Capstone: Wine Classification Problem

Alastair Humphrey

18/06/2020

Introduction

The following paper forms an exploration into using a number of machine learning techniques to solve two classification problems. The first classification problem will be to data mine the physicochemical properties of wines to classify the wine into a binary classification of white or red wine. The second classification problem will apply similar methods to the same data set to classify wine preferences as represented by expert pallets. The model performance demonstrates the complexity of modelling no-binary classification problems and the intricacies this adds to data mining processes.

The report is split into a data exploration and pre-processing section, methodologies and outcomes and then finally some conclusions to draw it all together. The output from the different models show that the more complex the task is, given by the number of classes to be predicted, the lower the accuracy of a model. As well as showing that it is quite possible to predict a wine's colour based solely on the physicochemical properties.

Classification problems are of real concern within the data science community as businesses gain increasingly more data and want to exploit its potential for better decision making. With more feedback data from customers, inter-connected technology with industry 4.0 (the internet of things) and evolving health data, being able to quickly categorise the data by class can aide in segmenting complicated data sets to find actionable outputs from the data mining processes.

Data exploration and pre-processing

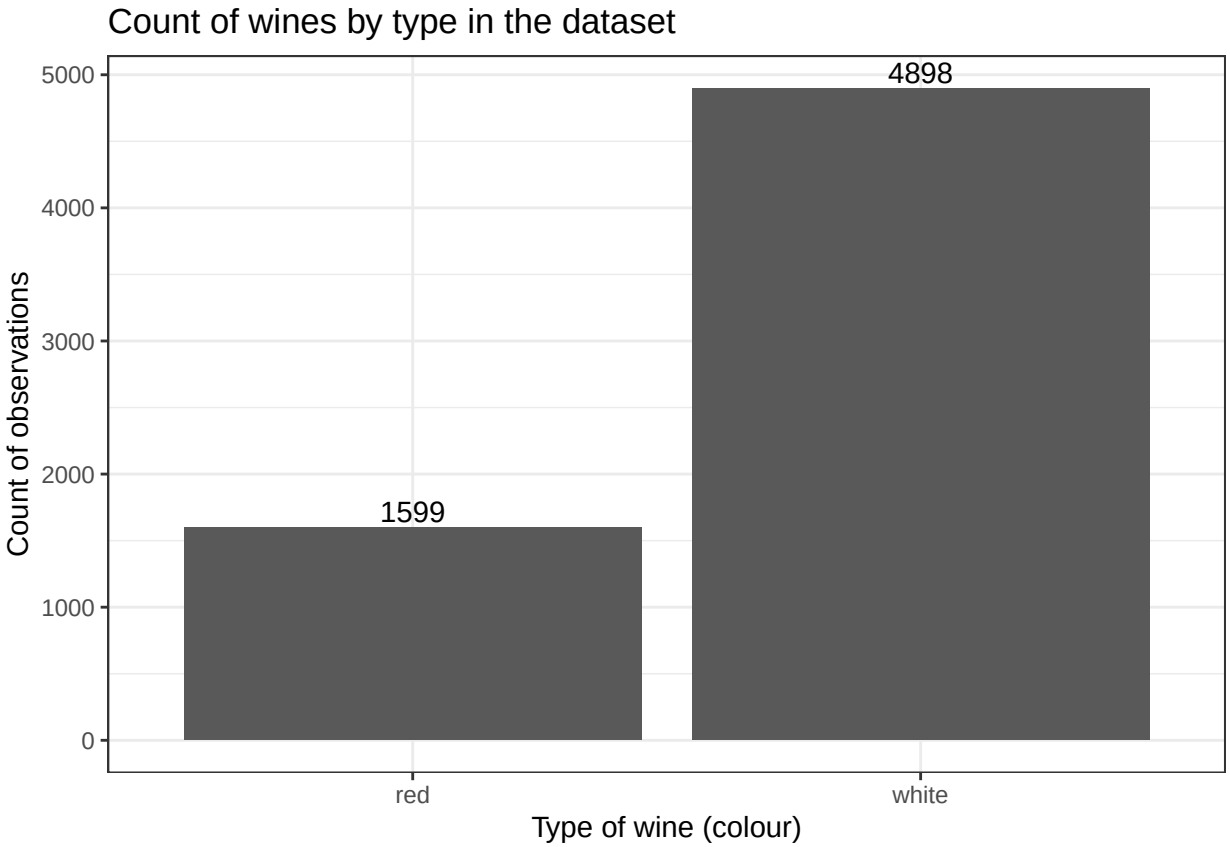
The data set includes 6497 observations of 13 variables. Each observation is a variant of the Portugese Vinho Verde wine. [1]

Column	Class	Notes
type	Categorical discrete	An observation of the colour of the wine.
chlorides	Continuous number	The measurement of the amount of chloride ions in a wine sample. Specifically used to find sodium chloride (salt) content in wine.
density	Continuous number	Wine density is a parameter used to help winemakers control quality and create consistency in their output.

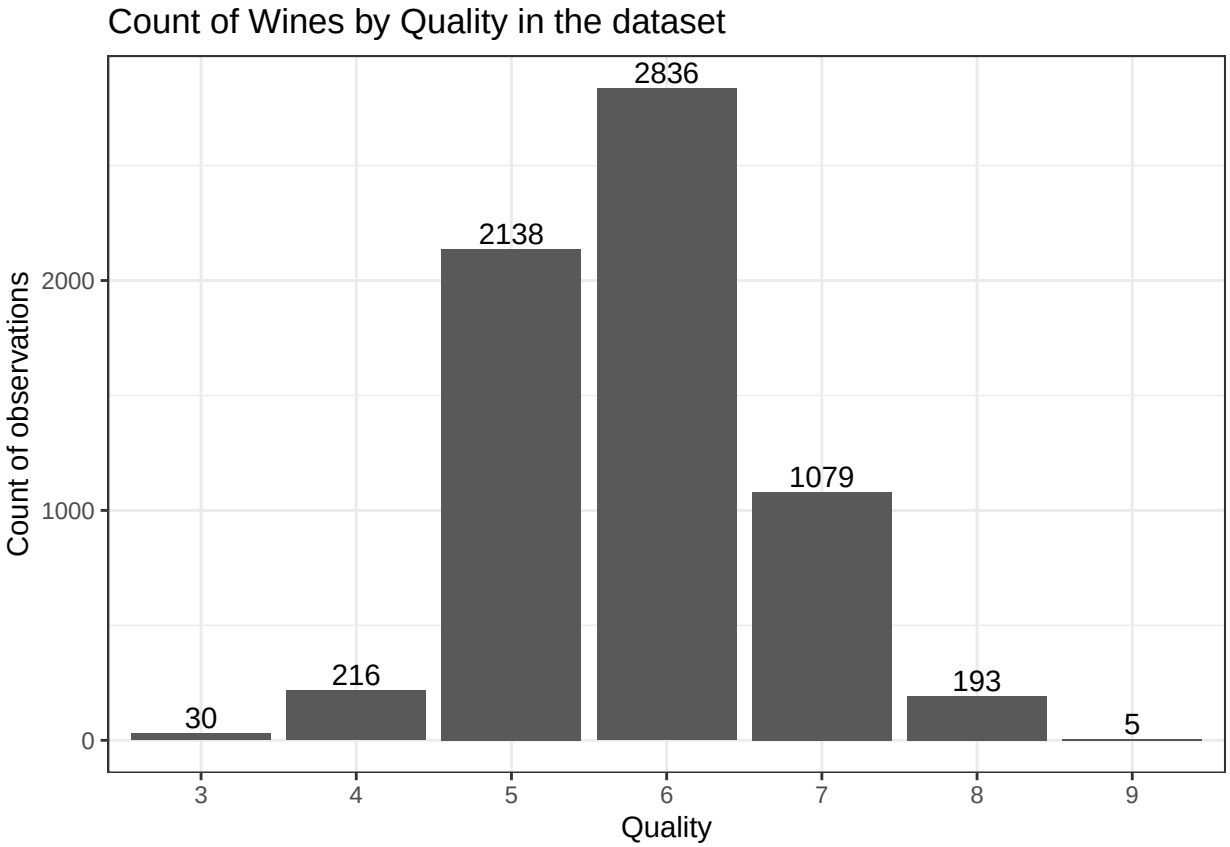
[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Column	Class	Notes
pH	Continuous number	pH is used by winemakers to discern ripeness versus acidity. Acidity in and of itself forms one of the 4 fundamental components of wine and gives it the tart or sour taste.
sulphates	Continuous number	Sulphates (principally sulphur dioxide SO2) are a food additivitie used in wine making to maintain the wine's flavour and freshness by slowing the oxidization process. Sulphates is the measure of the bound SO2 molecules which have become complexed with another compound.
alcohol	Continuous number	Created during the fermentation process by the yeast consuming the sugars to create ethanol, alcohol forms a central point of the four fundamentals of wine.
quality	Categorical discrete	A classification and rating of wines by 1 to 10 by wine experts. For each observation 5 experts' ratings were averaged to give the final quality catergory.
fixed acidity	Continuous number	Fixed acidity is a sub classification of the acids within wine. Predominantly, fixed acids in wine are tartaric, malic, citric, and succinic. All but succinic come from within the grapes rather than from the fermentation process. High acidity is often found in wines produced in cool climates and is currently out of favour with wine critics. These will often be treated with a neutralising agent or malo-lactic fermentation to reduce the sour taste.
volatile acidity	Continuous number	Volatile acidity refers to the steam distillable acids present in wine. Principally acetic acid but also lactic, formic, butyric, and propionic acids. Acetic acid (vinegar) in wine is usually considered a spoilage product although in some reds a small amount can be used to enhance flavour.
citric acid	Continuous number	Citric acid is a naturally occurring acid that can be found in minute quantities naturally in grapes. It is predominantly found in wine after being used as an additive. Often used to increase the acidity of a wine, in Europe it can only be used as a stabilising agent.
residual sugar	Continuous number	Residual sugar is the amount of sugar left over after the fermentation process. This is what gives wine a sweetness or conversely its absence is the driver of dry wines. Some nations (France and Germany) allow additional sugar to be added to wine during different stages of the wine making process.
free sulfur dioxide	Continuous number	Free sulphur dioxide measures the amount of SO2 that can be extracted via a manual aspiration or flow injection analysis. The free SO2 is the unbound compounds capable of exerting an antioxidant preservative effect.
total sulfur dioxide	Continuous number	Total sulfur dioxide is the total of free and bound SO2 within the observation.

