Data Science Project
STAT 1361
Brian Forristal

Is the Quality of Wine Purely Subjective?

Introduction

Provided with the chemical compositions of Portuguese wines, we intend to predict the quality of wine and infer which predictors are objectively relevant to a wine's subjective sense of quality. The data of interest can be downloaded from the UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. The dataset contains 6,497 observations with 12 predictors, 1 response; and no missing values. Originally, the data came in two sets: one for red wines and another for white wines. The data did need major reformatting, other than the addition of a categorical variable for the color of the wine. Otherwise, the data was tidy and ready for analysis.

- ***Response variable:*** Quantitative, discrete numerical ranging from 0 – 10 for wine quality. Subjective, sensory-based score.
- ***Predictor variables:*** 1 categorical with classes "red" or "white" for the color of wine, and 11 quantitative measures for psychochemical compositions or characteristics and toxicity levels.

The primary research question was "How well can predictions of a wine's quality be made from the chemical composition predictors in the data and which features are most influential in that relationship?" Additional areas of interest were what type of functional form is best in approximating the true relationship between the predictors and the response (i.e., wine quality); do flexible or inflexible methods make better predictions, and; if utilizing an algorithmic method, can variables of importance be satisfactorily identified when compared to random noise?

Methods/Results Overview

Data Import and Scaling: Prior to any modeling, our data was partitioned into training and test sets comprising of an 80/20 split. Models were built on the training set, then predictions made on the test set, and the model selection criteria was the test error, or test mean-squared error of the predictions. The lower the test error, the better the model accuracy. All quantitative variables were standardized, while `color` and `quality` remain unchanged.

Simple Linear Regression: We began our analysis with a Simple Linear Regression model with one predictor onto `quality` – our qualitative response variable. Fitting a model to each predictor/response pair allowed us to see the change in test error across all twelve models. As evident in the correlation matrix, with a strong, positive correlation between them, `alcohol` onto `quality` yielded the lowest test error of 0.618. This value was our initial benchmark to compare all further methods to. The next best model being `density` onto `quality` with a staggering increase in test error to 0.695. In regards to the model's interpretability, it can be inferred that as the alcohol content of a particular wine increases, so too does its quality score.

Multiple Linear Regression: The full model with all predictors was our multi-variable benchmark. By simply including all the predictors in the model, the test error dropped to 0.535. Performing backward, forward, and stepwise variable selection resulted in models with all predictors selected, thus, it is irrelevant to further comment on their individual model accuracies.

Ridge and Lasso Regression: Both methods utilized cross-validation techniques to determine the optimal value for Lambda, the shrinkage parameter. Given the results, test errors for Ridge and Lasso were 0.539 and 0.535. Despite the highly-inflexible regularization methods, neither yielded lower test errors than the full model with all predictors.

Polynomial: An 11th degree polynomial term for `alcohol` onto `quality` yielded a test error of 0.605; a meager improvement over the initial Simple Linear model. Although not the smallest test error thus far, it's important to reflect on the fact that a model with one predictor, fitted to an extremely high flexibility, yielded a test error not *much* worse than that of our Multiple Linear Regression model with all 12 predictors. This steers our belief that it is possible a highly flexible, complex model with multiple predictors may result in even better test error.

Regression and Smoothing Splines: A 12th degree natural regression spline slightly improved the test error to 0.603, yet we lost accuracy with a 16th degree smoothing spline at 0.604 test error. Going back to our previous statement, it's apparent the data are non-linear in nature thus a more complex model may yield even better results.

Generalized Additive Model (GAM): Even further beyond linearity, our Generalized Additive Model was fit with smoothing splines with 27 degrees of freedom for each predictor, except the binary `color` variable. The high degree of flexibility may approximate the functional form of the data much better than other linear and non-linear counterparts. The test error was 0.503. Reducing our model's bias for increased variance was worthwhile for lower test error.

KNN Regression: Resulting in a substantial reduction in test error, down to 0.468, the regression-version of K-Nearest Neighbors (KNN) makes continuous, numerical predictions in contrast to its classification counterpart. The optimal value of the tuning parameter was k = 10. Despite the momentous reduction in test error, there was a substantial loss in interpretability. There are no coefficient estimates or variable importance measures within the model to indicate which variables account for the greatest decrease in test error.

