Wine Data Project

Stephen Stewart

1666880

Introduction

The datasets being used are related to red and white variants of the Portuguese "Vinho Verde" wine. One data set is on red wine and has 1599 different varieties and the other is on white wine and has 4898 varieties. There are 11 physiochemical properties collected for each wine, which are continuous variables. There is also a quality property for wine which is an ordinal variable with possible ranking from 1 (worst) to 10 (best). The initial aim of this project is to analyse whether red and white wines share the same physiochemical properties using graphical methods and statistical tests. Then, to perform some unsupervised learning on the data (prior to modelling) to assess what properties may have an affect on the quality of wines. The main aim of this paper is to try to predict the quality of a wine based on its physiochemical properties using various regression and classification techniques. Finally, I conclude whether my initial inferences prior to modelling match those drawn from the data analysis stage, and choose a preferred model. (I have used *An Introduction to Statistical Learning* 2009 for PCA and for all data analysis inspiration).

Pre-Processing

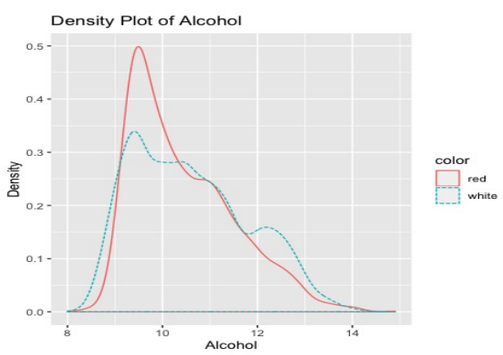
I merged the 1599 red and 4898 white wine observations into one data set of 6497 observations (which I called "winedata"), by adding a dummy variable to indicate whether an observation was a red or a white wine.

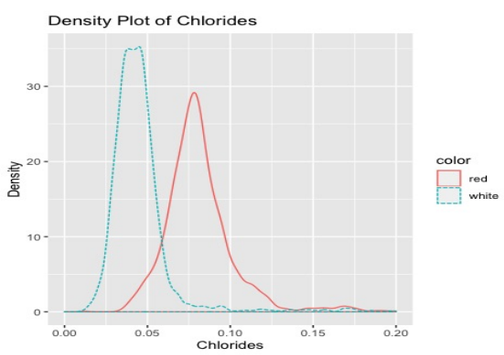
Data Quality Assessment

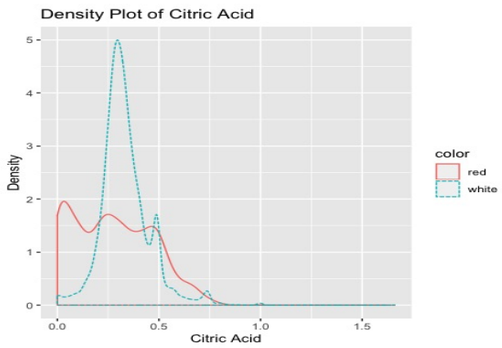
In the winedata set there were no missing values and 1177 duplicate observations which were removed, leaving 5320 unique observations (1359 red wine and 3961 white wine observations) for further analysis.

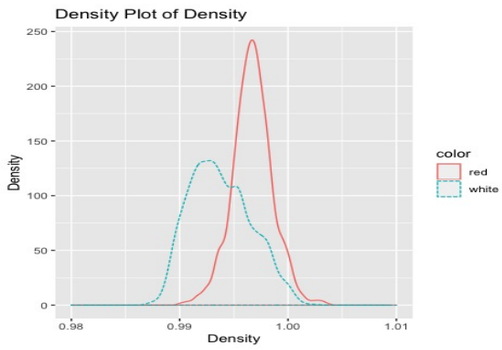
Feature Distribution

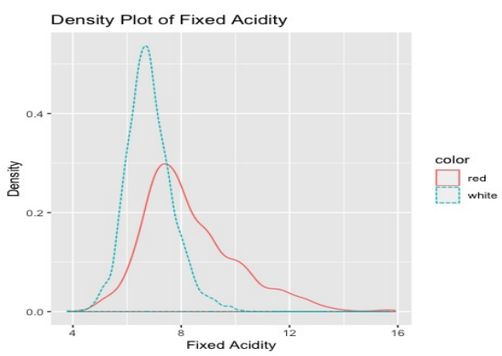
To check if each variable follows the same distribution for both red and white wines we assess if the variables follow normal distribution graphically using density plots:

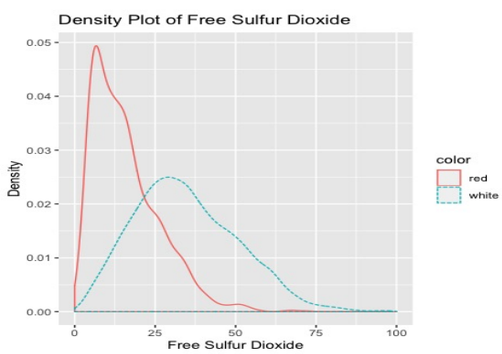


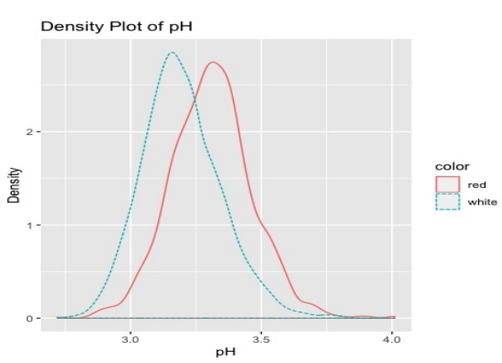


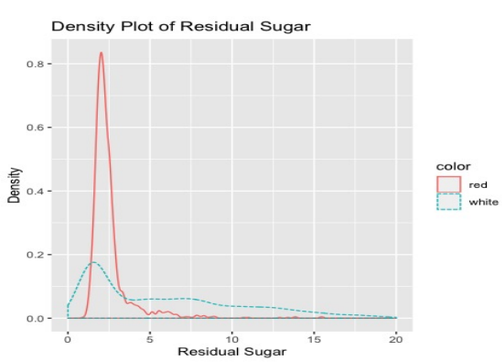












From the above density plots it can be seen that the chloride, citric acid, density, fixed acidity, free sulfur dioxide, pH, citric acid (for white wines) and residual sugar (for white wines) variables appear to follow normal distribution. For the next step I have assumed that all variables are normally distributed for both red and white wines. We must now check if the variances of the distributions are the same using an F-test. If the distributions have equal variance we must perform an unpaired two-sample t-test using the pooled variance, otherwise the Welch-Satterthwaite approximation to the degrees of freedom is used.

Results from tests

The results of the F-test showed that only variable with equal variance from both the red and white wine distributions was pH, all other variables had significantly different variances. The results of the t-test showed that the means from both distributions were significantly different for all variables. Hence red and white wines have **significantly different physiochemical properties**. Therefore I will fit separate models to predict white and red wine quality.

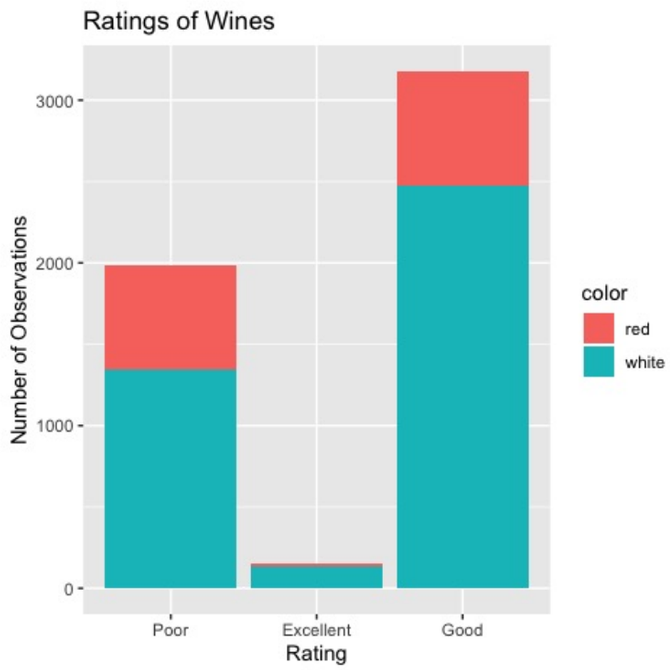
Standardising Variables

The numeric variables were scaled to have mean 0 and standard deviation 1 so that they are on a comparable scale. This is a necessary step before any classification or principal component analysis is undertook.

