Final Paper

Wine Quality Testing

Venkata Kanishk Chaganti

Sagar Paresh Vora

Content

1.Introduction

Now-a-days, it is very important for the food and beverage industry to maintain consistency in their products through carefully selected quality control procedures. Red wine is popular worldwide and is beneficial due to the presence and amount of its compounds. The tradition of winemaking and wine consumption has been known for many centuries. The ancient Romans knew the health benefits of wine and popularized it. In this paper, we mainly focus on the quality control of red wine and the possible modes of interaction between these 12 variables and red wine phenolics that lead to the necessary changes in the dataset.

The process for making good quality involves crushing the grapes, alcoholic fermentation (with sugar, yeast, alcohol, carbon dioxide), aging process, malolactic conversion, racking , filtration and finally into the bottles which are sold in the stores. The objective of this project is to anticipate the quality of wine using the ingredients which are used in the preparation process. This paper will describe the data set and which regression algorithms are useful for predicting the best quality of wine.

2.Data Description

- What is the source?

Paulo Cortez, University of Minho, Guimarães, Portugal, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal in the year 2009.

- What are the response and potential predictor variables?

The Predicted Variables are:

1 - fixed acidity, 2 - volatile acidity, 3 - citric acid, 4 - residual sugar, 5 - chlorides ,

6 - free sulfur dioxide, 7 - total sulfur dioxide, 8 - density, 9 - pH, 10 - sulphates,

11 - alcohol

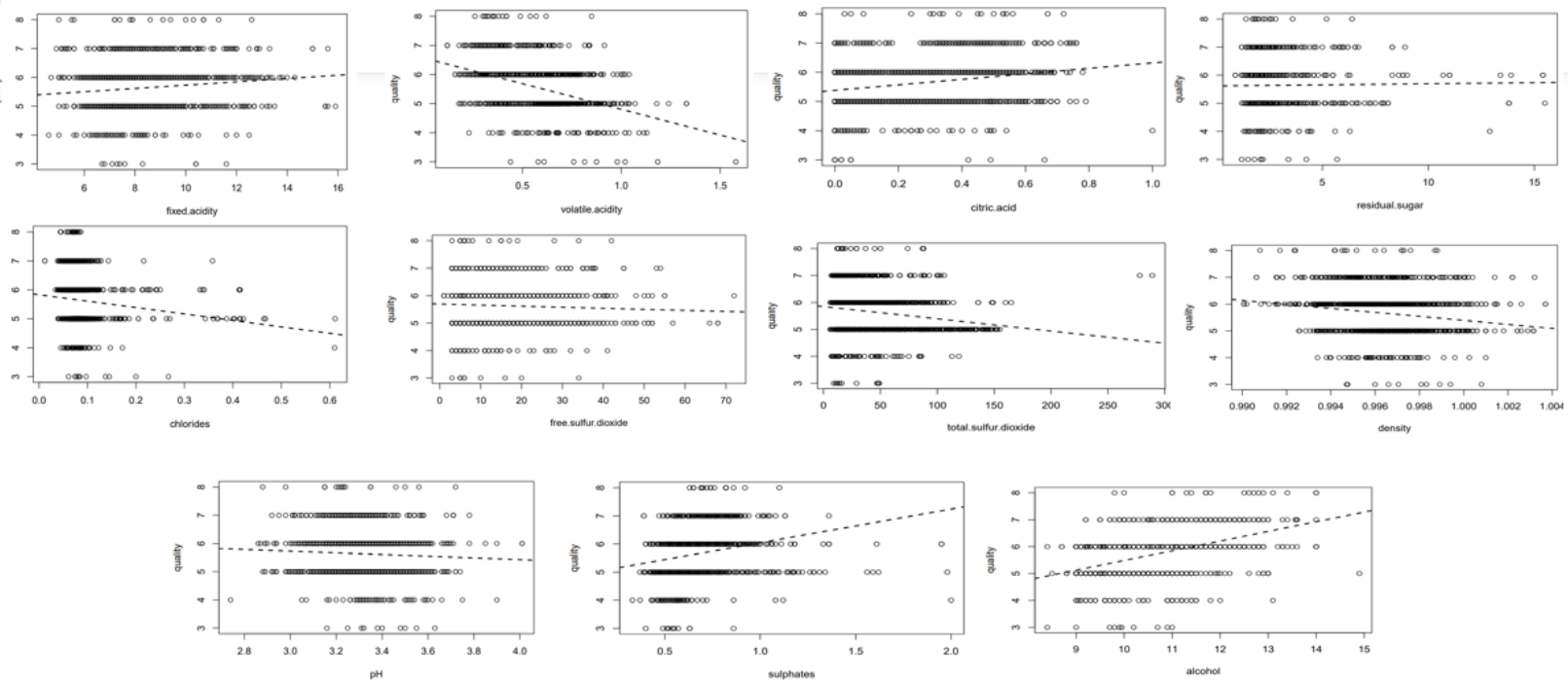
The Response variable is:

12 – quality

- What is the dimension?

There are in total 1599 rows and 12 columns in our dataset.

Figure 2.1



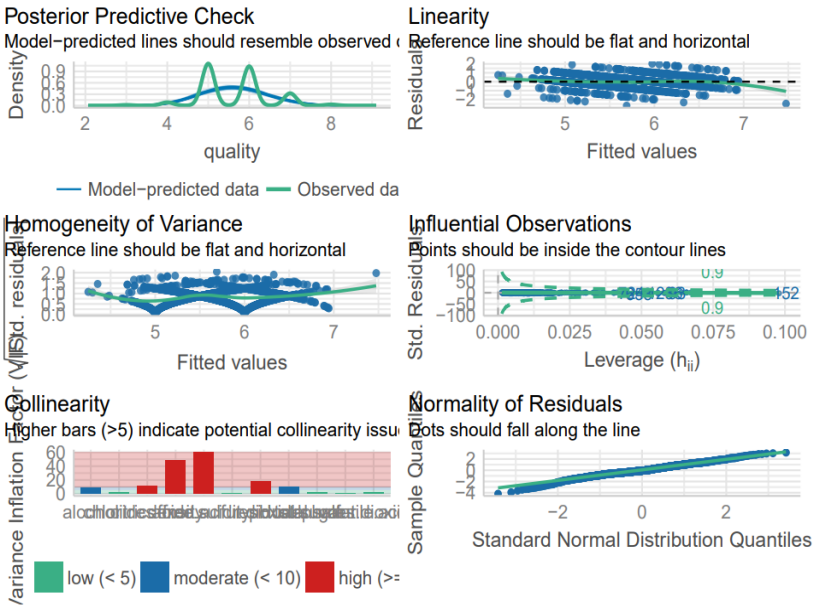
In the above figure 2.1, we can see that volatile.acidity, sulphates and alcohol appear to have the strongest relationship with quality because they have the best fit line as the line passes through most of the data and they also have a very high correlation coefficient .While free.sulfur.dioxide, pH and residual.sugar appear to have the weakest relationships with quality because they have a horizontal line which reflects that the relationship with the data is poor and they also have a very less correlation coefficient.

```{r}
table(red$quality)
```

```
  3   4   5   6   7   8 
 10  53 681 638 199  18 
```

From the figure 2.2, we can also see that there are six different types of quality present in our dataset. As we see that the 3rd class contains 10 variables and the 8th class contains 18 variables whereas the 5th, 6th and 7th class contains a lot of variables which makes this dataset imbalanced.

Figure 2.2.



From the figure 2.3, we found out that there were no such patterns observed so we can say that linearity is satisfied. There were few high leverage points but we can ignore that as the data is too big. We can say that there were few variables which had high collinearity, so we will take them out of the data-set.

Figure 2.3.

