Determining Wine Quality and Rating Based on Production Predictors

Group #13:

Daisy Gonzalez

Alejandro Perez Gonzalez

Erica K Hay

Gabriel Dau

Math 4322 - Introduction to Data Science & Machine Learning

1. **Introduction** (Gabriel)

With the increasing popularity and demand for quality wines, the wine industry has begun to integrate new technologies in the field of wine quality evaluation to ensure high standards and consistent product. Wine quality evaluation has become an imperative process in the wine industry not only to guarantee customer safety and satisfaction, but can also be used as an economic tool to gain an advantage against competition and increase profits. New advancements in the wine certification process have grown to include numerous physicochemical tests such as acidity, density, and pH levels, while other sensory nuances have been left to the attuned palates of wine experts to explain the intangibles that often accompany high quality wines. In the modern era of wine making, the collection and understanding of large datasets can and have been used to increase the successful production of quality wines.

Question:

Which predictors will lead to a higher quality rating for white wine?

About Our Data: (Gabriel)

Our dataset is from the University of California Irvine's Machine Learning Repository Center for Machine Learning and Intelligent Systems. Our dataset is a part of a pair of two datasets related to red and white variants of Portuguese "Vinho Verde" wines whose export has led Portugal to become an industry leading producer of wine.This data set focused on the quality of wine based off of 4,898 observations of rating wine quality rating that was based off of taste, noted as

sensory in the data set website.. For simplicity and conciseness, we have chosen to focus only on the white wine dataset. Using our dataset, we want to determine overall wine quality using 11 predictors: grams of tartaric acid per decimeter^3 (fixed.acidity), grams of acetic acid per Decimeter^3 (volatile.acidity), grams of citric acid per decimeter^3 (citric.acidity), grams of residual sugar per decimeter^3 (residual.sugar), grams of sodium chloride per decimeter^3 (chlorides), miligrams of free sulfur dioxide per decimeter^3 (free.sulfur.dioxide), miligrams of total sulfur dioxide per decimeter^3 (total.sulfur.dioxide), fluid density in grams per centimeter^3 (density), pH level (pH), grams of potassium sulphate per decimeter^3 (sulphates), and alcohol by volume % (alcohol).

Predictor Variables:

1. Fixed Acidity
2. Volatile Acidity
3. Citric Acidity
4. Residual Sugar
5. Chlorides
6. Free Sulfur Dioxide
7. Total Sulfur Dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

Response Variable:

1. Quality: Quality is the wine rating that was given to the wine based off of the predictors listed above. Our goal of this project is to understand which predictors are indicators of a good wine quality, a quality rated above a 5. As well as understanding which predictors are the most important to wine quality.

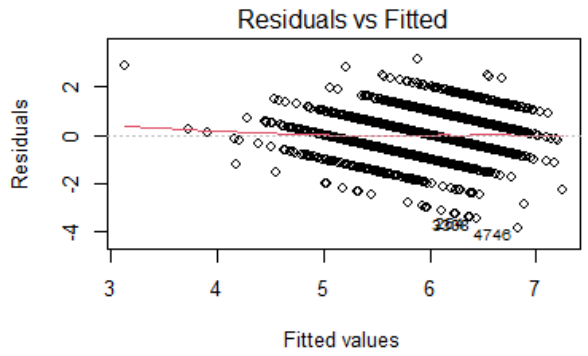
2. Methods

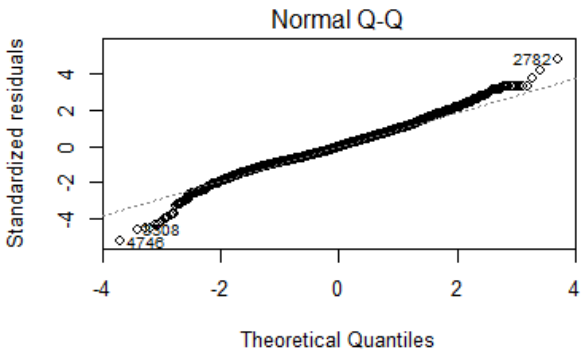
(Simple Linear Regression, Daisy)

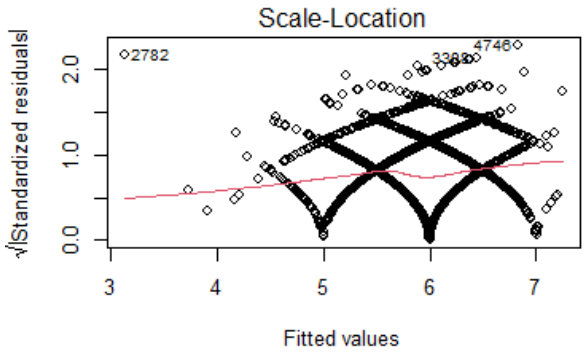
The response variable, quality, for the white wine data set is a quantitative variable. Therefore, the first approach used to analyze the data was linear regression. This approach is the easiest to implement because we are analyzing all 11 predictors and assuming that the quality result has a linear relationship with all the variables. The linear regression model also performs well for linearly separable data and is easier to interpret. The issue with assuming a linear relationship is that it is overfitting on the data because the model is too generalized on the data set. The linear regression model is also prone to multicollinearity, which may produce less reliable results.

