White Wine Quality Prediction

Abstract

The perceived quality of white wine is thought to be dependent on a variety of chemicophysical properties such as sugar, alcohol, etc. A dataset containing wine experts' scores of *vinho verde* white wine and their respective chemicophysical composition was analysed using multiple linear regression to find an reasonable model. This model used to predict the quality of white wine was evaluated with respect to standard measures of performance and stability. The possibility of using the model to predict the quality of red wine was also explored and rejected. The model also confirmed background research that suggested sugar plays an important role in wine quality.

1 Introduction

This report aims to predict the perceived quality of white wine using multiple linear regression based off of 11 chemicophysical properties. Background research[1] suggested that sugar generally plays an important part in wine quality, and so it is expected that sugar is significant in the final regression model. Further, the applicability of the model trained on white wine data is considered on red wine data.

2 Dataset

The choice to focus on the white wine dataset[2] (instead of the red wine) was motivated by its greater number of observations and consequent reliability. The *vinho verde* white wine samples come from the north of Portugal, where 11 continuous and quantitative chemicophysical properties (Appendix A.1) were measured. Wine quality was measured as the response variable based on the median score from 0 to 10 of at least 3 evaluations made by wine experts in 2009 who did not have any knowledge of the chemicophysical compositions.

3 Analysis

3.1 Assumptions: Independence and Linearity

To perform a valid multiple linear regression, independence of observations is required. Wine quality is an inherently subjective concept, so independence may be violated if the same judge had scored multiple wines. However, this issue was somewhat mitigated with wine quality defined as the median of scores by at least three judges, so the assumption of independence can be considered met.

Linearity between wine quality and each chemicophysical property is also required. A grid of scatter plots relating wine quality to each explanatory variable confirmed reasonable linearity as no plots showed clear non-linear shape.

3.2 Model Selection

A desirable linear regression model should keep only features that, according to the data, significantly affect white wine quality to maximise prediction accuracy and minimise overfitting. Standard forwards and backwards stepwise variable selection approaches were used to find model candidates, with high accuracy and low numbers of predictor variables, that locally minimise the AIC (Akaike Information Criterion).

Both approaches selected the model below with 8 predictor variables out of the 11 total:

$$\begin{aligned}\widehat{\text{quality}} = 154.1062 + 0.0681(\text{fixed. acidity}) \\ -1.8881(\text{volatile. acidity}) + 0.0828(\text{residual. sugar}) \\ +0.0033(\text{free. sulfur. dioxide}) - 154.2913(\text{density}) \\ +0.6942(\text{pH}) + 0.6285(\text{sulphates}) + 0.1932(\text{alcohol})\end{aligned} \tag{1}$$

A stable model can be loosely defined as a model which is still likely to be chosen by some process even when small changes are made to the data, and is an important consideration in science as models should be reproducible by other studies.

To produce a variable inclusion plot (Appendix A.3), the data is bootstrapped where each data point is given a random weight, a GIC-minimising model (generalised information criteria) is chosen, and an inclusion probability is assigned to each explanatory variable. This is repeated multiple times as the penalty multiplier λ (which penalises models with more explanatory variables), is increased. Explanatory variables whose lines are close or below the RV (random variable) line should not be considered as they are unlikely to appear in an 'op-

[1] Somm, G. (2019, May 29). Understanding the Role of Sugar in Wine. Retrieved November 13, 2021, from https://daily.sevenfifty.com/understanding-the-role-of-sugar-in-wine/

[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553. doi:10.1016/j.dss.2009.05.016

timal' model, which is the case for our model Equation (1).

The model stability plot (Appendix A.4) is produced by bootstrapping similar to the variable inclusion plot. A larger circle shows a higher probability of selection (given a fixed number of variables) which implies a more stable model. It was found that the model chosen by stepwise selection Equation (1) was the most stable model for 8 variables, and was also very stable in comparison to all other model combinations.

3.3 Homoskedasticity and Normality Assumptions

Homoskedasticity requires that the variance of the residuals is constant over the range of its fitted values. The residual points (Appendix A.2) above and below the zero line were roughly equally spread with no discernible pattern, so homoskedasticity is satisfied.

Normality requires that the error terms follow a normal distribution. We found that the standard normal QQ-plot of residuals roughly followed a straight line (which suggests normality) with some deviation in the tails. Regardless, the large number of observations (4898) means that the Central Limit Theorem can be relied on for valid inferences.

4 Results

To predict white wine quality, we arrived at the final model Equation (1). This model was cross-validated with the most stable 3-,4-, and 7-variable models as they showed the greatest comparative stability. The 3- and 4-variable models had much greater error (RMSE and MAE) and lower R^2. Both the 7- and 8-variable (Equation (1)) models were very similar in error, R^2, and stability, hence we chose the 8-variable model because it was also chosen by both forwards and backwards stepwise selection.

To interpret this model, a one-unit increase in each of the explanatory variables in Equation (1) will on average impact the wine quality score by its corresponding coefficient. For example, our model predicts that an increase of 1 gdm^{-3} in residual sugar results on average in a 0.0828 increase in wine quality.

