Assessing the Predictability of Wine Quality Measures using Linear Models

Group: Carolina Garza, Michael Saenz, Ray Thomas

Introduction:

For our exploration, we have chosen the "Wine Quality Data Set" from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality?ref=hackernoon.com). This data set comes with information about both red and white vinho verde wine samples, collected from the northern part of Portugal. For conciseness sake, we have chosen to focus on just the white wine data, which consists of 4898 observations. It has 11 input variables which were gathered from certain physicochemical tests done on the samples. The data set also consists of 1 output variable (Quality). The variables are as follows:

1. **Fixed Acidity -** Numerical, measured in g(tartaric acid)/dm3, range is [3.8, 14.2]
2. **Volatile Acidity -** Numerical, measured in g(acetic acid)/dm3, range is [0.08, 1.10]
3. **Citric Acid -** Numerical, measured in g/dm3, range is [0, 1.66]
4. **Residual Sugar -** Numerical, measured in g/dm3, range is [0.6, 65.8]
5. **Chlorides -** Numerical, measured in g(sodium chloride)/dm3, range is [0.009, 0.346]
6. **Free Sulfur Dioxide -** Numerical, measured in mg/dm3, range is [2, 289]
7. **Total Sulfur Dioxide -** Numerical, measured in mg/dm3, range is [9, 440]
8. **Density -** Numerical, measured in g/cm3, range is [0.9, 1]
9. **pH -** Numerical, range is [2.72, 3.82]
10. **Sulphates -** Numerical, measured in g(potassium sulphate)/dm3, range is [0.22, 1.08]
11. **Alcohol -** Numerical, measured in vol.%, range is [8, 14.2]
12. **Quality -** Numerical - discrete, score between 0 and 10, range is [3, 9]

The focus of our exploration will be in identifying which factors are most important toward determining the quality of wine. We also expect to assess the efficacy of general linear models in predicting the quality. In order to identify the important factors, we will utilize the correlation matrix, as well as stepwise regression. Furthermore, we will explore the relation between the factors by comparing linear models built with different combinations of the 11 input variables as well as the potential interaction effects that might play an important role in quality prediction.

Methodology:

In order to analyze the data, we first make sure to explore the distribution of the data. We also look at the correlation of variables within the data. We utilize a histogram and QQ-plot, which makes it possible to visualize the overall distribution of the data frequencies. Using a correlation matrix, we take note of which features are the most heavily correlated. We then graph those highly correlated features to visualize the relation for ourselves. In order to make sure we would have the best selection of what features to use for our model, we begin by creating a preliminary model using the full set of features. Then we utilize stepwise selection to get the best linear model. We compare the results from the backward, forward, and both directional selection

to evaluate the differences. The assessment metric for stepwise selection is based on the AIC of each model.

Once we have our model, we check the r-squared and adjusted r-squared values to evaluate linearity. Then we proceed to calculate prediction accuracy to determine the viability of using the model for real prediction purposes. We also build polynomial regression models of second and third degree, starting with full models and then utilize stepwise selection as above, to determine if this results in a better model due to the addition of interaction terms which are unaccounted for in the multiple regression model of degree-1.

Data Analysis:

In order to be able to tell which features most heavily impact the quality of white wine, we must first understand the type of distribution of our white wine quality. We first approach this by plotting the frequencies of the different white wine quality scores via a histogram. From this histogram, we are able to see that, for the most part, our data follows a bell-shaped curve, even though it is slightly skewed. To double check our findings, we then use the Q-Q plot, as well. This plot gives us similar findings.

