Final Report for Red Wine Analysis

Ruiqiang Chen, Michael DeWitt, David Williams, Alex Vannoy

7/28/2017

1 Introduction

The purpose of this document is to report the proposed statistical models for classification of red wine based on eleven predictors. The purpose of this analysis is to provide a model to the vintners in order for them to better predict the quality rating for their product. Analysis will be performed using both regression techniques and classification techniques.

2 Description of Data

The data set provided is the Wine dataset from UC Irvine of red *vinho verde* wine samples, from the north of Portugal [Cortez et al., 2009]. It consists of 1599 with a total of 11 physicochemical predictors and a response variable. These predictors include the following: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol with the quality feature being associated with the judgement of the individual wine's quality. Quality is the feature of interest for the dataset as the vintner is interested in judging the wine's quality through objective means rather than today's subjective method of averaging the one to ten point judgment of taste-testers. A summary of these measures as well as the response variable can be seen in Table 1.

Table 1: Summary Statistics for the Wine Dataset

Descriptions	min	median	mean	max
fixed acidity (g(tartaric acid)/dmˆ3)	4.60	7.90	8.32	15.90
volatile acidity (g(acetic acid)/dmˆ3)	0.12	0.52	0.53	1.58
citric acid (g/dmˆ3)	0.00	0.26	0.27	1.00
residual sugar (g/dmˆ3)	0.90	2.20	2.54	15.50
chlorides (g(sodium cloride)/dmˆ3	0.01	0.08	0.09	0.61
free sulfur dioxide (mg/dmˆ3)	1.00	14.00	15.87	72.00
total sulfur dioxide (mg/dmˆ3)	6.00	38.00	46.47	289.00
density (g/cmˆ3)	0.99	1.00	1.00	1.00
pH	2.74	3.31	3.31	4.01
sulphates (g(potassium sulphate)/dmˆ3)	0.33	0.62	0.66	2.00
alcohol (% vol.)	8.40	10.20	10.42	14.90
quality	3.00	6.00	5.64	8.00

The distribution of these different criteria can be seen below in the histograms in Figure 1.

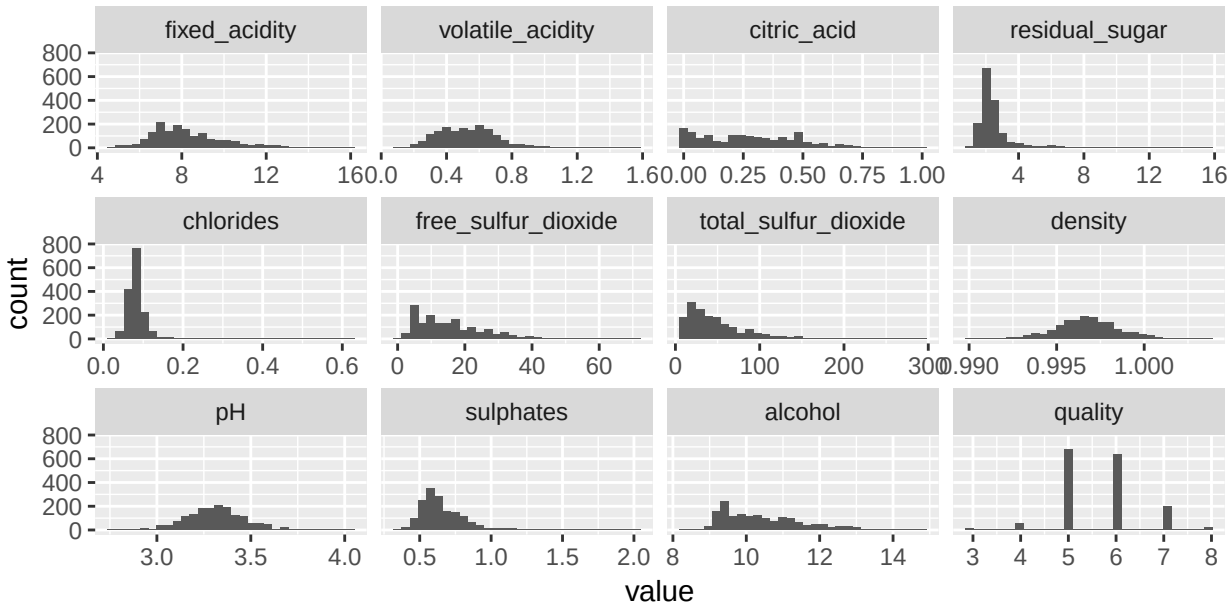

Figure 1: Histogram of all variables in the data set

The following are slightly right skewed: Fixed Acidity, Volatile Acidity, Citric Acid, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, and Alcohol. Residual Sugar and Chlorides are heavily right skewed with density and pH appearing more normally distributed. Completing a Shapiro Wilke normality test on the components indicates that all are non-normal. Reviewing the individual components there appears to be a slight irregularity with total free sulfur dioxide. This can be seen in the histogram of free sulfur dioxide values.

