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 bases on 11 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 performing 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, alcohol, quality 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 1-10 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)/dm3)	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.

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 this variable.

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.

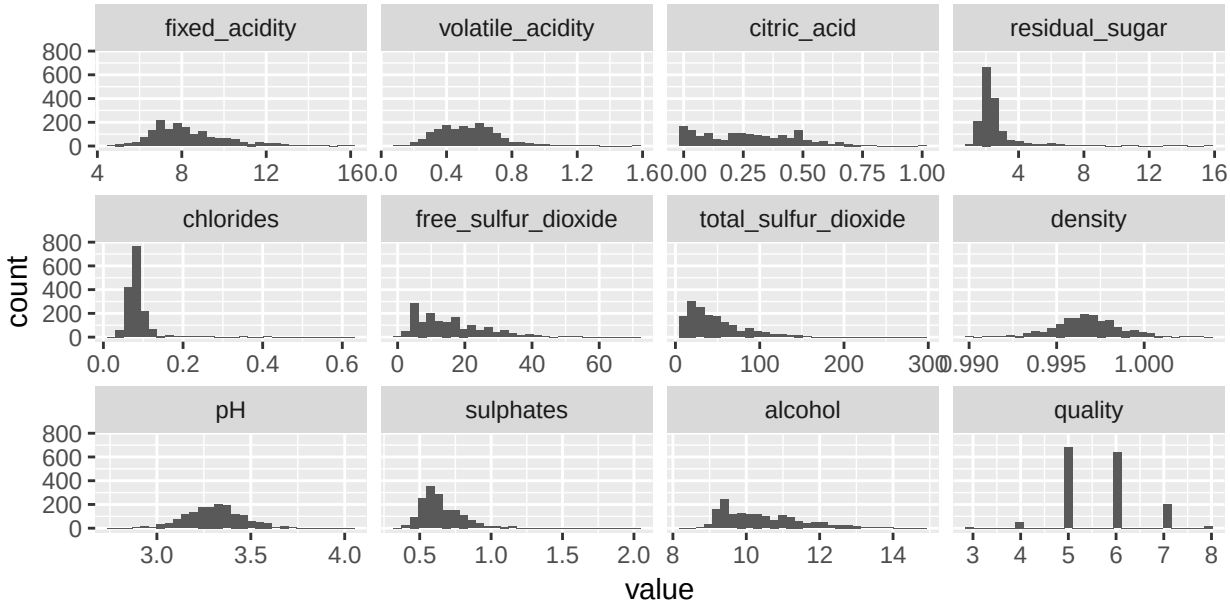

Figure 1: Histogram of all variables in the data set

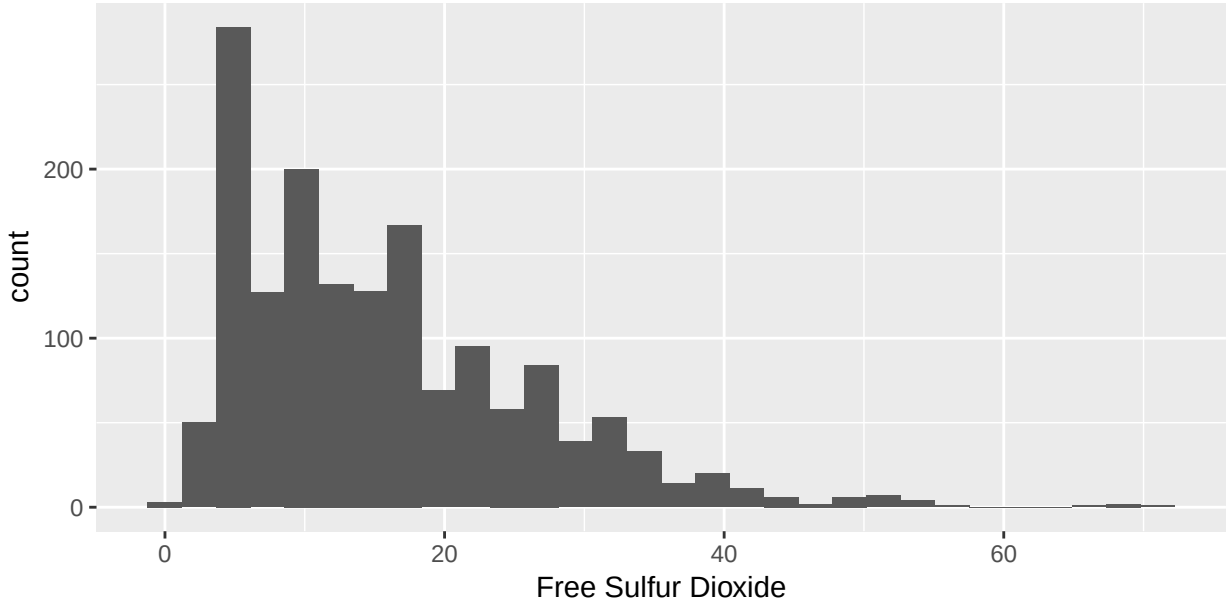

Figure 2: Histogram of Sulfur Dioxide Predictor

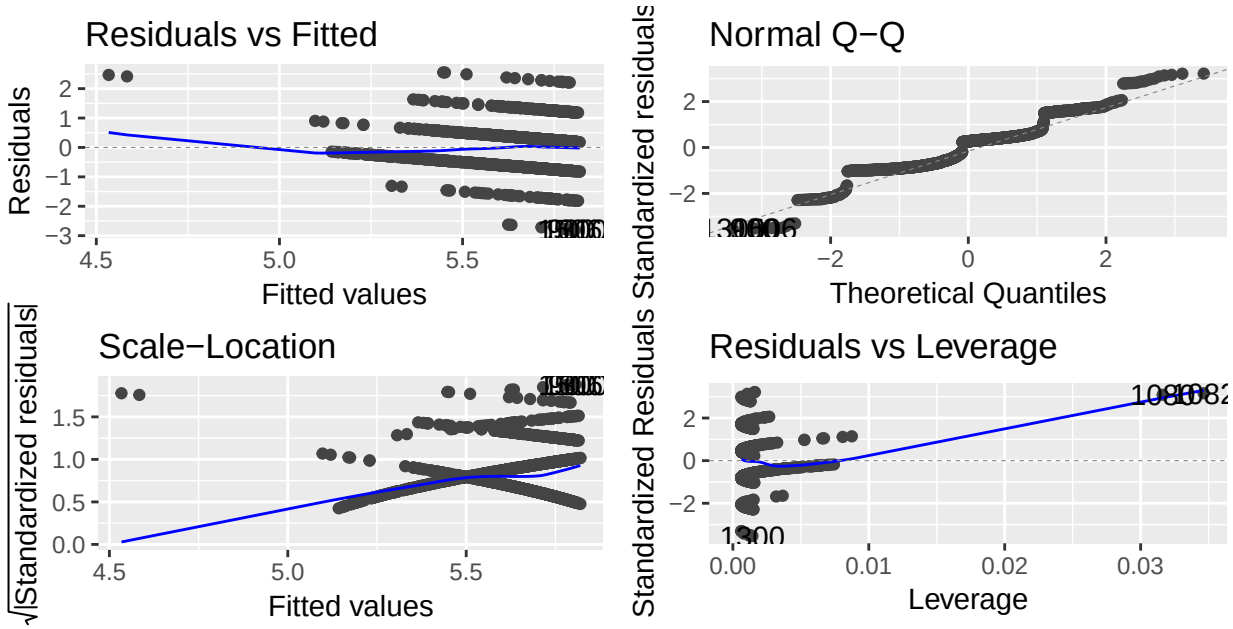

Figure 3: Residual Plots of Linear Model Predicting Quality with Total Sulfur Dioxide

These two wines have been removed from the clean dataset in order to be better predictors. The presence of these two wines may result in incorrect or inaccurate predictions. Ass 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 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 different modeling methods were produced. These include Least Squares Regression, Ridge Regression, Lasso Regression, Principle Components Regression, and Partial Least Squares Regression. The quality integer was the value that the model attempts to predict for each of these methods. The data was 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 different 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 predictor

after three variables were selected did not increase the accuracy of the model. The training data was then used to determine the best subset of the linear model with three predictors. The best subset included:

3.1.1.1 Residual Analysis

Here we need to make some plots against of the fit vs predictors and fit vs prediction to cross off that we considered our residuals

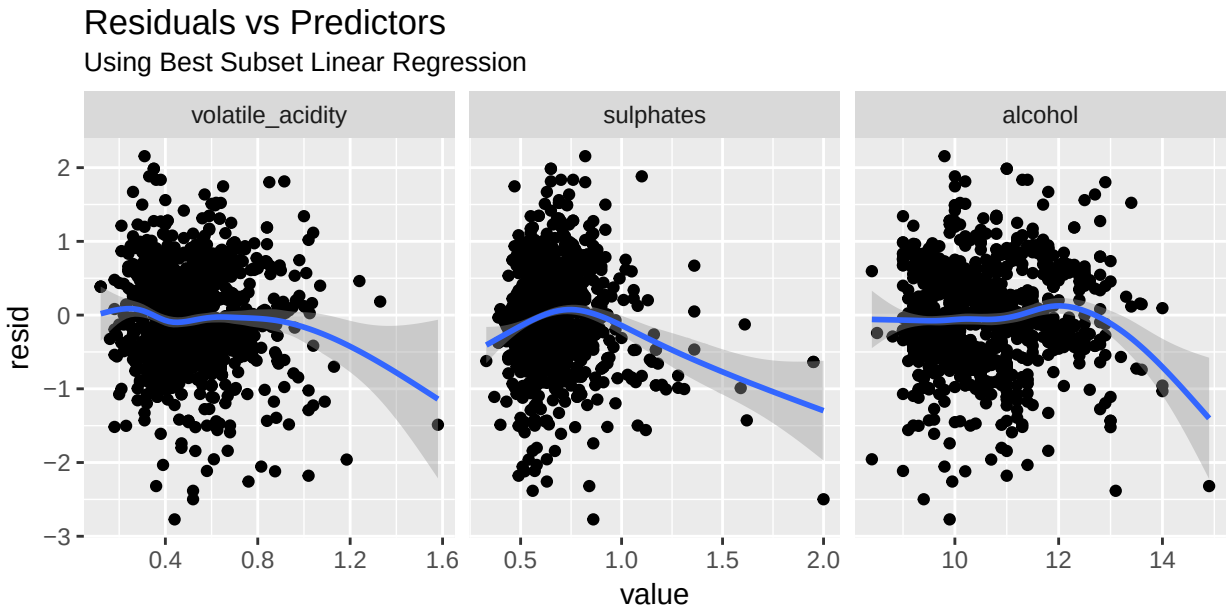

Figure 4: Plot of Residuals from Linear Regression

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 optimum 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 11 much smaller coefficients.

3.1.3 Lasso Regression

Lasso regression was used with cross-validation on the data set. Cross-validation was used to determine the best lambda which was 0.005. As a function of the lasso regression only pH was shrunk to zero with total sulphur dioxide and free sulphur dioxide being shrunk to near zero. Also similar to ridge regression, there appears to be one much larger coefficient in relation to the others.