Boosting: A boosting model optimized at an interaction depth of 11 with 2,000 trees produced a test error of 0.379. The decrease in test error was at the cost of severely increasing the ambiguity and "black-box nature" of our model; further degrading interpretability. Decision

Trees, and more specifically, Random Forests, offer even greater predictive accuracy while ultimately remaining somewhat intuitive.

Decision Trees: A single Decision Tree, optimally pruned, resulted in a test error of 0.585. Although the test error increased compared to the insanely flexible GAM, this method is quite interpretable. (See Appendix A for the plot). It appears that out of all twelve predictors, \`alcohol\` and \`volatile acidity\` were chosen to make predictions. The correlation between both variables is a miniscule at -0.038. The likelihood they are linear combinations of each other in relation to the response may be quite low. A Multiple Linear Regression with \`alcohol\` and \`volatile acidity\` onto \`quality\` revealed that the parametric version of the analysis produced a lower test error, 0.573, than of the Decision Tree.

Bagging: Compiling 1,000 bootstrapped Decision Trees with all 12 predictors resulted in a momentous drop in test error to 0.354. As with the single Decision Tree, \`alcohol\` and \`volatile acidity\` are the first and second highest ranked variables of importance based on their influence on the test error. The additional step in bootstrapping is to resample the data with replacement then build successive trees. Due to the high number of trees built, the test error is stabilized towards the expected value. We would be hard-pressed to find a better model, but Random Forests are known for just that.

Random Forest: As expected, the Random Forest algorithm with ntree = 500 and mtry = 4 out performed our bagged model with a decrease in test error to 0.351. Beginning with a Simple Linear Regression model with \`alcohol\` onto \`quality\`, we have successfully reduced our test error by 43.2% in utilizing a highly flexible Random Forest with all predictors.

Model Comparison, Variable Importance

Assessing All Models for Significance: In our follow-on permutation analysis, we built a function in R containing all the models previously analyzed thus far that randomly permutes the response variable to measure the change in each model's test error. (See Appendix B for the graphical results of part 1). The Decision Tree built a stump each time with one constant for the entire vector of response values, thus, it will not be considered in the next statement. The other models displayed an interesting pattern. The more complex models (Boosting in particular) had large variances, seen as covering more of the horizontal area above the x-axis coordinates, while the inflexible (Ridge and Lasso) had quite narrow variances. This appears to be the bias-variance trade-off in action. The graph helps illustrate our analysis thus far. Smaller red dot values indicate lower test error, with ensemble methods clearly in the lead, and the distance from those dots to the mounds of permuted test errors as anecdotal measures for overall model significance. Boosting is furthest by this last measure, but due to its high variance and challenging parameter tuning, we have opted to go with a much "simpler" Random Forest model.

Random Forest Variable Importance and Model Fine-Tuning: After assessing the performance of many types of models, we decided to proceed to the permutation phase of our analysis with a Random Forest with ntree = 500, and mtry = 4, and all predictors. First, we permuted the response variable to determine if the model's test error was due to noise. (See Appendix C for the graph). There are no permuted-test errors smaller than that of null of 0.352. The p-value of this is 1/101 = 0.0099. The distribution of permuted-test errors is distinctly normal; thus, the null test error is statistically significant.

Second, we permuted each variable individually 50 times, under the assumption that sampling over 30 observations will produce a normal plot of values, to see what influence the permutations had on the test error. (See Appendix D for the graphical output of the first iteration). A dotted-red horizontal line indicates the null test error of the original Random Forest. Variables displaying the largest impact on the prediction accuracy are furthest from the dotted line. It appeared as if `color` was no more significant than random noise as its distribution of test errors lay directly on the line. `Alcohol`, `free sulfur dioxide`, `sulphates`, and `volatile acidity` were all generally the same distance from the null; implying that they were equally significant in affecting the model. `Alcohol` was consistently cited as being the most important variable in a number of the other models: Linear Regressions had the largest coefficient estimates attributed to `alcohol`; Polynomial and Spline methods outperformed Multiple Linear Regression models with all predictors given the single predictor was `alcohol`; and ensemble methods' variable importance measures routinely listed `alcohol` at the top. Removing `color` from the data and refitting the Random Forest increased the test error to 0.352. Model complexity was slightly decreased at the cost of accuracy… a valuable trade depending on the business costs associated with measuring specific factors of production, quality control, etc.