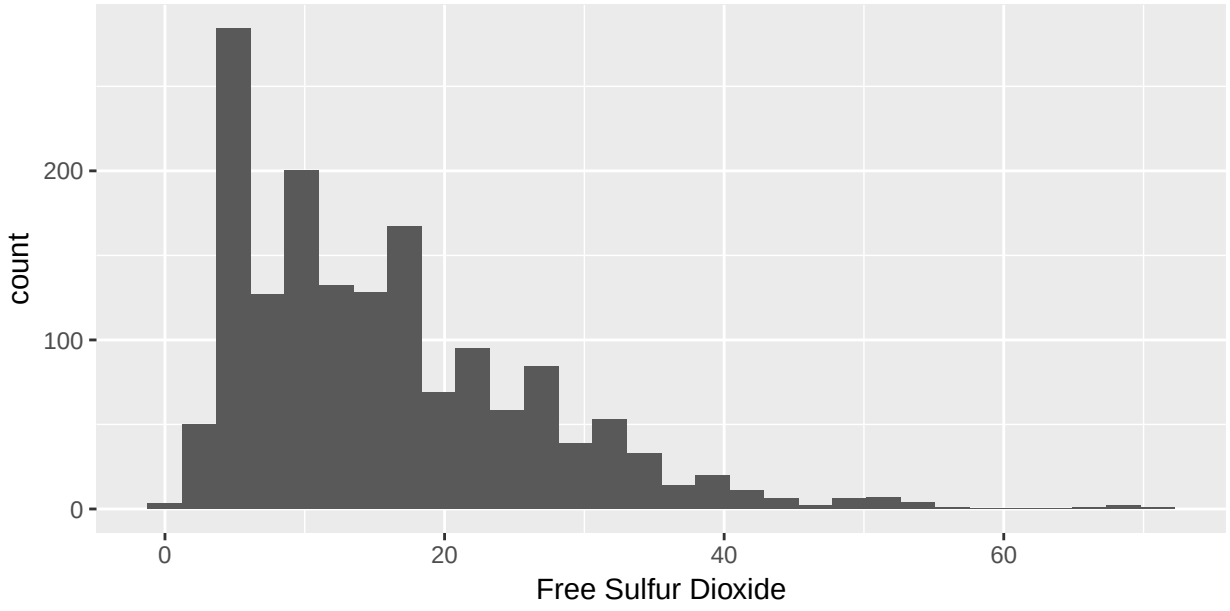

Figure 2: Histogram of Sulfur Dioxide Predictor

As well as the fit of free sulfur dioxide display high studentized residuals and leverage and thus should be considered for removal in the modeling process. These wines are 1080 and 1082.