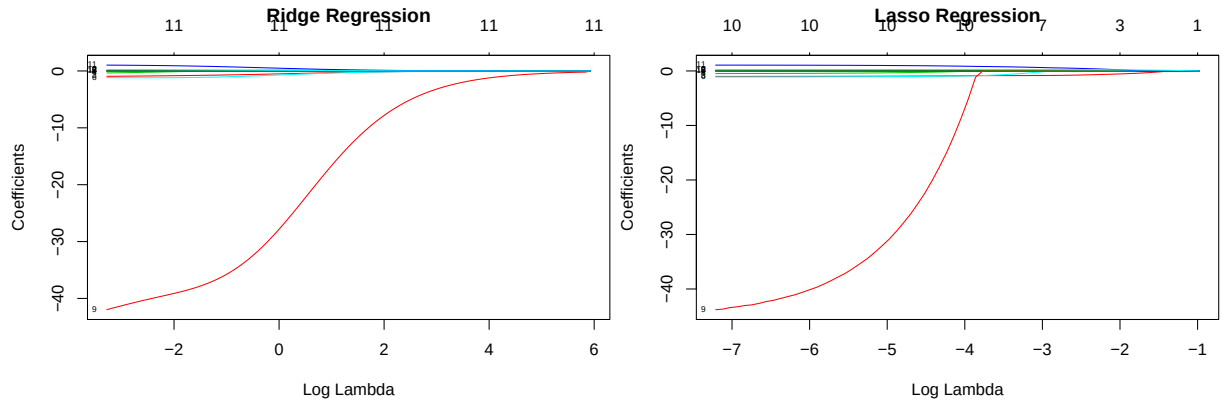

Figure 5: Plot of Lambda vs Coefficients for Ridge/Lasso Regression

3.1.4 Principal Components Regression

Principal components regression was used. Based on the analysis of the principal components, the first nine principal components were used to be trained on the training set. This was done because 90% of the variation could be explained by these first nine components. This is graphically displayed in the plot of principal components.

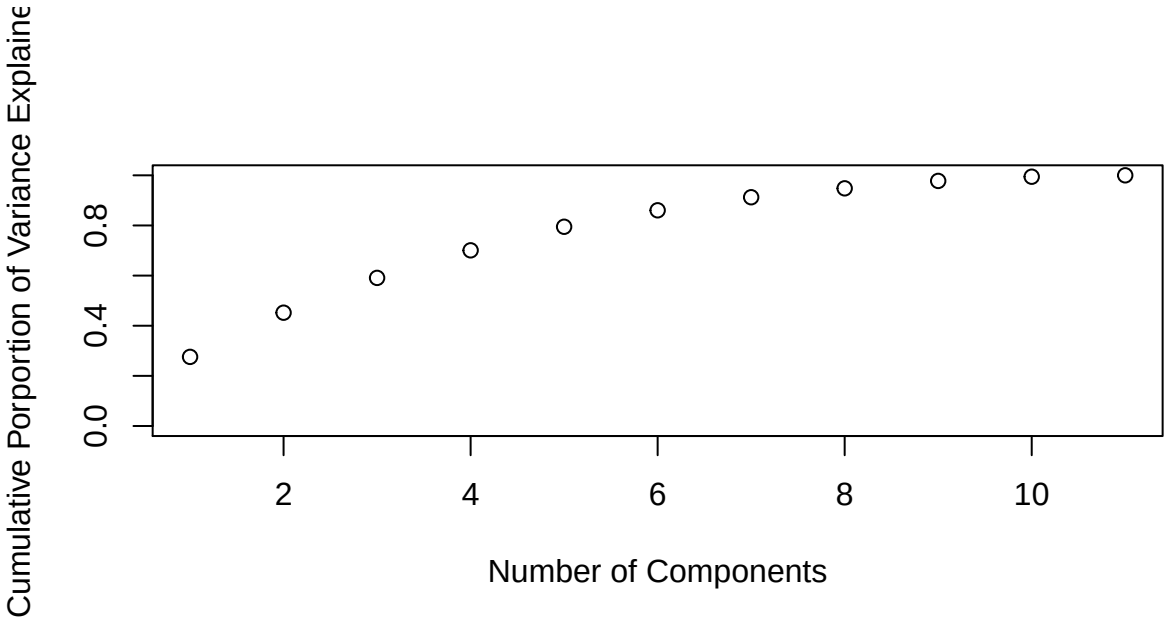

Figure 6: Variance Explained for Principal Components Regression

3.1.5 Partial Least Squares Regression

Partial least squares regression was used. However, the difference is that it uses quality response as supervision over the principal components. Using this method one can see from the plot of patial least squares components that after roughly 2-3 components, the model accuracy does not increase drastically.