Permuting each variable 50 more times measured the resulting test error without `color` in the data. To better identify which variables were performing better or worse, the mean permuted-test error was calculated for each variable and graphically displayed in contrast to the previous iteration. (See Appendix E for the graphical results). Notice `residual sugar` was the only variable to have decreased in importance in contrast to the previous iteration. Re-fitting the Random Forest without `residual sugar` dropped the test error back to the null value of 0.3512. Why might this be happening? Could `color` and `residual sugar` be linear combinations of each other and `quality`?

The final two iterations yielded quite profound results. The additional removal of `chlorides` reduced the test error to 0.347. The lowest yet! Consequently removing `density` increased it back to 0.352. Eight variables had almost the exact same test error as all 12 did. We choose to keep `density` in the data for lower test error rather than reduce the variable count. Through a time-consuming process, both in coding and computation, we have thoroughly verified which variables are truly important in predicting wine quality. If we solely relied on the supplied variable importance measure from the Random Forest package, we would not have gleaned these important insights. Here, we halt any additional alterations to the data and remark on the results. The final model with the lowest test error of 0.347 was a Random Forest with ntree = 500, mtry = p/3 = 3, and of the following formula:

"quality ~ alcohol + sulphates + pH + density + total.sulfur.dioxide + free.sulfur.dioxide + citric.acid + volatile.acidity + fixed.acidity."

Thoughts and Takeaways

How many models seemed to perform "best" in terms of predictive accuracy? How did you measure this? Relative to what the models are doing, does it make sense why they would perform similarly well or are they quite different? Do you have a sense of whether such models are actually "significantly" better than others?

The ensemble-class of non-parametric models performed best in terms of predictive accuracy. More specifically, the Random Forest with ntree = 500 and mtry = p/3 = 3 produced a trained model which predicted the test set's response output with the best accuracy. Bagging is a close second with a marginally worse test error of 0.354 but required significantly more computation time per model due to the increase to ntree = 1,000. The increase in test error may be partly due the decreased randomness of the Bagged model by limiting the number of variables chosen uniformly for consideration at each split of the tree.

Beyond ensemble methods, KNN-Regression and the highly complex GAM did particularly well. We believe, given the results of our analysis and the performance of particular models, that the data's underlying functional nature is non-linear and complex. This is evident in that the best models were flexible, non-parametric methods and more traditional linear methods underperformed by almost a factor of two.

The test error, or mean squared error (MSE) of the test set predictions, is the key metric for model selection. Of the entire dataset, 80% of observations were randomly partitioned into the training set, to which models were fit on, and the remainder as the test set; to which fitted models' prediction accuracies were measured – via the MSE. The lower the test error, the better the model. Training errors are not considered during model fitting and parameter tuning because it is irrelevant to have a model that does well on training data but poorly on the test data. Only tuned parameters that result in decreased test error are ideally suited for model comparison. By this performance measure, the Random Forest outperforms other models in its class; and humiliates linear/parametric methods.

To determine if a particular method is genuinely "significant", we randomly permuted the response variable, keeping all others constant and unchanged, rebuilt the model, then re-measured the test error. This was done for all models under consideration in this analysis. As discussed in the previous section, the permutation-testing of each variable fine-tuned the Random Forest model and reduced the number of required variables needed to make accurate predictions. Due to this, we believe that the final model is truly significant and reliable; given the provided data.

Among the top-performing models, which variables seemed most important? Are they mostly the same between models or are they quite different? Do you have any intuition as to why certain variables might appear more important in some models but not in others? Think about what those variables actually measure, what their general relationship to the response might be, and what kinds of models might do better or worse at picking up different kinds of effects.

Across most of the models, \`alcohol\` appeared to be the most important predictor. It has the highest correlation with wine quality. Of the three variables removed from the final Random Forest model (\`chlorides\`, \`residual sugar\`, and \`color\`), \`alcohol\` is more correlated with both \`chlorides\` and \`residual sugar\` than others. There may exist linear combinations of these three variables with the response that explain why, prior to permutation-testing, the Random Forest model had the lowest test error. Once removed, the absence of noise and collinearity brought the test error down.

Generally, across all other model types, \`alcohol\` stands out as the most important. Linear models estimate the coefficient with the highest magnitude. The same is true for regularization techniques of Lasso and Ridge regression. Moving beyond linearity, it is no longer possible to solely relying on coefficient estimates to measure importance. Variable importance measures within R objects were insufficient and not the single source of truth. Permuting each variable, measuring the change in test error, and removing noisy variables reduced the complexity and test error of our final model.