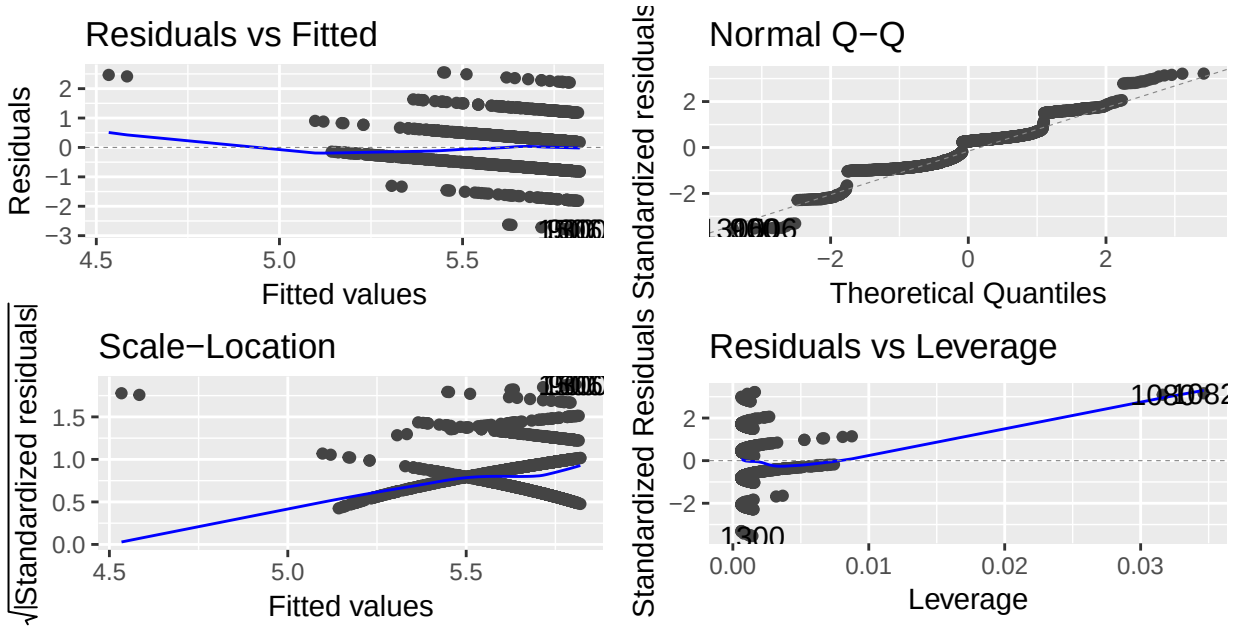

Figure 3: Residual Plots of Linear Model Predicting Quality with Total Sulfur Dioxide

These two data points have been removed from the clean dataset as they will influence the models more heavily than we would prefer. The presence of these two wines may result in incorrect or inaccurate predictions. As we did not gather this dataset, we do not know if this information was incorrectly captured or if these values are real. However, given the strong indication that these two points are outliers with high leverage we feel it is a good assumption to remove these two points.

3 Method

In order to estimate the test error of any of the generated models, the data was divided in testing and training data sets with which to train then models and then test and estimate the testing error. Seventy percent of the raw data was randomly selected and placed in the training set with the remaining thirty percent used as the testing data set.

3.1 Regression

In order to select the best fit regression model, several regression modeling methods were used: Least Squares Regression, Ridge Regression, Lasso Regression, Principle Components Regression, Partial Least Squares Regression, and Boosting. In each regression the response variable was the integer value of Quality. The data were divided into two sets, a training set to train the model and a testing set for model validation. We will now go deeper in the model generation process for each of these modeling types and methods

3.1.1 Least Squares

The least squares regression method that was tested was the best subset selection. The methodology used to determine the best subset model was to first run cross-validation on the training set in order to determine the best number of predictors to include in the model. This analysis indicated that any additional predictors after three variables did not increase the accuracy of the model. The training data were then used to determine

the best subset of the linear model with three predictors. The best subset included volatile acidity, sulphates, and alcohol.

3.1.1.1 Residual Analysis

The linear regression model requires the assumption that residuals from the model have constant, unrelated variances. The plots below show the fitted values versus the residual values and illustrate that we satisfy the assumptions required to conduct a linear regression.

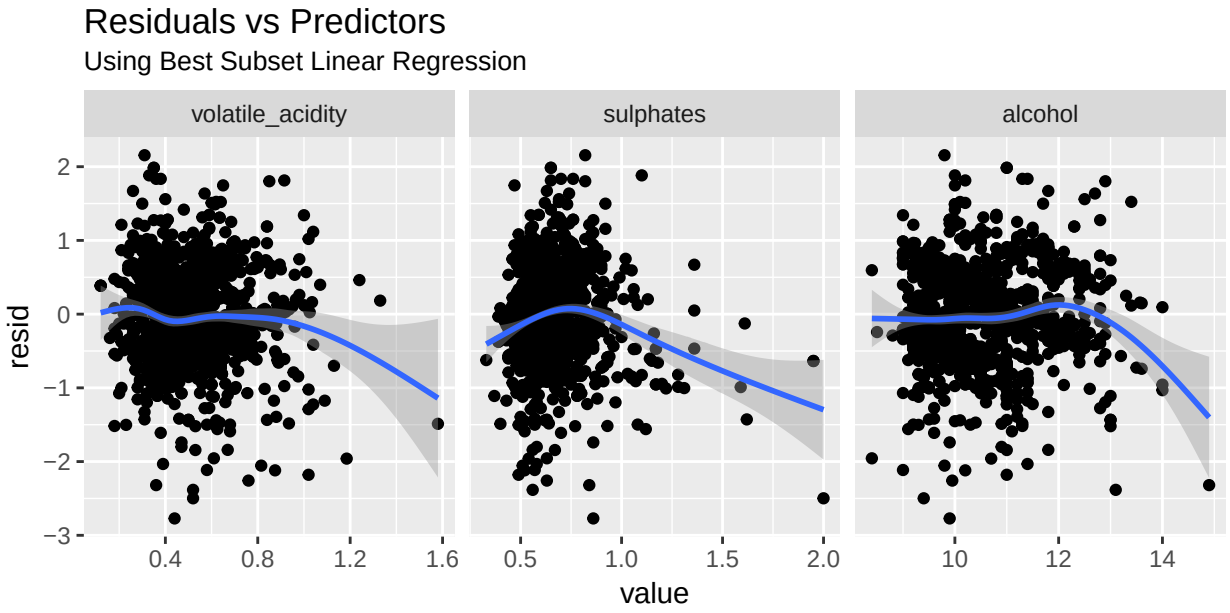

Figure 4: Plot of Residuals from Linear Regression

The residuals appear to have no distinct pattern which is a positive sign that there are not lurking relationships that have not been treated by the modeling.

