ST563 Final Project

Final Project Group 6

20 April, 2021

Contents

Introduction

There has been a push towards the implementation of different analytic solutions within the agriculture industry, especially in the realms of crop health and yield. These sort of analyses can help farmers reduce waste and improve profits. In particular, the viticulture industry stands to benefit from such methods, being that the production of wine is lengthy and multifaceted. Winemakers can work to optimize revenue by analyzing their input and output (crops and wines). Specifically, costs may be cut out if machine learning methods were able to predict the quality of wine based on different metrics, since wine producers would no longer have to hire expert wine tasters to determine wine quality. The main question the study looks to answer is can the quality of wine be predicted using a statistical model instead of taste testers.

The main objective is to model and predict the quality of wine using a variety of statistical learning techniques. This will be done by performing both regression and classification methods, with regression ranking the quality of the wines on a scale of zero to ten, with zero being the lowest quality and ten being the highest quality. The classification models will be classifying the wines into two categories, high and low, with wines in the low category having a score of zero to five, and wines in the high category having a quality score between six and ten. The assessment of the prediction accuracy will be performed by the following methods:

Table 1: ML Methods

Classification	**Regression**
Classification Tree	Multiple Linear Regression (MLR)
Random Forest	Random Forest

The research question of interest for this study is whether or not machine learning methods can accurately predict the quality of a wine, independent of using taste testers. In the classification setting, the team hypothesizes that the machine learning methods can correctly classify wines as "high" or "low" quality with 80% accuracy. In the context of regression, the team hypothesizes that the regression techniques used will be able to achieve a low root mean squared error. In both settings, though, the team expects the random forest to outperform the other methods.

Required Libraries

To run the code for the project, the following libraries are required:

- `tidyverse`: for all the data reading and wrangling
- `knitr`: for rmarkdown table outputs
- `kableExtra`: for making fancy tables
- `rmarkdown`: for output documents
- `corrplot`: for correlation structure visualization
- `leaps`: for best subsets selection

- `tree`: fitting the basic classification tree
- `randomForest`: fitting random forest models
- `gridExtra`: for stitching plots together
- `jtools`: for linear model summary output

Data

The data set used is the Wine Quality data set from the UCI Machine Learning Repository. The data set is comprised of 1600 observations of different variants of the Portuguese "Vinho Verde" red wine. For each wine, various physical and chemical features of each wine were measured, including a measurement of quality given by an expert wine taster. The objective is to use the different features to predict the quality of the wine using multiple machine learning techniques.

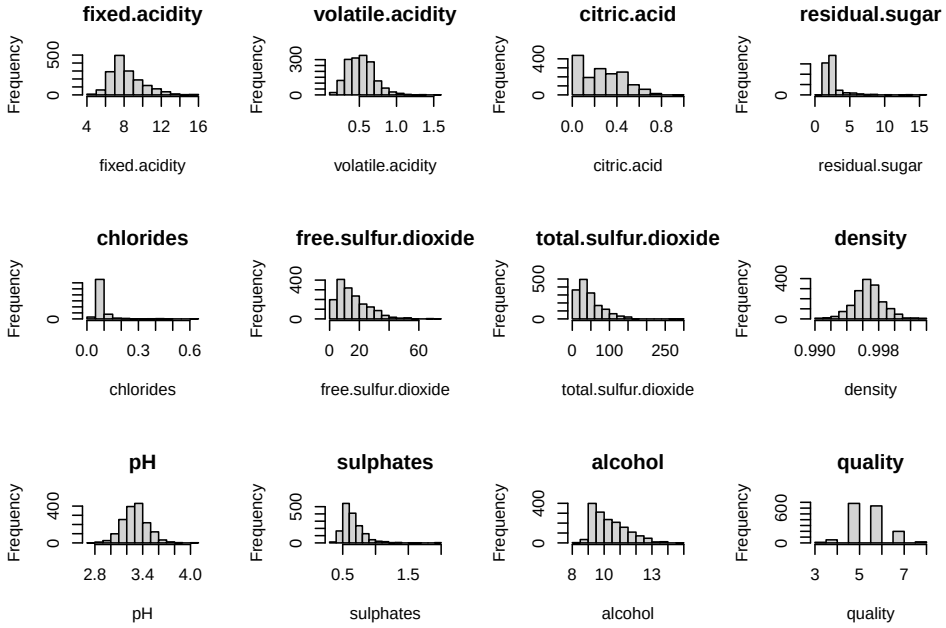
Variable Descriptions

Each entry in the data set represents the different metrics of the following attributes of a single type of wine.

