edX Vinho Verde Quality Prediction Project

James Cruikshanks

7/27/2021

Introduction

Using a public data set and R scripts, this project attempts to create a model which would be useful to identify wine quality with a suite of chemical analyses. The data set is available at the following URL: http://www3.dsi.uminho.pt/pcortez/wine/winequality.zip. Exploratory data analysis reveals some useful insights about a few of the predictors and the subjective quality ratings themselves. The goal of the project is refined based on these insights, with the problem being defined as one of classification. Two algorithms, Recursive Partitioning and Regression Trees (RPART) and random forest, are used to predict the quality classification of wines based on the data.

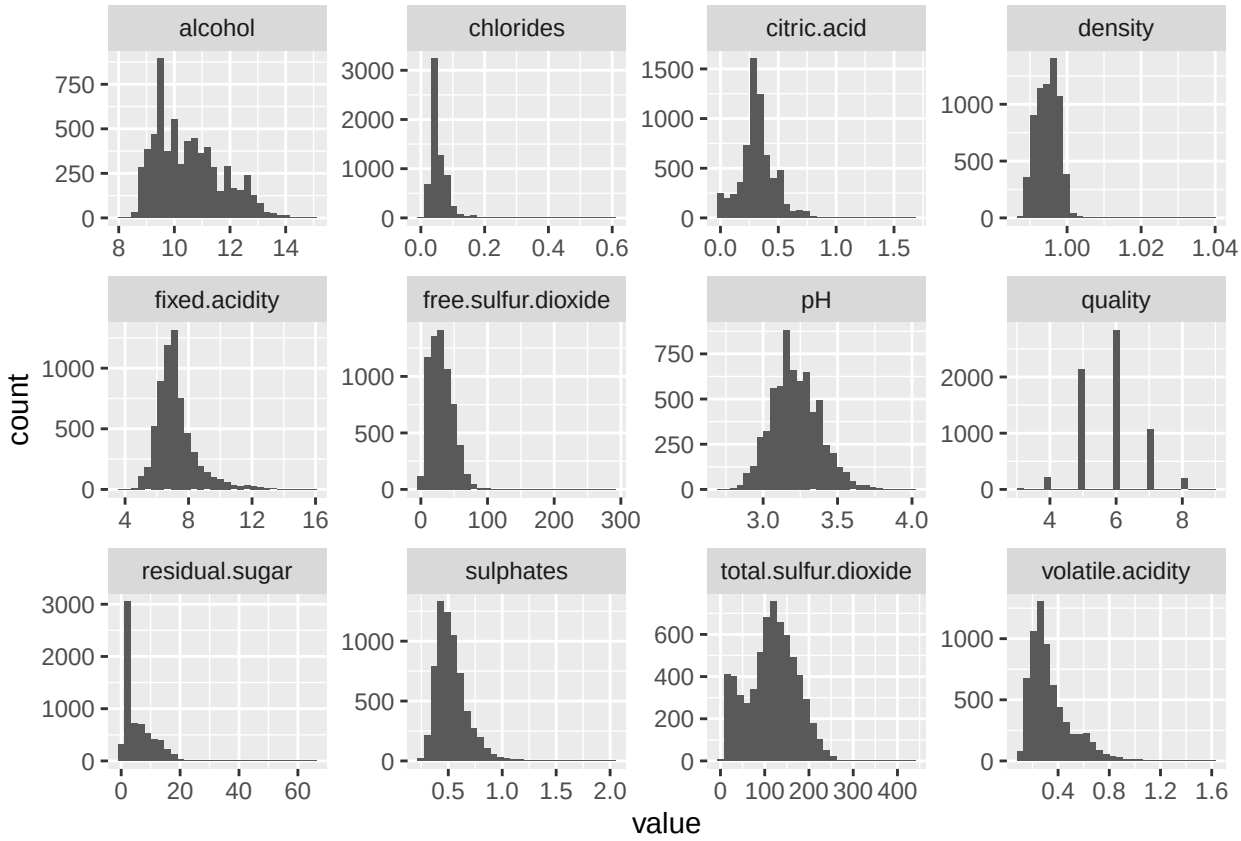
Analysis

Data Exploration

After downloading the tidy data set for each of the white and red wines, they are combined with an added factor for wine colour into a larger data set names 'wine quality'. We note the length (6497) and the features present in the data set. We can see the structure of the data by printing the first few rows of the data.

