ML Project - Wine

Christopher Wheatley

2022-10-24

Contents

Executive Summary **3**

Initial evaluation of Data **3**

Missing Values **4**

Data Partition **4**

Analysis **5**

Training Data Analysis 5

Correlation 5

Detailed Data Analysis 7

Dependent variable - 'quality' 7

Independent Variable 1 - 'fixed.acidity' 8

Dependent Variable 2 - 'volatile.acidity' 9

Dependent Variable 3 - 'citric.acid' 10

Dependent Variable 4 - 'chlorides' 11

Dependent Variable 5 - 'total.sulfur.dioxide' 12

Dependent Variable 6 - 'density' 13

Dependent variable 7 - 'sulphates' 14

Dependent variable 8 - 'alcohol' 15

Analysis Summary 16

Method **16**

Model Formulation 16

Classification And Regression Trees (CARTs) 16

Example Trees 16

Random Forests 17

Model Generation 17

Testing 20

Summary **20**

Acknowledgement **20**

Referencing **21**

Executive Summary

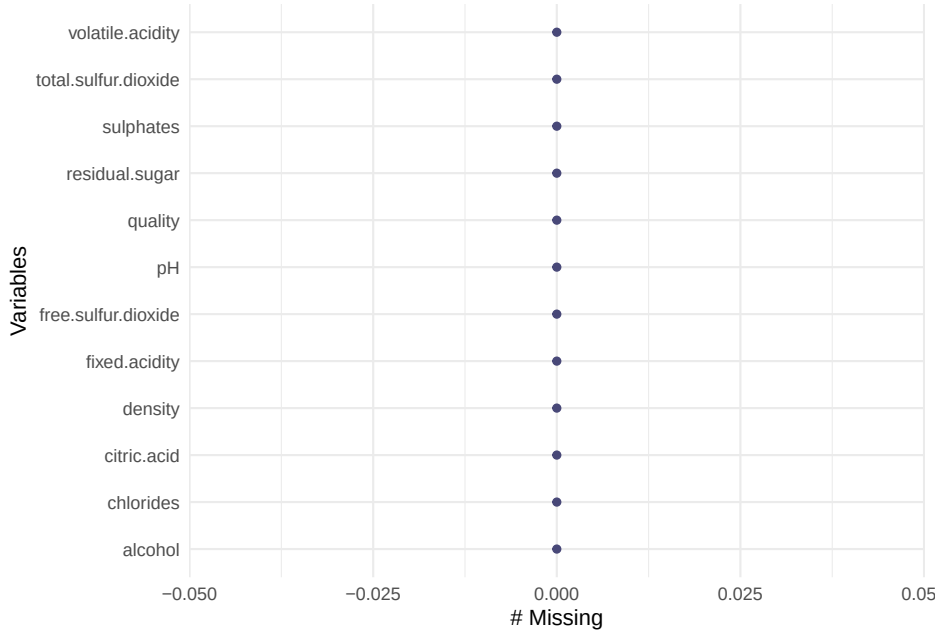
This project sets out to demonstrate sound theoretical understanding and the practical application of data analytics and machine learning through the 'R' programming language. Data pertains to testing information of red wine 'vinho verde' produced in Portugal. Wine from this region was sampled and tested using certification step analytical tests (Cortez et al., 2009). The data set structure has the dimensions of 1599 (n) observations with 12 variables (m). The data was partitioned into training and test sets (.8 and .2 respectively). The model's dependent variable 'quality' was identified and analysis was performed with relation to the others in training. Correlation was assessed between 'quality' and the remaining 11 independent variables. The three least correlated variables where removed from the analysis step for brevity. A Random Forest model was chosen for it's performance in regression and classification problems. From the training data a model was generated and refined in terms of optimal randomly selected variables and the number of trees for each forest. The training model was validated with test data. Producing an RMSE of .53 and an accuracy score of 70% (correct predictions / total predictions). I would like to acknowledge and reference the information contributors and authors from which this project sourced. Specific academic reference is provided at the end of this report.

Initial evaluation of Data

```
## Rows: 1,599
## Columns: 12
## $ fixed.acidity        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
## $ volatile.acidity     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
## $ citric.acid          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
## $ residual.sugar       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
## $ chlorides            <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
## $ free.sulfur.dioxide  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
## $ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
## $ density              <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
## $ pH                   <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
## $ sulphates            <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
## $ alcohol              <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.~
## $ quality              <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5, 7~
```

All variables, excluding 'quality' are of the class "dbl", which is an abbreviation for 'double-precision floating point number'. Meaning the information holds a numeric value able to contain decimal values. The parameter 'quality' is an integer. This will be our class variable for the supervised model. As such, we may have to factorize this parameter for ease in computation and modelling.

Missing Values



Nil missing values.. phew.

Data Partition

Partition data into training and test sets. Due to the relatively low magnitude of observations, we will set the test data to .2 of the total observations (321). This will prioritize validation accuracy over training accuracy, in turn reducing the risk of high variance or an overfit on a small data set.

