ST563 Final Project

Yuying Zhou, Aaron Brake
Kathryn Hill, Michael Evans

May 7, 2021

Contents

1 Introduction

The following report applies both regression and classification methods to the dataset containing information on the quality on red wine from the North of Portugal. Along with containing a measure of wine quality, the dataset contains physicochemical tests of the wine. The data comes from the UCI Machine Learning Repository.

The goal of this analysis is to create (1) a predictive model to determine quality based on chemical factors of the wine as opposed to quality determined by the arbitrary opinions of tasters and (2) an interpretable model to help wine makers have a better understanding of how to increase their wine quality.

1.1 Data Outline

The dataset on red wine originally contains 1,599 observations of 12 variables. All of the observations and variables are considered for this analysis. The twelve variables given in the dataset are 'Fixed Acidity', 'Volatile Acidity', 'Citric Acid', 'Residual Sugar', 'Chlorides', 'Free Sulfur Dioxide', 'Total Sulfur Dioxide', 'Density', 'pH', 'Sulphates', 'Alcohol', and 'Quality'. All of the variables, except for 'Quality', are continuous numeric variables. 'Quality' is an ordinal variable ranging from 1, indicating the worst wine, to 10, indicating the best wine, that is calculated as the median score of the wine given by three different tasters.

To conduct classification on this data, the 'Quality' variable needed to be collapsed into a factor. The factor was chosen to be:

$$\text{Quality}_{factor} = \begin{cases} High & \text{if } Quality > 5 \\ Low & \text{if } Quality \leq 5 \end{cases} \tag{1}$$

This encoding results in two fairly balanced categories. The 'High' category results in 855 observations, and the 'Low' category results in 744 observations. Thus, the no-information rate for this factor is $\frac{855}{(855+744)} \approx 0.535$.

1.2 Evaluating Models

In this analysis, the original dataset of 1,599 observations is split into a training and a test dataset. The training set contains a random selection of 85% of the original data. The remaining 15% is used as the test set to evaluate any predictive models generated using the training set. In order to compare models with regression and classification techniques, test mean squared error (MSE) and misclassification rates will be used, respectively.

Test MSE is the average of the distance between the actual and observed values squared and the most common evaluation method for regression models.

$$\text{Test MSE} = \frac{1}{n}\sum (y_i - \hat{f}(x_i))^2$$

The misclassification rate is the proportion of observations classified to the wrong category and is very popular for evaluating classification models.

$$\text{Misclassification Rate} = \frac{1}{n}\sum I(y_i \neq \hat{y}_i)$$

In both techniques, the optimal prediction model will have the lowest value of test MSE or misclassification. However, values of test MSE and misclassification are not directly comparable to one another. A lower value of test MSE than misclassification rate would not indicate a better model, for example. Each metric is applicable to its respective technique.

2 Exploratory Data Analysis

Table 1 presents a numerical summary of the data, including the five number summary, plus the mean:

	Min.	Q1	Median	Mean	Q3	Max.
Fixed Acidity	4.60	7.10	7.90	8.32	9.20	15.90
Volatile Acidity	0.120	0.390	0.520	0.528	0.640	1.580
Citric Acid	0.000	0.090	0.260	0.271	0.420	1.000
Residual Sugar	0.900	1.900	2.200	2.539	2.600	15.500
Chlorides	0.012	0.070	0.079	0.087	0.090	0.611
Free Sulfur	1.00	7.00	14.00	15.87	21.00	72.00
Total Sulfur	6.00	22.00	38.00	46.47	62.00	289.00
Density	0.990	0.996	0.997	0.997	0.998	1.004
pH	2.74	3.21	3.31	3.31	3.40	4.01
Sulphates	0.330	0.550	0.620	0.658	0.730	2.000
Alcohol	8.40	9.50	10.20	10.42	11.10	14.90
Quality (Ordinal)	3.00	5.00	6.00	5.64	6.00	8.00

Table 1: Numerical Summary of Data

Figure 1 is a graphical summary of each variable in the dataset:

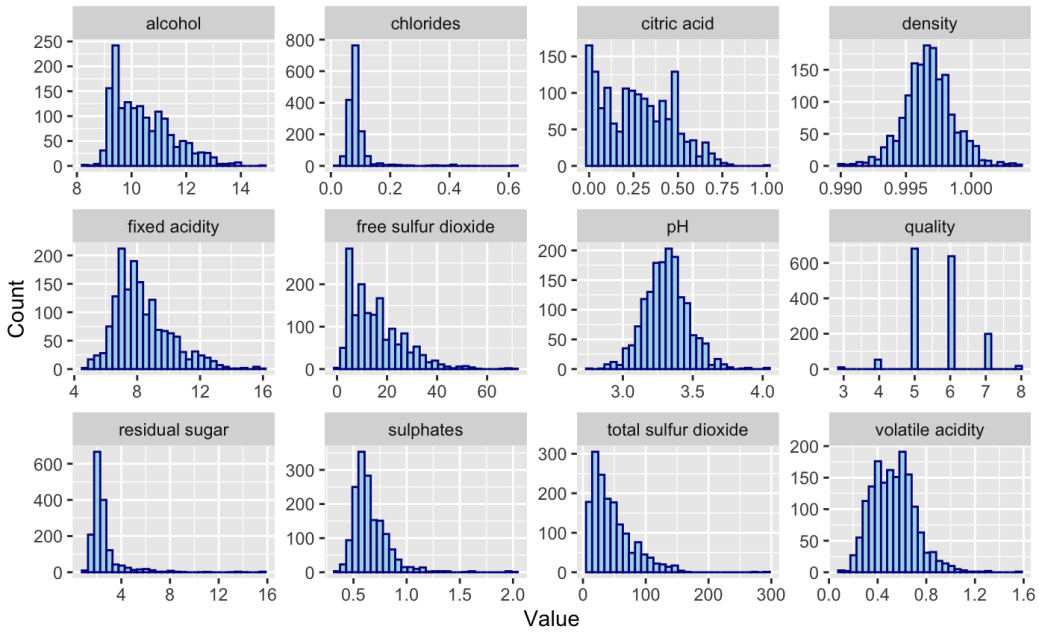

Figure 1: Graphical Summary of Data

We also investigated collinearity of the dataset since it may hurt the interpretability of the models. Figure 2 shows pairwise correlation for all variables:

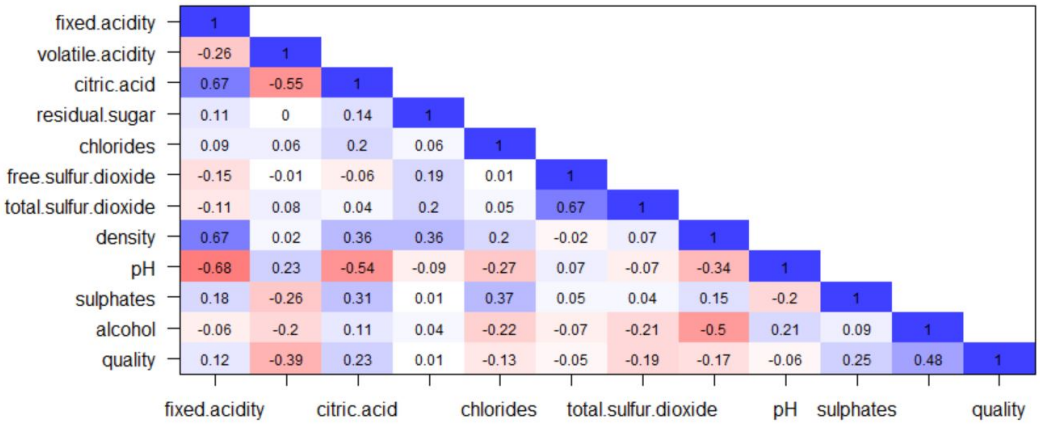

Figure 2: Correlation of Variables

The following findings were made from the summaries of the variables:

- Both 'Free Sulfur Dioxide' (1.00 to 72.00) and 'Total Sulfur Dioxide' (6.00 to 289.00) have large ranges. Without knowing too much about these chemical properties, it is hard to say what this means for the future analysis, but it is a rather large range of values.
- 'pH' ranges from 2.740 to 4.010 and averages 3.311, indicating the overall acidity of wine. Wine is generally known to be acidic, so this is a good finding to confirm in the data.
- The average 'Alcohol' is given as 10.42, which is generally known to be closer to 12 for wine. However, it is hard to say if this region of wine typically has different characteristics.
- The summary of 'Quality' suggests much of the data is a 5 or 6, as these values represent the first quarter, median, and third quarter values.
- A number of the variables (i.e., 'density' and 'pH') appear to be normally distributed. The majority of variables have right-skewed distribution.
- There may be linear relationship between a few variables, like 'Fixed Acidity' and 'Citric Acid' (67%), 'Fixed Acidity' and 'Density' (67%), and 'Alcohol' and 'Quality' (48%), which would lead to investigating multicollinearity in future methods.