Next, we take a look at the distributions of the features by plotting histograms of each:

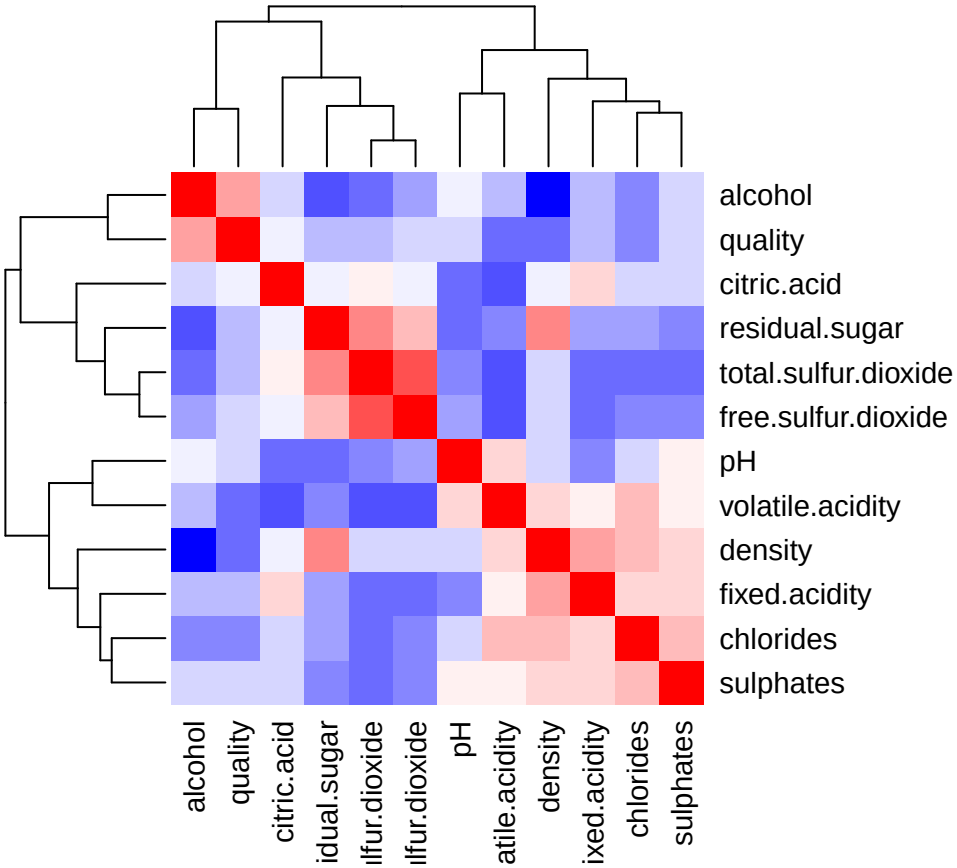
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	colour
7.4	0.70	0.00	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5	red
7.8	0.88	0.00	2.6	0.098	25	67	0.997	3.20	0.68	9.8	5	red
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	red
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	red
7.4	0.70	0.00	1.9	0.076	11	34	0.998	3.51	0.56	9.4	5	red
7.4	0.66	0.00	1.8	0.075	13	40	0.998	3.51	0.56	9.4	5	red



We note a few interesting observations:

- Quality is only rated at integers below 3 or above 9
- Most wines rated either 5 or 6
- Total sulfur has a bi-modal distribution
- The density is very uniform across all observations
- long tails on many of the distributions

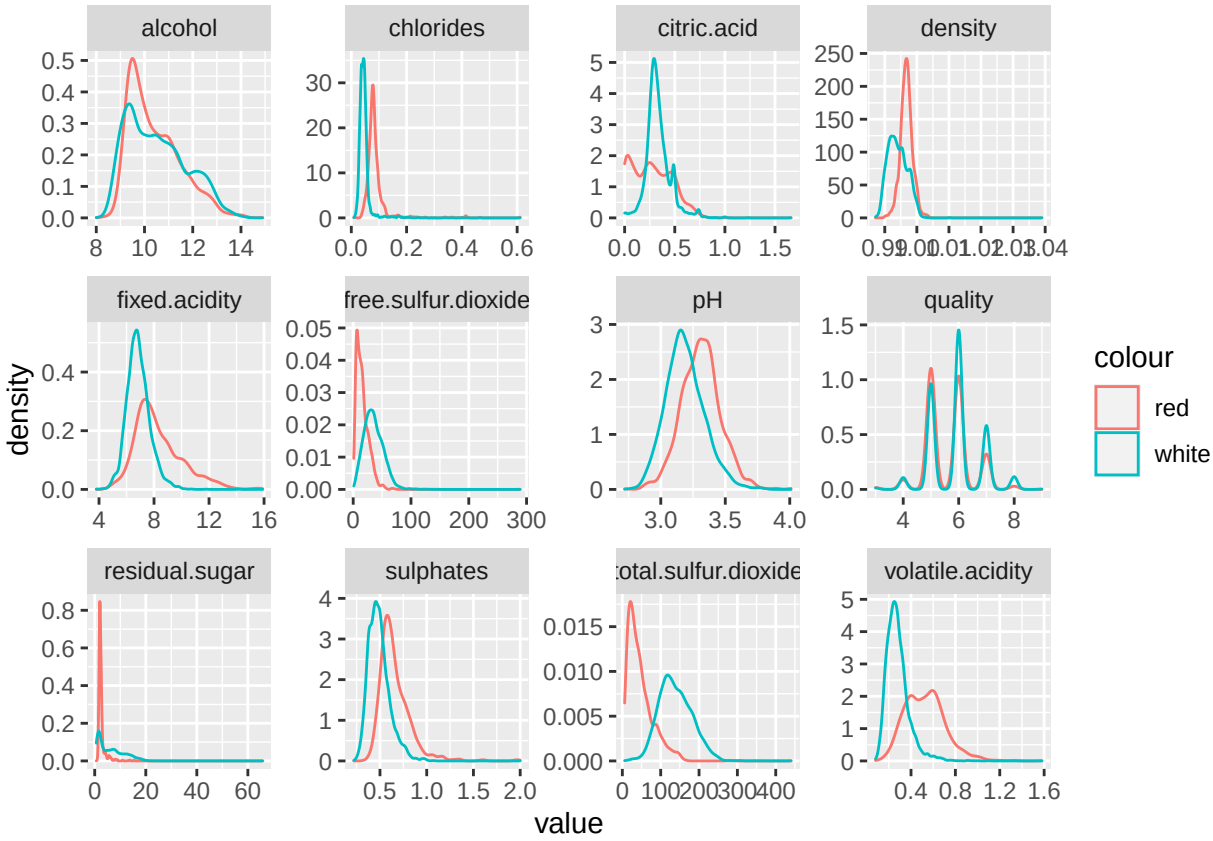
Next, we check for correlations between the features using a heatmap with red being positive and blue being negative.