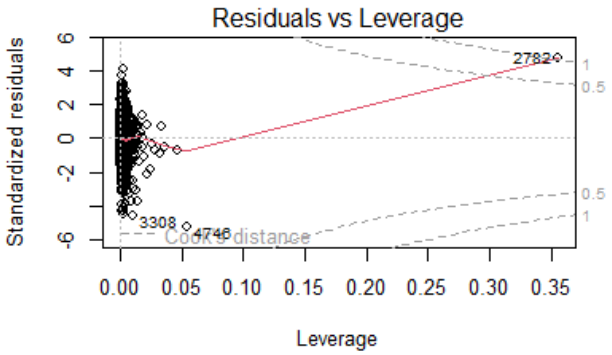
$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_{11}X_{11} + \varepsilon$$

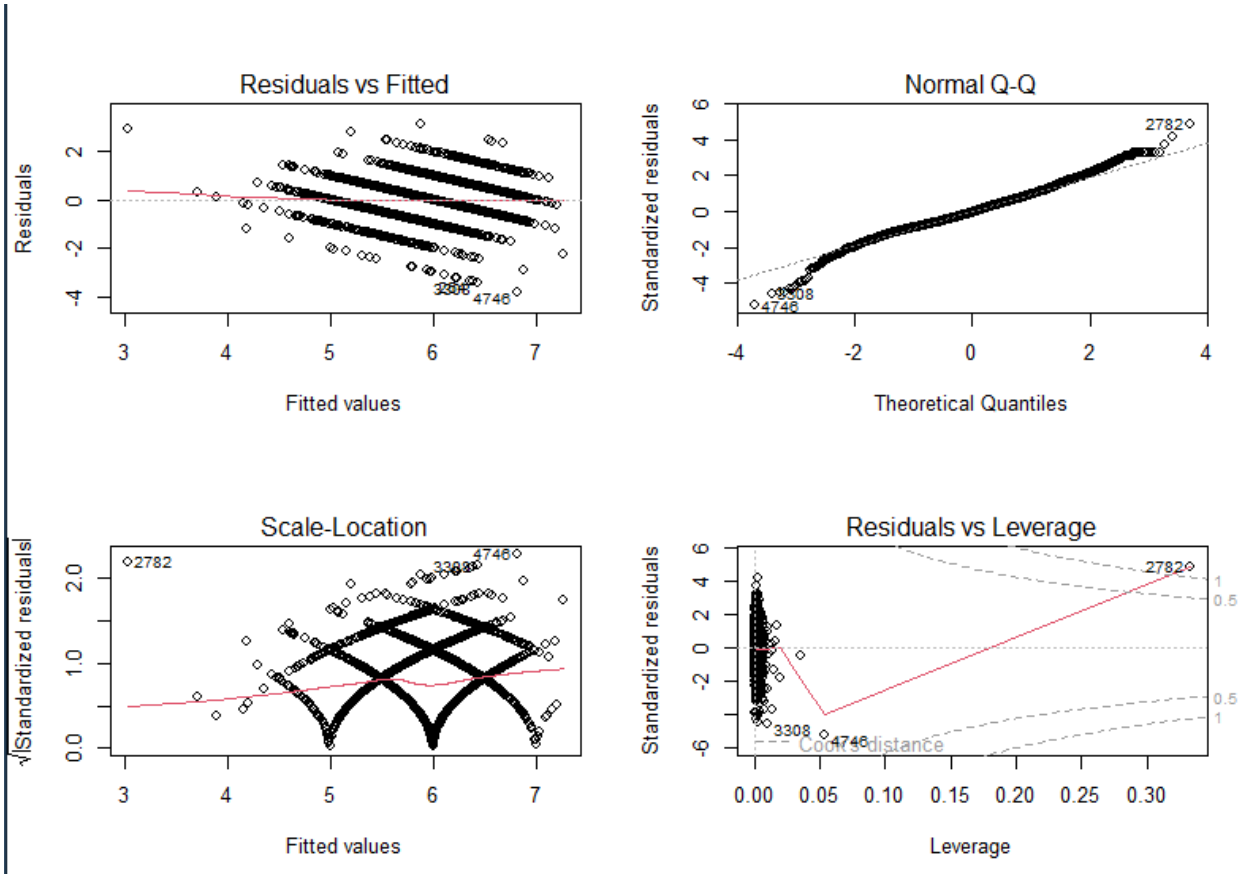
The linear regression model shown above compares all 11 predictors against quality. The formal Y will represent the quality, B0 will represent the intercept, Xi for $i = 1, 2, 3, \ldots, 11$ will represent the predictors: *Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Density, pH, Sulphates,* and *Alcohol,* B1 is their respective coefficients that we will later figure out through linear regression.