3 Regression Methods

3.1 Linear Regression

Linear regression, one of the most well-known applications of supervised learning, is a method utilizing one or more variables to predict a quantitative variable of interest by attempting to find a linear relationship between the variables. As a simpler method, it typically carries results with higher interpretability when compared to other methods. Many other statistical methods are based on linear regression.

As the exploratory data analysis indicated correlation between some variables, multicollinearity was investigated for the linear regression model. A variance inflation factor (VIF) score was applied

to the response (quality) to all predictors. The VIF is a ratio of the coefficient for a variable under the full model to the coefficient for the same variable under a simple linear model. Higher values of VIF, typically greater than 5, indicate problematic collinearity among variables.

Two predictors, 'Fixed Acidity' with a score of 7.7 and 'Density' with a score of 6.4, were considered high, while all other predictors scored within acceptable ranges (less than 5). To address the multicollinearity in the linear regression model, 'Fixed Acidity' was removed as predictor and the VIF score were acceptable ranges, including 'Density' with a new score of 2.3. The test MSE for the linear regression model was 0.350. As shown in Figure 3, 'Alcohol' and 'Sulphates' are positively associated with wine quality and the associations are statistically significant at significance level of 5%. 'Volatile Acidity', 'Chlorides', and 'Total Sulfur Dioxide' are negatively associated with wine quality. It was also noted that the adjusted R square was 0.348, which implied the linear model may not be a good fit for the data.

	quality							
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*
(Intercept)	1.40	14.71	0.00	0.02	-27.47 – 30.27	-0.04 – 0.04	0.10	0.924
volatile.acidity	-1.13	0.13	-0.25	0.03	-1.39 – -0.87	-0.31 – -0.19	-8.57	**<0.001**
citric.acid	-0.21	0.15	-0.05	0.04	-0.51 – 0.08	-0.12 – 0.02	-1.42	0.157
residual.sugar	0.01	0.01	0.02	0.03	-0.02 – 0.04	-0.03 – 0.07	0.64	0.523
chlorides	-1.90	0.44	-0.11	0.03	-2.77 – -1.04	-0.16 – -0.06	-4.33	**<0.001**
free.sulfur.dioxide	0.00	0.00	0.04	0.03	-0.00 – 0.01	-0.02 – 0.10	1.41	0.159
total.sulfur.dioxide	-0.00	0.00	-0.12	0.03	-0.00 – -0.00	-0.18 – -0.06	-3.70	**<0.001**
density	2.99	14.66	0.01	0.03	-25.77 – 31.75	-0.06 – 0.07	0.20	0.838
pH	-0.46	0.15	-0.09	0.03	-0.75 – -0.17	-0.14 – -0.03	-3.13	**0.002**
sulphates	0.89	0.12	0.19	0.03	0.65 – 1.13	0.14 – 0.24	7.22	**<0.001**
alcohol	0.29	0.02	0.38	0.03	0.25 – 0.34	0.32 – 0.45	11.93	**<0.001**
Observations	1359							
R^2 / R^2 adjusted	0.352 / 0.348							

Figure 3: Linear Regression Model

3.2 Tree-Based Methods

Tree-based methods were adopted since they fit non-linearity and interactions naturally. They can be used for both regression and classification methods.

3.2.1 Regression Trees

Regression trees are created by splitting the predictor space into distinct regions and assigning a prediction value for every observation falling into that region. Regression trees are highly interpretable, which is shown by Figure 4.