- **fixed acidity**: the quantity of fixed acids found in the wines. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic, all of which come from the grapes except for the succinic acid, which comes from the yeast in fermentation.
- **volatile acidity**: the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. These acids generally come from the fermentation process.
- **citric acid**: added to the wine as a natural preservative or for acidity and tartness.
- **residual sugar**: the quantity of sugar left in the wine after the fermentation process.
- **chlorides**: the amount of salt in a wine.
- **free sulfur dioxide**: the amount of SO_2 that is not bound to other molecules.
- **total sulfur dioxide**: total amount of SO_2 in the wine. Sulfur Dioxide is used throughout all stages of the winemaking process to prevent oxidation and bacteria growth.
- **density**: density of the wine.
- **pH**: pH of the wine.
- **sulphates**: quantity of sulphates in the wine. Sulphates are used as a preservative.
- **alcohol**: alcohol content by volume.
- **quality**: score of quality of the wine given by expert tasters (score between 0 and 10).

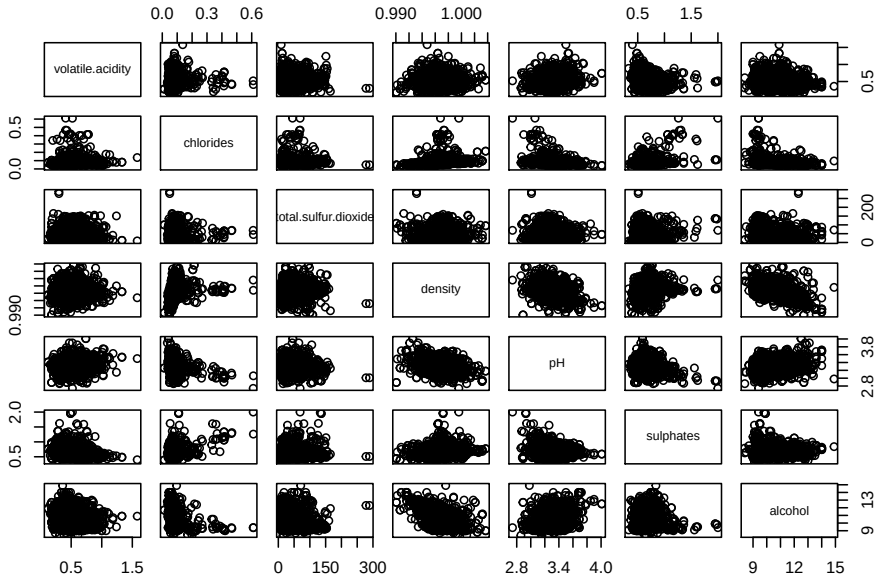
Data Exploration

After reading the data, the first step was to create and view histograms of all of the variables. In doing so, one can get a feel for the distributions of the individual variables.