3.1.2 Ridge Regression

Ridge regression was performed on the dataset as well. Cross-validation was performed on the training data set to determine the optimal value for lambda for the ridge regression. This lambda, 0.079 was then used in a ridge regression model with the testing dataset. Further, there seems to be a very large coefficient with eleven much smaller coefficients.

3.1.3 Lasso Regression

Next Lasso regression was performed with cross-validation. We conducted cross-validation to determine the optimal value for lambda for the Lasso regression. This value was 0.005. As a function of the Lasso regression, only pH was shrunk to zero; total sulfur dioxide and free sulfur dioxide were shrunk to near zero. As was the case with ridge regression, there appears to be one large coefficient compared to the others.

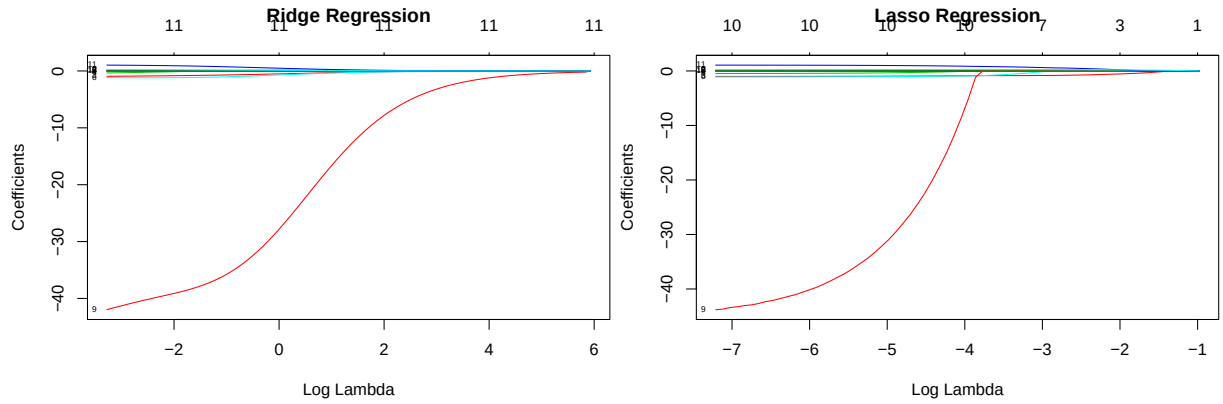

Figure 5: Plot of Lambda vs Coefficients for Ridge/Lasso Regression

3.1.4 Principal Components Regression

Next we performed Principal components regression. Analysis of the principal components revealed that 90 percent of the variation in the response could be explained by the first nine components, so we used these first nine principal components for model training. This is graphically displayed in the scree plot of principal components.

