
       - Alcohol and Sulphates had positive impacts on wine quality. Wines with higher alcohol content and sulphate levels are associated with higher quality.
       - Volatile Acidity and Density had negative impacts on wine quality. This shows that higher levels of these variables are associated with lower wine quality.
       - Citric Acid and Chlorides seemed to be less important predictors in this model because of higher p-values. **These were removed.**

See the output below

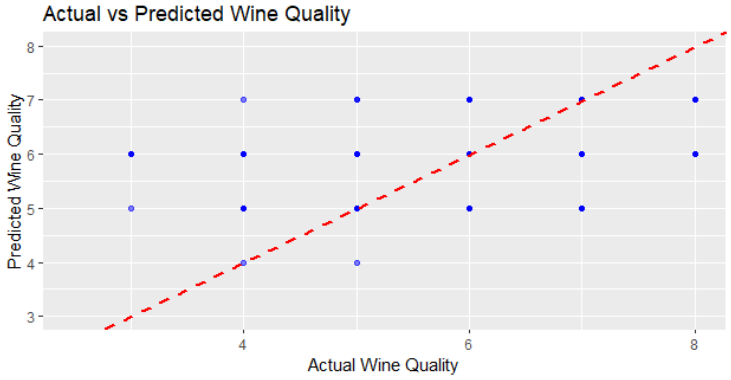
```
> options(scipen = 999)
> summary(model)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + alcohol + type, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6277 -0.4707 -0.0439  0.4574  3.0090

Coefficients:
                       Estimate   Std. Error t value             Pr(>|t|)
(Intercept)          118.5453446  15.4446562   7.675   0.0000000000000196 ***
fixed.acidity          0.0926649   0.0176113   5.262   0.0000001485009281 ***
volatile.acidity      -1.4843780   0.0899371 -16.505 < 0.0000000000000002 ***
citric.acid           -0.0477824   0.0895369  -0.534              0.593599
residual.sugar         0.0688707   0.0065520  10.511 < 0.0000000000000002 ***
chlorides             -0.5112783   0.3760184  -1.360              0.173979
free.sulfur.dioxide    0.0042520   0.0008490   5.008   0.0000005683725255 ***
total.sulfur.dioxide  -0.0013056   0.0003621  -3.606              0.000314 ***
density             -117.7734093  15.6707932  -7.515   0.0000000000000664 ***
pH                     0.5298898   0.1010175   5.246   0.0000001620311600 ***
sulphates              0.7063052   0.0845799   8.351 < 0.0000000000000002 ***
alcohol                0.2060831   0.0197734  10.422 < 0.0000000000000002 ***
type                  -0.3883555   0.0626639  -6.197   0.0000000006182533 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7319 on 5186 degrees of freedom
Multiple R-squared:  0.2947,    Adjusted R-squared:  0.2931
F-statistic: 180.6 on 12 and 5186 DF,  p-value: < 0.00000000000000022
```

Linear regression with all variables

```
> summary(model)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
    sulphates + alcohol + type, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6159 -0.4693 -0.0442  0.4580  3.0097

Coefficients:
                       Estimate   Std. Error t value             Pr(>|t|)
(Intercept)          121.4930036  15.3279394   7.926 0.00000000000002744 ***
fixed.acidity          0.0932138   0.0171128   5.447 0.00000005357965742 ***
volatile.acidity      -1.4813263   0.0845680 -17.516 < 0.0000000000000002 ***
residual.sugar         0.0702820   0.0064879  10.833 < 0.0000000000000002 ***
free.sulfur.dioxide    0.0042274   0.0008488   4.981 0.00000065422700 7094 ***
total.sulfur.dioxide  -0.0013173   0.0003603  -3.656              0.000259 ***
density             -120.8822970  15.5434172  -7.777 0.00000000000000008892 ***
pH                     0.5604484   0.0991083   5.655 0.00000016427032119 ***
sulphates              0.6824631   0.0831868   8.204 0.00000000000000291 ***
alcohol                0.2060024   0.0196253  10.497 < 0.0000000000000002 ***
type                  -0.3811467   0.0619841  -6.149 0.00000000837520232 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7319 on 5188 degrees of freedom
Multiple R-squared:  0.2944,    Adjusted R-squared:  0.293
F-statistic: 216.4 on 10 and 5188 DF,  p-value: < 0.00000000000000022

>
```

Linear regression with without citric.acid and chlorides

- **Performance Evaluation**
  - Scatterplot comparing actual vs predicted wine quality – From the scatterplot below, we can conclude that the model seems to predict the wine quality with good performance because most of the points are close to gradient line. However, there is as aspect that the model predicts better higher wine qualities in comparison to l when predicting wines with lower quality scores.



  - **Evaluation of the performance on test data -** To evaluate the model's performance, I used several metrics namely
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R-squared
    - Adjusted R-squared
    - Residual analysis
  - From the evaluations found Mean Squared Error (MSE) of 0.6379, Root Mean Squared Error (RMSE) of 0.7987 , an R-squared of 0.1840 and an Adjusted R-squared of 0.1777.
  - From these results I note that the model seems to be fairly accurate since it has low MSE and RMSE values. Additionally, with low R-squared and adjusted R-squared, the model has limited explanatory power and may not be adequately capturing the factors influencing the wine quality. While the model provides some insights into the relationship between the predictors and wine quality, the model could be

improved by having additional predictors, transforming existing features, or using nonlinear models.

- The same sentiment is supported by the residuals plot using a histogram below which shows a non-normal distribution with slight right skewness. This shows that a few larger positive residuals may be present. But generally, the histogram plot suggests that the linear regression model provides a reasonable fit to the data.