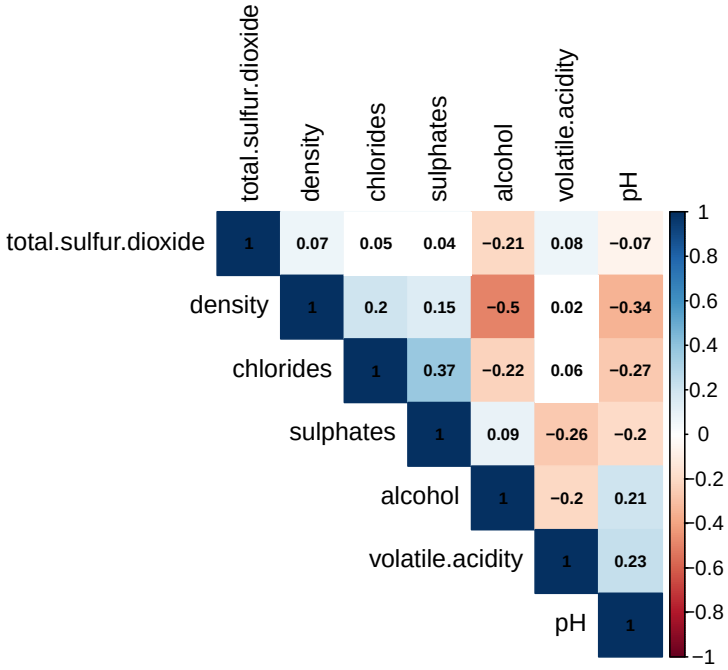


Most wines are classified as moderate wines, with less wines falling in the excellent and poor categories. The density and pH of the wine seem to be normally distributed, while the sulfur, acid, and sulphate content appear skewed to the right. The density varies very little; looking at the scale on the distribution, one can see that it is approximately normally distributed around 0.997 with a standard deviation of about 0.002. The distribution of alcohol content came as no surprise; an average bottle of wine will range from about 10-12% alcohol by volume, with some exceptions. Beyond 16%, the wine is then considered fortified wine, and red wines don't usually have alcohol content below about 8%.

Next, pairwise scatterplots of the variables were created and evaluated. This method allows one to obtain a greater understanding of how the variables relate to one another, as well as if there is any correlation structure. If the variables are related, there may be problems moving forward, so it is important to know these relationships before the analysis is performed.



Based on the pairwise scatterplots, one may conclude that pH may be linearly correlated with density, sulphates, and alcohol. This shows that there may be some relationship between the chemical properties of the wines. There also appears to be a relationship between density and alcohol. The following correlation plot gives a numeric summary of the linear relationship between the variables.



The only linear correlations that are notably high are between alcohol and density. This trend is logical because the density of alcohol is slightly less than water, meaning that as the concentration of alcohol gets higher, the density will naturally decrease. Similarly, the

sulphates and chlorides seem to have a slightly positive linear relationship, while pH and density have a slightly negative linear relationship.

Good Wine Qualities v. Bad Wine Qualities

After a split in the data was performed to pull out the wines classified as high quality and the wines classified as low quality, further analysis was performed on each group to see if any of the variables were significantly different between the two groups.

Table 2: Summary Output for Good Wines

	Min.	Q1	Median	Mean	Q3	Max.
fixed.acidity	4.70000	7.100000	8.0000	8.4740351	9.650000	15.60000
volatile.acidity	0.12000	0.350000	0.4600	0.4741462	0.580000	1.04000
citric.acid	0.00000	0.115000	0.3100	0.2998830	0.460000	0.78000
residual.sugar	0.90000	1.900000	2.2000	2.5359649	2.600000	15.40000
chlorides	0.01200	0.067000	0.0770	0.0826608	0.087500	0.41500
free.sulfur.dioxide	1.00000	7.000000	13.0000	15.2725146	20.500000	72.00000
total.sulfur.dioxide	6.00000	20.000000	33.0000	39.3520468	50.000000	289.00000
density	0.99007	0.995185	0.9964	0.9964666	0.997685	1.00369
pH	2.86000	3.210000	3.3100	3.3106433	3.400000	4.01000
sulphates	0.39000	0.590000	0.6600	0.6926199	0.770000	1.95000
alcohol	8.40000	10.000000	10.8000	10.8550292	11.700000	14.00000
quality	0.00000	0.000000	0.0000	0.0000000	0.000000	0.00000

Table 3: Summary Output for Bad Wines

	Min.	Q1	Median	Mean	Q3	Max.
fixed.acidity	4.60000	7.10000	7.800000	8.1422043	8.9000	15.90000
volatile.acidity	0.18000	0.46000	0.590000	0.5895027	0.6800	1.58000
citric.acid	0.00000	0.08000	0.220000	0.2377554	0.3600	1.00000
residual.sugar	1.20000	1.90000	2.200000	2.5420699	2.6000	15.50000
chlorides	0.03900	0.07400	0.081000	0.0929892	0.0940	0.61100
free.sulfur.dioxide	3.00000	8.00000	14.000000	16.5672043	23.0000	68.00000
total.sulfur.dioxide	6.00000	23.75000	45.000000	54.6451613	78.0000	155.00000
density	0.99256	0.99612	0.996935	0.9970685	0.9979	1.00315
pH	2.74000	3.20000	3.310000	3.3116532	3.4000	3.90000
sulphates	0.33000	0.52000	0.580000	0.6185349	0.6500	2.00000
alcohol	8.40000	9.40000	9.700000	9.9264785	10.3000	14.90000
quality	0.00000	0.00000	0.000000	0.0000000	0.0000	0.00000