The training data set has the following dimensions (nrow, ncol). 1278, 12

The test data set has the following dimensions (nrow, ncol). 321, 12

Analysis

Training Data Analysis

```
## 'data.frame':    1278 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 .
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

All variables within our data set are defined as either numeric or an integer. As such; let's evaluate the correlation between our independent variables and the dependent variable 'quality'. Correlations with lesser statistical relation, will be filtered out of our analysis to expedite this project and ensure the project is submitted prior to the advertised deadline.

Correlation

```
##                          Quality
## fixed.acidity         0.10928379
## volatile.acidity     -0.41271180
## citric.acid           0.23213775
## residual.sugar        0.02447357
## chlorides            -0.13762942
## free.sulfur.dioxide  -0.03598658
## total.sulfur.dioxide -0.16576392
## density              -0.20584579
## pH                   -0.06145488
## sulphates             0.24521016
## alcohol               0.49012743
```

Adjust values to absolute and arrange values descending.

```
##                         Quality
## alcohol              0.49012743
## volatile.acidity     0.41271180
## sulphates            0.24521016
## citric.acid          0.23213775
## density              0.20584579
## total.sulfur.dioxide 0.16576392
## chlorides            0.13762942
## fixed.acidity        0.10928379
## pH                   0.06145488
## free.sulfur.dioxide  0.03598658
## residual.sugar       0.02447357
```

Filter out the 3 least correlated:
- residual.sugar - free.sulfur.dioxide - pH

Detailed Data Analysis

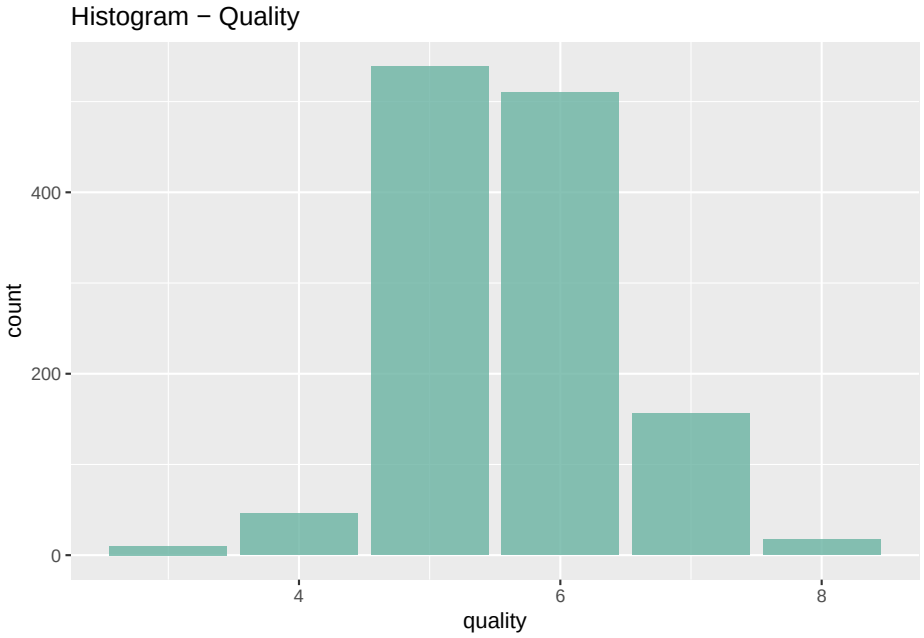
Dependent variable - 'quality'

This variable refers to the subjective quality of our wine samples. Values are represented as a median score provided from at least three separate human / 'sensory' testers. Testing was conducted using a blind tasting method. Quality was to be graded by each tester in the range of 0(very bad) to 10(excellent).

Sample:

```
##  [1] 5 5 5 6 5 5 5 7 7 5
```

Histogram:



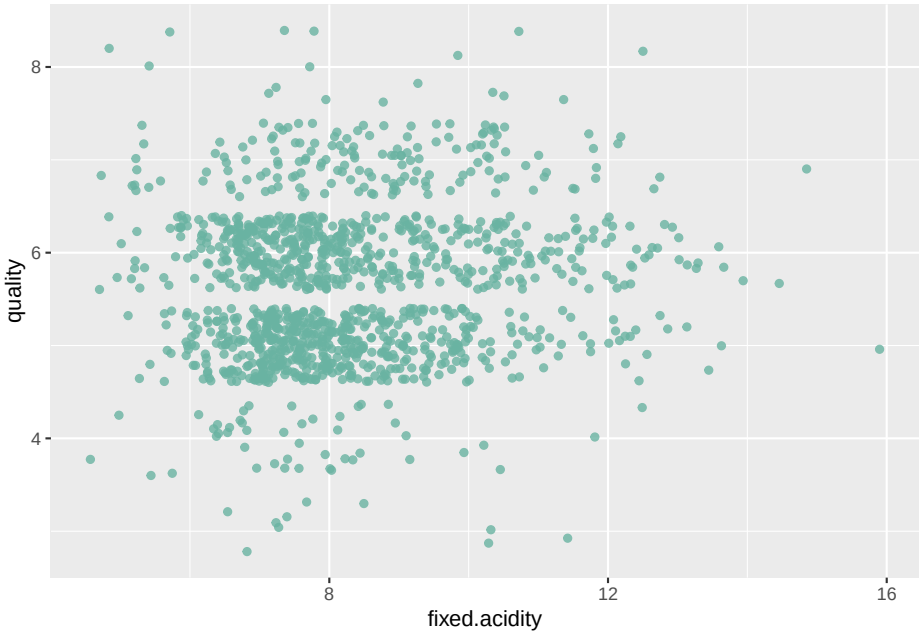
5 and 6 are the highest reported quality scores from our training data set.

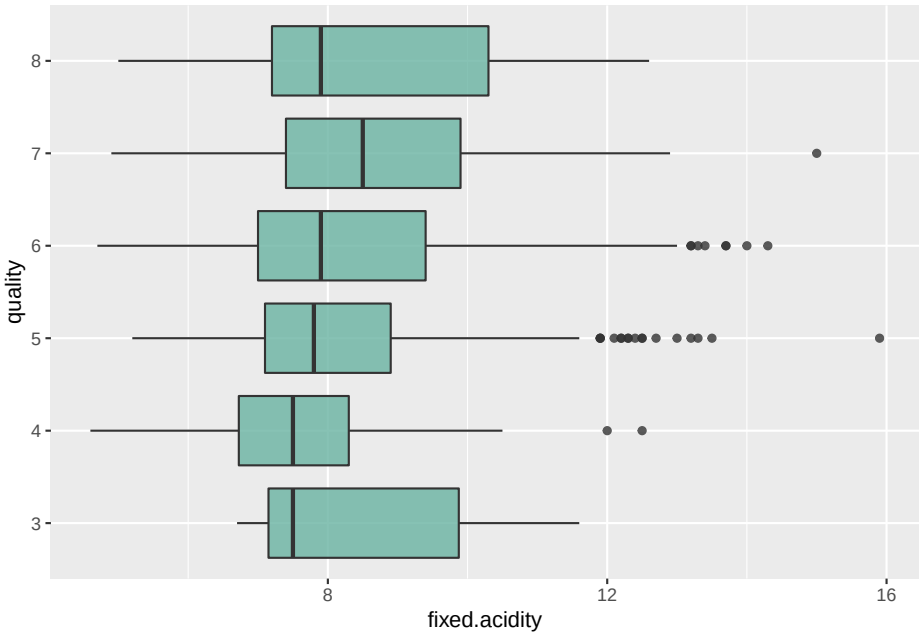
Frequency tabulated:

```
## # A tibble: 6 x 3
##   quality count proportion
##     <int> <int>      <dbl>
## 1       5   539       0.42
## 2       6   510       0.4
## 3       7   156       0.12
## 4       4    46       0.04
## 5       3    10       0.01
## 6       8    17       0.01
```

Independent Variable 1 - 'fixed.acidity'

When it comes to taste, it is widely known that acid plays a key role. Food or beverages low in acidity are often characterized as dull or dampened, where as high acidity is often described as sour or tart. As such, acidity in wine is no different and is believed to be a strong predictor in quality assessments. Fixed acids; loosely meaning they are produced in the fermentation process. The most commonly occurring fixed acids are; tartaric, malic, citric and succinic.

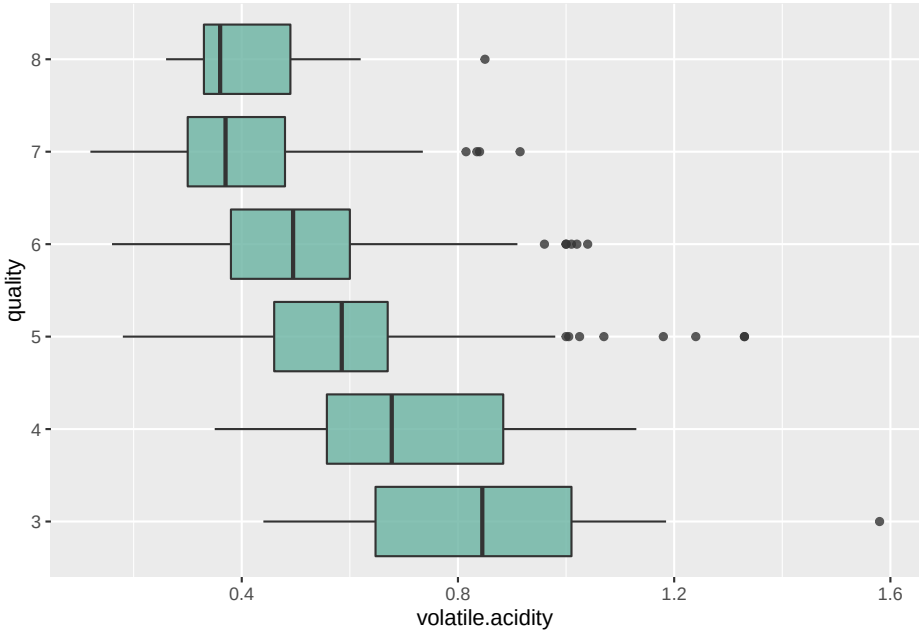




The box plots of fixed.acidity against quality shows minimal variability and no clear relationship. Fixed acids within the range of 7 and 11 are most common across each of the quality values.

Dependent Variable 2 - 'volatile.acidity'

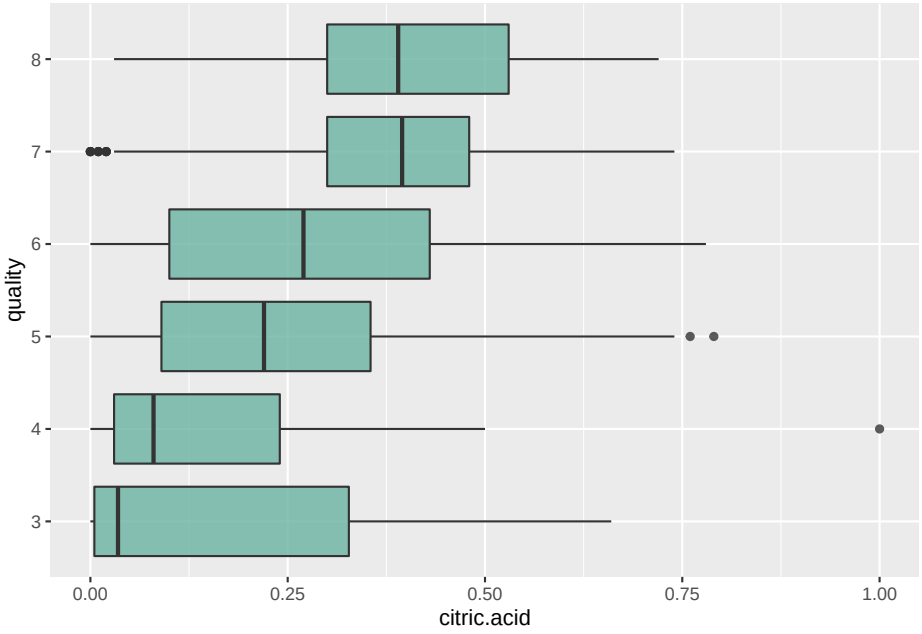
Volatile acids are often related to aroma as opposed to taste. They are describes as gas like acids within wine. Since our smell is a key contributor to taste, i would think there will be a relationship here.



Lesser volatile.acidity values infers a higher quality score.. interesting.

Dependent Variable 3 - 'citric.acid'