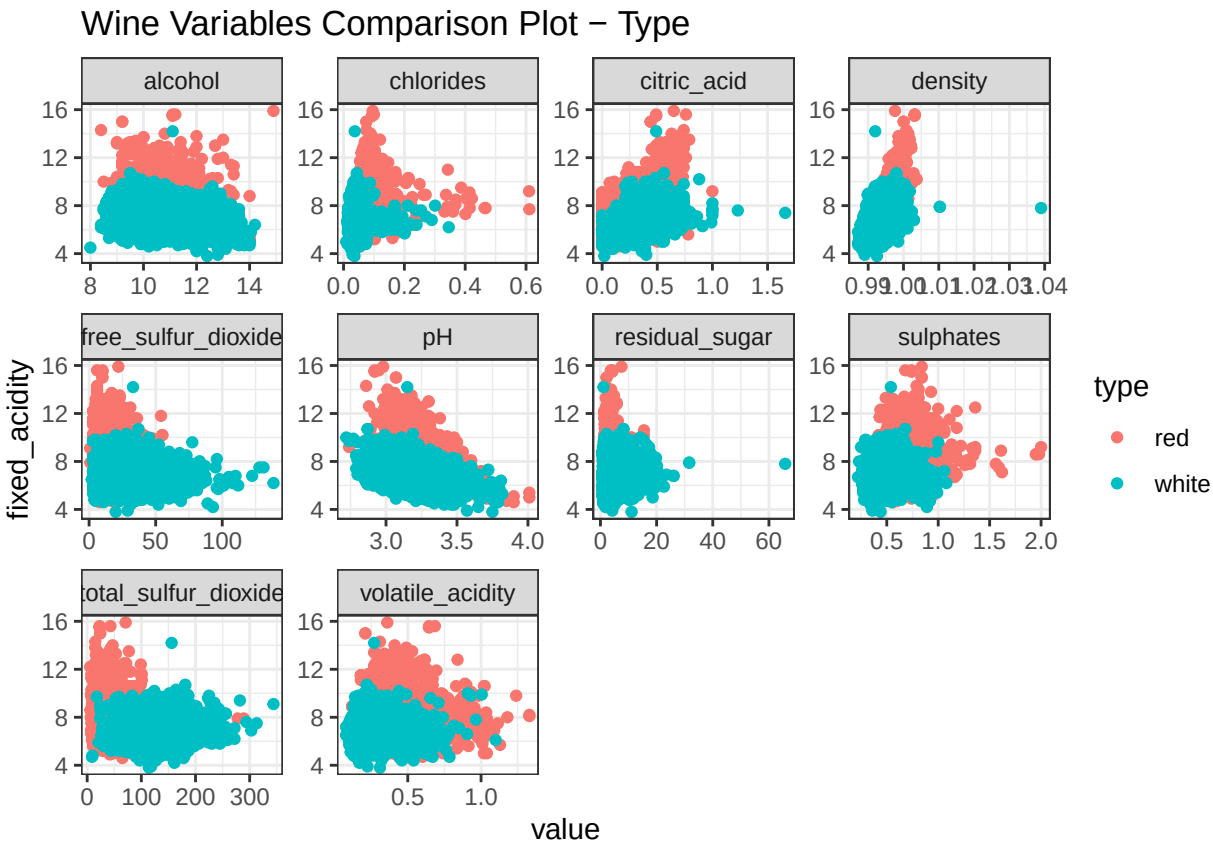
The data set has a large imbalance in the number of red wine versus white wine, with over 3 white wines for every red observation.