The two groups were similar in most of the variables. A notable difference is that the wines classified as high quality had a lower mean value, 39, compared to that of low quality wines,

54.6, for total sulfur dioxide. Another dissimilarity is the difference in the mean for alcohol. The high quality wines had a mean alcohol content at 10.8, while the low quality wines had a lower mean alcohol content around 9.9. Differences also existed in fixed acidity, with the lower quality wines having a lower mean. There were also several variables that do not seem to change between the two groups including pH, residual sugar, chlorides, and density.

Data Processing

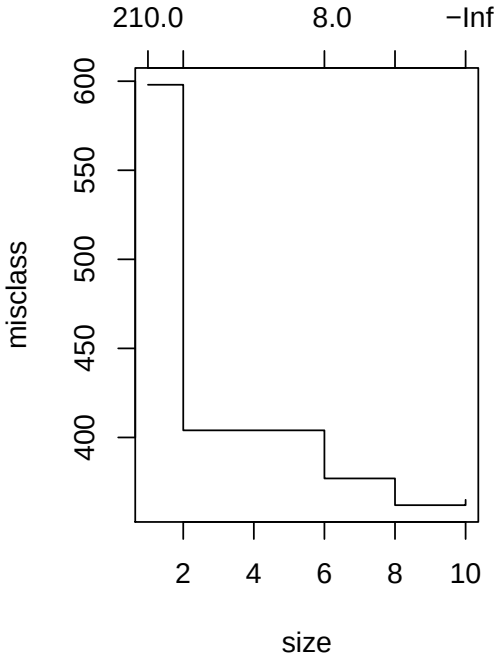
Fortunately, the data set was very tidy; there were no missing values. The data was first split into a training and test data set. The training data set uses 80% of the observations and the test set contains the remaining 20%. The models were built on the training set and evaluated on the test set. This is important so that the models are not overfit. A variable was then created for classification, transforming the quality variable into a binary "low" or "high" response as mentioned in the Introduction.

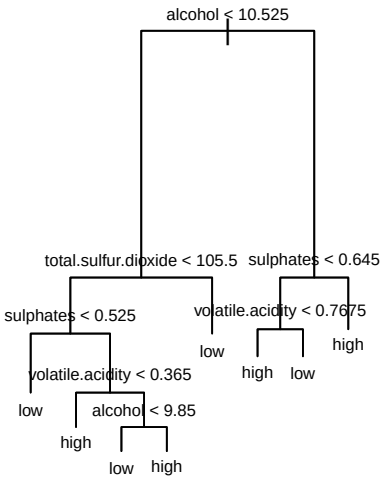
Classification Models

Classification Tree

The basic classification tree is based on partitioning the data into subgroups using simple binary splitting. Initially, all objects are considered a single group. Then, the group is repeatedly split into two subgroups based on the criteria of a certain variable. In the classification setting, an observation is classified in a specific region with majority vote. Trees are grown as big as possible and then pruned back using cost-complexity pruning. This is done to ensure the data is not being overfit. The trade-off here is that pruning increases the bias.

Pruning is performed on the tree using cross validation. A tree is built many times with a different number of splits and the tree that produces the smallest error rate is selected. Cross validation on this model shows that 8 is the optimal number of splits to use, as compared to the first tree which produced 10 splits. The resulting pruned model will now have 8 terminal nodes and uses four variables: Alcohol, Total Sulfur Dioxide, Sulphates, and Volatile Acidity.






```
## [1] 0.2447224 0.3000000
```

The pruned tree produces a test error rate smaller than the original tree with a misclassification error of 0.3. The advantage of using a basic tree is that it is easy to understand and has a good interpretability, which is not something that translates over to ensemble methods. Additionally, the classification tree is not very computationally expensive, unlike ensemble classification methods. That being said, using ensemble methods like the random forest can often improve the error rate of a simple classification tree, since it is an average of many classification trees. While the test MSE was small in the classification tree, a random forest model was built next, in hopes that it may improve the accuracy of the predictions.