Adding Classification Factor

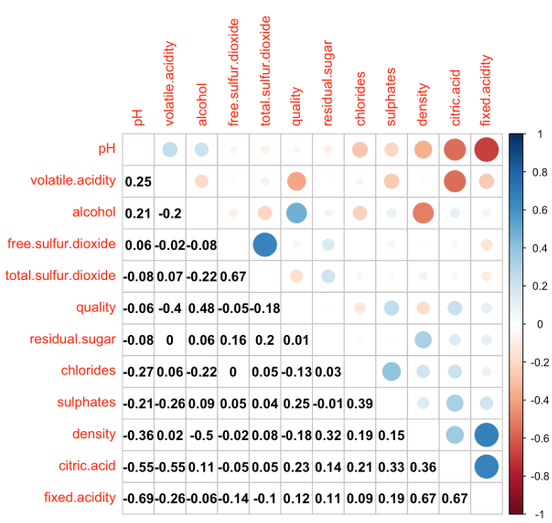
I have decided to add a qualitative factor to the data to denote whether a wine is poor, good or excellent using the following rating system.

Quality	Rating	Number of Observations
1-5	Poor	640 red, 1348 white
6-7	Good	702 red, 2477 white
8-9	Excellent	17 red, 136 white

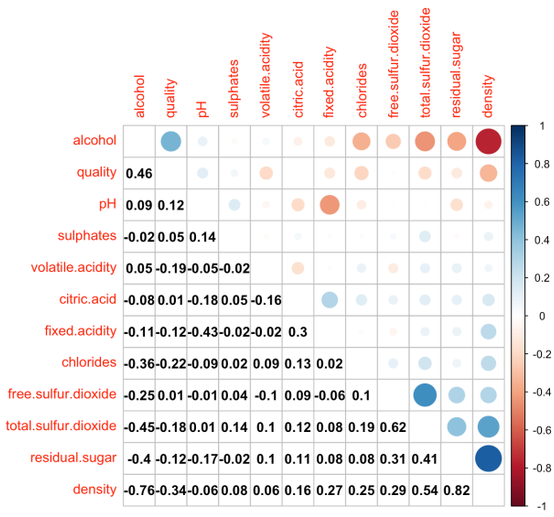


Correlation of Features

The correlation matrix of the correlation coefficients between variables for both red and white wines are given below.



(a) Red Wine



(b) White Wine

We obtain the following significant correlations for red wine features,

Correlation	Feature
Moderate positive	Total sulfur dioxide and free sulfur dioxide (0.67), fixed acidity and citric acid (0.67), density and fixed acidity (0.67)
Moderate negative	pH and fixed acidity (-0.69), citric acid and pH (-0.55), citric acid and volatile acidity (-0.55), alcohol and density (-0.5)

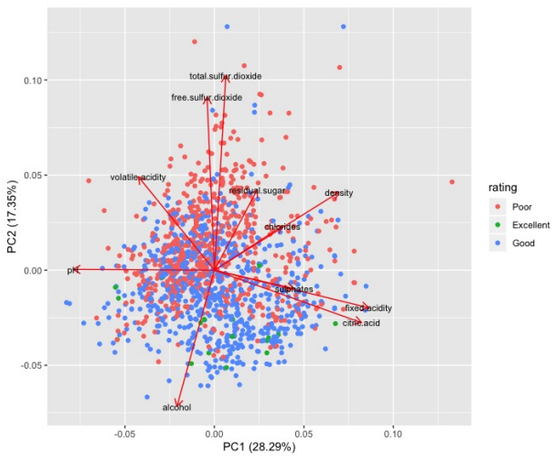
and the following significant correlations for white wine features,

Correlation	Feature
Strong positive	density and residual sugar (0.82),
Moderate positive	density and total sulfur dioxide (0.54), total sulfur dioxide and free sulfur dioxide (0.62)
Strong negative	Alcohol and density (-0.76)