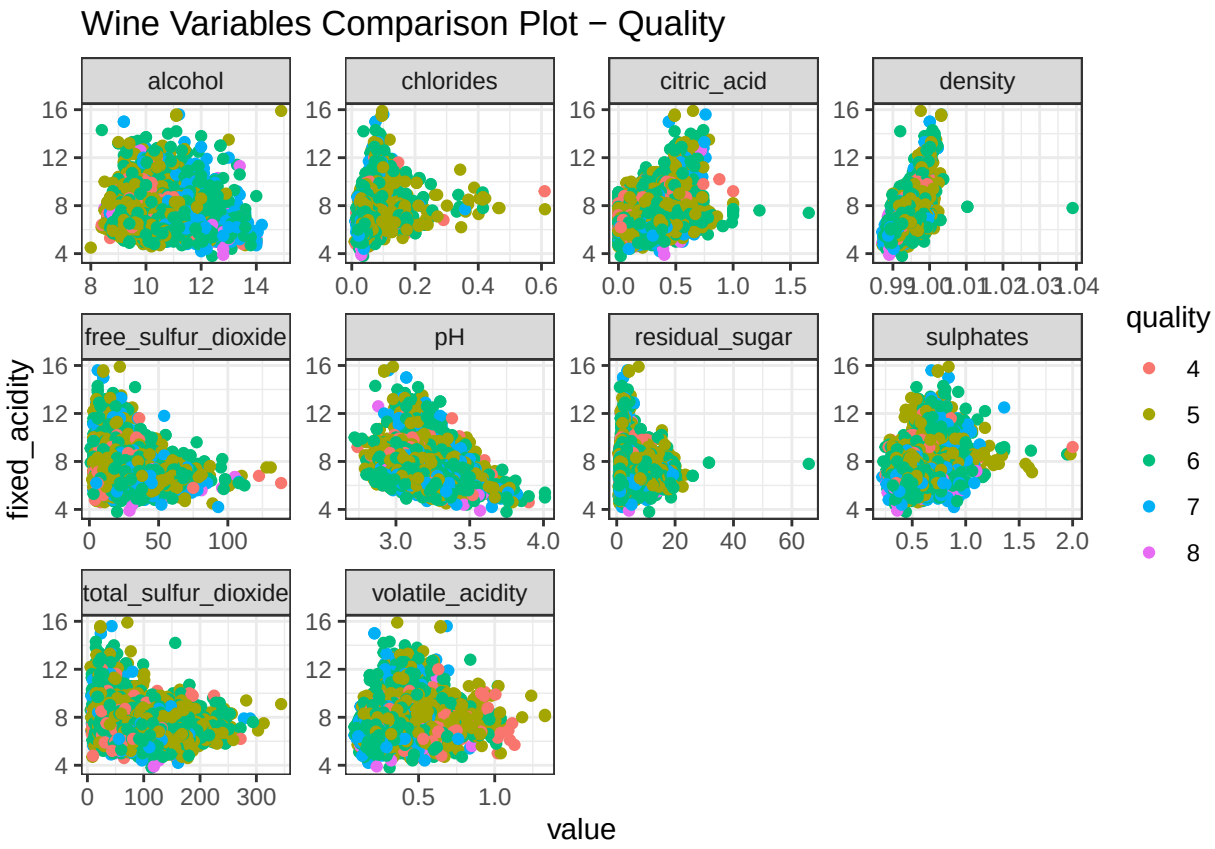


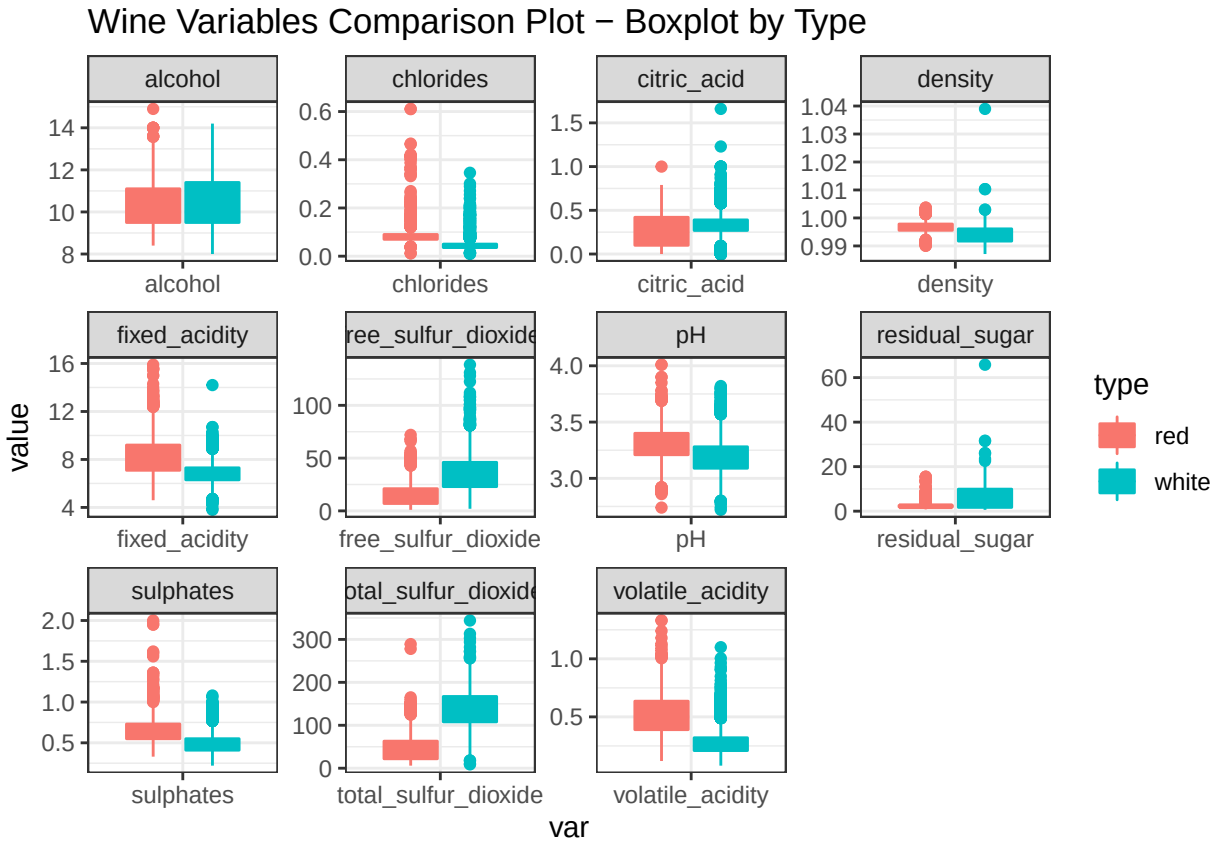
There is a normal distribution of quality ratings, the lowest rating present of 3 and the highest rating of 9 have very few observations within the data set. This is likely caused by the fact that the quality ratings are an average of 5 experts' ratings. To achieve an above or below average rating (5) there would need to be general consensus among the experts, this likely helps to explain why there is a significant bias towards an average rating. It could also be hypothesised that there are generally very few exceptional wines or very poor wines.

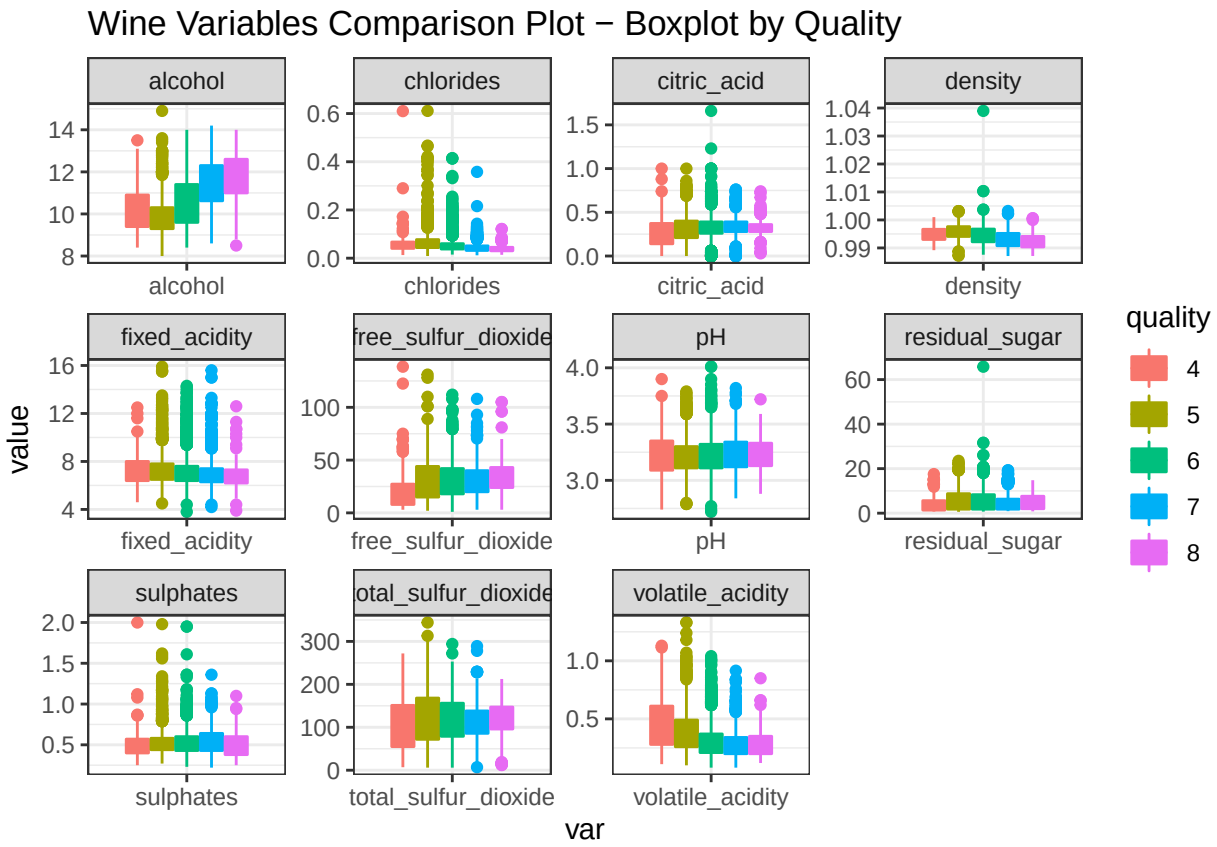


Within the context of the data mining challenge and the two classification problems there are not enough data points for the models being used to predict the quality 3 or 9 wines. As such, these observations are being removed from the data set as part of the pre-processing. Further data exploration can be undertaken to try and understand if there are any clear distinctions within the data set that would assist in classifying either the wines type or its quality. Two plots can aide in this. A plot of each variable against another (fixed acidity) and a bar plot per variable. These can be stratified by both the classification problems.


Wine Variables Comparison Plot − Type
alcohol
chlorides
citric_acid
density
free_sulfur_dioxide
pH
residual_sugar
sulphates
total_sulfur_dioxide
volatile_acidity
fixed_acidity
value
type
red
white


Wine Variables Comparison Plot − Quality
alcohol
chlorides
citric_acid
density
free_sulfur_dioxide
pH
residual_sugar
sulphates
total_sulfur_dioxide
volatile_acidity
fixed_acidity
value
quality
4
5
6
7
8


Wine Variables Comparison Plot − Boxplot by Type
alcohol
chlorides
citric_acid
density
fixed_acidity
free_sulfur_dioxide
pH
residual_sugar
sulphates
total_sulfur_dioxide
volatile_acidity
value
var
type
red
white



The data exploration highlights that there is no clear distinctions between the red and white wine types. However, there appears to be less obvious distinctions between the different qualities of the wine. Due to the visual analysis it can be expected that the data mining techniques will be more effective for the binary classification problem (type) versus the multiple class problem (quality).