An intuitive explanation as to why most models find the same variables important may be within the complex chemical interactions of chemistry; which is beyond the scope of this analysis. Particular variables may have highly complex interactions with each other, which may be detected by highly flexible models (Random Forest, KNN-Regression) not biased towards any type of underlying function of the data, but more inflexible models, under the assumption of linearity in the data, may be suitable for detecting the main effects of these interactions more precisely. This may account for the decrease in test error as the models became increasingly more complex and less biased.

When comparing the graphs in Appendix D and E, there is a pronounced difference in the variation between each permuted set of test errors. When each variable is permuted in the full model with all 12 variables, their mean test error values appear quite varied. After \`color\`, \`residual sugar\` and \`chlorides\` are removed, the variation is lessoned. Removing the variables decreased the model's test error but appears to have also reduced the overall variability, or noise, in the data.

What were the most challenging aspects of working with your particular dataset? Were you able to mitigate these issues or do you feel that your final results are less certain as a result of them? Perhaps most importantly, do you really trust your "best" results at the end of the day? If you were in a position where you were personally held liable for any negative outcomes associated with implementing your model, how worried would you be?

The data are particularly challenging in that the predictors are quantitative-psychochemical compositions, while the response is a qualitative, sensory-based scoring. We could not reconcile the biases in the scoring process. Another challenge with the dataset is that we only assessed the main effects of predictors on the response. Modeling higher-order interactions with a Random Forest may result in substantially decreased test error at the expense of increasing computation time and model complexity. This would be appropriate for a follow-on analysis and more research.

We took the appropriate steps, given the provided data, to mitigate these issues and ensure any connections and patterns existing could be modelled and tested. Segmenting the data into training and test sets ensured the performance of each model could be properly measured on simulated "real-world data". Models were trained and tested by a variety of classes: Linear Regression, Multiple Linear Regression, Lasso, Ridge, KNN-regression, Smoothing Splines and Polynomial Interaction models, Decision Trees, complex ensemble algorithms, and more. As we traded variance for bias between the different models, we ultimately settled on Random Forest Regression; the trade-off was such that our prediction accuracy increased. We then were able to measure each variables' importance by permutation-testing. This resulted in the smallest test error of our analysis with the added benefit of removing three noisy variables.

Upon completion of this analysis, we identified particular aspects of the data that have yet to be tested – the scorer bias and higher-order interaction terms – but given our thorough analysis thus far, we are confident in our findings, such that they provide a basis for future analysis. Prior to making any serious suggestions to business operations or wine producers, additional variables regarding both areas of concern must be collected and implemented in future analyses. Until then, we aren't fully confident in our findings. The added information may completely change our results and models, rendering them obsolete.

Imagine that you had to present a summary of your findings to a non-technical decision-maker. In other words, this person is going to take some kind of action based on what you report to them but they lack the technical expertise to really "check your work" or even understand how you arrived at any of your conclusions. What would you report to them (i.e. what kind of recommendations would you give)? How would you present it to them? Imagine you had the ability to request more and/or different data. What might you want to request? More observations? More (or different) variables included?

Key stakeholders within the business have enlisted the in-house analytics team to provide value-added insights towards optimizing and understanding the production of our wine. The primary goal of the analysis was to uncover which psychochemical agents, if any, would be suitable to use in predicting the quality score. The insights may help maximize production quality while controlling for costly ingredients that negatively affect the score. Given the available data, our team was successful in revealing several key insights: The dataset contained sufficient observations for an adequate analysis; the best model for predicting quality was highly flexible and produced low test error rates; techniques for identifying which chemicals were important did reduce the variable count to a minimum of 9 out of 12 while not sacrificing the accuracy of the predictions; and a foundational understanding of our production was gained which will benefit future analyses.

The best model identified in our analysis is called a Random Forest. It is a bit of a "black-box", but you can imagine it as literally a forest of 500 trees where the leaves on a branch are predictions. Each tree in the forest is grown by randomly choosing 1/3 of the available variables and building an optimal tree from them. Predictions for a specific response values are the average across all the trees. That is how predictions are made. The randomness in choosing the variables is what experts believe aids the algorithm in find patterns in the data that otherwise would be obscured by one or more influential variables more strongly correlated with the response variable. Correlation is the tendency for two variables to follow the same pattern.