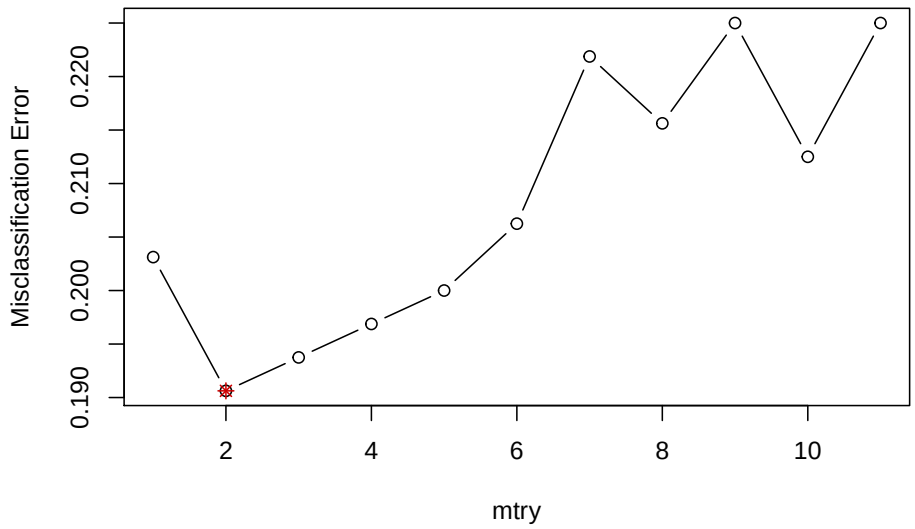
Random Forest for Classification

The random forest model is an ensemble tree-based method, which creates multiple trees from bootstrap samples and averages the results from all the fitted trees. Many bootstrap samples are created with replacement and then a classification tree is fitted on each bootstrap sample with a random subset of the predictors. Once a prediction has been made by the bootstrapped trees, the final classification is based on majority vote of the bootstrap predictions. Similarly, the prediction in the regression context is an average of the predictions of the bootstrapped trees.

The advantage for using an ensemble method over a regular classification or regression tree is that because there are many bootstrap samples being averaged together, there is less variance. This is similar to how the variance of the sample mean goes down as the sample size increases. Although it will probably increase our prediction accuracy, the random forest

loses the interpretability that the basic trees have. Furthermore, the algorithm that is used to fit the random forest is very computationally expensive.

The random forest model requires tuning of the number of randomly selected features at each split using cross validation. Here the model is built many times with the number of predictors included from one to all predictors. The test MSE is evaluated for each and the smallest value is selected. A plot of mtry as a function of the misclassification error is included below.



From cross validation, the model that produced the smallest error rate was the model built with 2 randomly-chosen variables at each split. This model produced a test error of 0.197, which is significantly lower than the test error for the basic classification tree (0.3). The random forest was selected as the best method for predicting if a wine were to fall into the "high" or "low" quality designation.

Regression

Multiple Linear Regression

Multiple Linear Regression is the extension of simple linear regression to a set of multiple predictor variables. The goal of multiple linear regression is to model the linear relationship between a continuous response and two or more predictor variables. Like simple linear regression, coefficient estimates are found by minimizing the sum of the squared errors. The formula for a multiple linear regression model with p predictor variables is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i,$$

where observation $i = 1, ..., n$

Y_i = dependent variable (response)