5 Discussion

A coefficient of determination measures the percentage of variation explained by the regression model, and is used to measure in-sample strength of our model. Our model Equation (1) had a coefficient of 0.282, and explains 28% of the observed variation in wine quality. The low coefficient of determination is actually quite reasonable in this field because quality is a very subjective response variable, and was measured discretely.

Although the model represents a function in which variable inputs will produce a predictor value for wine quality, our model is reflective only of data that has been inputted into the model. The model should not extrapolate wine quality scores computed using chemicophysical values the range of data inputted into the model. We argue the range of plausible input values should be within 1.5×IQR from the 1st and 3rd quartiles respectively, which eliminates anomalous values. Additionally, the plausible value lower bound must be ≥ 0.

Considering our white wine model Equation (1) on the red wine data, we found the RMSE to be 1.03. As a baseline comparison, we considered a null model from the red wine data, which uses the mean of all red wine qualities as the predictor. It's RMSE was 0.807, which shows that the white wine model Equation (1) is not suitable for predicting red wine. This was not unexpected as we found that red wine has very different chemicophysical properties, even outside our plausible value ranges.

Another limitation of our analysis was that `geom_jitter()` was used throughout this report for visualisations to randomly offset the points in scatter plots both vertically and horizontally. This was done because wine quality takes only integer values between 0 and 10. As suggested in lectures and on Ed, wine quality was considered as a discretisation of some underlying continuous scale.

Finally, we found that residual sugar did appear in the final model, and had a p-value < 0.0001 indicating that sugar does in fact play an important role in wine quality.

6 Conclusion

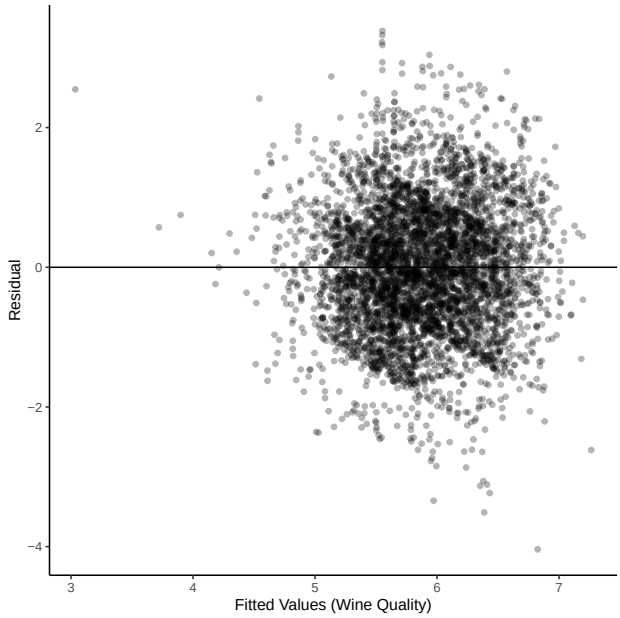
To conclude, we found that an 8-variable multiple regression model Equation (1) was most appropriate to predict the quality of white wine with respect to AIC stepwise variable selection, model stability, and out-of-sample performance. Additionally, we found that the white wine model was not appropriate to predict the quality of red wine, and that sugar did indeed play a strong role in white wine quality as expected.

A Appendices

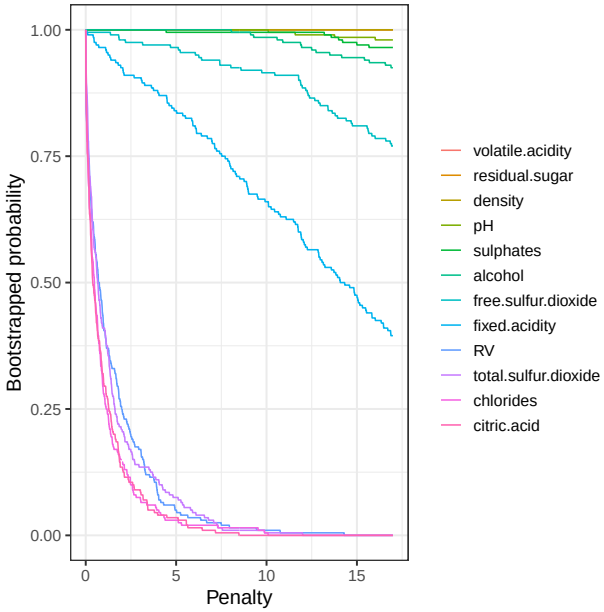
A.1 Explanatory Variables

Explanatory Variables	Units
Fixed acidity	g(tartaric acid)·dm^{-3}
Volatile acidity	g(acetic acid)·dm^{-3}
Citric acid	g·dm^{-3}
Residual Sugar	g·dm^{-3}
Chlorides	g(sodium chloride)·dm^{-3}
Free Sulfur Dioxide	mg·dm^{-3}
Total Sulfur Dioxide	mg·dm^{-3}
Density	g·cm^{-3}
pH	pH
Sulphates	g(potassium sulphate)·dm^{-3}
Alcohol	vol.%

A.2 Homoscedasticity



A.3 Variable Inclusion Plot



A.4 Model Stability Plot