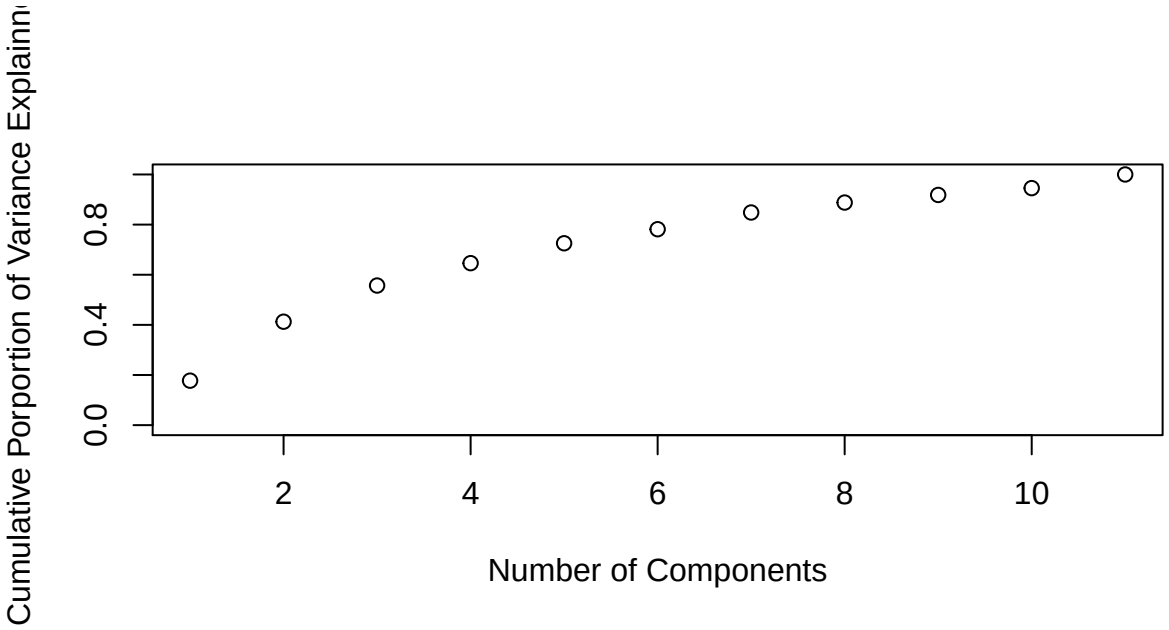

Figure 7: Variance Explained for Partial Least Squares Regression

3.1.6 Boosted Regression

Boosted regression was also used. 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.