3. Methods and Results

3.1 Multiple linear regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Formula:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + \epsilon$$

MLR equation for quality of red wine.

$$quality = \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 citric.acid + \beta_4 residual.sugar + \beta_5 chlorides + \beta_6 free.sulfur.dioxide + \beta_7 total.sulfur.dioxide + \beta_8 density + \beta_9 pH + \beta_{10} sulphates + \beta_{11} alcohol + \epsilon$$

```
##
## Call:
## lm(formula = quality ~ ., data = red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## 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
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Figure 3.1.1

In figure 4.1.1 we can observe that all the β values of every predictors. The probabilities of some variables are not significant enough for our model, so we will be using the AIC step() function to determine which variables are significant enough to predict the quality of red wine.

3.1.1 Akaike Information Criterion (AIC):

AIC measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size. The AIC value can be given by:

AIC = n log(RSS/n) + 2(p + 1)

Smaller the AIC value, better the model.

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.4300987  0.4029168  10.995  < 2e-16 ***
## volatile.acidity    -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides           -2.0178138  0.3975417  -5.076 4.31e-07 ***
## free.sulfur.dioxide  0.0050774  0.0021255   2.389    0.017 *
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                  -0.4826614  0.1175581  -4.106 4.23e-05 ***
## sulphates            0.8826651  0.1099084   8.031 1.86e-15 ***
## alcohol              0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Figure 3.1.2

After performing backward step function we can see that fixed.acidity, citric.acid, residual.sugar, density variables are not significant for quality. Wecan also say there is almost 36% variability in the response is explained by the predictors in the model.

Reduced MLR model:

$$quality = 4.4300987 - 1.0127527\, volatile.acidity - 2.0178138 chloride$$

$$+ \ 0.0050774 \ free.sulfur.dioxide \ - \ 0.0034822 total.sulfur.dioxide$$
$$- \ 0.4826614 pH \ + \ 0.8826651 sulphate \ + \ 0.2893028 \ alcohol$$

The above equations states that the quality of red wine decreases by 1.0127527 with respect to $volatile.acidity$, 2.0178138 with respect to $chloride$, 0.0034822 with respect to $total.sulfur.dioxide$, 0.4826614 with respect to pH.

The quality of red wine increases by 0.0050774 with respect to free.sulfur.dioxid, 0.882665 with respect to sulphates and 0.2893028 with respect to alcohol.

3.1.3 Box-Cox Transformation :

Box-cox transformation provides us with λ value which help us with which transformation is needed.
If **λ =1,** then no transformation is needed.
If **λ =** 0, then log transformation is needed.
If **λ** =0.5, then square root transformation is needed.

```
summary(powerTransform(lm))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.9138           1        0.683       1.1447
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 64.24689  1 1.1102e-15
##
## Likelihood ratio test that no transformation is needed
##                            LRT df   pval
## LR test, lambda = (1) 0.532163  1 0.4657
```

Figure 3.1.3

By performing power transformation, it provides us with **λ =1.** As we got 1 as **λ** value, no transformation is not needed to be performed.

3.1.4 F- Test

First we will test if atleast one of the predictors are significant to determine the quality of red wine.

H0: β1 = β2 = ...β11 = 0

None of the predictors are significant for determining the quality of red wine.

Ha: at least one $\beta j \neq 0$

At least one of the predictors is significant for determining the quality.

From figure 4.1.4 P-value is 2.2e-16*** with significance and we can conclude that we reject the null hypothesis and at least one of the predictors are useful for determining the quality of wine.

```
Analysis of Variance Table

Model 1: quality ~ 1
Model 2: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + alcohol
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   1598 1042.17
2   1587  666.41 11    375.75 81.348 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.1.4

Next, we perform a partial F-test to check if variables removed in a reduced model are not significant for the model.

H0 : $\beta 1 = \beta 3 = \beta 4 = \beta 8 = 0$

Variables fixed.acidity, citric.acid, residual.sugar, density are significant for determining the quality of wine.

Ha :$\beta 1 \neq 0$ or $\beta 3 \neq 0$ or $\beta 4 \neq 0$ or $\beta 8 \neq 0$

Variables fixed.acidity, citric.acid, residual.sugar, density are significant for determining the quality of wine.