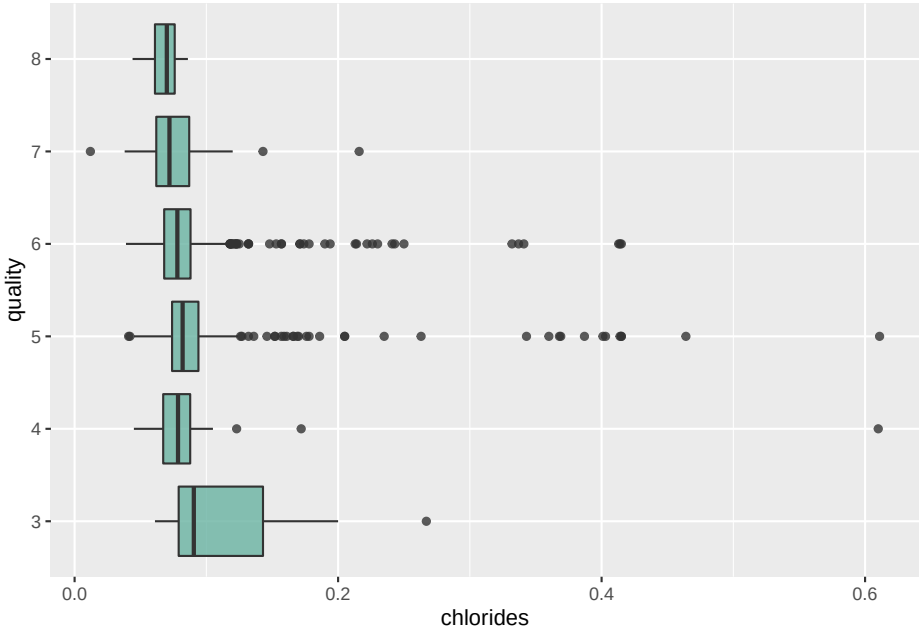
Citric acid is utilized in wine production as a preservative or supplement and is often related to flavors which are tart or sour.



Higher citric.acid values infers an increase in quality score.

Dependent Variable 4 - 'chlorides'

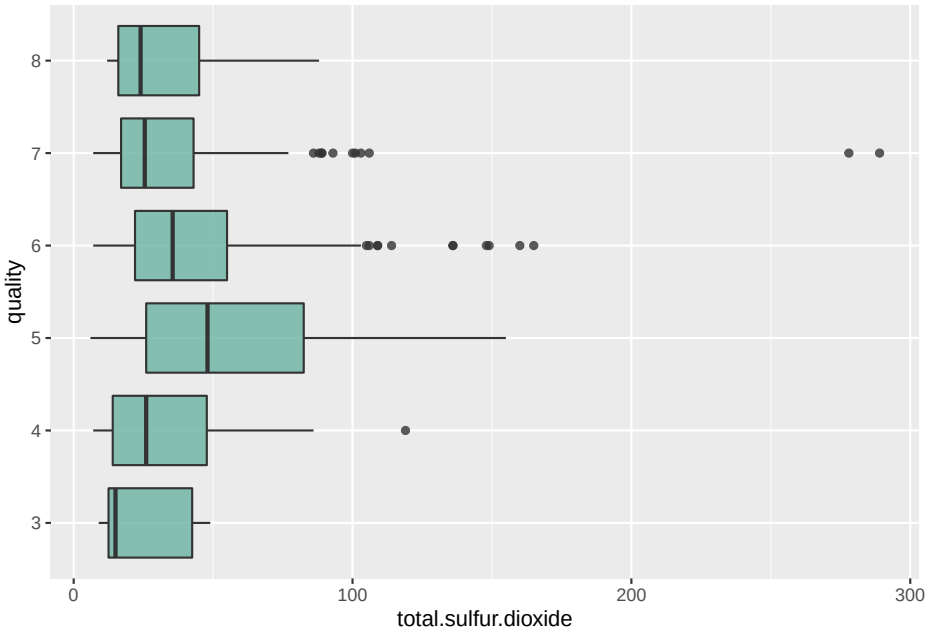
Chloride is used as a measure of salt in the wine.

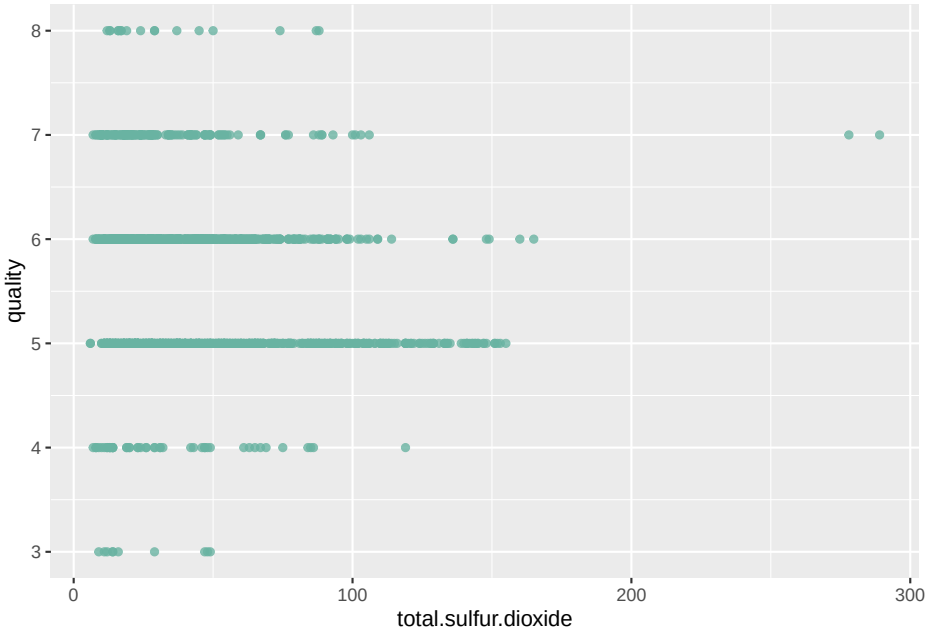


Increased salt infers a lower quality scored wine.

Dependent Variable 5 - 'total.sulfur.dioxide'

Is a preservative added to wines in the bottling process, to reduce the likelihood of spoiling.

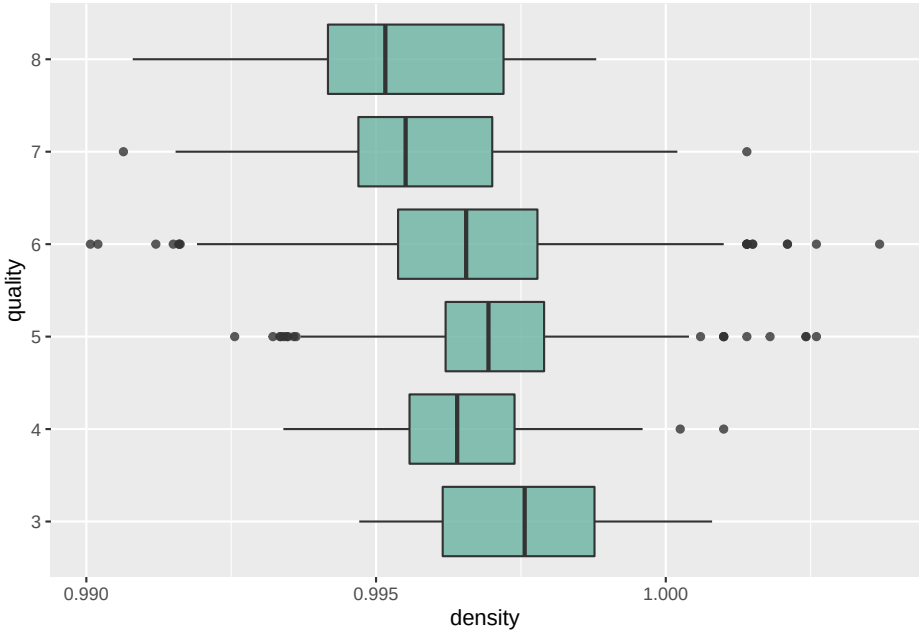




The above plots highlight a potential non-linear relationship and a slightly symmetrical distribution. Higher levels of the sulfur dioxide is attributed to consistency / average quality scores. Which may highlight it's effectiveness in controlling volatility and stabilizing quality around the mean.

Dependent Variable 6 - 'density'

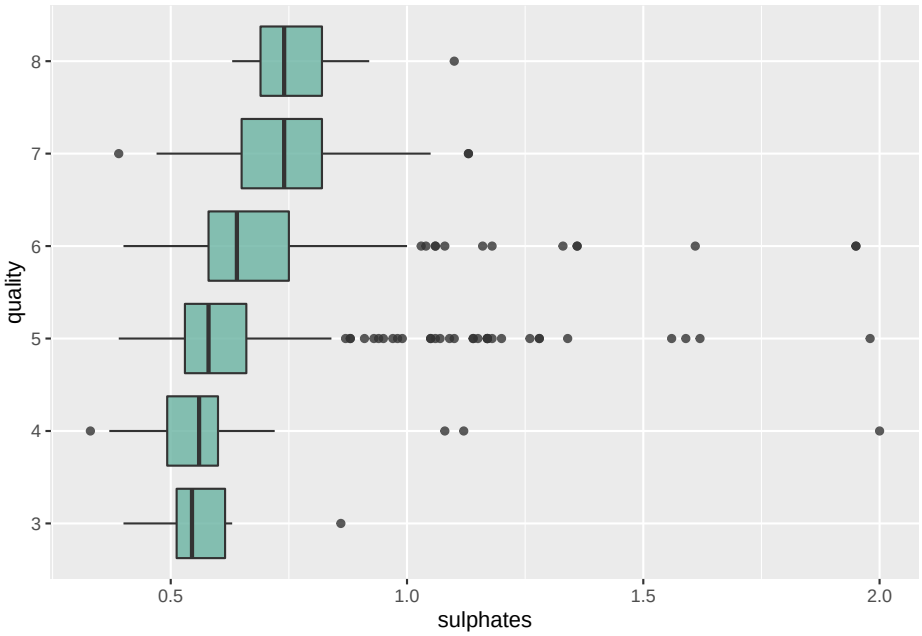
Relates the specific gravity of the wine, or weight / mass. Measured in grams per milliliter (g/mL). Density is used to describe the proportional composition of sugars, alcohol and other dissolved solids.



Less dense samples infer higher quality scores.

Dependent variable 7 - 'sulphates'

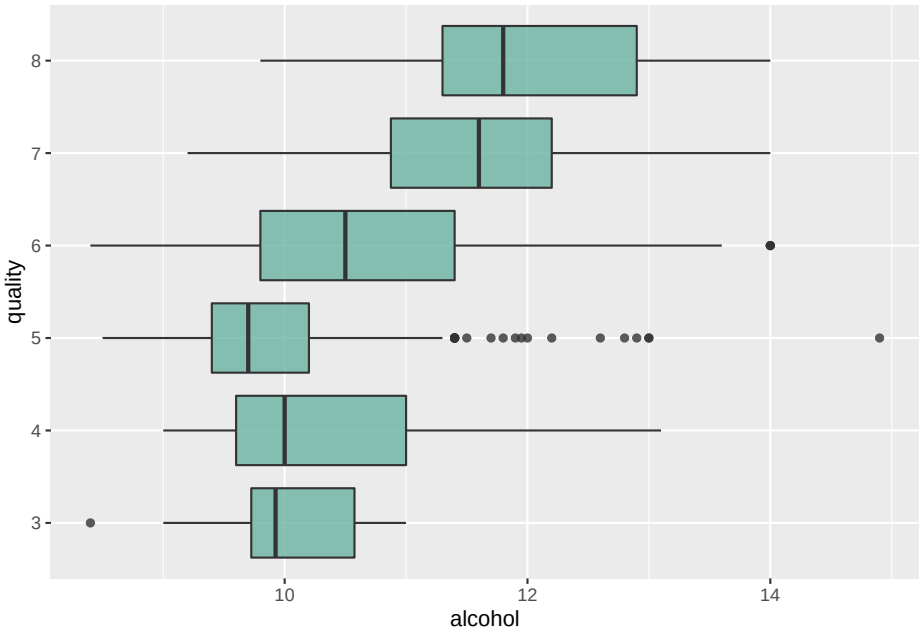