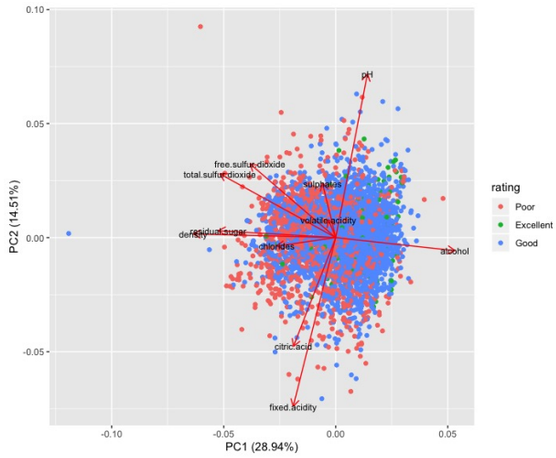
From these correlations I expect that when I fit a model later on that only one (out of the two) correlated features will be needed to predict the quality of wines, especially for the strongly correlated features.

Principal Component Analysis

Below are the graphs of the first and second principal components for the red and white winedata sets. The loading vectors are the red arrows. In doing this we can visualise in two dimensions what variables may have influence on the quality of wines, before fitting any models.



(a) PCA: Red Wine



(b) PCA: White Wine

From the above figures, the higher the alcohol content, the more good wines there appear to be for both red and white wines. For red wines, the higher the total sulfur dioxide, the more poor wines there appear to be.

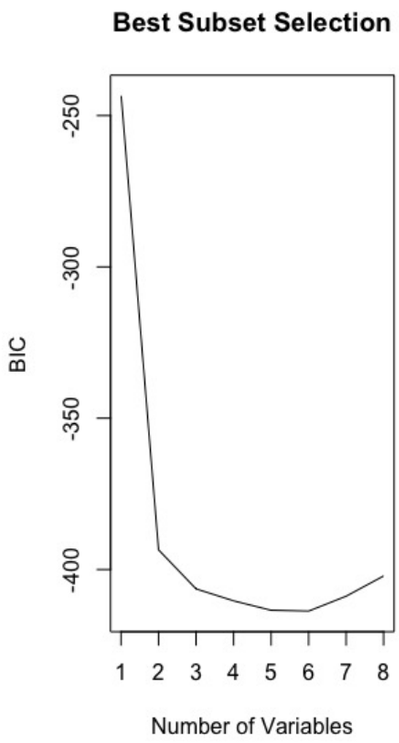
Partitioning Data

I randomly partitioned 80% of the winedata set into a training data set and used the remaining 20% as a test data set.

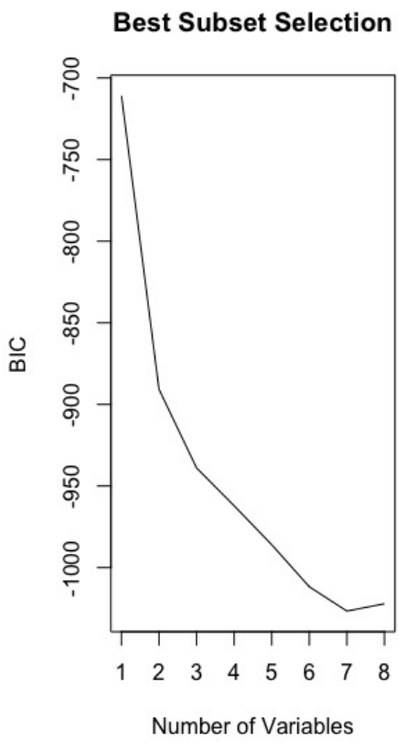
Data Analysis

Multiple Linear Regression

We would like to fit choose the best linear regression model to predict quality of red and white wines that has the lowest training error rate. Best subset selection method applied to all predictor variables yielded the following Bayesian Information Criterion (BIC) curves:

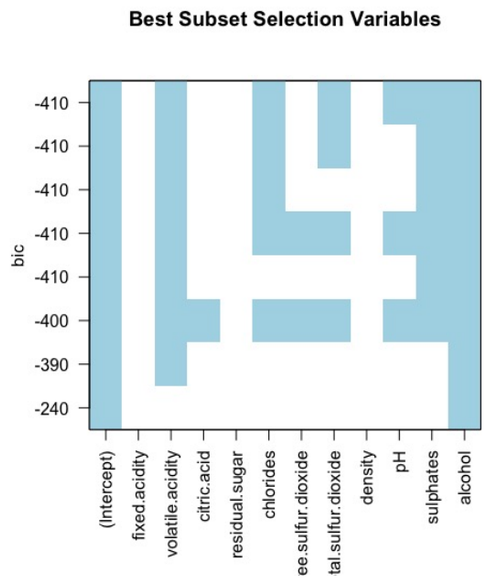


(a) Red Wine

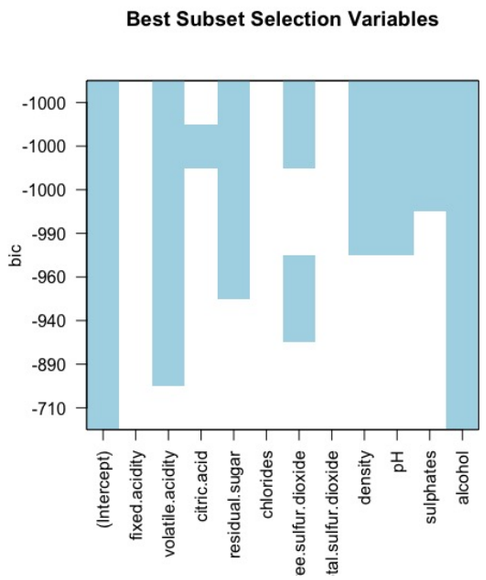


(b) White Wine

The lowest BIC implies the optimal model. For red wine 6 predictor variables are optimal, however there is minimal reduction in the BIC between 5 and 6 variables, so I will use 5 predictors in this model. For white wine, 7 predictor variables are optimal in the linear model. To choose the optimal variables for our model we can plot them for their given BIC level:



(a) Red Wine



(b) White Wine

The blue boxes along the top row indicate features that are present in the model with lowest BIC. Using this information I will fit a linear model to predict the quality of red wines

using the volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol features. I will also fit a linear model to predict the quality of white wines using the volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates and alcohol features. This is a very interesting result as it shows that the significant features in predicting the quality of red and white wines are quite different, which is not surprising due to the results from the feature distribution section.

The Model

Below are the resulting optimal multiple linear regression models to predict the quality of wines.

	quality		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.71	5.60 – 5.82	**<0.001**
volatile.acidity	-0.22	-0.26 – -0.18	**<0.001**
chlorides	-0.06	-0.09 – -0.02	**0.001**
total.sulfur.dioxide	-0.11	-0.18 – -0.04	**0.002**
sulphates	0.11	0.07 – 0.15	**<0.001**
alcohol	0.32	0.27 – 0.37	**<0.001**
Observations	1077		
R^2 / R^2 adjusted	0.345 / 0.342		

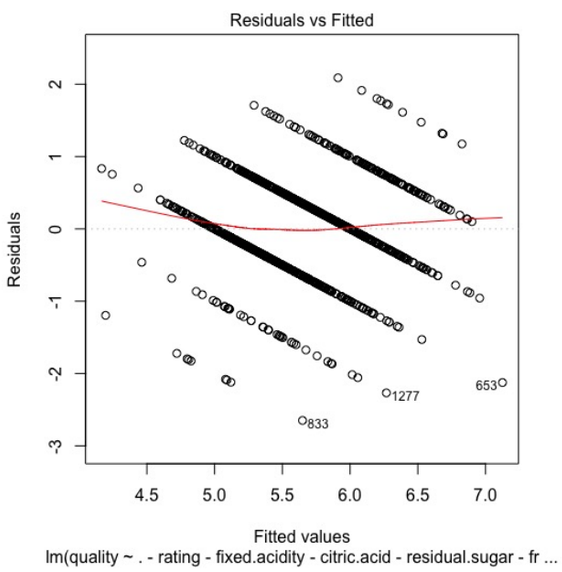
(a) Red Wine

	quality		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.64	5.60 – 5.68	**<0.001**
volatile.acidity	-0.30	-0.34 – -0.25	**<0.001**
residual.sugar	0.26	0.21 – 0.32	**<0.001**
free.sulfur.dioxide	0.07	0.04 – 0.10	**<0.001**
density	-0.30	-0.39 – -0.21	**<0.001**
pH	0.09	0.06 – 0.12	**<0.001**
sulphates	0.10	0.06 – 0.14	**<0.001**
alcohol	0.30	0.25 – 0.36	**<0.001**
Observations	3179		
R^2 / R^2 adjusted	0.291 / 0.289		