Data Partitioning

To aide in understanding the capability of the produced models the data set needs to have a validation data set to test the final model on. Thus the data set is partitioned into a Validation and training data set. All test models will only use the training dataset to ascertain which model produces the best fit as well as to optimise the model. To further aide in this, the training dataset is further split into a test and train data set. Each partition uses a 0.8 train / 0.2 test split.

Methodology

The methodologies used in the capstone project are those described in the HarvardX Data Science course. The details of this can be found in Rafael A. Irizarry's 'Introduction to Data Science'.[2]

[2]Introduction to Data Science, Data Analysis and Prediction Algorithms with R, By Rafael A. Irizarry, SBN 9780367357986, Published November 19, 2019 by Chapman and Hall/CRC

Base notation for machine learning

In a machine learning problem there are two components:

a) an outcome

b) the features that will be used to predict the outcomes

These are denoted as Y for an outcome and $x_1, x_2, x_3, ..., x_p$ for features. In the context of a categorical outcome, the number of categories is denoted as K. In the case of a binary outcome, such as the wines type, then $k = 0, 1$. However, when there are more than two categories of outcome data, then the notation becomes $k = 1, ..., K$.

The data mining problem has an unknown outcome that is wanted to be classified and K features to classify with.

outcome	feature 1	feature 2	feature 3	feature p
?	X_1	X_2	X_3	X_p

To create these predictions, the training data with known outcomes is used:

outcome	feature 1	feature 2	feature 3	feature p
y_1	$x_1, 1$	$x_1, 2$	$x_1, 3$	x_1, p
y_2	$x_2, 1$	$x_2, 2$	$x_2, 3$	x_2, p
...	...	...	...	...
y_n	$x_n, 1$	$x_n, 2$	$x_n, 3$	x_n, p

The classification model will be a decision rule for identifying which of the K classes should be predicted. For the quality of the wine each class of k, $f_k(x_1, x_2, x_3, x_p)$. For binary classification models, such as the wines type will create a desicion rule $(f_1(x_1, x_2, x_3, x_p) > C \rightarrow K_1) \wedge K_2$ with C being a pre-determined cut-off.

Overall Accuracy

To measure the performance of the models the accuracy will be used. Accuracy is simply the proportion of predictions which are correct, $\hat{Y} = y$ over the total number of predictions n. The confusion matrix, specificity, sensitivity, prevalence and ROC will also be used in helping to decide the optimal model. As accuracy is the main measure being used to describe the models, the other measures will not be described. [3]

Basis for the models: Conditional Probabilities

The outputs of the models can be represented as the probability of a class k being the correct class. Thus for quality categories, probability can be denoted as:

$$p(\mathbf{x}) = Pr(Y = k|\mathbf{X} = \mathbf{x}), for\ k = 1, ..., K$$

Where:

[3] For an in detail explanation of each of the model performance measures, please see: Ibid: https://rafalab.github.io/dsbook/introduction-to-machine-learning.html, accessed on: 16/06/2020

$p(\mathbf{x})$ is a conditional probability functions of the predictors.

The prediction of the class for any given $\mathbf{x}$ is the class with the highest probability, denoted as:

$$\hat{Y} = \max_k \hat{p}_k(\mathbf{x})$$

In the instance of a binary categorisation problem such as the wines type problem, the predicition of the class is the average of all the predictors 1,0. In effect, the model is a prediction of the proportion of the 1s at a given $\mathbf{X} = \mathbf{x}$ within the training dataset. This can be denoted as:

$$E(Y|\mathbf{X} = \mathbf{x}) = Pr(Y = 1|\mathbf{X} = \mathbf{x})$$

Models

The following models use the `caret` package due to its efficient coding style.

Linear Regression

For the binary classification wine type problem, a good baseline for model building can be in linear regression. A linear regression model in the context of a classification problem using just a single predictor can be denoted mathematically as:

$$p(x) = Pr(Y = 1|X = x) = \beta_0 + \beta_1 x$$

Based on this linear model, a decision rule can be calculated using $C = 0.5$. Using this model, K_1 is white wine. Therefore, the model becomes:

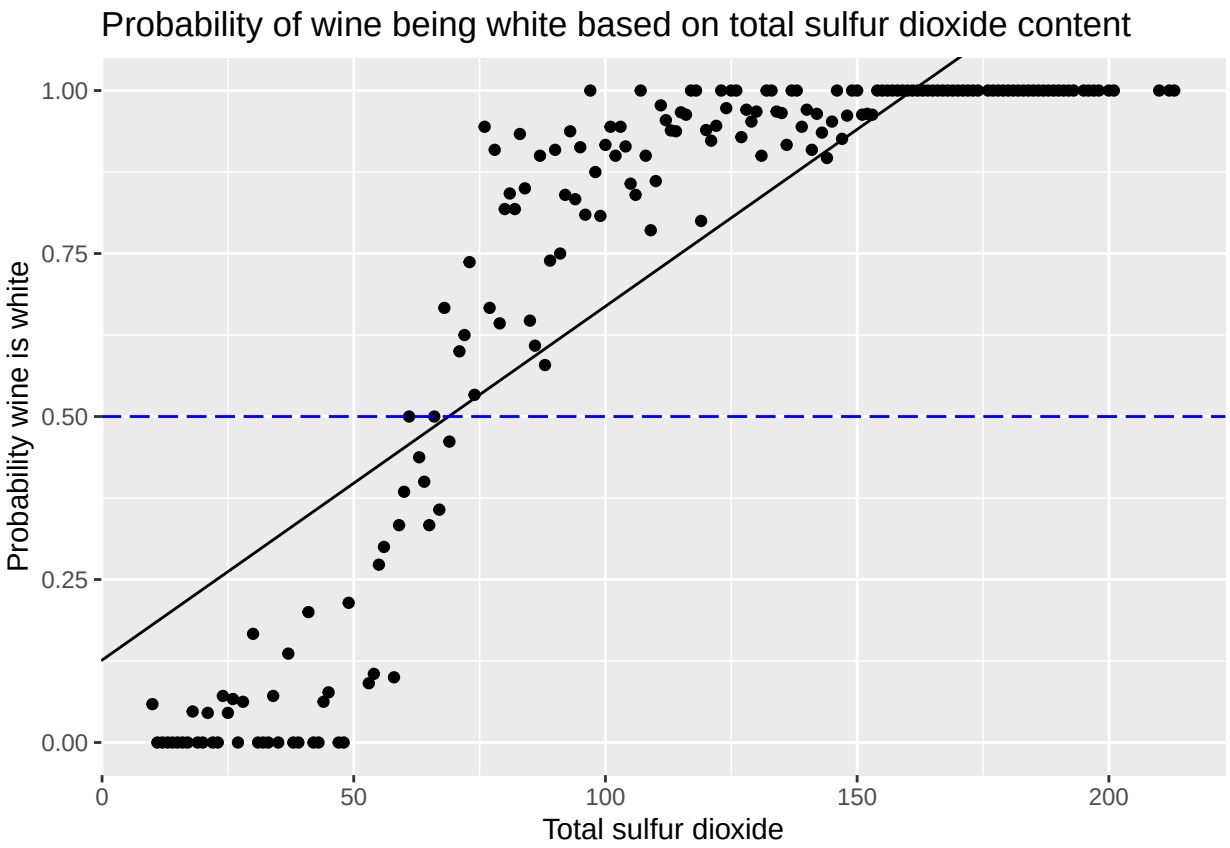
$$\hat{p}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$(\hat{p}(x) > 0.5 \rightarrow white) \wedge red$$

With the 11 predictors, it would be hard to judge which variable could present the best model. Based on this, all 11 predictors were run with a linear model and the output performance can be seen in the below table.

	Variable	Accuracy
Accuracy	fixed acidity	0.8223938
...2	chlorides	0.7731660
...3	density	0.7403475
...4	pH	0.7480695
...5	sulphates	0.7905405
...6	alcohol	0.7548263
...7	volatile_acidity	0.8638996
...8	citric_acid	0.7548263
...9	residual_sugar	0.7548263
...10	free_sulfur_dioxide	0.8030888
...11	total_sulfur_dioxide	0.9237452

Based on this model, the total sulphur dioxide is the best predictor **x**. Using this, a plot of the probability by total sulphur dioxide level can be plotted along with the linear model.