From figure 4.1.5 p-value is 0.6124. Since the p-value is not significant we can declare that we fail to reject the null hypothesis and Variables fixed.acidity, citric.acid, residual.sugar, density are not significant for determining the quality of wine.

```
anova(lm2,lm)
```

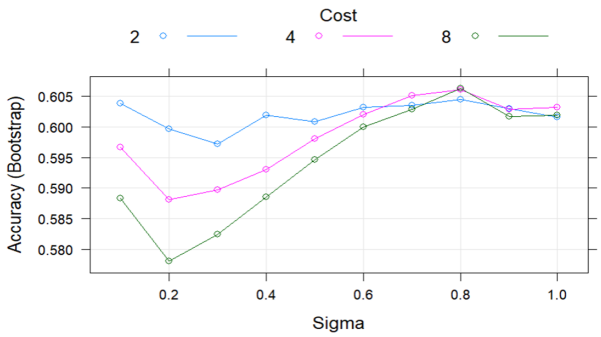
```
## Analysis of Variance Table
##
## Model 1: quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
## Model 2: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1591 667.54
## 2   1587 666.41  4    1.1264 0.6706 0.6124
```

Figure 3.1.5

3.2 Support Vector Machine.

Figure 3.2

```
Confusion Matrix and Statistics

          Reference
Prediction   3   4   5   6   7   8
         3   0   0   0   0   0   0
         4   1   0   3   2   0   0
         5   1   7 152  66  11   0
         6   0  10  53 114  17   3
         7   0   0   0   9  27   1
         8   0   0   0   0   1   1

Overall Statistics

               Accuracy : 0.6138
                 95% CI : (0.5685, 0.6576)
    No Information Rate : 0.4342
    P-Value [Acc > NIR] : 2.208e-15

                  Kappa : 0.3685

 Mcnemar's Test P-Value : NA
```

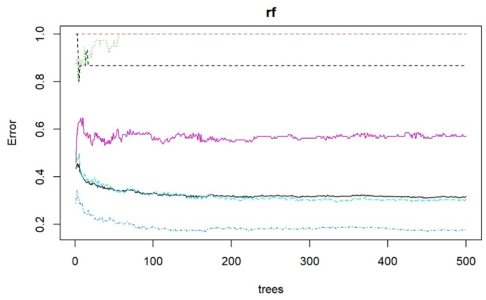


Support Vector Machine is a linear model for classification and regression problems. In Figure 4.2.1, we can observe that at Sigma 0.8 cost 8 we get the best accuracy which comes out to be 61.38% respectively. Also, we have a kappa value of 0.3685 which says that we have an average model so we can use a different model to check which model is best for our dataset.

3.3 Random Forest

Random forest can be used for both classification and regression problems . We performed random forest on a red wine dataset to see how well it performs to determine the quality of wine. Default number of 500 trees were used.

```
Call:
 randomForest(formula = factor(quality) ~ ., data = Train, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 31.55%
Confusion matrix:
  3 4   5   6  7 8 class.error
3 0 1   7   0  0 0   1.0000000
4 1 0  25  13  0 0   1.0000000
5 0 0 406  83  4 0   0.1764706
6 0 1 104 302 26 1   0.3041475
7 0 0   4  69 56 1   0.5692308
8 0 0   0   6  7 2   0.8666667
```




```
## Overall Statistics
##
##                Accuracy : 0.6576
##                  95% CI : (0.6132, 0.7001)
##     No Information Rate : 0.4342
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4413
##
```

Figure 3.3

By performing random forest we can observe that there is 31.55% average error for each observation using predictions from the trees that do not contain the same OOB samples. At the 95% confidence interval the values stay under (0.6132, 0.7001) and provide us with the accuracy of 66%. The plot between trees and errors as shown in figure 4.2.2 represents how the error reduces then the number of trees increases.

4. Conclusion

•As the accuracy of random forest is much greater than any of the other regression models, we can conclude that it works best for predicting the quality of red wine.

5. Code Appendix

The R code for the wine quality testing project can be found in : https://github.com/chvkanishk/project