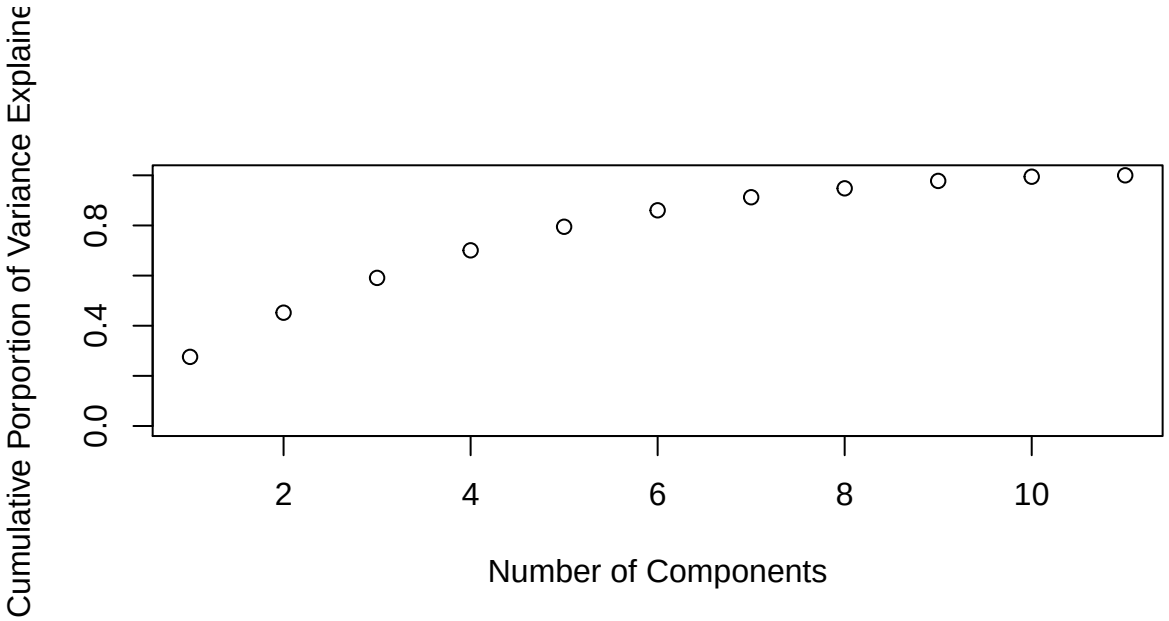

Figure 6: Variance Explained for Principal Components Regression

3.1.5 Partial Least Squares Regression

Next we performed partial least squares regression using the response variable Quality as supervision over the principal components. Using this method, we see from the plot of partial least squares components that after roughly 2-3 components, model accuracy no longer substantially increases.

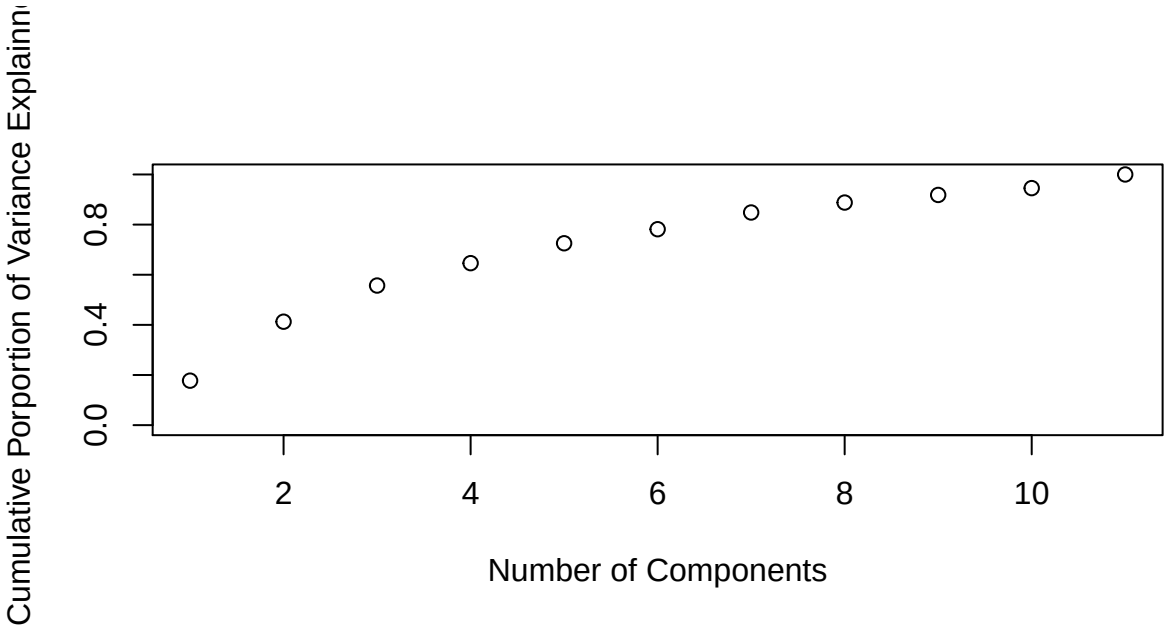

Figure 7: Variance Explained for Partial Least Squares Regression

3.1.6 Boosted Regression

Finally we performed Boosted regression. The interaction depth was limited to four in order to reduce the likelihood of over-fitting the data. The model was trained on 5,000 different trees and we assumed a Guassian distribution. This should be safe enough as we have a relatively large data set.