As the plot clearly shows, the base linear model is very efficient at predicting the wine type, achieving a 92% accuracy. However, we can see visually that the probability seems to have a non-linear format. Thus a linear model does not fit this data perfectly. Further to this, the prediction method requires an output between 0 and 1, however the linear model can take a value less than 0 and greater than 1. The blue dashed line shows the $C = 0.5$ decision rule.

Logistic Regression (GLM)

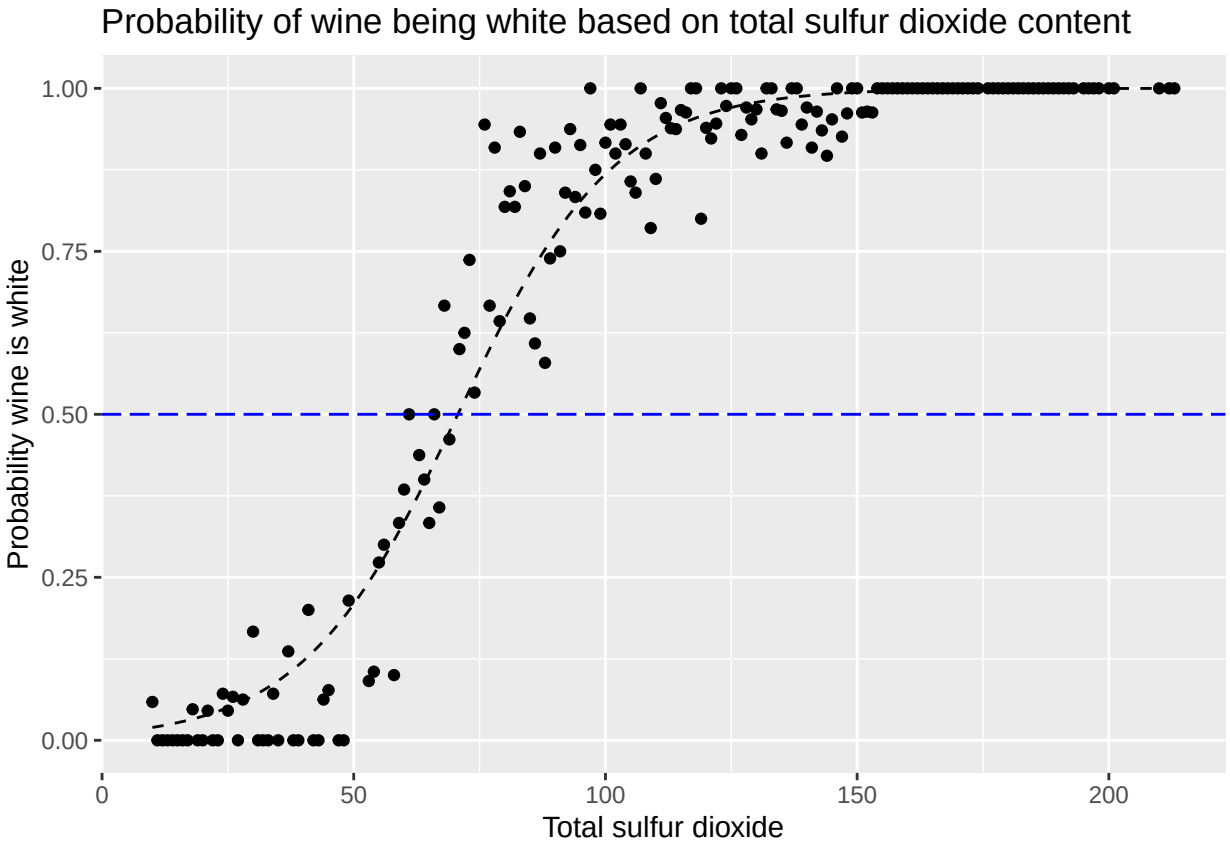
Generalised linear models is an extension of the linear regression model. The model creates a function g that $g(Pr(Y = 1|X = x))$ can be modelled as a linear combination of predictors. Logistic regression is the most common GLM model where the estimate of $Pr(Y = 1|X = x)$ is between 0 and 1. This uses a logistic transformation denoted as:

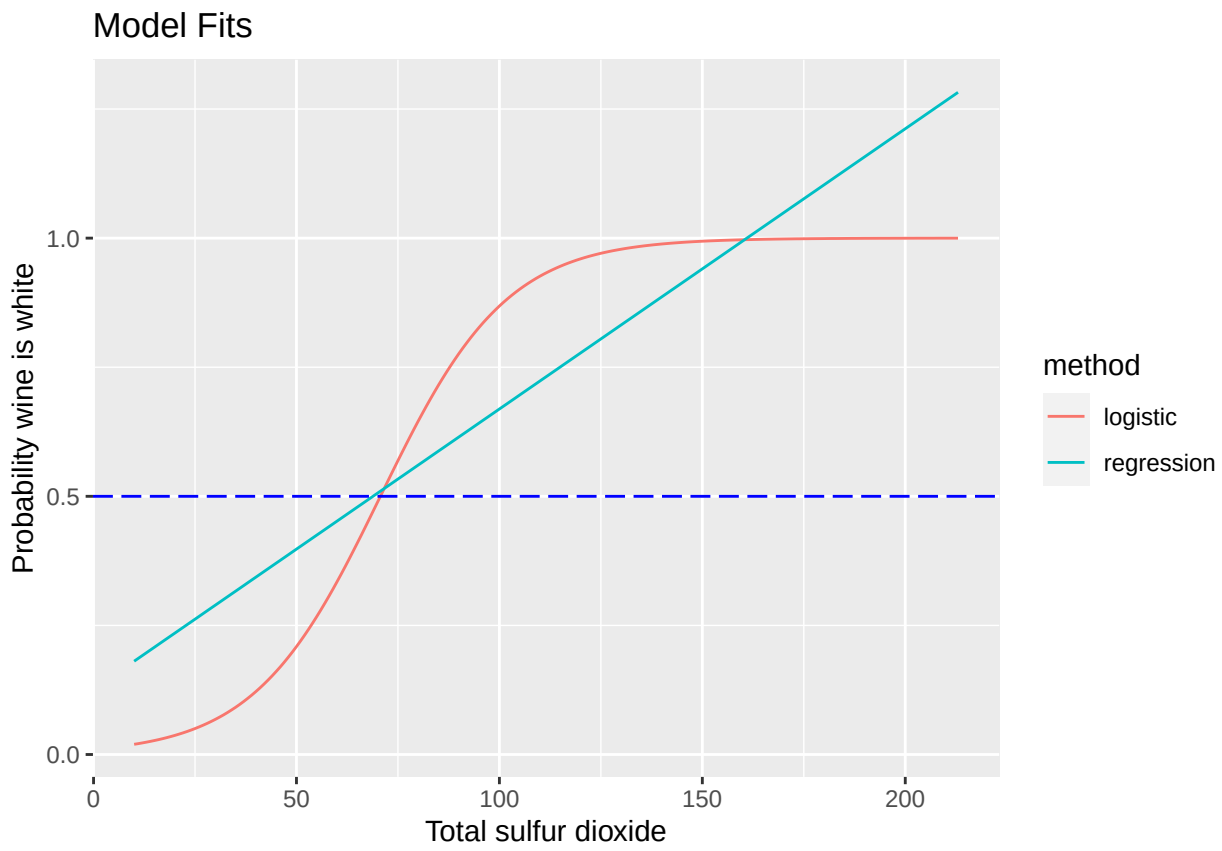
$$g(p) = \log \frac{p}{1-p}$$

Based on this, the logistic regression model for a single predictor, the model becomes:

$$g\{Pr(Y = 1|X = x)\} = \beta_0 + \beta_1 x$$

Using the total sulphur dioxide variable as the single predictor, the following plots illuminate the GLM model as well as both models plotted together.


Probability of wine being white based on total sulfur dioxide content
Probability wine is white
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Total sulfur dioxide



Logistic regression using multiple predictors can be denoted as:

$$g\{x_1, .., x_p)\} = g\{Pr(Y = 1|X_1 = x_, ..., x_p)\} = \beta_0 + \beta_1 x_1, ...\beta_p x_P$$

With $g(p) = \log \frac{p}{1-p}$. Used in the binary categorisation model, this model with all 11 of the variables used as predictors creates a highly accurate model.

K-Nearest Neighbours (KNN)

K nearest neighbours is a smoothing technique that uses the relative distance between observations to group them into clusters. If a two predictors example is used, then for all observations (x_1, x_2) can have their distance measured for all other points in the two dimensional space. The clustering uses the K nearest neighbours to create its probability estimate for $p(x_1, x_2) = Pr(Y = 1|X_1 = x_1, X_2 = x_2)$. In the actual model run, the distance is calculated in p dimensional space between all points, a slightly more complex concept but following the exact same principal.