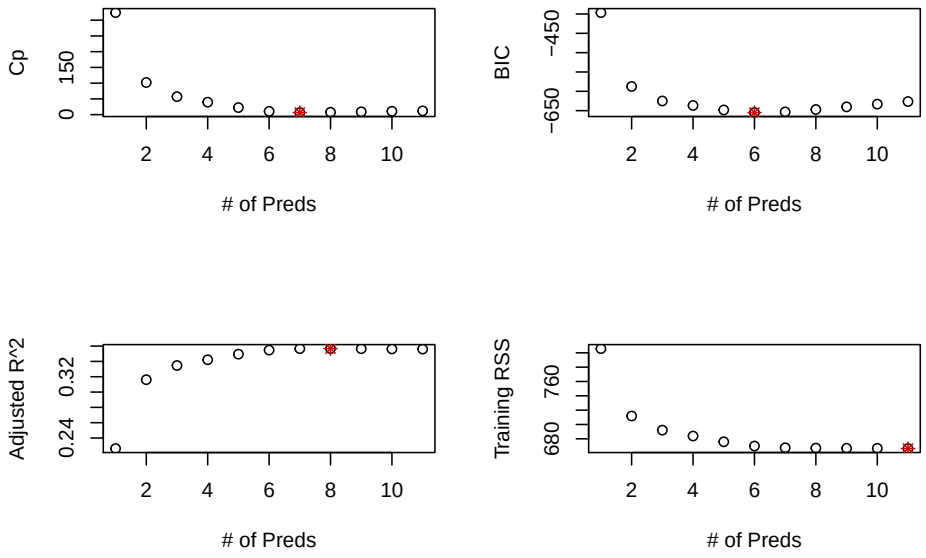
X_{ij} = independent variables for $j = 1, ..., p$

β_0 = y-intercept

β_j = slope coefficients for each variable

ϵ_i = residuals (error term of model)

Although it is possible to use every predictor in the MLR model, Best Subsets Selection selected the variables that best explain the variation in the response. The `leaps` library and the `regsubsets()` function in R produces comparisons of all possible models from the given set of predictors. Best Subset Selection compares all the possible combinations of predictors to select the best model that produces the smallest test error. Three different model criteria were evaluated to compare models of different amounts of predictor variables: Mallow's Cp, BIC, and R^2_{adj}.



Although each of the three model criteria suggested a different number of predictor variables, they were all within the same region (between 6 and 8 predictors). All three produced a similar error rate, so the theory of Occam's Razor was implemented and the simplest, six-variable model was selected. This was the model favored by BIC, which gives the greatest penalty for complexity in models. The six variables used in the regression context are then `volatile.acidity`, `chlorides`, `total.sulfur.dioxide`, `pH`, `sulphates`, and `alcohol`. All of the variables from the classification selection appear with the exception of density. In the regression context, `pH` was the variable that proved to be a rather unintuitive inclusion, because it is similar to density in that it is not easily detected by human senses. The inclusion

of these two variables in the model portray the importance of the chemical properties of a bottle of wine when it comes to quality.

Observations	1279
Dependent variable	quality
Type	OLS linear regression

F(6,1272)	122.539
R^2	0.366
Adj. R^2	0.363

	Est.	S.E.	t val.	p
(Intercept)	4.379	0.454	9.652	0.000
volatile.acidity	-1.062	0.113	-9.394	0.000
chlorides	-1.759	0.463	-3.803	0.000
pH	-0.507	0.132	-3.833	0.000
total.sulfur.dioxide	-0.002	0.001	-4.317	0.000
sulphates	0.981	0.127	7.741	0.000
alcohol	0.299	0.019	15.649	0.000