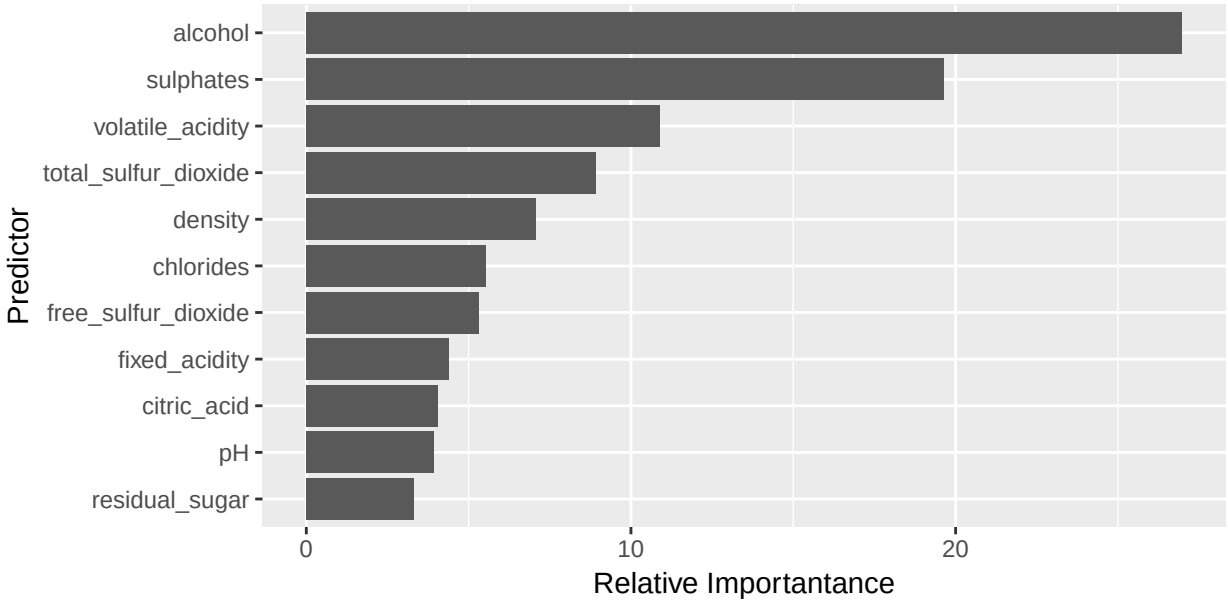

Figure 8: Relative Importance from Boosted Regression

3.1.7 Model Selection

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

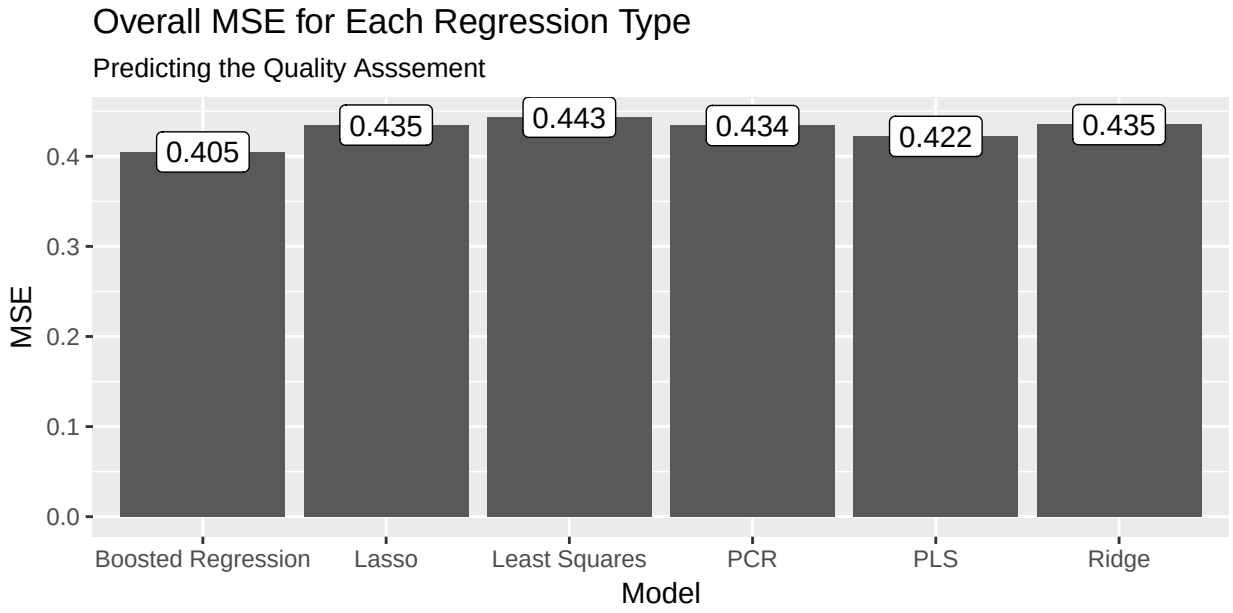

Figure 9: Plot of Results of Different Regression Techniques

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

Several different classification models were used in this analysis given the new variable added to the data set. The methods used were K-Nearest Neighbors, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and tree classification. These different models were trained on the training data set and then applied to the testing dataset to estimate the accuracy. It is important to note that greater accuracy was achieved by scaling values for the K-Nearest Neighbors approach as this approach uses Euclidean distances and thus is sensitive to scale differences. The phenomena can be seen as with the unscaled values the validation algorithm found that 17 were used versus 64. The larger number of neighbours makes for a much more global model, less sensitive to immediate neighbours in the bias versus variance trade off. The tree classification model was trained first through cross-validation and then pruned to six leaves in order to reduce the impact of over-fitting in the bias variance trade off.