The regression tree included 11 terminal nodes. Only four predictors, which were 'Alcohol', 'Sulphates', 'Volatile Acidity', and 'Total Sulfur Dioxide', were used in constructing the tree. The tree was not pruned, since the cross validation indicated that the most complex tree resulted in the lowest deviance. The test set MSE associated with the regression tree is 0.427, indicating this model leads to test predictions within 0.654 of the true wine quality value.

Based on the regression tree, 'Alcohol' is the most important factor in determining wine quality and 'Sulphates' is the second most import factor in predicting. It was also noticed the worst 'Quality' predicted was 4.300 and the best 'Quality' predicted was 6.442, which did not result in a significant difference in tasting.

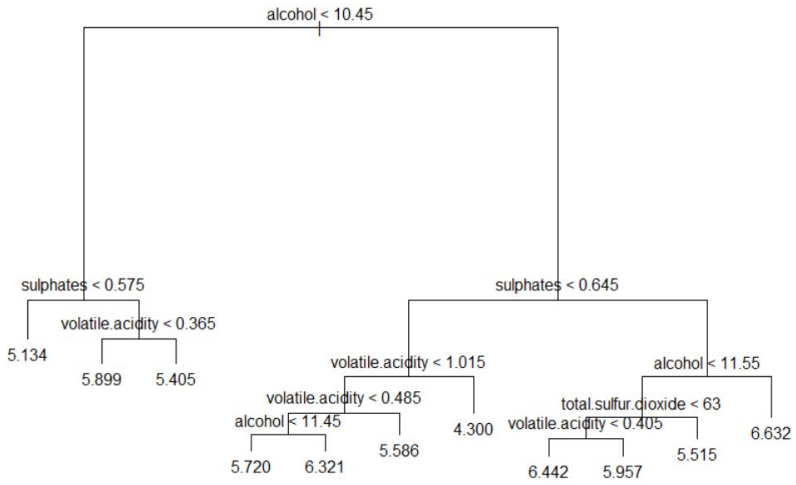

Figure 4: A Regression Tree for Predicting Wine Quality

3.2.2 Random Forest

The random forest method is an application of decision trees. It is similar to the bagging method, except at each split of the decision tree, only a random sample of a specified number of predictors are considered. In bagging, all of the predictors are used. By using a random subset of predictors, the trees are decorrelated, which reduces variability.

Given trees do not, generally, have a high level of predictive accuracy and can be very non-robust, we used random forest to improve the predictive performance. For the random forest method, we tried different number of variables (i.e., from 1 to 11) at each split. The model with four variables at each split resulted in lowest test set MSE (i.e., 0.3162). The result indicated this model leads to test predictions within 0.5623 of the true wine quality value. Random forest methods improved the accuracy of predictions compare to regression trees and the linear models but sacrificed the intrinsic interpretability. It is impossible for vineyards to gain insights from the random forest models in terms of how to improve wine quality based on chemical factors.

The importance results indicated that across all of the trees considered in the random forest, 'Alcohol' and 'Sulphates' are the most important variables in determining wine quality. %IncMSE is based on the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. IncNudePurity is a measure of the total decrease in node impurity resulting from splits over the variable, averaged over all trees. 'Alcohol' has %IncMSE of 62.85 and IncNodePurity of 182.78. 'Sulphates' has %IncMSE of 53.60 and IncNudePurity of 116.06. The remaining variables had much lower values for these two measures of variable importance.

3.2.3 Boosted Trees

The boosted trees method is another application of decision trees in which the model learns more slowly as compared to bagging. The model uses information from previous trees to generate the current tree. The tuning parameters for this method include the number of trees, shrinkage parameter, and depth of tree.

In this study, we tried several combinations of the shrinkage parameter (0.001, 0.002, 0.01, 0.02) and depth of tree (1, 2, 3, 4) and used 5,000 trees. The model with the shrinkage parameter of 0.02 and depth of tree of 2 resulted the lowest test set MSE (i.e., 0.359). The test set MSE is slightly worse than random forests, but superior to the regression trees.

Based on the relative influence statistics, 'Alcohol', 'Volatile Acidity', and 'Sulphates' are the three most important variables in predicting wine quality. As the partial dependence plots show in Figure 5, 'Quality' is increasing with 'Sulphates' and decreasing with 'Volatile Acidity'. 'Quality' increases and reaches a peak as 'Alcohol' increases to approximately 13 and then decreases as 'Alcohol' continues to increase.