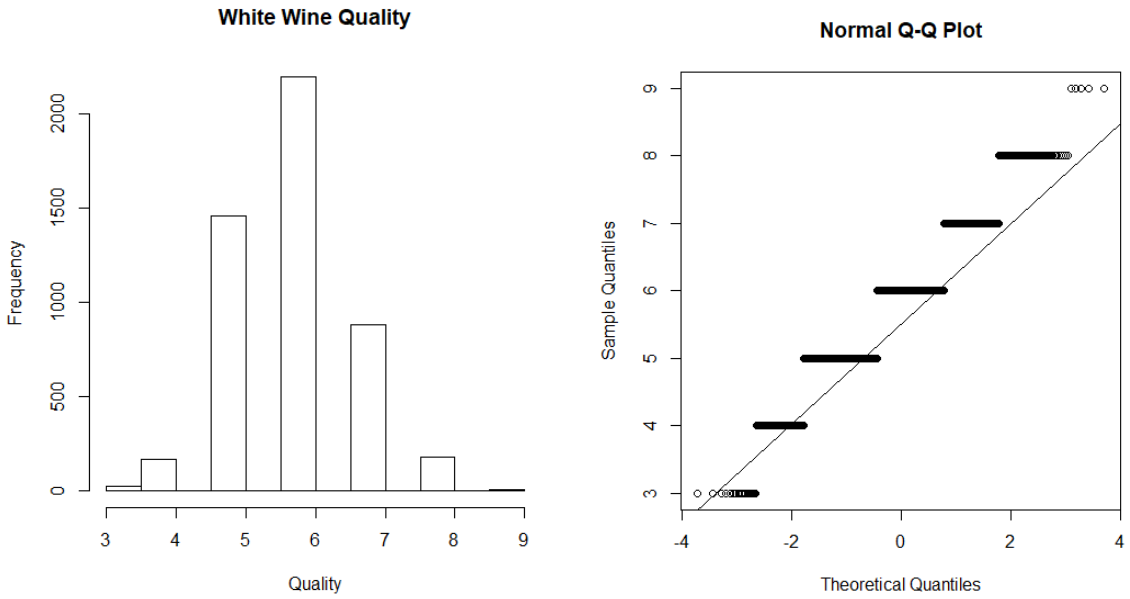

Figure 1, 2: Histogram and normal Q-Q plot for white wine quality

From our correlation matrix we can see the most heavily correlated pairs. By plotting these pairs together, we can also observe this correlation graphically. These pairs are Residual Sugar vs Density, Alcohol vs Density, and Free Sulfur Dioxide vs Total Sulfur Dioxide. We would assume that our best model would only contain one of the features from each pair, as having both would be redundant and may lead to bias.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile.acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07	-0.19
citric.acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08	-0.01
residual.sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45	-0.10
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36	-0.21
free.sulfur.dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	0.00	0.06	-0.25	0.01
total.sulfur.dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45	-0.17
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78	-0.31
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0.00	0.00	-0.09	1.00	0.16	0.12	0.10
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02	0.05
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00	0.44
quality	-0.11	-0.19	-0.01	-0.10	-0.21	0.01	-0.17	-0.31	0.10	0.05	0.44	1.00