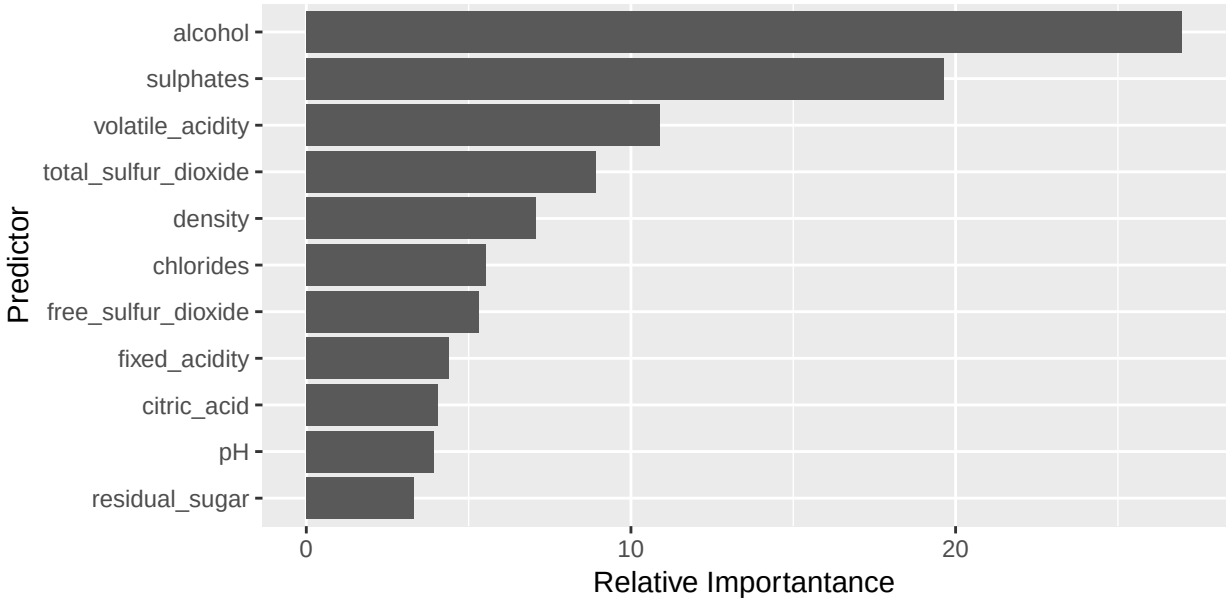

Figure 8: Relative Importance from Boosted Regression

3.1.7 Model Selection

The resulting mean squared errors for each regression method were tabulated in order to determine the superior model.

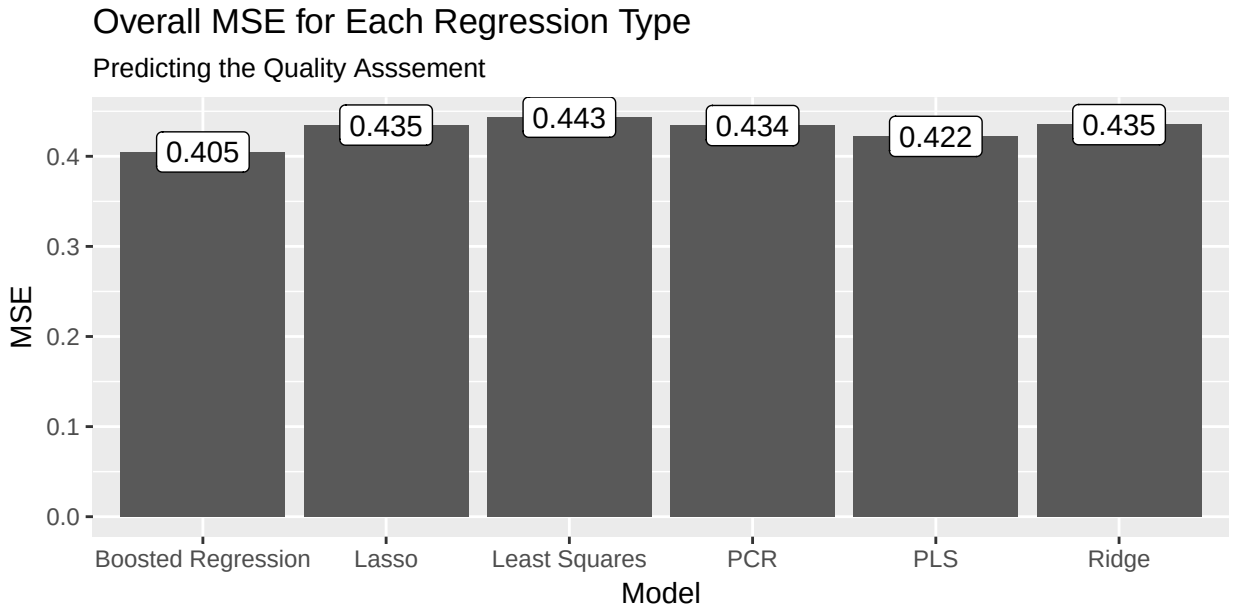

Figure 9: Plot of Results of Different Regression Techniques

3.2 Classification

For classification purposes the wines were segregated in to three different classes. These classes include "good" ($quality > 7$), "medium" ($quality\ between\ 4\ and\ 7$) and "poor"($quality < 4$).