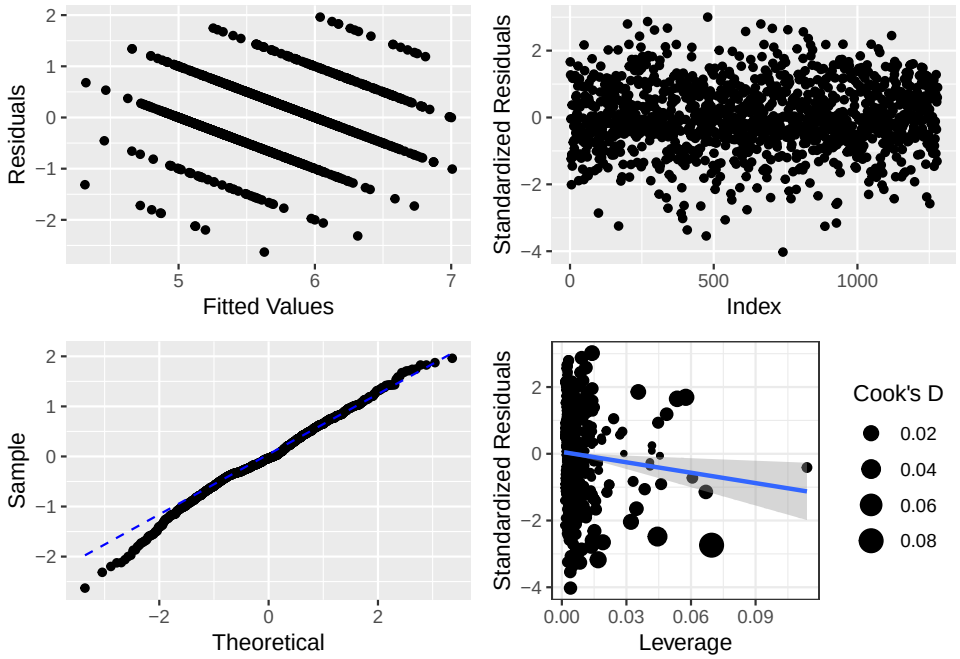
Standard errors: OLS

The adjusted R-squared value 0.363 suggests that these six predictors combined explain about 36.3% of the variance in `quality` of a wine.

The multiple linear regression model is subject to the following assumptions about the data:

- Linear relationship between response and predictors
- No collinearity exists between predictor variables
- Y_i's are independent and identically distributed
- $\epsilon_i \sim N(0, \sigma^2)$

To assess these assumptions, various diagnostics plots were evaluated:



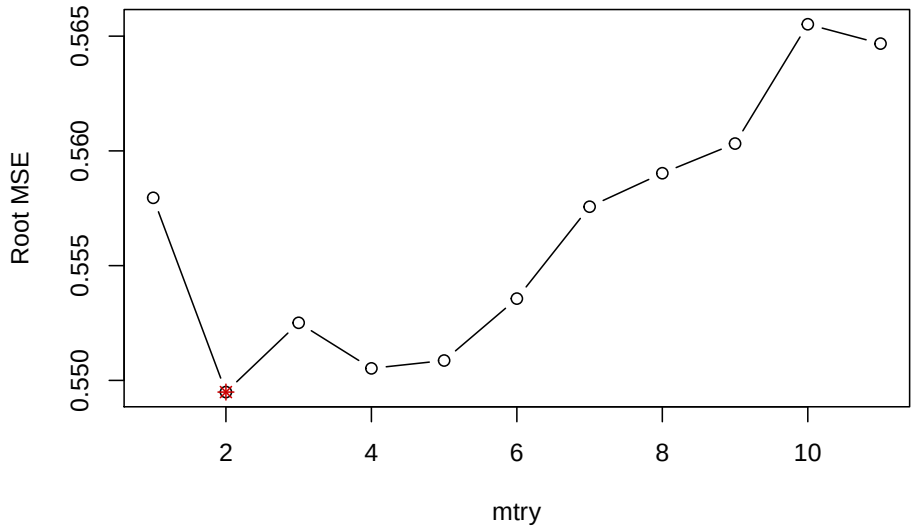
Residuals indicate that the MLR assumptions are mostly satisfied. The residuals vs. fitted values plot (top left) has unusual stratified lines, but this is caused by the observed responses being integer values of 4, 5, 6, 7, etc. rather than continuous values. The plot does not show heteroscedasticity and residuals are spread evenly around 0 over the entire range of fitted values. When residuals are standardized, as in the top right plot, one would expect about 95% to be inside the [-1.96, 1.96] interval; the actual proportion is 0.947. The QQ-plot in the bottom left compares the normality of residuals against a theoretical distribution. Except for a slight downward bias of the lowest 5% of residuals at 2 standard deviations below 0, the data fits a normal distribution almost perfectly. On the bottom right plot, high leverage points (those with leverage ≥ 0.03) are not dramatically biased, being spread both above and below the 0 residual. Cook's Distance tops out at about 0.05, well below the generally accepted threshold of 1 that is considered highly influential. These plots taken together suggest that the data generally fulfills the requirements for modeling using MLR.

The largest benefit to using a multiple linear regression model is its interpretability and ease of use. The coefficients can be computed efficiently and they have clear and easy interpretations. It is also possible to make inference on any of the variables or linear combination of variables. The most significant drawback of using the multiple linear regression model is that it is subject to distributional assumptions, meaning that if the assumptions are violated, the estimates and inference will be unstable at best and invalid at worst. Based on the residual plots, though, it appears that the model does not violate any of the model assumptions.