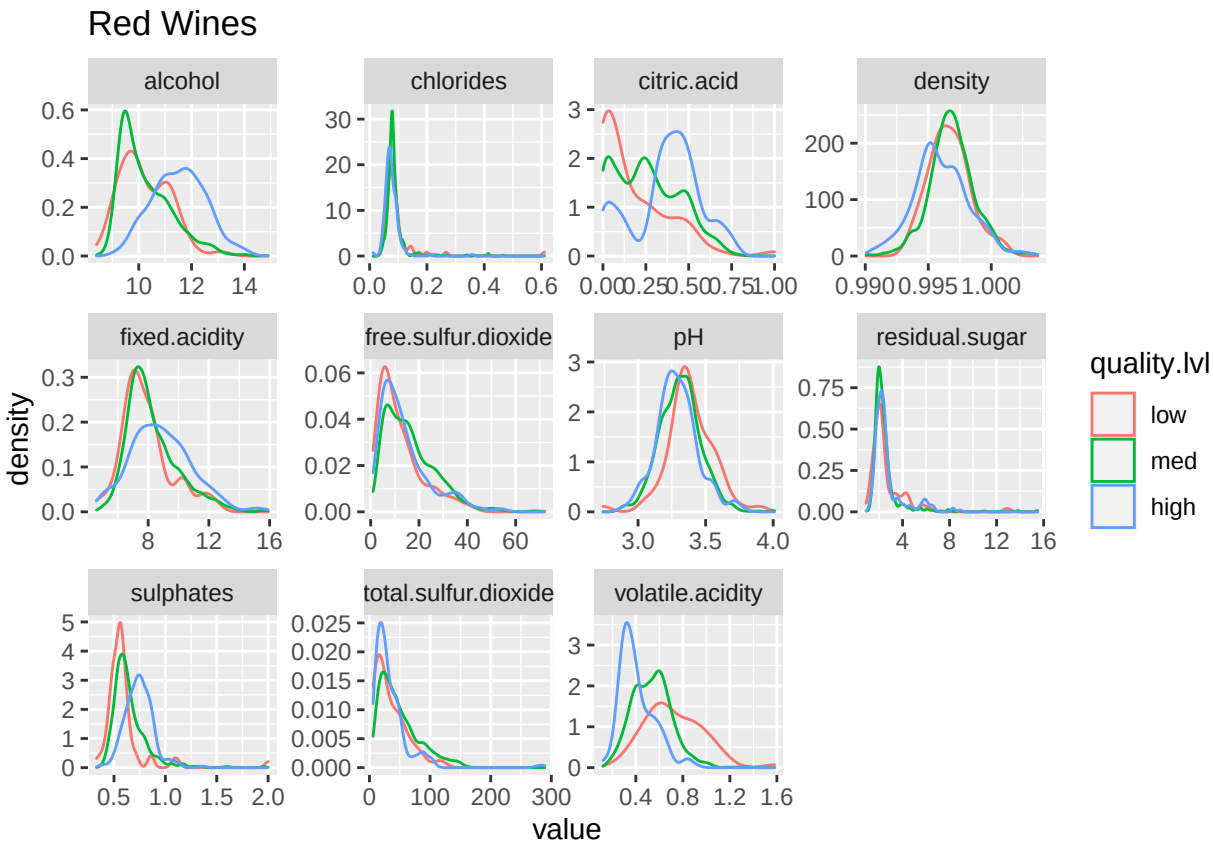
Alcohol, density, and volatile acid appear most correlated with quality. We also see correlations of density with alcohol and free sulfur dioxide with total sulfur as may be expected. Alcohol is less dense than than the water which makes up most of the remainder of the wine, and free sulfur dioxide is a component of total sulfur dioxide.