Sulphates are another preservative that protect the wine from over oxidization or bacterial spoilage. Derived from sulphur dioxide, levels are controlled here in Australia due to safety regulations. Negative physiological reactions to wine are often attributed to sulphate levels. Which may be a miss conception, as the more common reaction is due to a histamine response. Sulphate levels reduce over time, so if you're sensitive to this compound go for older wines.



Higher sulphates infer higher quality scores.

Dependent variable 8 - 'alcohol'

Alcohol is a natural by product of the fermentation process.



Higher alcohol infers higher quality scores.

Analysis Summary

From our observations of the numeric data. It can be argued that statistical relationships exist between our outcome (dependent) variable and the selected parameters (independent variables) in training. As such a Random Forest model will be generated due to their strengths in regression and classification tasks.

Method

Model Formulation

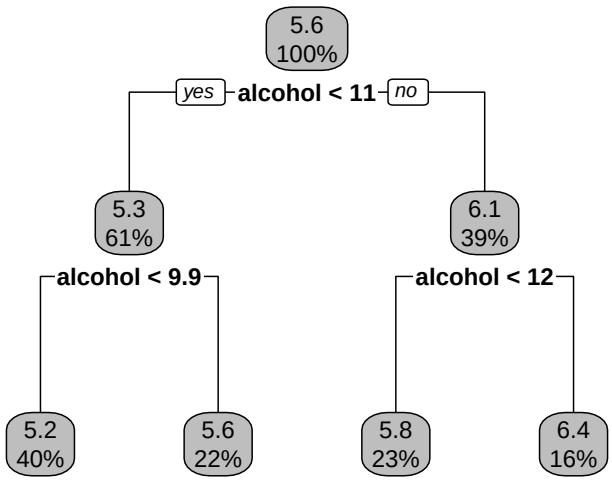
In order to understand random forests, one must first understand the principle of Classification And Regression Trees (CARTs). This is due to the fact that Random Forests form an aggregate value produced by numerous randomly generated CARTs.

Classification And Regression Trees (CARTs)

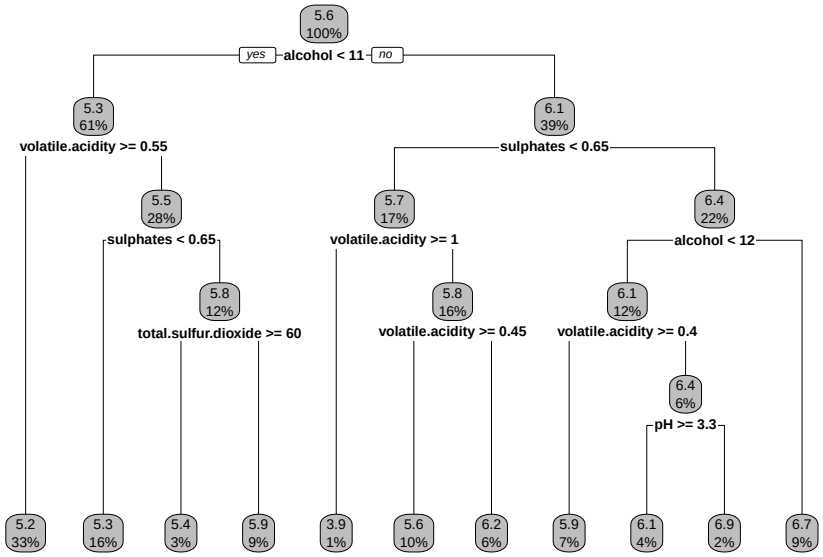