3.3 Comparison of Models

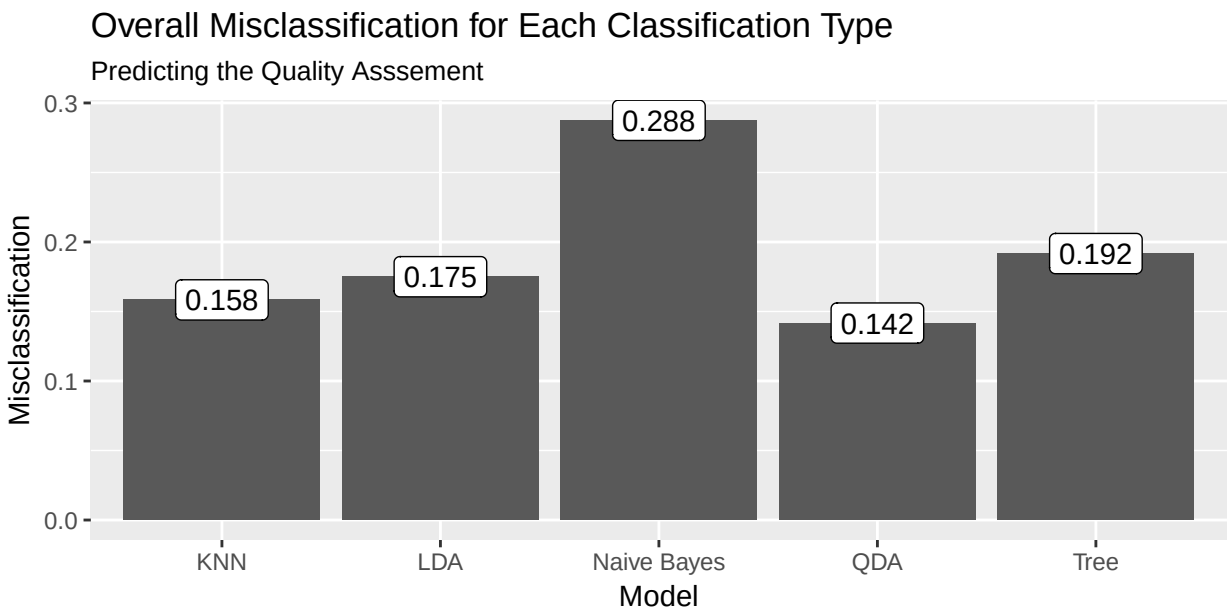

Figure 10: Plot of Results of Different Classification Techniques

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 regression resulted in the highest accuracy of all regression models; however, this accuracy comes at a cost of interpretability. Because the boosted algorithms have little intrepreation this accuracy is more 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 superior model for inference with high accuracy is characterized by the below equation:

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

Thus from 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 instances one can see that sulfate content appears to have a stronger positive influence on the wine quality while wines with additional residual sugars reduce the quality score. In the hands of the vintner, these relationships can be explored or potentially exploited to produce a higher quality wine more consistently.

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.

All of the methods for both classification and regression tended to indentify similar "influential predictors." Lasso, ridge, PCR, and PLS did a superior job at indentifying and treating potential collinearities in the predictors. For example acidity, citric acid and pH (which is a measurement of acidity) appear to be correlated in the corelation matrix. These models were able to minimize the impact of these collinearities where the least squares regression was not. This is why the decision was made to remve pH from the best subset selection predictors.

There were some issues with the dataset. As we used a publically available dataset without indentifying details (wine name, winery name, etc) we could not do further analysis into the cause of potential outliers. Had this information been available we could have been more confident with the elimination of outliers and other discrerning 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

This analysis shows that for regression the boosted regression resulted in the highest accuracy of all regression models; however, this accuracy comes at a cost of interpretability. Because the boosted algorithms have little intrepreation this accuracy is more 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.

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. The latter is of great importance as it has a much higher accuracy of predicting lower order classifications into a higher class. This is of critical importance for the reputation (and perhaops costing) of these wines.