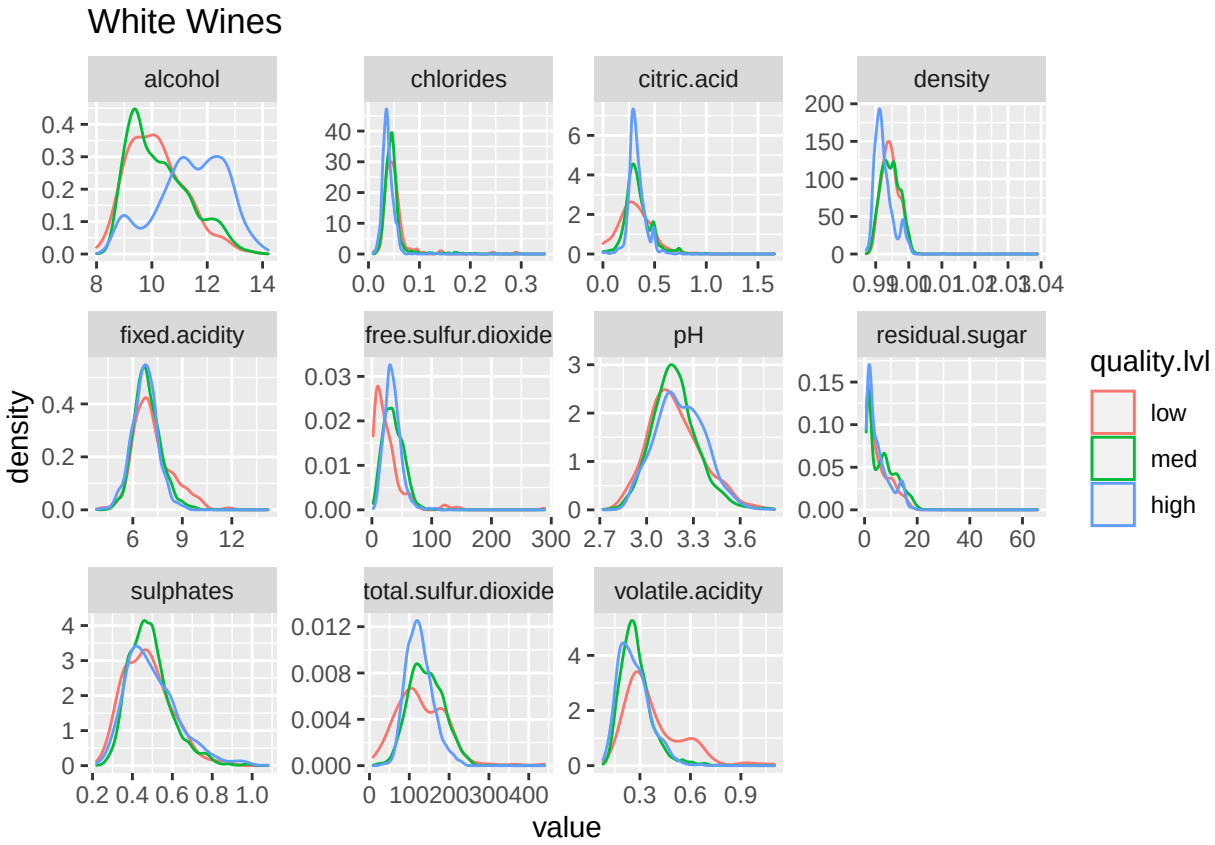
Next, to better achieve the goal of the analysis, we group the wine quality ratings. 5 and 6 are grouped into 'med' (very prevalent); 3 and 4 to 'low'; and 7, 8 and 9 into 'high' (both much less prevalent). This allows for better visualization and more practical classification. A wine maker is likely most interested in avoiding low quality product or achieving high quality product.

First we visualize the differences between reds and whites using density plots:



Here we note the reason for the bimodal total sulfur distribution, with whites containing higher levels in general. Similarly, reds contain more volatile acidity. What is not observed is a marked difference in quality through this data set.