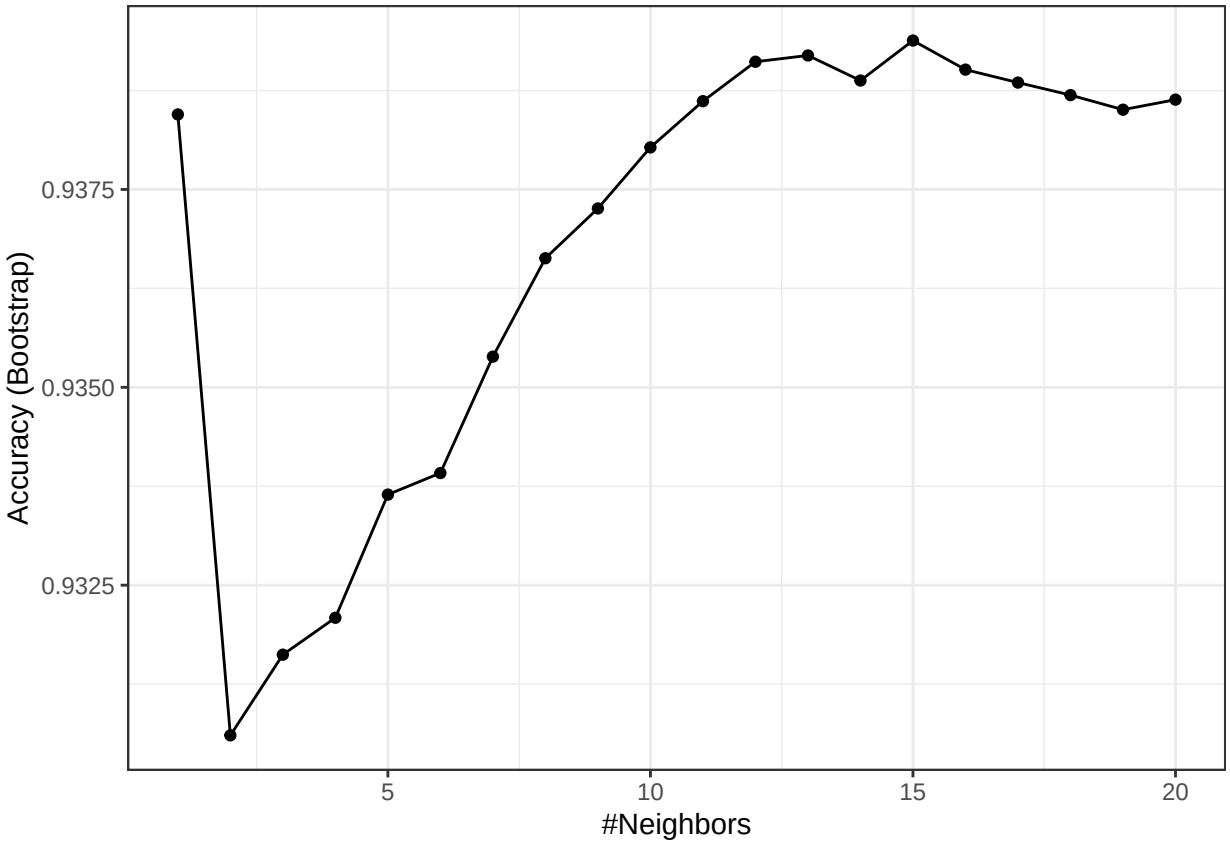
Model cross-validation Picking the number of surrounding points to include in the calculation (k) becomes important. At $k = 1$ the model will just predict based on single data points, likely to lead to a highly over fitted model that will perform poorly on new datasets. When k is large and approaches the total number of observations in the dataset, then there is next to know predicative value in the method. Thus finding the optimal level requires using a technique call cross validation.

In cross-validation, the training data set is split multiple times into test and training datasets. The model is then run against those datasets for a sequence of values of k. Based on all these models the errors of the predictions versus the actual values are calculated. By minimising Mean Squared Error (MSE), the best

value for k can be selected. As the training data set is only a sample of the total data the true error cannot be calculated, thus just the predicted MSE (apparent error) is calculated. This is denoted as:

$$\hat{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

The plot shows the change in accuracy values associated with the different k values. The best value from the cross validation is $k = 11$ for the binary classification problem.



Naive Bayes and Quadratic discriminant analysis

Using Bayes theorem, $p(x)$ can be re-written as:

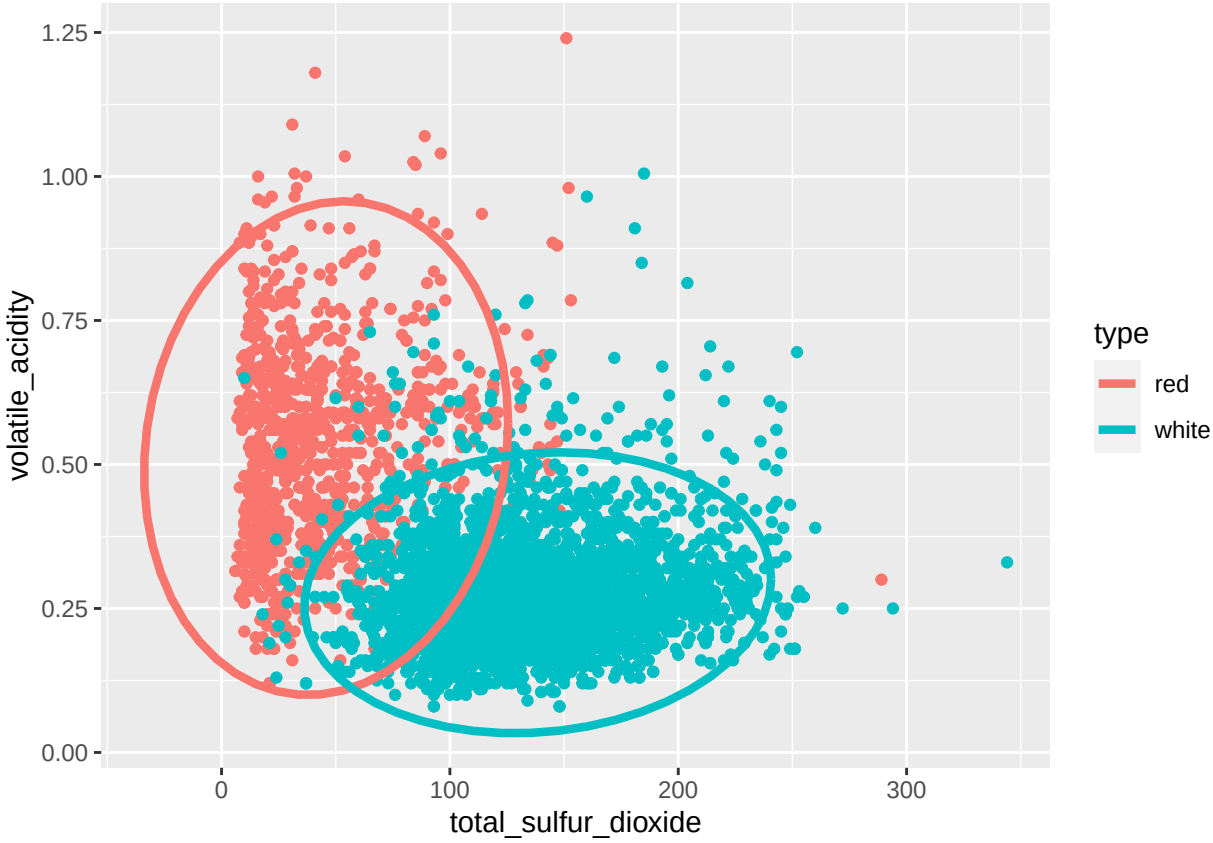
$$p(x) = Pr(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{f_{X|Y=1}(\mathbf{x})Pr(Y = 1)}{f_{X|Y=0}(\mathbf{x})Pr(Y = 0) + f_{X|Y=1}(\mathbf{x})Pr(Y = 1)}$$

With $f_{X|Y=1}$ and $f_{X|Y=0}$ being the distributions of the predictors $\mathbf{X}$ for the classes $Y = 1$ and $Y = 0$. The inference from the reformation is that if the distribution of the predictors can be accurately modelled then a decision rule can be derived that will have a high accuracy for prediction of the two classes. However, the reliance on modelling of the distribution of the predictors is not often possible, particularly when there are a large number of predictors.