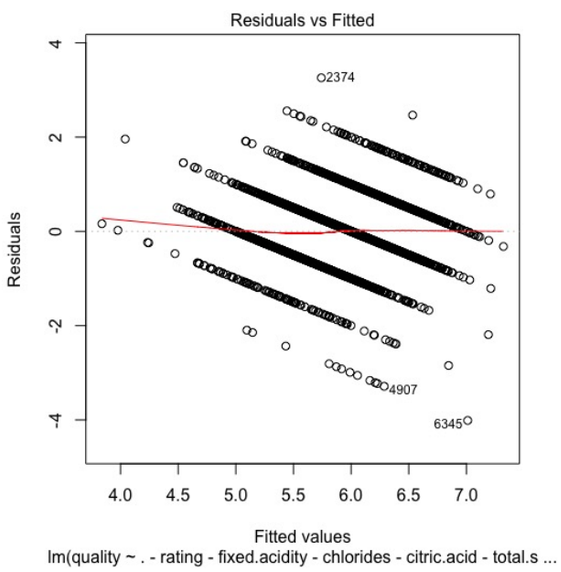
(b) White Wine

Linearity Assumptions

Do the assumptions for linear regression hold, mainly, do the error terms have constant variance. By plotting the fitted values against their corresponding residuals we can examine if heteroskedasticity exists (or not).



(a) Red Wine



(b) White Wine

There is no evidence in (a) or (b) for heteroskedasticity or a non-linear relationship. Hence, the errors have constant variance as required. The diagonal bands correspond to the different discrete responses for the quality of the wines, this would imply that performing clustering would be useful for further analysis.

Quality of Model

Next we check the quality of the parameter estimates in the model by comparing the standard errors of the parameter estimates above with the estimates from 100 bootstrap samples (using the entire data set).

Original Standard Deviation's

(Intercept)	*volatile.acidity*	*chlorides*	*total.sulfur.dioxide*	*sulphates*	*alcohol*
0.0557	0.0198	0.0169	0.0348	0.0197	0.0239

(a) Red Wine

Original Standard Deviation's

(Intercept)	*volatile.acidity*	*residual.sugar*	*free.sulfur.dioxide*	*density*	*pH*	*sulphates*	*alcohol*
0.0212	0.0223	0.0294	0.0147	0.0443	0.0149	0.0183	0.0262

(b) White Wine

Bootstrap Standard Deviation's

intercept	*volatile.acidity*	*total.sulfur.dioxide*	*sulphates*	*alcohol*	*chlorides*
0.0546	0.0234	0.0315	0.0230	0.0236	0.0169

(a) Red Wine

Bootstrap Standard Deviation's

intercept	*volatile.acidity*	*residual.sugar*	*free.sulfur.dioxide*	*density*	*pH*	*sulphates*	*alcohol*
0.0259	0.0231	0.0331	0.0235	0.0615	0.0143	0.0165	0.0358

(b) White Wine

The original standard errors are close to the bootstrap estimates which would imply that the model is accurate, however I think this can be improved using classification.

Conclusions from Model

For both wines; the quality has a strong positive relationship with alcohol content (i.e. higher alcohol content implies higher wine quality), a strong negative relationship with volatile acidity (i.e. higher volatile acidity implies lower wine quality) and a positive relationship with sulphates (i.e. higher sulphates content implies higher wine quality). Apart from these three variables all other features that predict white and red wine qualities are different. Most notably for white wines the quality has a strong positive relationship with the residual sugar feature. It is possible to assess the strength of the relationships between the quality and features based on the estimates of the feature coefficients because all features have been standardised.

Accuracy of Model

Model	**Training MSE**	**Test MSE**
Red Wine	0.4326001	0.453370
White Wine	0.5527832	0.5798496

The test error rate is only slightly larger than the training error rate for both models which is good.