3.2.1 Model Selection

We conducted our analysis for this new variable of quality classes using several classification models: K-Nearest Neighbors, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and tree classification. We trained each model using the training data set and then applied the resulting model to the testing dataset to estimate the model's test error. It is important to note that we achieved greater accuracy by scaling values for the K-Nearest Neighbors approach, which is sensitive to scale differences across variables. Using unscaled values, the validation algorithm was best using 17 variables; using scaled values, 64 were best. The larger number of neighbours makes for a much more global model: it is less sensitive to immediate neighbours in the bias-variance trade-off. The tree classification model was trained first through cross-validation and then pruned to six leaves to reduce the impact of over-fitting in the bias-variance trade-off.

3.3 Comparison of Models

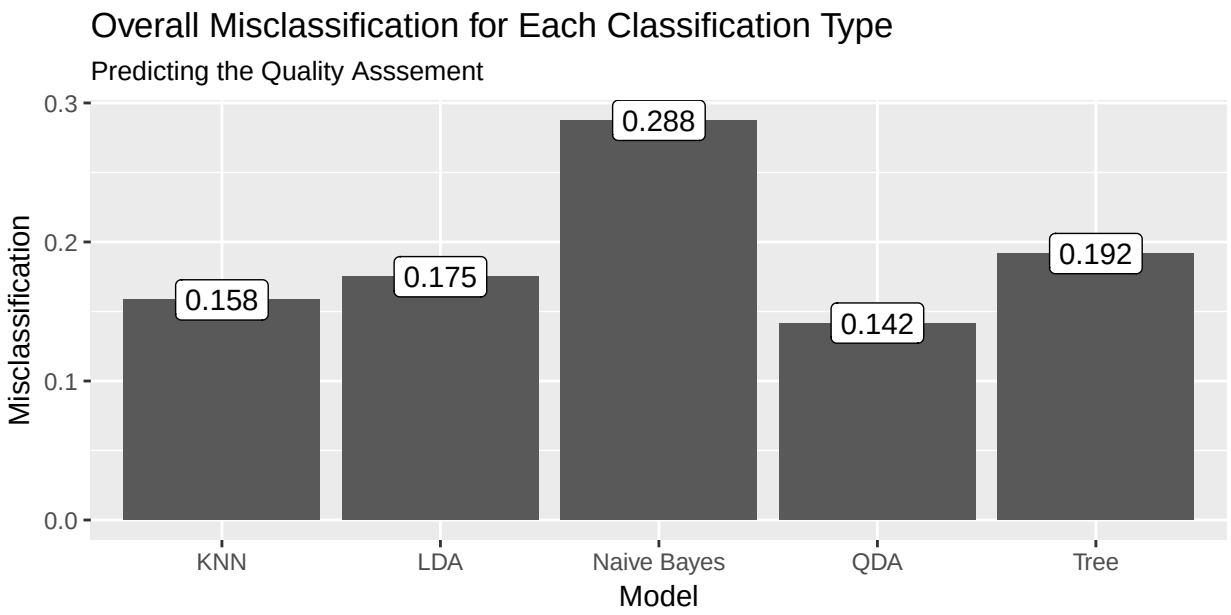

Figure 10: Plot of Results of Different Classification Techniques

All of these models seek to maximize the global accuracy of the model. More interesting for the vintners is the ability to detect each of the three different classes of the wines.

Table 2: Detailed Classification Accuracy

Method	Good	Medium	Poor
KNN	0.24	0.98	0.00
Naive Bayes	0.52	0.77	0.06
LDA	0.43	0.93	0.06
QDA	0.61	0.90	0.89
Tree	0.36	0.92	0.00

4 Discussion

This analysis shows that for regression the boosted model resulted in the highest accuracy of all regression models; however, this accuracy comes at a cost of interpretability. Because the boosted algorithms have little intrepretation this accuracy is more beneficial for prediction than inference. If inference is the goal for the vintner and horticulturalists who seek to understand the properties that make good wines, the model with higher interpretability and the second highest accuracy is the Lasso regression model. While the PLS is more accurate, again it suffers from ease of interpretation.