Quadratic discriminant analysis is a version of the naive Bayes theorem where the assumed distributions of $p_{\mathbf{x}|Y=1}(x)$ and $p_{\mathbf{x}|Y=1}(\mathbf{X})$ are multivariate normal. Using two predictors, estimates of two averages, two standard deviations and the correlations for each case of $Y = 1$ and $Y = 0$ can be calculated. Using these

figures, distributions $f_{X_1,x_2|Y=1}$ and $f_{X_1,x_2|Y=0}$ can be approximated. A plot of the QDA model can be used on the binary classification problem with the circles describing 95% of the points in each prediction.

type	avg_1	avg_2	sd_1	sd_2	r
red	45.79577	0.5283612	32.52032	0.1748433	0.099008
white	138.45684	0.2772224	41.75033	0.0995770	0.098228



Linear discriminant analysis

As alluded to in the explanation of QDA, having more predictors quickly becomes quite cumbersome as an average, standard deviation is required for each predictor. With more than 2 predictors, the number of calculations becomes quite unmanageable and slow to compute based on the formula $K \times p\frac{(p-1)}{2}$.

Using linear discriminant analysis (LDA) simplifies this by assuming that the correlation structure between all of the classes is the same.

Classification (decision) trees

As has been alluded to previously in this investigation the curse of dimensionality creates problems for data-mining. As previously pointed out there is a conceptual issue when working with p dimensions but it also houses a problem for computation speed. Methods such as regression and KNN do not face this challenge, but LDA and QDA do. In the binary wine type classification problem, there would need to be 110 parameters calculated for the LDA model and 550 for the quality classification problem.

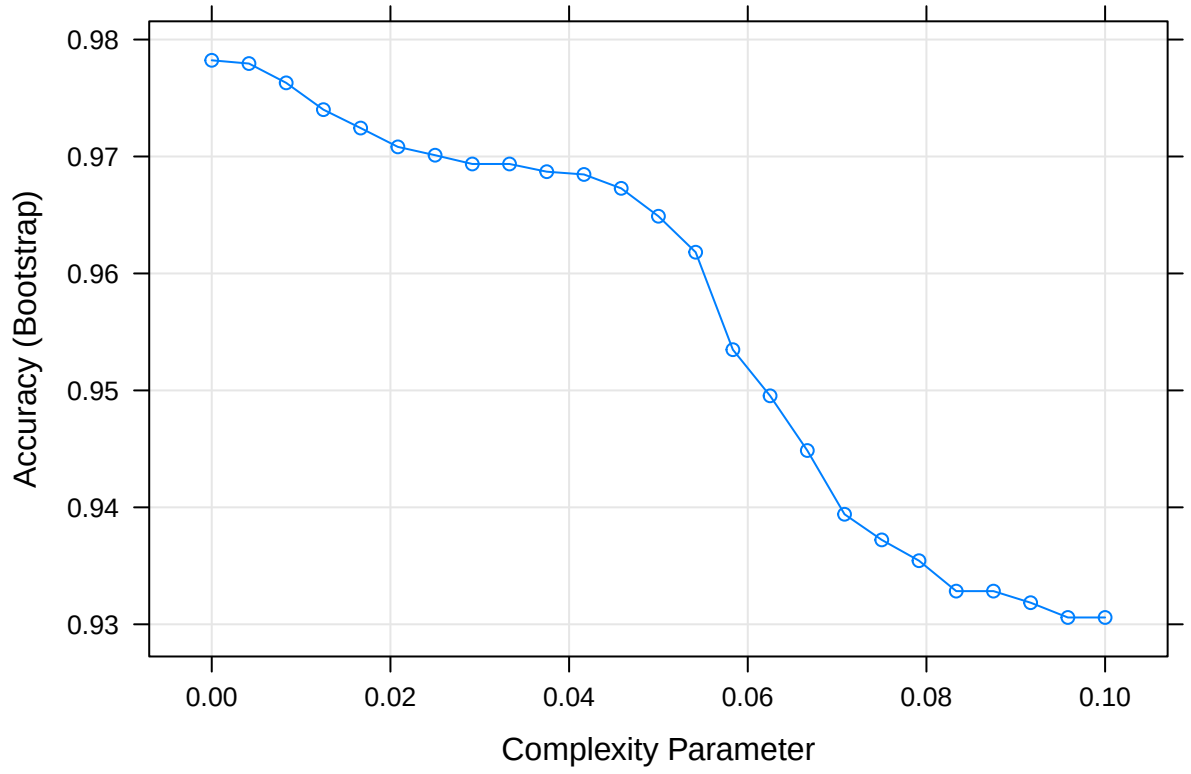
The principal behind a classification tree is to create a set of decisions that leads to a prediction by segmenting the data set, resulting with each node having a prediction $\hat{Y}$. The data is segmented into J non-overlapping regions, $R_1, R_2, R_3, ..., R_J$ and then for each x that falls within R_j region compute the number of each class to form the predicted class for that node.

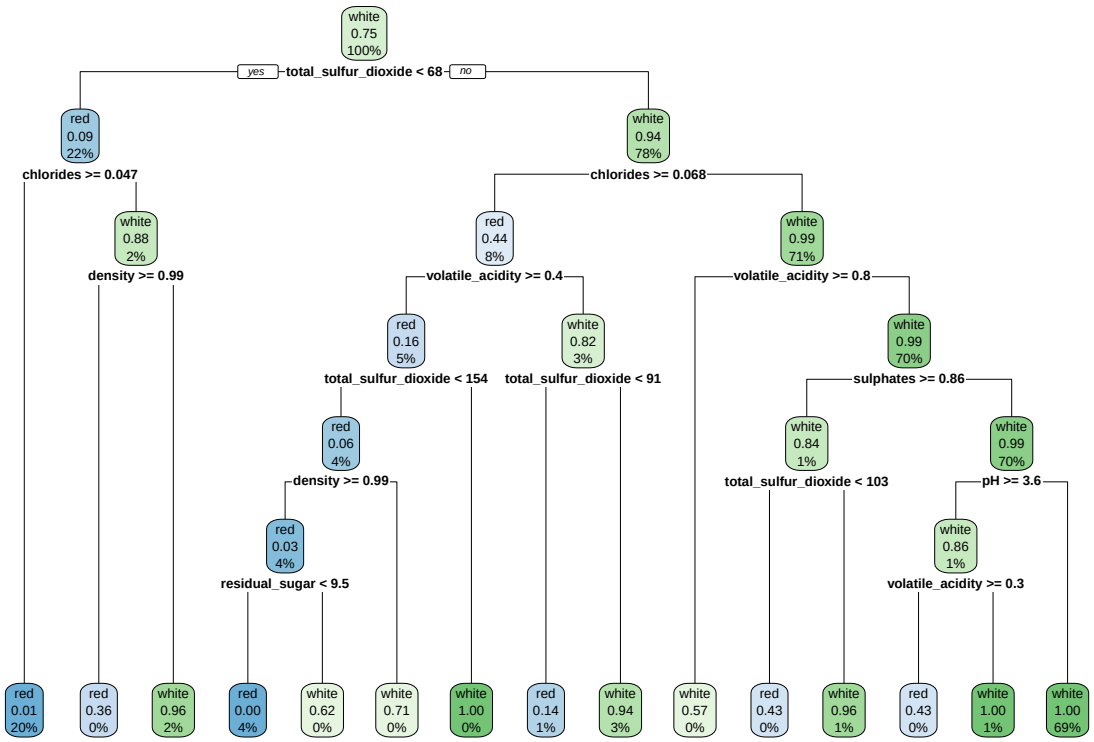
The task thus becomes defining the partitions. If the perfect partitions are a partition filled with only one class, then the partitions would be capable of producing perfectly accurate classifications. The metric to calculate the partition points in this instance is the Gini Index, in the perfect partition the Gini index would be 0, for every deviation from all classes being the same the Gini index increases.

With $\hat{p}_{j,k}$ as the proportion of observations in partition j belonging to class k. Thus:

$$Gini(j) = \sum_{k=1}^{K} \hat{p}_{j,k}(1 - \hat{p}_{j,k})$$

To define when a new partition should be created the concept of complexity parameter (cp) is used. The gini index must reduce by more than a set amount to create a new partition. Within the caret package, the cp can be set as a tuning parameter to find the best cp to maximise the accuracy of the model. This is presented for the binary classification problem in the next graphic with the tuning parameter cp being selected. The second graphic shows the binary classification decision tree model.





Classification trees are extremely visual and as the plot shows, can be very simply interpreted to follow the decision rules. The upsides of the model are offset by its rigidity and potential to over train using a single data set.