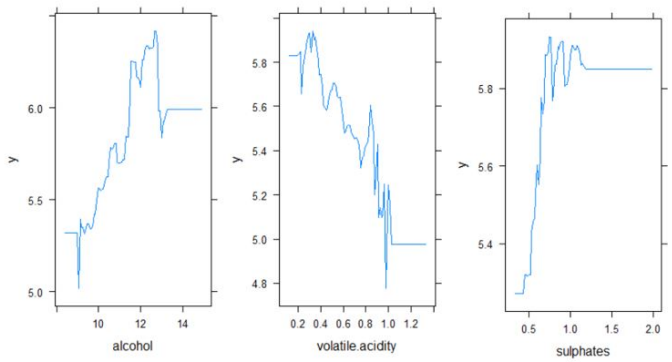

Figure 5: Partial Dependence Plots for Alcohol, Volatile Acidity, and Sulphates

4 Classification Methods

4.1 Logistic Regression

Extending the basis of linear regression, logistic regression seeks to predict the probability that a given response variable belongs to a certain class.

The classification methods were conducted based on the 'Quality$_{factor}$' variable, and the 'High' category was used as the reference group. A logistic regression model was conducted, using the all predictors, since it combines three desirable qualities: interpretability, generally good predictability, and generally good inference qualities. Two different link functions were used for the full logistic regression, the usual logit link, which maps the log odds, and the probit link, which maps to the CDF of the Normal distribution. Table 2 shows the test misclassification error rate for these methods, as well as all of the other methods discussed in this section. The test misclassification error for both link functions was almost identical. Note, the full logistic regression beats the no-information rate, so these methods are providing quality predictive results. However, there are obvious improvements to be made to the full model, such as doing variable selection or shrinkage of regression estimates.

Before any other logistic regression methods are considered, the assumptions of the logistic regression should be checked. The assumptions for the logistic regression model are:

1. Observations is independent across rows
2. Large sample size in each category
3. No extreme outliers in the data
4. Linear relationship between the data and the link function
5. Predictors are independent (i.e. no multicollinearity)

For this model, it is important there are no structural dependencies between the observations, which is the case for the wine data. A large sample size is needed for each outcome in logistic regression. The count for both 'high' and 'low' is $\gg 10$, so this assumption is validated. To test

for the linear regression of the data, the predictors can be plotted against either the log-odds or the normal CDF (whichever corresponds to the appropriate link function). For sake of space these plots are not included, but every predictor appeared to have at least adequate linearity for both link functions.

Outliers can be detected using the Cooks Distance statistic. Figure 6 gives plots of the Cooks distances for the full models with their respective link functions. Observation 653 appears to have a slightly worrisome Cooks Distance for both models. When checking this data, it has both the maximum 'Fixed Acidity' and 'Alcohol' values, which may be the reason for this relatively large value. However, after checking this observation there are no obvious signs of data entry errors, so this observation is just noted but not removed since out rightly deleting suspected outliers is bad practice. For the logit link model, observation 93 has a borderline value. The only value for this observation near the endpoints of the observed values was 'Sulphates', but again no obvious errors were found so it is just mentioned but not removed.

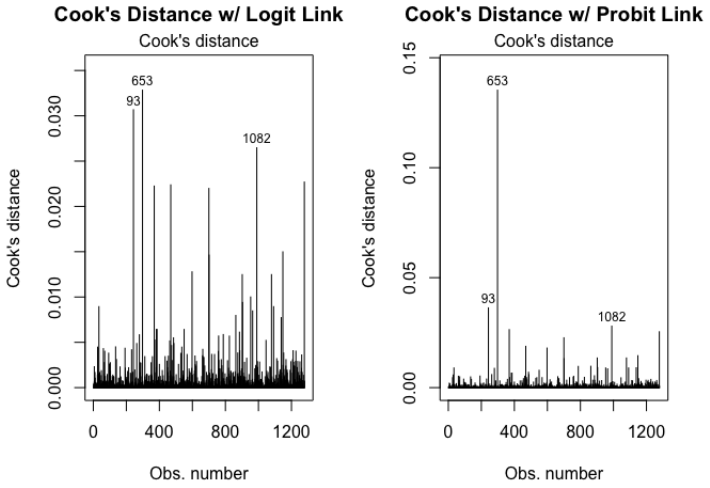

Figure 6: Cooks Distance for the Full Logistic Regression

To address the multicollinearity assumption, the VIF's of the two models were calculated (using the *car* package in **R**). The 'Fixed Acidity' and 'Density' predictors both had $VIF's > 5$, which indicates some level of multicollinearity, meaning this assumption cannot be fully validated. However, just because these two predictors may violate the multicollinearity assumptions doesn't mean they should just be dropped from the model, since they may still provide useful information.