To see which variables are important: shuffle one column of data, refit the model, and measure its performance. Do that a number of times per variable. This is similar to removing the column entirely, but is more robust. The graph provided (Appendix D) shows just how important each variable is. See `color`, it's on the dotted-red line. That means `color` may not be as important as we previously thought, which is not intuitive as most people think color is very important. To proceed, we'll remove `color` from the data set and redo the previous steps until no more variables look like `color` did.

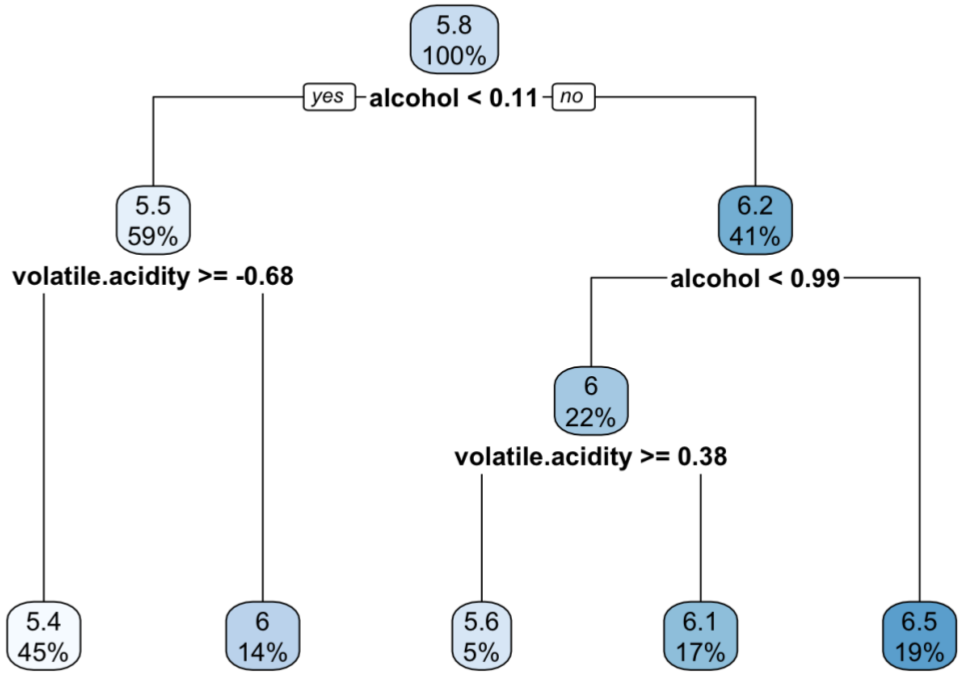
In the future, as we collect more and more data, the Random Forest model takes longer and longer to calculate. Right now, a single model takes about 20 seconds. Do that for each of the 12 variables, 50 or more times; that'd take my laptop 3.33 hours. This is a simple example, but if we were to expand our dataset by thousands of variables and millions of observations, a "big data" dataset, our computation time would be incredibly large and take weeks or months to complete. Another item identified by the analytics team that we are going request access to is Amazon AWS Cloud Services to perform larger computations on. In the cloud, we can take advantage of many simulated computers and reduce the overall computation time to a fraction of that on a single computer.

In addition to faster computers, we recommend several adjustments to the current data collection procedures. First, we suggest that demographics of the wine scorers be collected: age, gender, expertise-level, etc., which would be anonymized to protect their personal-identifiable information before analysis. This additional information may be helpful in reconciling discrepancies in our models. Subjective, sensory-based scoring of wine may be subject to human error and bias. Accounting for that additional variability, or change, in the scoring of wine quality is important; and may improve our predictions.