Red Wines
alcohol
chlorides
citric.acid
density
fixed.acidity
free.sulfur.dioxide
pH
residual.sugar
sulphates
total.sulfur.dioxide
volatile.acidity
density
value
quality.lvl
low
med
high

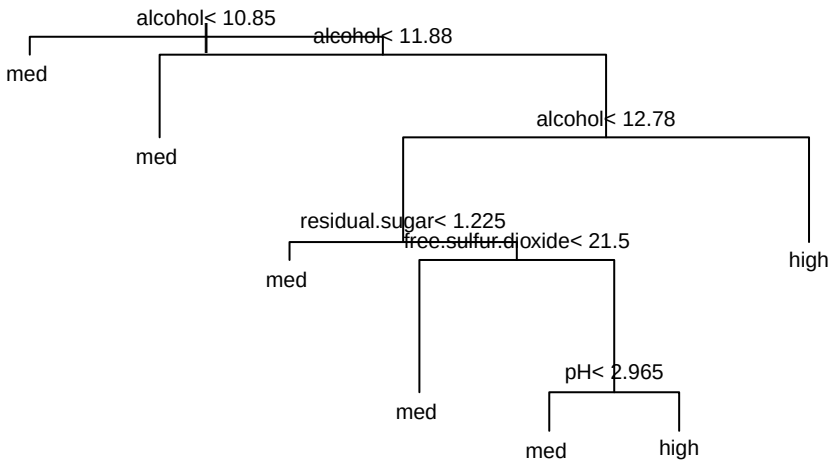


Using these plots we see the same features predicting quality as found using a general correlation. Higher alcohol, lower density, and lower volatile acidity wines appear to be rated more highly.

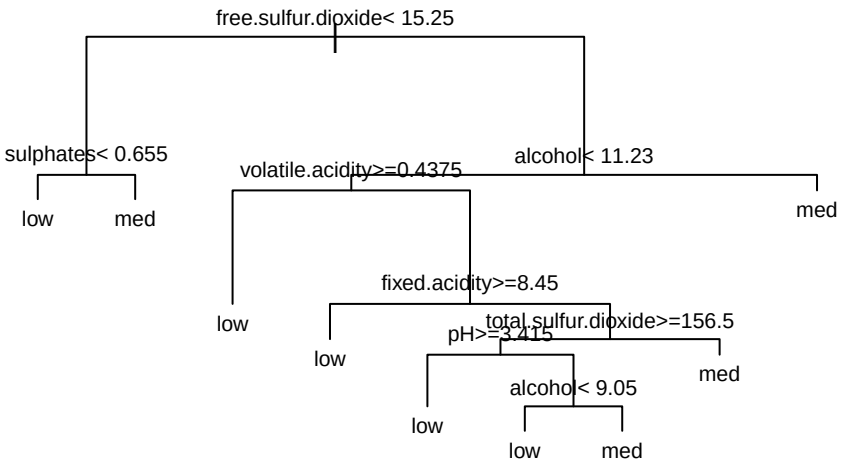
Modeling

Firstly, the training data set 'wine quality' is partitioned into a 'train' set and a 'test' set for various models to be trained while avoiding overfitting. The caret function 'createDataPartition' is used to stratify the data by quality level to maintain the prevalence of high and low quality wines. 80% of the data, or 5199 randomly selected ratings make up the training set and the remaining 1298 ratings are used for model testing. 80% is used for training to avoid over fitting with the moderate number of predictors.

The first classification algorithm used is RPART. The benefit of this algorithm is the easy interpretability although it is rarely the most accurate. The 'caret' package is used alongside the 'rpart' package to train the model. the complexity parameter, that is the minimum improvement in model accuracy to keep a new branch of the decision tree, is Kappa. This metric is better for the 'wine quality' data where high and low quality wines are rare. The metric compares the model accuracy to that of a random selector. The resulting classification model is shown here:

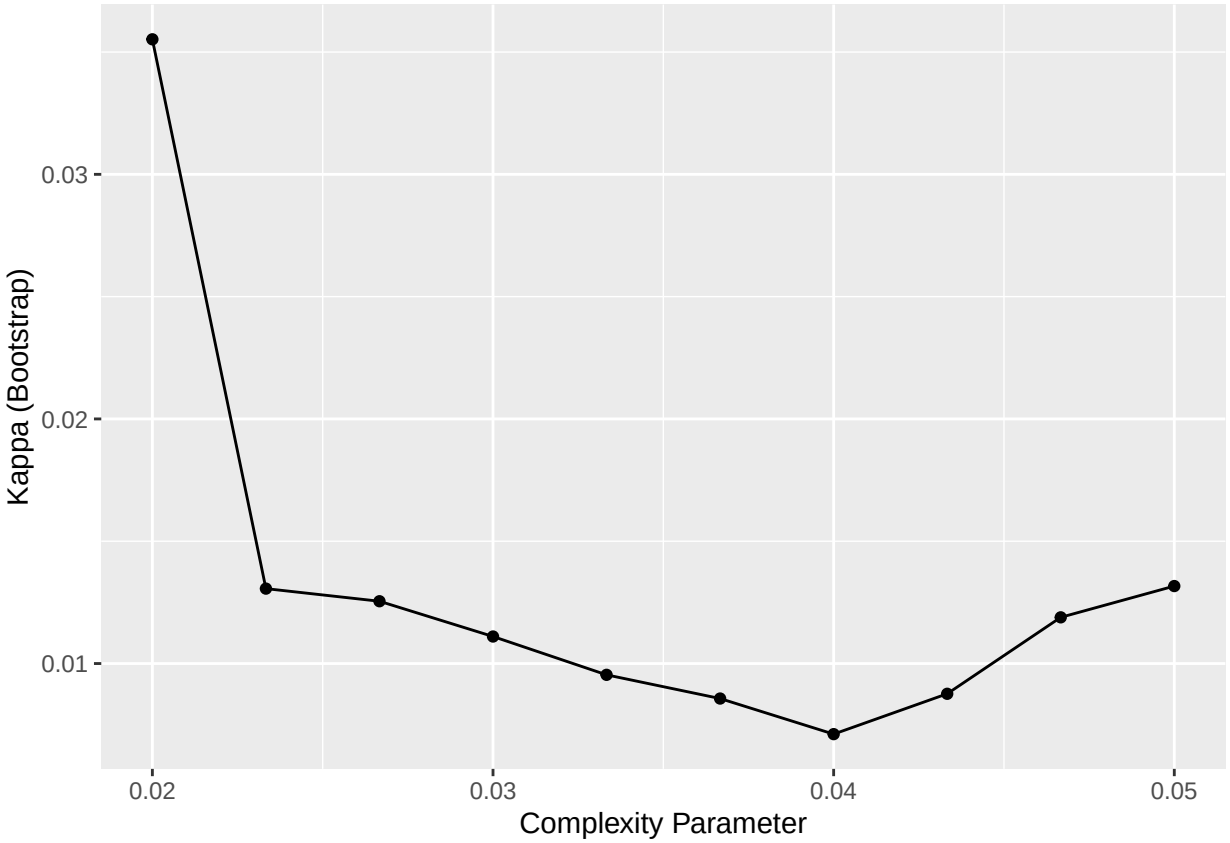


Alcohol, which we've seen is most strongly correlated to wine quality, is unsurprisingly the driver of the model. There are issues with the model, however, as no low quality wines are ever classified. This limits the usefulness of the model for any winemaker interested in avoiding low quality wine. To overcome this, we weight observations of low quality wines higher in the recursive partitioning process and check the model output:

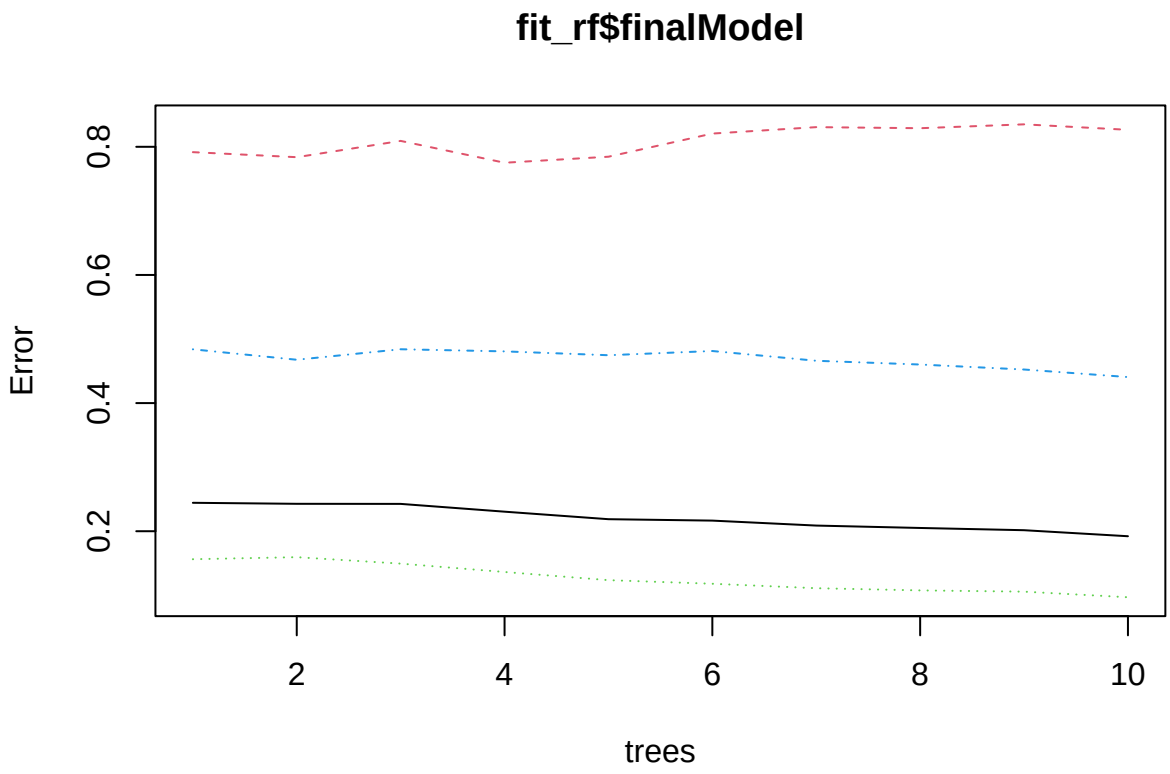


This model would be more useful for a winemaker looking to avoid poor product, but not as useful for one interested in a premium product. We can see the accuracy of the model across a grid of tuning parameters, in this case the complexity parameter 'cp'. We see an increase in Kappa, which is interpreted as only a small advantage over a random selection, increasing as 'cp' is lowered. This is likely due to overfitting the training data, a weakness of the RPART algorithm. The F1 score of the model is also included. F1 is chosen as it balances selectivity with sensitivity.