4.1.1 RIDGE

One of the best methods to deal with multicollinearity in predictors is to use the RIDGE penalty. The usual *glmnet* package can handle the logistic regression. Ten fold cross validation was used to select the λ parameter. Again, the two link functions were both run. For each link function, the 1-standard-error cross validation λ was used. Table 2 shows the test misclassification errors for the 1-se λ for each link function. The error rates show a very slight improvement over the unconstrained model. This slight improvement, coupled with the benefits of multicollinearity make the Ridge model more desirable than the full model.

4.1.2 LASSO

To both perform variable selection, and to deal with the multicollinearity issues, LASSO regression was conducted. The λ constraint was, as in the Ridge, selected using the 1-se, 10 fold cross validation.

This value of lambda corresponded to a model with 4 variables, for both link functions. The variables selected were: 'Volatile Acidity', 'Total Sulfur Dioxide', 'Sulphates', and 'Alcohol'. This model not only performs variable selection, but also deals with the multicollinearity, since the problematic variables were not selected. However, the test misclassification errors were not great, and were worse than both the full logistic and Ridge methods. So, the multicollinearity is better, but the predictive performance is worse.

4.1.3 Best Subsets

Since the *leaps* package in **R** only conducts OLS regression, the *bestglm* package was used to conduct the best subsets. Both AIC and BIC were used to determine the "best" model, using only the probit link. AIC chose a model size of 8 and BIC chose a model size 5. Both of these models do not see an improvement in test misclassification error, however both the AIC and BIC models do not have any VIF problems indicating the multicollinearity issue from the full logistic model has been reduced.

Method	Link/Selection Criterion	Misclassification Error
Full	Logit Link	0.256
Full	Probit Link	0.259
LASSO	Logit Link	0.284
LASSO	Probit Link	0.275
RIDGE	Logit Link	0.250
RIDGE	Probit Link	0.253
Best Subsets	AIC / Probit Link	0.256
Best Subsets	BIC / Probit Link	0.272

Table 2: Test misclassification error rate for each method, with link function and/or selection criterion used

All of the logistic regression methods performed fairly similar, with all the methods falling within a 3% range of test misclassification error. However, the full model is not recommended because of the multicollinearity problem, among a few other assumption issues. Both the best subsets and LASSO regression help mitigate these issues, but have worse predictive performance. Since the Ridge model both effectively deals with the multicollinearity, has the best test misclassification error, and also retains a large amount of interpretability, it is the recommended logistic regression model. Both link function perform similarly, so one is not recommended above the other. However, as will be discussed further below, this classification method is limited in the amount of information it is able to convey (see section 4.2.1).

4.2 Tree-Based Methods

4.2.1 Random Forest

Under classification, the random forest method is done similarly to how it was used for regression. However, instead of predicting a 'Quality' value between 1 and 10, the method predicts a class, 'Low' or 'High' in this case.

Here, the misclassification rate was found for the number of predictors being 1 all the way up to 11, the maximum number of predictors for this dataset. If 11 predictors is the best model, this suggests the bagged model is the best.

With the random forest method, the best results come from the simplest application of the model using only 1 predictor. The misclassification rate for this model is 0.144. For more predictors, the misclassification rate increases, however it stays around 0.150 and 0.160.

The overall performance of the random forest model is pretty good. However, the limitations of the model should be considered. To start, the model is only predicting two classes of the variable. While it is good this model is able to only misclassify 14% of the test set, it is important to consider that it is only giving information regarding a wine being 'Low' or 'High' quality. If, for example, there was a significant difference in the quality between a wine ranked 7 and 8, this model would not capture that.

