Individual Assignment 5

Ali Rafieepouralavialavijeh, 20900871

Problem statement and data used:

Dear manager, I will analyze two datasets related to red and white Vinho Verde wine samples from the north of Portugal in this analysis. The goal is to understand the difference between white and red wine. These datasets, when merged, have 6497 observations in total and have a series of physicochemical test results as features to model the wine quality. The question this analysis answer is "***how to use the physicochemical test results to predict whether a wine is red or white?***". As such, the dataset is cleaned and added a new column specifying the type of wine.

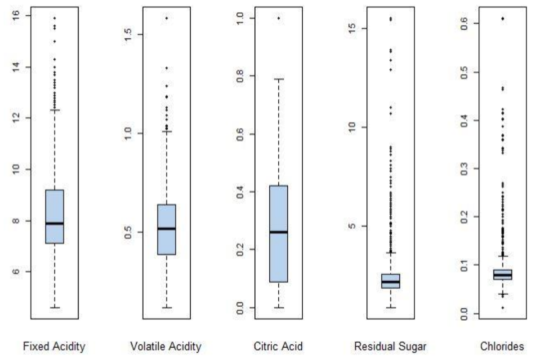


Figure 1 Distribution of physicochemical test results for red wine

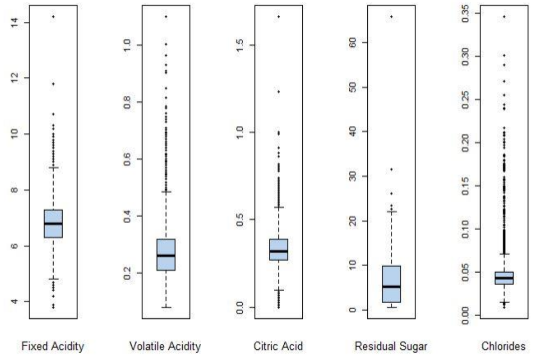


Figure 2 Distribution of physicochemical test results for white wine

Planning:

In this analysis, I am going to use logistic regression to develop a model to identify the type of a particular wine with its physicochemical characteristics, as such the outcome variable is "color," and the predictor variables are "*Fixed Acidity*," "*Volatile Acidity*," "*Citric Acid*," "*Residual Sugar*," "*Chlorides*," "Alcohol," "Total sulfur dioxide." These predictors are selected such that most assumptions of logistic regression would be satisfied, and the model result is acceptable (please see more details in the appendix). The first step in building a model is to check its basic assumptions, and as such, I first check two assumptions, and then after making the model, I review another three assumptions.

- **incomplete information**: need full combinations of variables (*Inspection of the merged dataset shows that all set of combinations for both type of wines is considered*.)
- **complete separation**: logistic regression fails if our data does not overlap (*Visual inspection of the dataset shows that the two outcome results are not separated, you could see an example of this in the appendix*)

The next step is to make the model and test for three other assumptions as stated below:

- **Linearity**: linear relationship between continuous predictors and the logit of the outcome variable (*this assumption is violated. For this, new variables that reflect the interaction between each predictor and the log of that predictor are added and the results demonstrated that they are* *significant. This means that the predictors do not have linear relationship with the logit. Please see the results of the analysis in more details in the appendix*)

- **Independence of errors**: this is same as the regression model (*Durbin-Watson test of independence failed to provide enough evidence on independency of errors*)
- **No perfect multicollinearity:** predictor variables should not correlate highly (*I inspected the VIF (Variance Inflation Factor) to investigate multicollinearity. The largest VIF was 1.345027, less than 10. The average vif was 1.191421, close to 1)*

Analysis:

Now that the model is built, the model's confidence intervals show that all predictors are significant because they do not include 1. The results of the model demonstrate a high correlation between the predictors and the outcome variable. However, since these relationships are non-linear, we cannot necessarily judge how many times each predictor increases the chance for a wine to be either red or white. The intercept coefficient shows whether we consider other predictors or not it tends to be oriented towards one wine color. This is a clear indication that the dataset is imbalanced (4898 white wine observations versus 1599 red wine observations). Of the features, results show that residual sugar plays a more important role in comparison to other features in determining the wine type. After that, total sulfur dioxide ranks second, and then quality and fixed acidity follow. The chlorides play the least essential effect on determining the wine color.