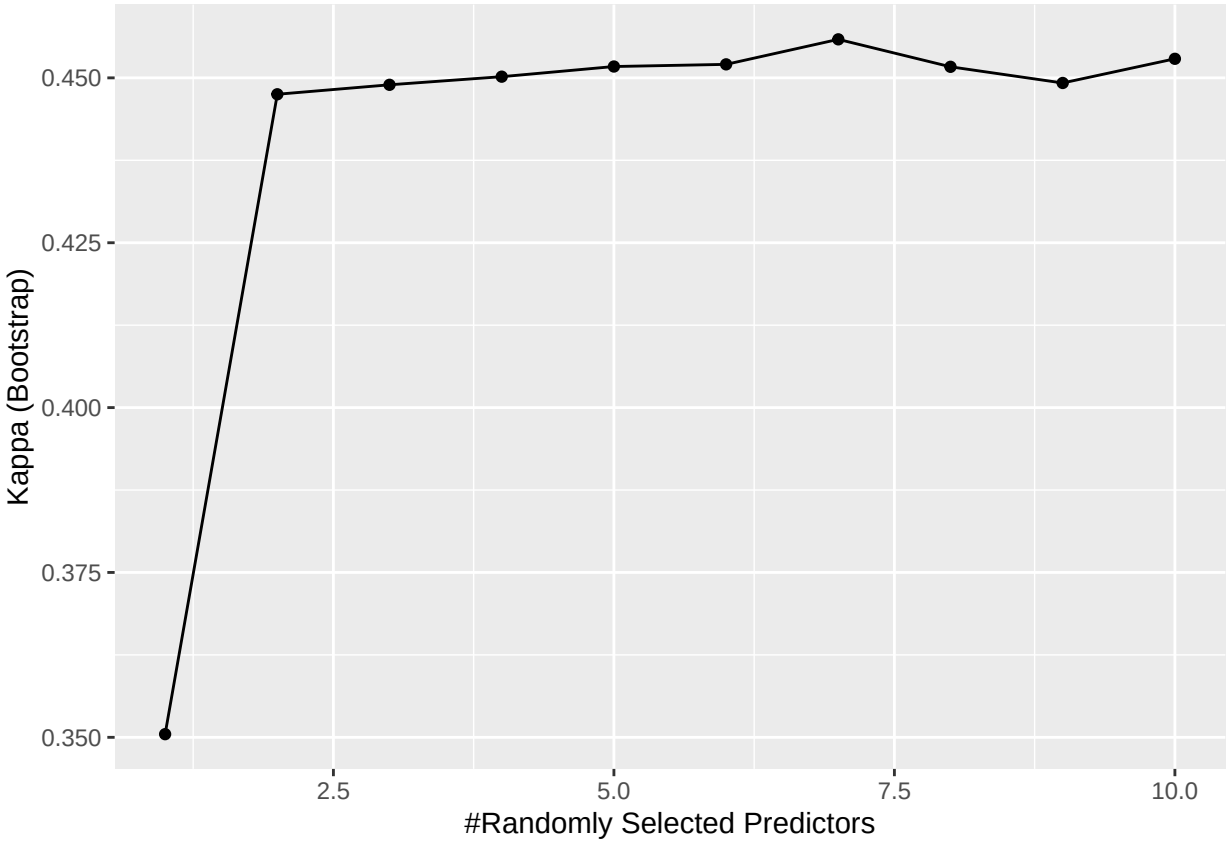
	F1 Score
Class: low	0.15102
Class: med	0.67531
Class: high	NA



Model performance can certainly be improved with other, more flexible, algorithms. Random forest is run next to classify both high and low quality wines with a higher F1 score. We use the same Kappa metric and reduce the number of trees to 100 after an initial run with the default 500 trees that shows a plateau in the error metrics, to reduce computation time.