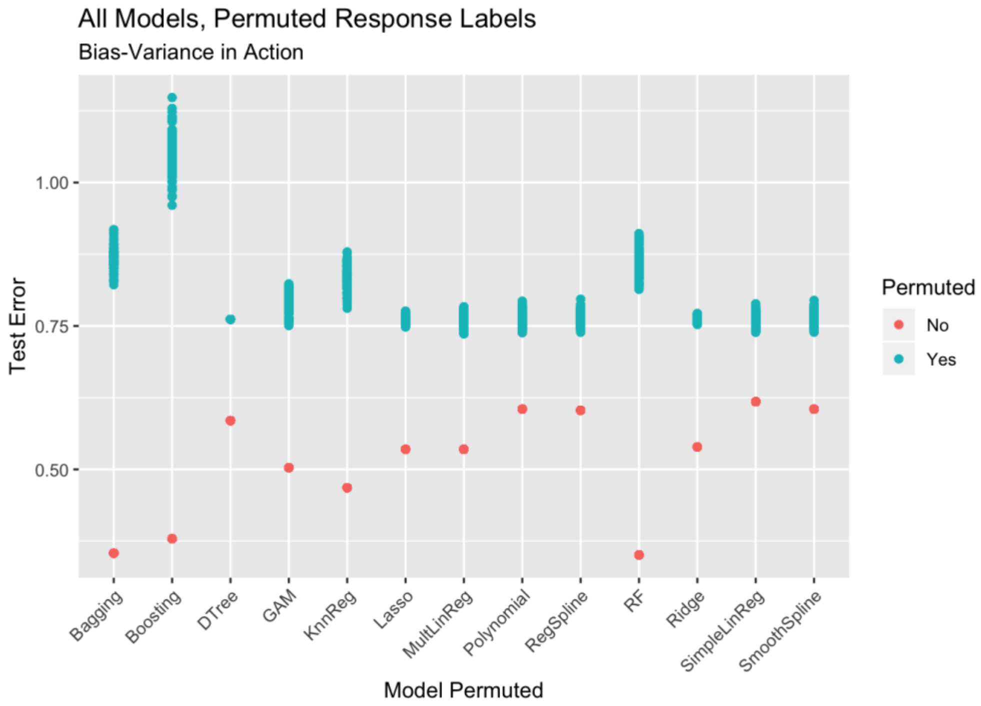
Secondly, we suggest that the sourcing of ingredients for the production of wine may be heavily influenced by batch, distributor, or location it is sourced from. Are all the grapes sourced from the same geographic region? If not, this may influence our model. Was a batch of grapes, or indeed the wine in the bottle, subjected to higher than normal temperatures during transport? This may become obvious in the analysis of variance of the batches and may affect chemical interactions. We suggest all recommendations be implemented prior to the next analysis to see what affect is had on the models.

In summary, given the provided data and a Random Forest model, we have narrowed down the list of important variables in our wine production; ultimately reducing it to 9 out of 12. According to this preliminary analysis, alcohol content is the most important variable in predicting the quality score a particular wine will receive. And, lastly, we have identified several key opportunities in the data collection process that may benefit future analyses.