Figure 3: Correlation Matrix of the white wine data

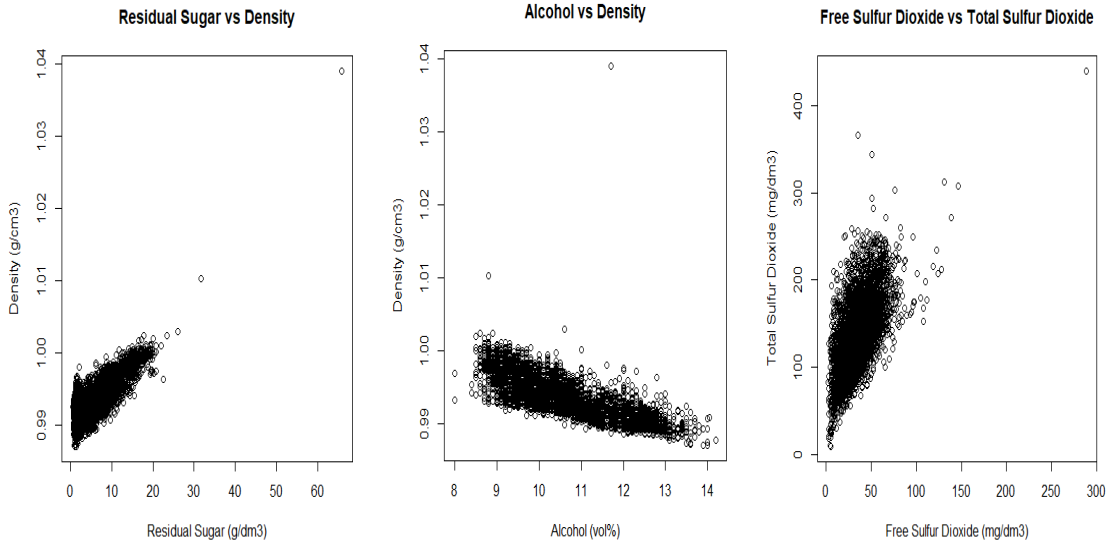

Figure 4-6: Scatterplots of features with the highest correlations

In order to find which features have the most impact on wine quality, we use stepwise selection to find the best linear model. Using the backwards method, we find that our model with the lowest AIC contains the following features: Fixed Acidity, Volatile Acidity, Residual Sugar, Free Sulfur Dioxide, Density, pH, Sulphates, and Alcohol. This is interesting, because although it only keeps Free Sulfur Dioxide from our previously mentioned correlation pair, this model keeps both features from the other most highly correlated pairs: Residual Sugar, Density, and Alcohol. Running stepwise selection with the 'forward' parameter also leads to an identical model as the with the backward method.

```
> summary(step.model)

Call:
lm(formula = whiteWine$Quality ~ Fixed.Acidity + Volatile.Acidity +
    Residual.Sugar + Free.Sulfur.Dioxide + Density + pH + Sulphates +
    Alcohol, data = whiteWine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8246 -0.4938 -0.0396  0.4660  3.1208

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
Fixed.Acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
Volatile.Acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
Residual.Sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
Free.Sulfur.Dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
Density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
Sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
Alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7512 on 4889 degrees of freedom
Multiple R-squared:  0.2818,    Adjusted R-squared:  0.2806
F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

Figure 7: Summary produced in R for our backwards stepwise selection

Now that we have our chosen features for our linear model, we can test the accuracy of its wine quality predictions. Since our wine quality values are discrete, we round the predictions to the closest integer, and calculate the prediction accuracy from this new rounded matrix.

Diving deeper into correlations between predictors and wine quality yields an intriguing discovery. Oftentimes, both high and low quality wines would have similar attributes when looking at one feature at a time. This evidence strongly suggests lack of a linear relationship, and also strongly suggests that an interaction term should be added. Comparing the models with and without interaction terms yields significantly different results.

After deciding that the data is not fit for a purely linear model to describe, we square the model and measure the results. The adjusted r-squared value without an interaction term is 0.2818 while the model with the interaction term yields an adjusted r-squared value of 0.346. This is a ~25% increase in explained variance, which is a notable improvement. This confirms our previous hypothesis. Additionally, we verify the outputs of a cubed model to create additional and more complex predictors. This marginally improves explained variance but the time to run the model is excessive, and the model is needlessly complex for such a minor improvement.