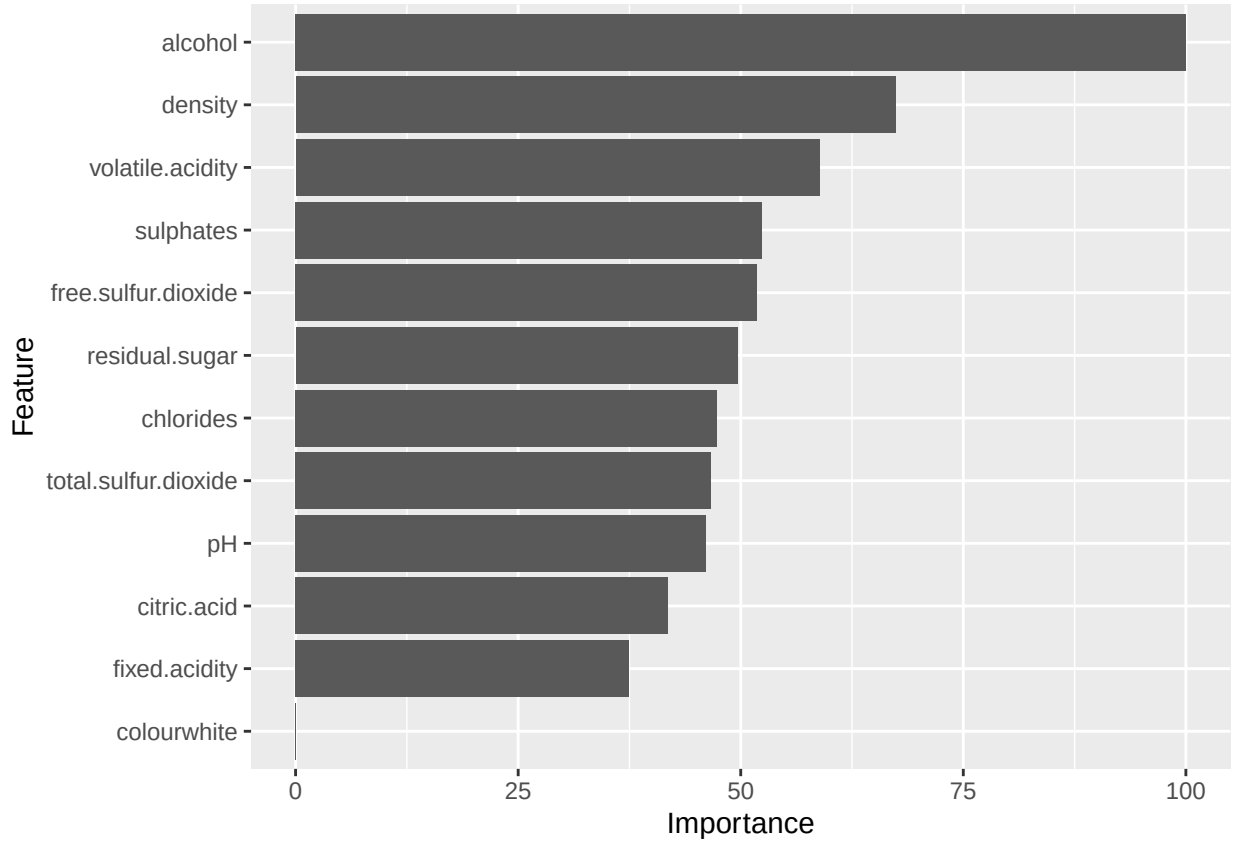


The number of predictors selected for each decision tree in the random forest can be a useful tuning parameter for the algorithm. In this case we see little improvement beyond 3 predictors per tree but slightly higher accuracy with 8, which is selected for use in the final model.



Next we plot the variable importance of each predictor to see which are most important for modeling, below. We notice the most important predictors are those which were most highly correlated with quality in the original data: alcohol, density, and volatile acidity. A winemaker might cut analysis costs by forgoing lower importance predictors.

	F1 Score
Class: low	0.34286
Class: med	0.90495
Class: high	0.66379



Results

Using the random forest model, we predict the quality level of the wines observed in the test set and print the confusion matrix and F1 scores.

```
##           Reference
## Prediction low med high
##       low   12   9    0
##       med   34 933  101
##       high   3  52  154
```

Here we see that the model is much improved from the RPART weighted decision tree, with high quality wines predicted, nearly double the F1 score on low quality wines, and over 20% higher F1 score for medium quality wines. This reinforces the model performance seen during training with Kappa values of 0.48, indicating that the model was close to 50% better than chance at identifying the rare high and low quality wines. High quality wines are more easily identified than lower quality wines with the final model, and the model would be useful in practice.

Conclusions

After downloading a data set of wine ratings, a number of statistical insights were gathered to inform a modeling approach. Examining the distribution of quality ratings it is clear that the project goals were suited to a classification model. A decision tree model was inflexible yet helpful to interpret the decision making taking place. A random forest model improved upon the decision tree model, providing flexibility at the cost of interpretability. In the end a model was developed which could be used by a winemaker to automatically estimate quality based mostly upon alocohol content, density, and volatile acidity. The model was developed with both sensitivity and selectivity as evaluation metrics. It can identify low quality wines with an F1 score of 0.28, medium quality wines with an F1 score of 0.91, and high quality wines with an F1 score of 0.67. The model could be improved by inclusion of many more observations, and perhaps in consultation with subject matter experts to determine which chemical analyses would influence taste beyond those included in this data set. Other classification models beyond the scope of the course may also be considered, such as neural networks.