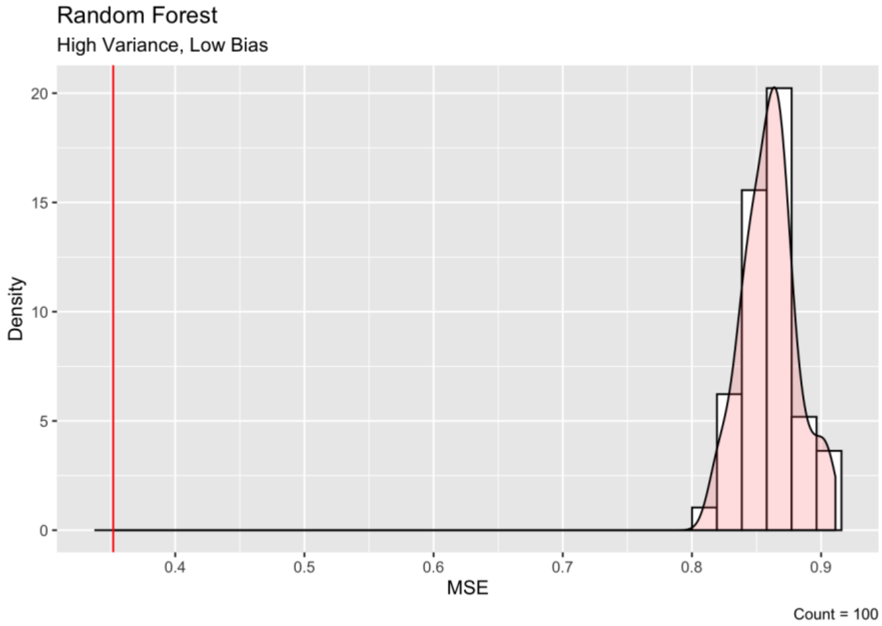
Appendix A



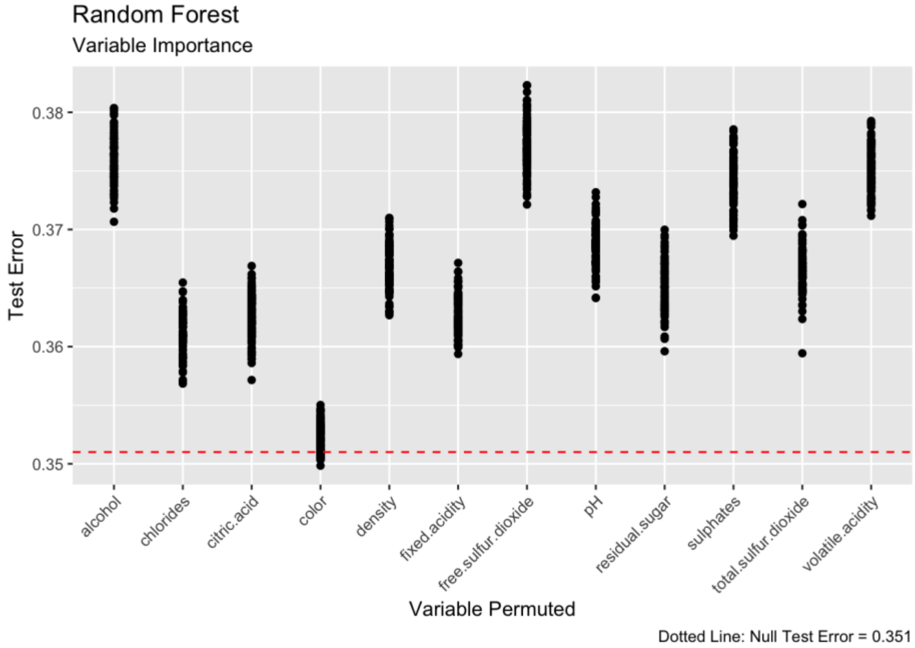
Appendix B



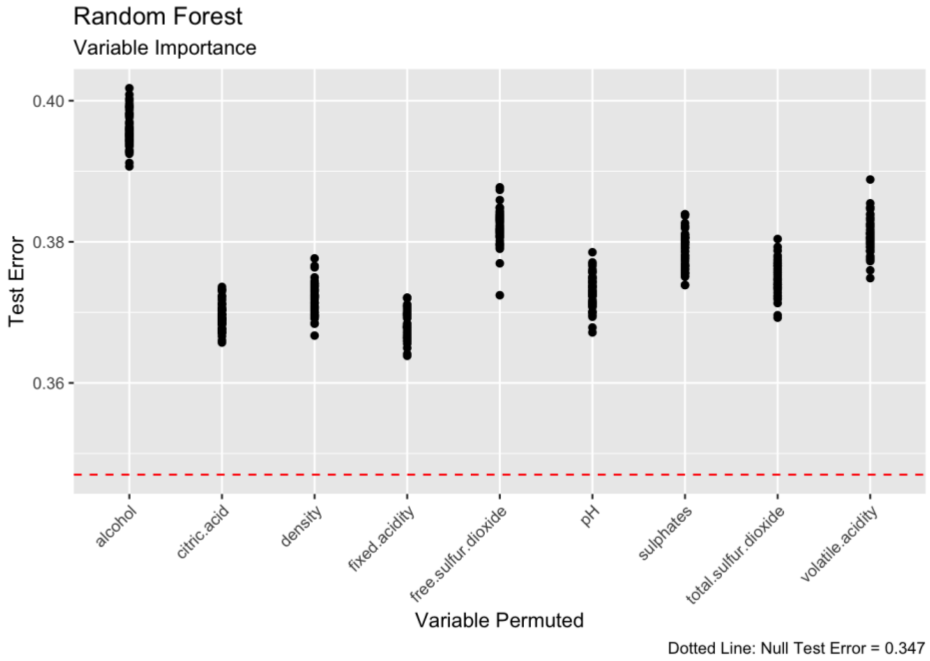
Appendix C



Appendix D



Appendix E


Random Forest
Variable Importance
Test Error
0.40
0.38
0.36
alcohol
citric.acid
density
fixed.acidity
free.sulfur.dioxide
pH
sulphates
total.sulfur.dioxide
volatile.acidity
Variable Permuted
Dotted Line: Null Test Error = 0.347