Thus with this in mind, the best model for inference with high accuracy is characterized by the below equation:

$$quality = 39.37 + 0.0823 * fixed\ acidity - 0.981 * volatile\ acidity - 0.405 * citric\ acid$$
$$-0.013 * residual\ sugar - 1.075 * chlorides + 0.006 * free\ sulfur\ dioxide \quad (1)$$
$$-0.002 * total\ sulfur\ dioxide - 37.09 * density + 1.032 * sulphates + 0.256 * alcohol$$

Whereas the best model for prediction is the boosted regression model. An example tree is show below:

Table 3: Example Boosted Regression Tree

	SplitVar	SplitCodePred	LeftNode	RightNode	MissingNode	ErrorReduction	Weight	Prediction
0	10	10.4500000	1	5	12	32.927684	240	-0.0000021
1	9	0.6350000	2	3	4	9.971645	141	-0.0003125
2	-1	-0.0005283	-1	-1	-1	0.000000	85	-0.0005283
3	-1	0.0000152	-1	-1	-1	0.000000	56	0.0000152
4	-1	-0.0003125	-1	-1	-1	0.000000	141	-0.0003125
5	7	0.9952450	6	7	11	11.203147	99	0.0004400
6	-1	0.0008572	-1	-1	-1	0.000000	39	0.0008572
7	5	5.5000000	8	9	10	10.837500	60	0.0001687
8	-1	-0.0006813	-1	-1	-1	0.000000	12	-0.0006813
9	-1	0.0003812	-1	-1	-1	0.000000	48	0.0003812
10	-1	0.0001687	-1	-1	-1	0.000000	60	0.0001687
11	-1	0.0004400	-1	-1	-1	0.000000	99	0.0004400
12	-1	-0.0000021	-1	-1	-1	0.000000	240	-0.0000021

Applying the equation 1 the vintner can examine each of the variables independently and provide some degree of inference regarding the chemical levels that influence the quality of the red wine. For instance we see that sulfate content appears to have a strong positive influence on wine quality while having additional residual sugars reduces wine quality (as measured by the subjective wine quality score). In the hands of the vintner, these relationships can be explored or potentially exploited to produce a more consistent, higher quality wine.

Turning to the classification method, the best overall classification method was Quadratic Discriminate Analysis. This is seen in both the overall accuracy as well in its ability to accurately classify each subcategory. While the other methods have lesser abilities to detect the good and poor quality wines, the QDA method showed the best accuracy in these two fields, which is very important for vintners when it comes to pricing and selling. The penalty of misclassifying a good wine as medium or a poor wine as medium/ good is severe as this may damage the reputation of the winery. From this analysis it is clear that QDA is the superior method for classification of red wines given this dataset. The priors and group means for the QDA are shown in the preceding figures.

Table 4: Priors used in Quadratic Discriminate Analysis

good	0.140
medium	0.823
poor	0.038

Table 5: Group Means used in Quadratic Discriminate Analysis (continued below)

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar
good	8.56	0.409	0.325	2.592

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar
medium	8.306	0.527	0.263	2.487
poor	7.694	0.68	0.129	3.061

Table 6: Table continues below

	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density
good	0.076	14.88	37.28	0.996
medium	0.09	16.6	48.43	0.997
poor	0.072	8.333	26.94	0.997

	pH	sulphates	alcohol
good	3.307	0.749	11.48
medium	3.304	0.655	10.25
poor	3.379	0.539	10.31

We would also like to address the fact that we have excluded two data points from the original data set. These were observations 1080 and 1082. These data points were excluded due to their being response outliers. We know this because their residual to leverage ratio is very large. If these points were included, the fitted models would be drastically different. For example, the flexible models would likely do better whereas the nonflexible models would do more poorly. Furthermore, we would also like to point out that several variable are highly correlated. For example fixed acidity, citric acid, and pH are all highly correlated. This would make sense to many people with a chemical background and to some mathematicians in hindsight. For instance pH is a measure of acidity. This collinearity is likely why LASSO shrinks several variables to zero and many more closely to zero. It is also likely why the PCR, PLS and ridge regression models produced similar processes while the full linear regressions suffers greatly.

As discussed, there existed some issues with the dataset. As we used a publically available dataset without indentifying details (wine name, winery name, etc) we could not further explore the cause of potential outlying values. Had this information been available we could have been more confident with the elimination of outliers and other discerning information.

The predictors, while loosely normal were not normal as evidence of a Shapiro Wilke test on all the predictors (all p values < 0.01). Log, square root and polynomial transforms were attempted to normalize the data but these transforms did not improve the normality of the predictors. Because of this fact we could have performed an advanced transformation to these values like a Boxcox transform, but this would have made interpretability of the resulting models much more difficult. As such these transforms were not used for sake of this analysis.

5 Conclusion

The best method for prediction is the boosted regression model. The best method for inference is the Lasso regression model. The best method for classifying into high, medium, and low quality wine is the quadratic discriminant analysis technique. Two data points were found to be response outliers and were removed. The distributions of the predictor variables do not seem to have a serious effect on the fitted models. The highly correlated variables do have an effect on the fit of the models, depending on the strengths and weaknesses of the various model fitting processes.

6 Index

All code is available at the following GitHub location https://github.com/medewitt/wine_analysis.