Multinomial Logistic Regression

A multinomial logistic regression model is used to predict the probabilities of the different possible outcomes of a categorical variable, given a set of independent variables. I fitted a multinomial logistic regression to predict the rating of white and red wines separately using the variables chosen by best subset selection previously. In the resulting model for red wine, the parameter for the chloride feature was statistically insignificant (p-value of 0.327 for Excellent response) so I fitted the logistic regression model without this feature.

The Model

		rating		
Predictors	*Odds Ratios*	*CI*	*p*	*Response*
(Intercept)	0.00	0.00 – 0.02	**<0.001**	Excellent
volatile.acidity	0.35	0.15 – 0.83	**0.016**	Excellent
total.sulfur.dioxide	0.12	0.02 – 0.83	**0.032**	Excellent
sulphates	2.21	1.34 – 3.62	**0.002**	Excellent
alcohol	8.08	4.15 – 15.76	**<0.001**	Excellent
(Intercept)	0.99	0.66 – 1.47	0.948	Good
volatile.acidity	0.57	0.49 – 0.66	**<0.001**	Good
total.sulfur.dioxide	0.53	0.41 – 0.67	**<0.001**	Good
sulphates	1.27	1.12 – 1.43	**<0.001**	Good
alcohol	2.86	2.36 – 3.48	**<0.001**	Good
Observations	1077			
R^2 Nagelkerke	0.367			

(a) Red Wine

		rating		
Predictors	*Odds Ratios*	*CI*	*p*	*Response*
(Intercept)	0.03	0.02 – 0.04	**<0.001**	Excellent
volatile.acidity	0.32	0.21 – 0.48	**<0.001**	Excellent
residual.sugar	3.27	1.92 – 5.56	**<0.001**	Excellent
free.sulfur.dioxide	1.49	1.22 – 1.81	**<0.001**	Excellent
density	0.36	0.16 – 0.84	**0.018**	Excellent
pH	1.58	1.25 – 1.99	**<0.001**	Excellent
sulphates	1.35	1.04 – 1.76	**0.024**	Excellent
alcohol	5.90	3.59 – 9.69	**<0.001**	Excellent
(Intercept)	1.29	1.12 – 1.49	**<0.001**	Good
volatile.acidity	0.37	0.32 – 0.44	**<0.001**	Good
residual.sugar	1.88	1.55 – 2.29	**<0.001**	Good
free.sulfur.dioxide	1.17	1.06 – 1.29	**0.002**	Good
density	0.52	0.38 – 0.71	**<0.001**	Good
pH	1.23	1.11 – 1.35	**<0.001**	Good
sulphates	1.35	1.18 – 1.53	**<0.001**	Good
alcohol	2.46	2.04 – 2.97	**<0.001**	Good
Observations	3179			
R^2 Nagelkerke	0.324			

(b) White Wine

Quality of Model

Once again, like we did for regression, we check the quality of the parameter estimates in the model by comparing the standard errors (SE) of the parameter estimates above with the estimates from 100 bootstrap samples (using the entire data set).

	(Intercept)	volatile.acidity	total.sulfur.dioxide	sulphates	alcohol
Excellent	1.8370808	0.43710165	0.9800508	0.25330930	0.3405452
Good	0.2026903	0.07452967	0.1276109	0.06176323	0.0995231

Original SE: Red Wine

Bootstrap Standard Deviation's

E:intercept	*G:intercept*	*E:volatile.acidity*	*G:volatile.acidity*	*E:total.sulfur.dioxide*	*G:total.sulfur.dioxide*	*E:sulphates*	*G:sulphates*	*E:alcohol*	*G:alcohol*
1.6739	0.1602	0.3382	0.0616	0.8733	0.0942	0.1887	0.0612	0.3437	0.1075

Bootstrap SE: Red Wine (E:= Excellent, G:= Good)

	(Intercept)	volatile.acidity	residual.sugar	free.sulfur.dioxide	density	pH	sulphates	alcohol
Excellent	0.25494697	0.20927555	0.27108697	0.10100508	0.4258974	0.11980092	0.13404301	0.25328444
Good	0.07317575	0.08104615	0.09961296	0.05011192	0.1582644	0.04994668	0.06573297	0.09628856