Confidence intervals of predictor variables		
	2.5 %	97.5 %
(Intercept)	1.647e+04	7.881e+05
Residual sugar	1.174e+00	1.362e+00
Chlorides	1.143e-19	1.642e-15
Total sulfur dioxide	1.052e+00	1.063e+00
Quality	5.330e-01	7.766e-01
Fixed acidity	3.528e-01	4.760e-01
Volatile acidity	1.036e-06	1.348e-05

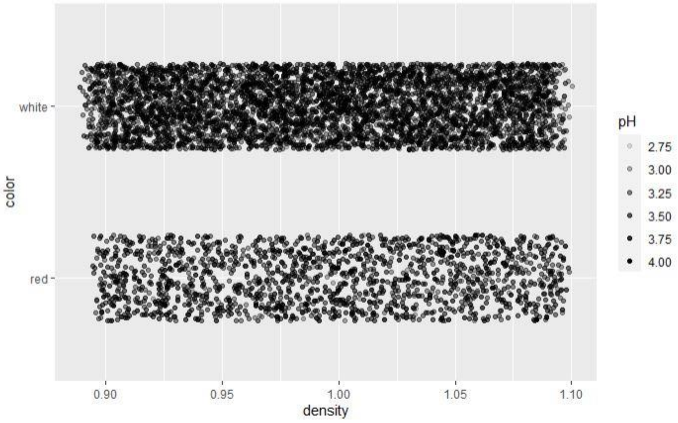
Conclusion:

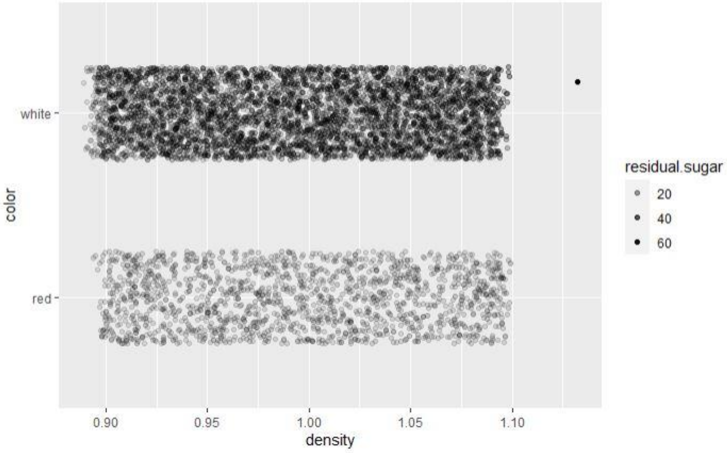
In this analysis, an imbalanced dataset is studied for important wine characteristics that help me distinguish between wine types. In general, since some of the model assumptions are violated, this model would not generalize well. This highly affects the model interpretation of confidence intervals as the relationships do not follow linearity. As a result, I could not conclude how many times a feature is vital in identifying a wine type. I noticed alcohol, sugar residuals, total sulfur dioxide playing the most critical roles, and chlorides, soleplates, and volatile acidity the least influential factors.

I propose using more sophisticated models to identify non-linear relationships between the outcome and predictors. Also, since the dataset is imbalanced, a possible approach to finding more reliable results could be to augment or penalize some data for tackling bias in the model.

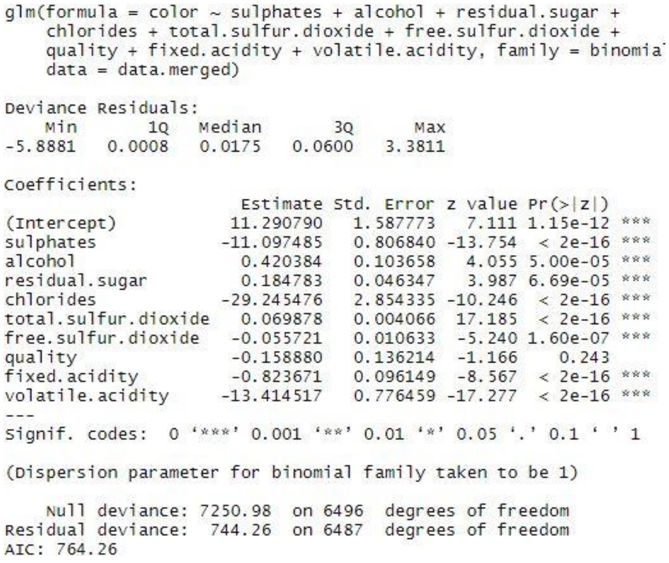
Appendix:

Example images showing the imperfect separation of the data:



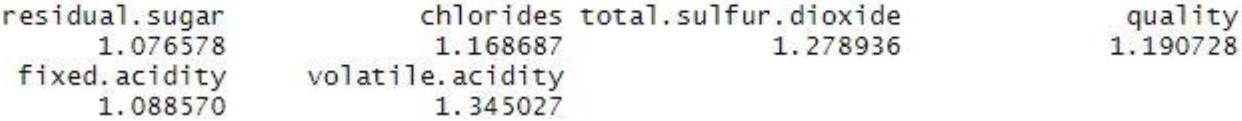


Model details:



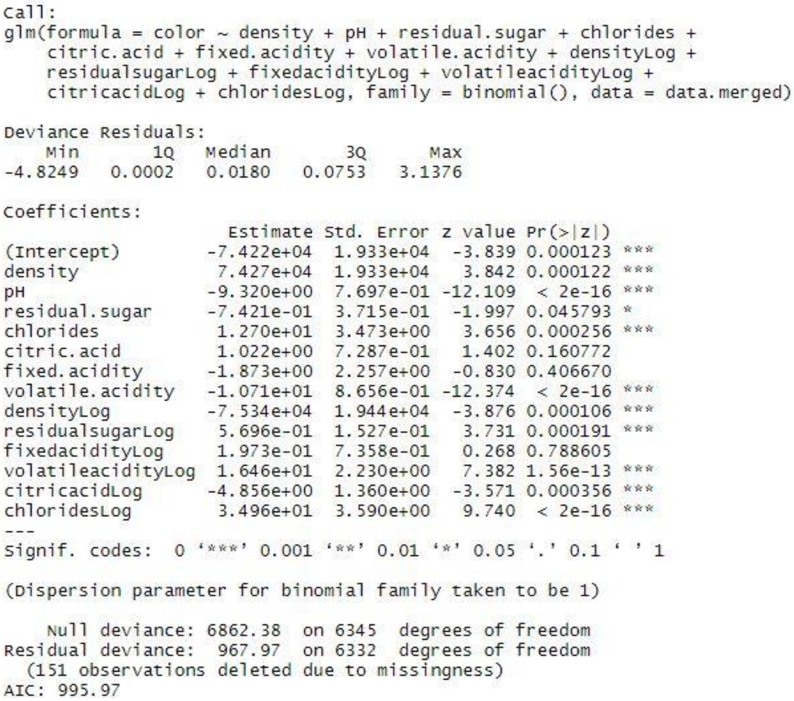
```
glm(formula = color ~ sulphates + alcohol + residual.sugar +
    chlorides + total.sulfur.dioxide + free.sulfur.dioxide +
    quality + fixed.acidity + volatile.acidity, family = binomia
    data = data.merged)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8881   0.0008   0.0175   0.0600   3.3811

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           11.290790   1.587773   7.111 1.15e-12 ***
sulphates            -11.097485   0.806840 -13.754  < 2e-16 ***
alcohol                0.420384   0.103658   4.055 5.00e-05 ***
residual.sugar         0.184783   0.046347   3.987 6.69e-05 ***
chlorides            -29.245476   2.854335 -10.246  < 2e-16 ***
total.sulfur.dioxide   0.069878   0.004066  17.185  < 2e-16 ***
free.sulfur.dioxide   -0.055721   0.010633  -5.240 1.60e-07 ***
quality               -0.158880   0.136214  -1.166    0.243
fixed.acidity         -0.823671   0.096149  -8.567  < 2e-16 ***
volatile.acidity     -13.414517   0.776459 -17.277  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7250.98  on 6496  degrees of freedom
Residual deviance:  744.26  on 6487  degrees of freedom
AIC: 764.26
```

Testing for multicollinearity:



```
residual.sugar            chlorides total.sulfur.dioxide               quality
      1.076578             1.168687             1.278936              1.190728
 fixed.acidity     volatile.acidity
      1.088570             1.345027
```

Testing for linearity:



```
Call:
glm(formula = color ~ density + pH + residual.sugar + chlorides + 
    citric.acid + fixed.acidity + volatile.acidity + densityLog + 
    residualsugarLog + fixedacidityLog + volatileacidityLog + 
    citricacidLog + chloridesLog, family = binomial(), data = data.merged)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.8249   0.0002   0.0180   0.0753   3.1376  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -7.422e+04  1.933e+04  -3.839 0.000123 ***
density             7.427e+04  1.933e+04   3.842 0.000122 ***
pH                 -9.320e+00  7.697e-01 -12.109  < 2e-16 ***
residual.sugar     -7.421e-01  3.715e-01  -1.997 0.045793 *  
chlorides           1.270e+01  3.473e+00   3.656 0.000256 ***
citric.acid         1.022e+00  7.287e-01   1.402 0.160772    
fixed.acidity      -1.873e+00  2.257e+00  -0.830 0.406670    
volatile.acidity   -1.071e+01  8.656e-01 -12.374  < 2e-16 ***
densityLog         -7.534e+04  1.944e+04  -3.876 0.000106 ***
residualsugarLog    5.696e-01  1.527e-01   3.731 0.000191 ***
fixedacidityLog     1.973e-01  7.358e-01   0.268 0.788605    
volatileacidityLog  1.646e+01  2.230e+00   7.382 1.56e-13 ***
citricacidLog      -4.856e+00  1.360e+00  -3.571 0.000356 ***
chloridesLog        3.496e+01  3.590e+00   9.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6862.38  on 6345  degrees of freedom
Residual deviance:  967.97  on 6332  degrees of freedom
  (151 observations deleted due to missingness)
AIC: 995.97
```