CARTs break apart a given predictor space at a specific value; into sub partitions. Thus, placing all data points into either one or the other partition. The model then measures the 'homogeneity' / similarity of classes within each partition. Adjusting the partition value until achieving the most accurate homogeneity. The above steps then start again with a new partition, only if the resulting homogeneity of class becomes better.

Example Trees

Single variable Tree



Multi variable Tree



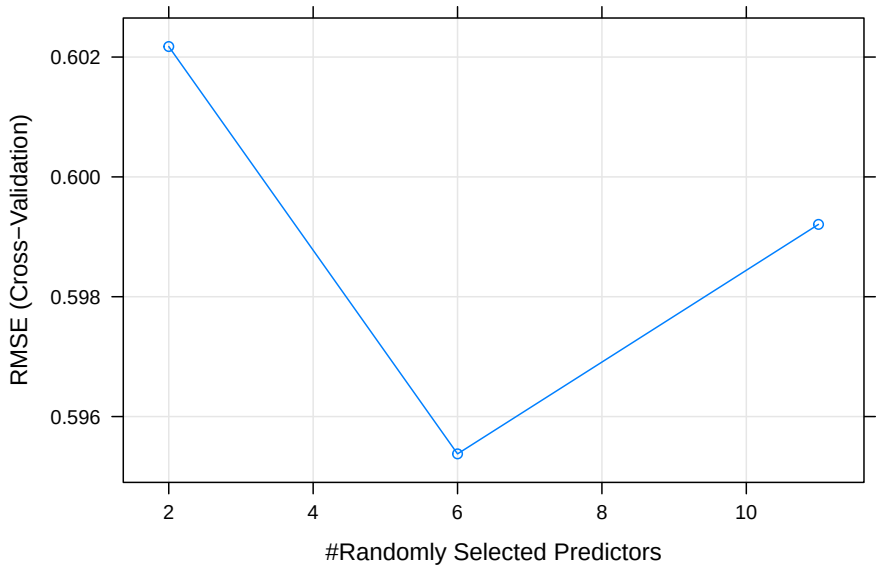
The effectiveness of CARTs is also their disadvantage, as they often create models that are highly accurate, but "overfit" the training data and when shown new information produce greater than acceptable error. Such error relative to training performance is known as High Variance. However, a solution to this variance can be found through randomness. Enter the Random Forests.

Random Forests

Randomly select (with replacement) rows from our data, ensuring the row total is equal to our original data. Then randomly select a subset of predictor variables. Generate CARTs from the selected predictor variables and aggregate results - to choose the highest performing combinations | permutations. Random Forests also provide a means to review variable importance. Meaning which independent variables elicit the greatest performance change in our model.

Model Generation

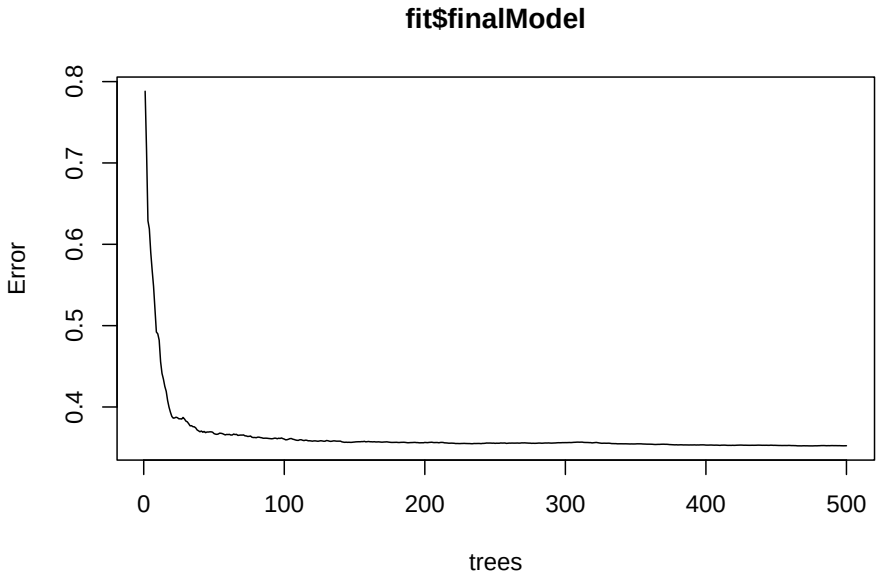
```
#Add a control feature, to utilize cross validation five fold for our training data.