Original SE: White Wine

Bootstrap Standard Deviation's

E:intercept	*G:intercept*	*E:volatile.acidity*	*G:volatile.acidity*	*E:residual.sugar*	*G:residual.sugar*	*E:free.sulfur.dioxide*	*G:free.sulfur.dioxide*	*E:density*	*G:density*	*E:pH*	*G:pH*	*E:sulphates*	*G:sulphates*	*E:alcohol*	*G:alcohol*
0.2140	0.0831	0.1947	0.0833	0.2557	0.1105	0.1064	0.0497	0.3651	0.2115	0.1046	0.0492	0.1156	0.0614	0.2269	0.1255

Bootstrap SE: White Wine (E:= Excellent, G:= Good)

The original SE's of the parameter estimates are closer to the bootstrap SE's than in the linear regression case (for both red and white wine models). This would suggest that the multinomial logistic regression model is a better model than the previous linear regression model.

Conclusions from Model

The parameter estimates would indicate that the same relationships between quality and properties can be drawn as in the multiple linear regression models.

Model Accuracy

To measure accuracy of the model I have calculated the probability of misclassification (which is what I am referring to when I use error rate in the below tables). The red wine model (a) misclassified 252 good, 148 poor and 0 excellent wines from the training set and 34 good, 31 poor and 0 excellent wines from the test set. The white wine model (b) misclassified 285 good, 637 poor and 0 excellent wines from the training set and 155 good, 59 poor and 0 excellent wines from the test set.

Model	**Training Error Rate**	**Test Error Rate**
Red Wine	27.7623%	23.04965%
White Wine	29.00283%	27.36573%

Interestingly the red wine model (a) can predict the rating of 77% of the wines in the test set correctly using only 4 out of 11 predictor variables. The training error rate is higher than the test error rate (especially for the red wine model) due to the difference in the observations in the training and test data set (i.e. test set much smaller that training set). However, this may also be a good sign that the model generalises well.

Conclusion

I was not correct in my hypothesis that when two properties were highly correlated, only one was needed to predict quality (after performing best subset selection). The multinomial logistic regression model is preferred over the linear regression model. Generally, the quality of red and white wines increases as alcohol and sulphates content increases and decreases as volatile acidity increases. Apart from this, the physiochemical properties that influence red and white wine quality differ completely. Most interestingly, the number of properties needed to predict red wine quality (with 77% accuracy) is only 4. 7 properties are needed to predict white wine quality with a similar degree of accuracy. In either case no excellent wines were predicted incorrectly. This is due to the fact that there are very few excellent wines (for both red and white) within the data set. The quality of analysis would be improved if there was equal reflection on both ends of the scale of quality. Overall, a test error rate of 23% and 27% for the chosen red and white wine models is still large and therefore the quality of wines cannot be boiled down to only their physiochemical properties.

If I performed the same analysis again I would change validation and test error sizes so that the test error is more reflective of the true error rate for both models. I would also perform hierarchical clustering techniques to try to see if groups of wines emerge that reflect the qualities of wines.

References

An Introduction to Statistical Learning (2009). [Online; accessed April 3, 2020]. URL: `http://faculty.marshall.usc.edu/gareth-james/ISL/`.

Extension

Linear Discriminant Analysis (LDA)

In linear discriminant analysis model we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to "flip" these around into estimates for $\Pr(Y = k|X = x)$. A linear discriminant analysis model was fitted to predict the ratings for both red and white wines using the same variables as in logistic regression to see if there is any improvement on accuracy.

Model Accuracy - Using Test Error Rate

The red wine model misclassified 34 good, 33 poor and 0 excellent wines, yielding a probability of misclassification of $\frac{67}{282} = 0.2375887$. The white wine model misclassified 154 good, 58 poor and 0 excellent wines, yielding a probability of misclassification of $\frac{213}{782} = 0.2710997$. There was no noticeable improvement of LDA.

Quadratic Discriminant Analysis (QDA)

Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. A QDA model was fitted to predict the ratings for both red and white wines using the same variables as in LDA to see if there is any improvement on accuracy.

Model Accuracy - Using Test Error Rate

The red wine model misclassified 34 good, 28 poor and 0 excellent wines, yielding a probability of misclassification of $\frac{62}{282} = 0.2198582$. The white wine model misclassified 141 good, 78 poor and 5 excellent wines, yielding a probability of misclassification of $\frac{224}{782} = 0.286445$. There is a no noticeable improvement of QDA.