Random Forest for Regression

As mentioned earlier in the Classification section, Random Forest is an ensemble tree-based method, which creates multiple trees from bootstrap samples and averages the results from all the fitted trees. This model will be built exactly in the same way as the classification

model, except the tree will not be predicting if the wine is a part of the "high" or "low" class, but rather predicts the quality score associated with each wine. The predicted value of the quality score will be an average of all of the generated regression trees. As before, the model will first be built using cross validation to determine the number of variables that should be used in order to produce the smallest test RMSE.



Based on cross validation, the random forest model for regression using two randomly-selected variables considered at each split was selected. This model produced a test error rate of `rf.reg.test.err`.

Results

Table 4: Regression Model Results

	Training RMSE	**Test RMSE**
Multiple Linear Regression	0.8764	0.6280
Random Forest	0.5803	0.5533

Table 5: Classification Model Results

	Training Error	**Test Error**
Classification Tree	0.2447	0.3000
Random Forest	0.1798	0.1969

Based on the output, the best model for both classification and regression was the random forest. This supports the hypothesis that the random forest would outperform the other methods. This was not necessarily a surprise, since the ensemble methods tend to outperform regular methods. As an average of many classification trees, ensemble methods generally improve prediction accuracy and reduce overall variance. Other ensemble methods, such as bagging or boosting, may also be able to yield higher prediction accuracy.

In the regression setting, the values of root mean squared error (RMSE) for both models were relatively low. The RMSE for the random forest was 0.5533 and the RMSE for the multiple linear regression model was 0.628, further supporting the hypothesis that the random forest would outperform the other models.

One of the main disadvantages to using ensemble techniques like the random forest is the fact that they tend to be very computationally intensive. The basic classification tree and multiple linear regression model took significantly less time to run than the random forest, but they were less accurate in their predictions. If an ensemble method were run on a very large data set with a lot of features, the model would take much longer to run and would be incredibly difficult to interpret.

Conclusion

With the recent push for quantitative solutions in industries like agriculture, there has been an emphasis on using machine learning to solve a variety of problems requiring prediction and inference. In particular, the viticulture industry stands to gain a lot from such methods, since historically it has not interacted much with the world of analytics. In this study, various machine learning techniques were performed that assessed the prediction accuracy in both classification and regression contexts.

Based on the results of the study, prediction of if a given wine has high or low quality was achieved with about 70-80% accuracy. Despite not being able to achieve 80% prediction accuracy with the basic classification tree, the random forest model showed an improvement and predicted with a higher accuracy. Although this is a *good* result considering the non-information rate of 53%, it is by no means excellent. These results indicate that machine learning algorithms could be a good technique when trying to predict wine quality, but that there is room for improvement. With a more formal, industry-expert-guided variable selection process, the team believes that higher prediction accuracy can be achieved. Machine learning models could potentially be combined with other techniques to add value to winemakers.

Limitations

The multiple linear regression model is subject to certain distributional and model assumptions, which were not explicitly tested or checked. Since the objective of this project was to predict rather than make inference, model assumptions were not explicitly checked.

All wines in the data set came from wineries in Portugal, although they are from different wineries, one may worry that the results may not be generalizable to wine production globally.

Similarly, will these prediction results hold when it comes to white wine, rose, or sparkling wine? Given this variability over location and type of wine, what can be said about the wine-making industry as a whole?

Some of the wine regions have more strict standards for the designations of wine quality. For example, regions like Bordeaux and Champagne in France have some of the highest standards in the world for their wines, which are much different than somewhere like Greece or New Zealand. How accurate will the predictions be when there are different standards and how might one account for that variation?

Further Research

Since this study solely looks at the prediction accuracy of machine learning techniques for red wines, the natural extension of the analysis would be to verify that the results hold constant for white wines as well. The University of California, Irvine Machine Learning database has data on white wines, so that would be a good place to start.

The characteristics of any given wine are often derived from different geological and/or climatological conditions. For example, a wine that comes from a region close enough to the water to get a sea breeze may have a hint of saltiness in its aroma or taste. Similarly, a grape coming from a relatively young grove might give the wine different features than if the grove were hundreds of years old. It would be interesting to see how different weather conditions, terroirs, plant ages, and slopes affect the outcome of the wine in terms of quality.