These findings support the idea that the true nature of the quality of wine is far from linear and greatly dependent on multiple variables interacting with each other at varying rates.

Although this is the case, our model is able to capture enough of the complexity to predict the wine qualities better than chance with varying degrees of success depending on the complexity of the model.

	R-Squared	Adjusted R-Squared	Prediction Accuracy
First-Degree	0.2818	0.2806	51.92%
Second-Degree	0.353	0.3475	53.82%
Third-Degree	0.4123	0.3939	55.37%

Figure 8:Table of outputs for the final models of each degree

After iterating through models to find the best one, we look at the R-squared value to see how much of the variance in the quality of the wine could be explained using our model. The best model has an adjusted r squared value of 0.35. Clearly, this is an extremely small number and prompts many more questions. The way each data point of quality for a given wine was collected was by the median of 3 wine tasters' opinions. This procedure was intended to reduce the variance of the assigned qualities of the wine in the data set and heavily concentrated the values near the mean (6). The data used to build the model was unbalanced to overrepresent average wine qualities.

Conclusion:

The features that have the greatest impact on wine quality are Fixed Acidity, Volatile Acidity, Residual Sugar, Free Sulfur Dioxide, Density, pH, Sulphates, and Alcohol. Through the method of using our linear models, we were not able to predict wine quality efficiently. One of the issues we had in prediction using our linear models was that our response variable was discrete, and not continuous. If we were to use a model more focused on classification, such as Random Forests or SVM (as the original researchers found to yield the best results), this would not have been as much of an issue. Furthermore, our data contained some extreme outliers, as can be seen in our correlation plots and box plots of each feature. These outliers were not the same for each feature, and some features, such as Alcohol had none. This made it difficult for us to choose whether or not to remove them.

Another shortcoming of being able to predict wine quality is that taste is very subjective. Not only is taste the most mysterious of the human senses, but the study of the relationship between physicochemical properties and sensory perception are still not fully understood to this day. Using real world knowledge about wine tasting data, these results are not out of the ordinary. It is established that a significant portion of one's perception on the quality of the wine is purely based on knowing the price of the wine before drinking it. A reasonable conclusion to draw from this is that taste is by and large a subjective opinion.

References:

Cortez, Paulo, et al. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” ScienceDirect, Decision Support Systems, 9 June 2009, https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub.

Appendix:

	Feature							
	Fixed Acidity	Alcohol	Free Sulfur Dioxide	Volatile Acidity	Sulphates	pH	Residual Sugar	Density
Number of Outliers	119	0	50	186	124	75	7	5

Figure 9:Table of outliers for the final features

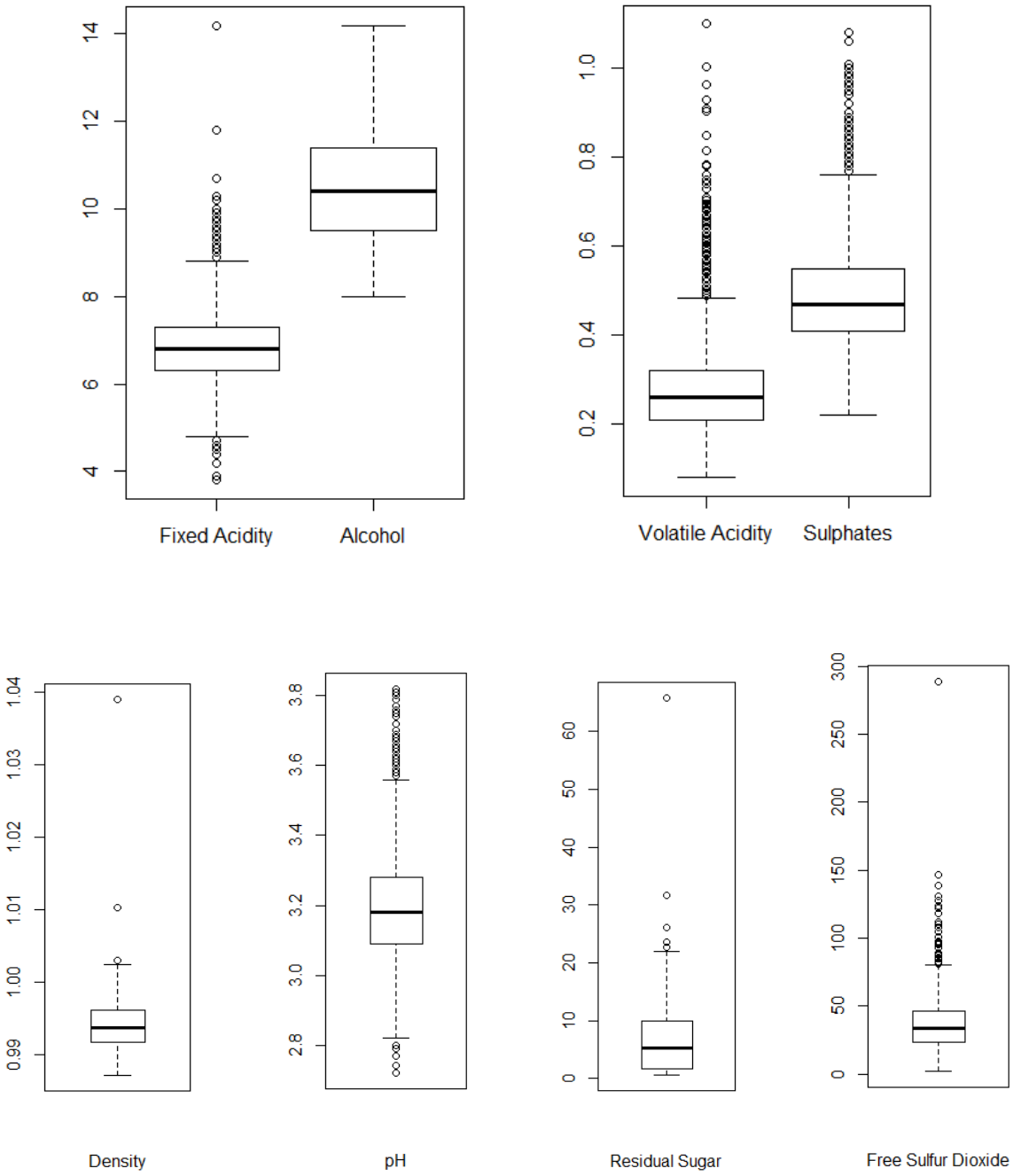

Figure 10: Boxplots for all of the final features