control <- trainControl(method = "cv", number = 5)

fit <- train(quality ~ ., data = train, method = "rf", trControl = control)

plot(fit)
```




```
fit$bestTune$mtry
```

```
## [1] 6
```

```
plot(fit$finalModel)
```

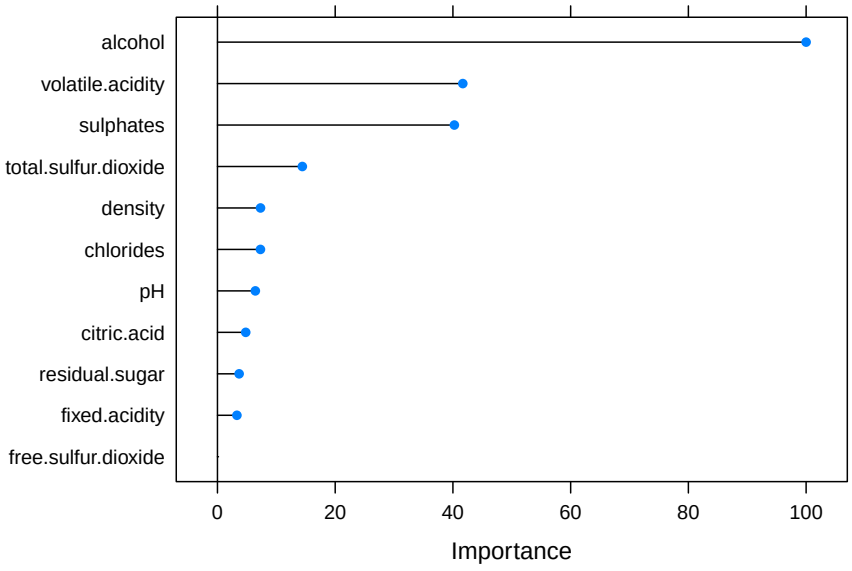


The model stabilizes with the least error after approx. 150 trees. Therefor we will reduce computational effort, setting our maximum number of trees to 150. I.e. ntree = 150. The optimal number of randomly sampled variables for our model is 6. I.e. Mtry = 6.

```
fit <- train(quality ~ ., data = train, method = "rf", trControl = control, ntree = 150)

#Demonstrate the Random Forest (caret) variable importance function.

varImp(fit)
```

```
## rf variable importance
##
##                       Overall
## alcohol               100.000
## volatile.acidity       41.679
## sulphates              40.243
## total.sulfur.dioxide   14.420
## density                 7.316
## chlorides               7.305
## pH                      6.435
## citric.acid             4.804
## residual.sugar          3.682
## fixed.acidity           3.304
## free.sulfur.dioxide     0.000
```

```
plot(varImp(fit))
```




```
#As expected our results are similar to the correlation figure earlier in the project.
```

Testing

```
y_hat <- predict(fit, test[,1:11])

y_hat[1:5]
```

```
##       12       13       23       29       56
## 5.094000 5.299111 5.248444 5.198667 5.107667
```

The resultant RMSE for regression fit is..

```
resultRMSE <- RMSE(y_hat, test$quality)

resultRMSE
```

```
## [1] 0.5338534
```

The resultant classification accuracy (Rounding predictions to the nearest integer).

```
y_hat <- y_hat %>% round(., digits = 0)

resultClassification <- mean(as.numeric(y_hat) == as.numeric(test$quality))

resultClassification
```

```
## [1] 0.7009346
```

Summary

Through data analysis and the generation of a Random Forest Machine Learning model the project has achieved a test RMSE of .53 and an overall accuracy of 70% (correct prediction / total predictions). Thus demonstrating that wine testing data in the form of our independent variables is a reasonable input to produce quality predictions subjective to our testers. To generalize our predictions I'd suggest the 'quality' variable be increased in terms of tester per sample or the number of tests to which the median value is derived, this would reduce bias. Given a large enough data set. Future projects could also generate recommender systems between wine and consumers.

Acknowledgement

Thankyou edX and HarvardX for an incredible professional certificate. This is just the beginning of my journey into Data Science. Your efforts are greatly appreciated.

Referencing

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J., 2009.
  Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47 (4), 547-553.