Random Forests

Random forests improve the predictive performance and increases stability by averaging multiple decision trees, a proverbial forest of random models. The system for building a random forest is to:

1: Build B decision trees using the training data set. Denoted as $T_1, T_2, ..., T_B$.

2: For each observation in the test set form a prediction $\hat{y}_j$ using T_J

3: Predict $\hat{y}$ with the most frequent class in $\hat{y}_1, ..., \hat{y}_T$

To ensure randomness in each of the different trees, the following steps are used. Where N is the number of observations in the training set and to create T_J, $j = 1, ..., B$:

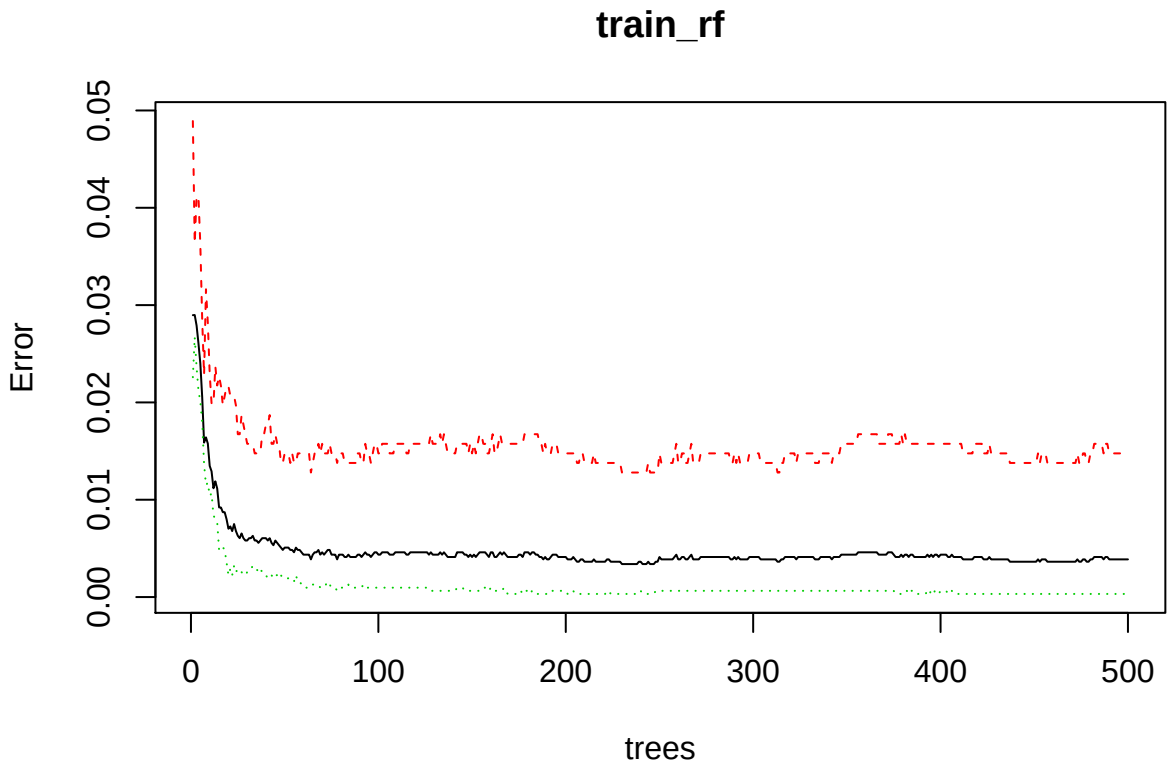
1: Create a bootstrap by sampling N with replacement

2: Randomly selecting features to be included within each forest to reduce the chances of overfitting

The random forest can be plotted to show the improvement in the error rate by adding more trees. It is also possible to improve the models through cross tuning parameters reflecting the smoothness of the model.

Table 4: Binary Classification: Type Models

	Model	Accuracy
Accuracy...1	Linear Regression	0.9237452
Accuracy...2	Logistic Regression - 11 Predictors	0.9903475
Accuracy...3	Knn	0.9276062
Accuracy...4	Quadratic Discriminant Analysis	0.9835907
Accuracy...5	Linear Discriminant Analysis	0.9942085
Accuracy...6	Classification (Decision) Tree	0.9816602
Accuracy...7	Random Forest	0.9942085
Accuracy...8	Random Forest - Tuned Params	0.9932432



Overall Model Performance

Based on the different models run on the test and train set, the best performing model in both classification problems was the random forest model. Based on this, the random forest model will be run on the full training data set against the validation training set.

Table 5: Multiple Class Classification: Quality Models

	Model	Accuracy
Accuracy...1	K Nearest Neighbours	0.5222008
Accuracy...2	Quadratic Discriminant Analysis	0.4430502
Accuracy...3	Linear Discriminant Analysis	0.5337838
Accuracy...4	Classification (Decision) Tree	0.5405405
Accuracy...5	Random Forest	0.6698842

Table 6: Final model results

	Model	Accuracy
Accuracy...1	Random Forest - classification (wine type) problem	0.9922720
Accuracy...2	Random Forest - classification (quality type) problem	0.7047913

Results interpretation

The different models performed very well in the binary classification problem. This wasn't unexpected as the two different types of wine have clearly distinct characteristics that showed through the models to help achieve an extremely high accuracy with the final model predicting the correct wine type 99% of the time.

What is more interesting is that the taste preferences of wine experts in the quality categorisation problem was understandably more difficult to predict. The fact that the model was able to predict quality with a 70% accuracy means that there are some common underlying features to higher quality wines and lower quality wines. This is compounded by the fact that the quality ratings were an average of 5 expert opinions.

Future steps for improving the models

For the binary classification of wine type, it is unlikely that a model could become much more accurate. At present, the tuned random forest model would only mistake 7 wines out of every 1000. For the wines quality analysis, a number of techniques could be attempted to see what impact they may have on the results.

Technique	Notes
Normalise Predictors	Using a transformation, normalise the values in the predictors so that they are closer in distance. This can negate the impact of any underlying bias' in the data set for large values in one of the predictors becoming overweighed. It can also limit the impact of any outlier predictors. Transformations that could be used in this scenario are, among others, a square root transformation or a log transformation.
Seperate the types	For the quality model, separating the wines type into white and red before data mining might make classification of the quality more accurate. As the binary classification task showed, there are clearly distinct elements to a different wine type and logic would dictate a hypothesis that what makes a quality red wine may not be a positive quality in a white wine.
Treat quality as a continious scale	Within this model, the quality of the wine was used as distinct categories. The models all used classification techniques to work out the probability of an observation being of that class. If instead the quality of the wine is treated as a continuous scale then the models can predict the likely rating of a wine. This would create a score which could use the root mean squared error to determine the best fitting model.

Technique	Notes
Ensemble	An ensemble could be used to find an improved result by averaging all of the predictions across all of the models. The `caretEnsemble` package contains the resources to undertake this.

Conclusion

Summary

This capstone project was an exploration into various data mining techniques applied to a data set with two classification problems. One of a binary nature and one with multiple classes. The different techniques used aimed at re-enforcing understanding of the topic and grasping the concept behind the techniques. The actual modelling showed the difference between a binary and multi class problem and the efficacy of the models in those situations. As expected, the larger number of classes K reduces the accuracy of the models due to increasing complexity. Another observation from this work is that for a subjective categorical class, there appears to be a strong bias towards the average grade with very few outliers at either end, thus an approximately normal distribution in wine quality. Based on this, the models performed slightly worse than would be expected if there are consistent standards to wine quality assessment. The inference being that wine preference is less based on quantifiably discernible elements and more upon subjective opinion.

Potential Impact

This work can be used in two ways:

i) This work could be used to inform which components have apparent largest impact on a quantifiably higher quality wine vs a lower quality wine;

ii) The techniques and methods used can be re-deployed to other classification problems - (this concept is explored more in future opportunities)

Limitations of this report

The methods used in this report have been kept to a relatively simple level choosing to follow the caret package. A broad range of other methods have been developed within the machine learning community that may be able to perform far better analysis of the data problem set. The report is also focused on a contextually small data set. With only 6,462 observations this is a simple challenge in comparison to many others such as the 10 Million observations MovieLens data set. The report also has scope to improve the models using the techniques highlighted in the future steps for improving the models section.

Learning outcomes

Given the context of the capstone project, this report has given the learner a great opportunity to embed some of the techniques covered in the HarvardX data science course. Many of the models have application in broad fields and can move data science work forwards at a strong pace. Applications being used in a professional setting by the learner include using smoothing models to better predict processing times during gamma irradiation processes for sterilisation processes, customer segmentation based on the data rather than a subjective category as well as predicting when a customer is leaving the business for a substitute.