Additionally, the random forest model lacks interpretability. While the model classifies well, it is a black box method, meaning it is given data and creates results, but seeing how the model achieves those results is difficult. Evaluating the importance results from a random forest method allow some insights to the model. For classification, the importance results are similar to that found in regression. The 'Alcohol' and 'Sulfates' variables are found to be the most important variables for predicting wine quality.

When looking at the multicollinearity that has appeared as an issue in other aspects of the analysis, it is noted that this only matters for inference, not predictability. Assuming similar structure in the training and test set, multicollinearity should not impact prediction. Thus, it does not need to be accounted for methods such as random forests.

4.2.2 Boosted Trees

Similar to random forests, the application of boosted trees between regression and classification only varies by the type of prediction.

For the classification portion of this analysis, combinations of shrinkage parameters (0.001, 0.002, 0.01, 0.02) and depth of tree (1, 2, 3, 4) with 5,000 tree were used to create different models. The model with the lowest misclassification rate of 0.184 had a shrinkage parameter of 0.01 and depth of 3. Other combinations of shrinkage and depth have misclassification rates between 0.190 and 0.240. Still, the overall best result is obtained from the random forest method.

Similar limitations of the random forests method apply to the boosted tree method. It is also considered a black box method. However, relative influence statistics can be used to determine the most important variables, which are the same as found in the regression application of boosted trees ('Alcohol', 'Volatile Acidity', and 'Sulphates'). The partial dependence plots show the same trends as Figure 5.

5 Conclusion

5.1 Results

This analysis set out to create (1) a predictive model to help better determine quality and (2) an interpretable model to help understand specific ways to improve wine. Both regression and classification methods were leveraged to answer these questions.

5.1.1 Predictive Model

Overall, the best predictive models were found with the random forest method for both regression and classification methods. The random forest model resulted in a test set MSE and test misclassification rate of 0.3162 and 0.144, respectively. For regression, the test MSE value indicates the model leads to test predictions within 0.5623 of the true wine quality value. Since quality is an ordinal variable with rating from 1 to 10, the random forest model may predict one level higher or lower than the actual rating level, on average, which is considered a fairly accurate predication. However, as previously noted, Random Forests is considered a black box method. If a vineyard was only interested in determining the quality of its wine without understanding what is causing the score,

this would be an appropriate technique. However, if a vineyard is looking to gain insights as to what it can do to improve its wine, it should look to a more interpretable method.

5.1.2 Interpretable Model

When looking for an interpretable method, a great solution is the boosted trees model. While this model performed a little worse for both regression and classification techniques, the information offered by relative influence statistics and the partial dependence plots helps make the model more interpretable. The relative influence statistics help identify the most important variables in the model, and the partial dependence plots help to illustrate the the marginal effect of an individual variable on the response variable. Similar results were found in boosted tree method for both regression and classification methods. 'Alcohol', 'Volatile Acidity', and 'Sulphates' are the three most important variables in predicting wine quality. 'Quality' is positively associated with 'Sulphates' and negatively associated with 'Volatile Acidity'. 'Quality' increases and reaches a peak at 'Alcohol' of approximately 13 and then decreases as 'Alcohol' continues to increase. The combination of this information helps create a better understanding of the findings of the model, while the model's highly comparable level of predictability makes it competitively useful.

The interpretable model would be helpful for identifying specific ways for a vineyard to improve their wine. For example, from the partial dependence plots resulting from the boosted model, it was identified that increasing 'Sulphates' results in higher quality wine. Similar findings for other variables could be applied to help a vineyard improve the quality of their product.

In choosing between applying these methods via regression or classification, one should consider personal motivations for the analysis. If one is simply content knowing if the wine is 'Low' or 'High' quality, then the classification method would be appropriate. However, if one is looking to get a better idea of what the exact quality of their wine is, perhaps regression methods should be considered.

5.2 Extensions

Further analysis of this dataset could be completed after gaining a further understanding of the chemical properties presented and how each interacts. This understanding could likely come from a chemist or someone in the wine industry with broader knowledge of these properties. With this information, perhaps the number of variables used in the analysis could be reduced to account for potential multicollinearity among variables.

Another extension of the analysis completed in this report could be attempting to broaden the classification techniques to a classifier of more than two classes. This would help increase the practicality of the results found, as the model would ideally be able to identify each of the levels of 'Quality'.