Residuals vs Fitted
Residuals
Fitted values
4746


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2782
4746


Scale-Location
√|Standardized residuals|
Fitted values
2782
4746


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
2782
3308
4746



After plotting both the original linear regression model and the second linear regression model with the most important variables it was evident that there was not a direct relationship between the predictors and the quality of wine. Both plots showed that there was not a linear relationship between the predictors and the quality of white wine. These results supported a possible theory that estimating the quality of wine would be difficult as everyone has varying preferences for wine which will make them rate the quality of wine based on personal preference rather than a standard definition of wine quality. This can be supported by noting that there is not equal variance in this linear model and that the quality of wine is a result of simple random sample.

Regression tree model (Alejandro, Erica):

We will implement a tree based model because of the ease implementation and interpretability. However, small changes in the data may cause a large change in the estimated tree and we might encounter the risk of overfitting the data. For that reason, we will also apply pruning, bagging, random forests and boosting. Implementing each of these methods will improve the predictive performance of our tree. Specifically, tree pruning will prevent overfitting the data by removing the nonsignificant nodes, making our subtree less complex and easier to interpret.

Formula: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides ++ free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol

3. Results (Erica, Daisy)

Model of Linear Regression:

```
Call:
lm(formula = quality ~ ., data = white_wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8348 -0.4934 -0.0379  0.4637  3.1143

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
citric.acid           2.209e-02  9.577e-02   0.231  0.81759
residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819,  Adjusted R-squared:  0.2803
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

The summary of the linear regression model reveals that multiple predictors are significant in predicting wine quality, such as volatile acidity, free sulfur dioxide, density, and alcohol. Volatile acidity, density, and alcohol produced the lowest p-values in the model, ranging from 2e-16 to 4.04e-15, indicating that they have a stronger relationship with the quality of white wine.

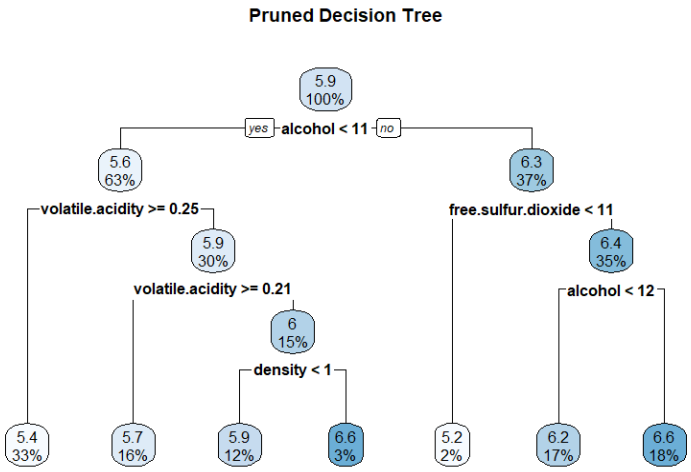
```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates + alcohol,
    data = white_wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8246 -0.4938 -0.0396  0.4660  3.1208

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7512 on 4889 degrees of freedom
Multiple R-squared:  0.2818,    Adjusted R-squared:  0.2806
F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

Regression Tree Model:



The pruned decision tree concludes that alcohol, volatile acidity, free sulfur dioxide, and density are the most significant predictors that determine wine quality. The predictor space got segmented into seven regions (terminal nodes) that determine their predicted wine quality:

R_1 = alcohol < 11 & volatile acidity >= 0.25 = 5.4
R_2 = alcohol < 11 & 0.21 <= volatile acidity <= 0.25 = 5.7
R_3 = alcohol < 11 & volatile acidity <= 0.21 & density < 1 = 5.9
R_4 = alcohol < 11 & volatile acidity <= 0.21 & density > 1 = 6.6
R_5 = alcohol > 11 & free sulfur dioxide < 11 = 5.2
R_6 = alcohol > 11 & free sulfur dioxide > 11 & alcohol < 12 = 6.2
R_7 = alcohol > 11 & free sulfur dioxide > 11 & alcohol > 12 = 6.6

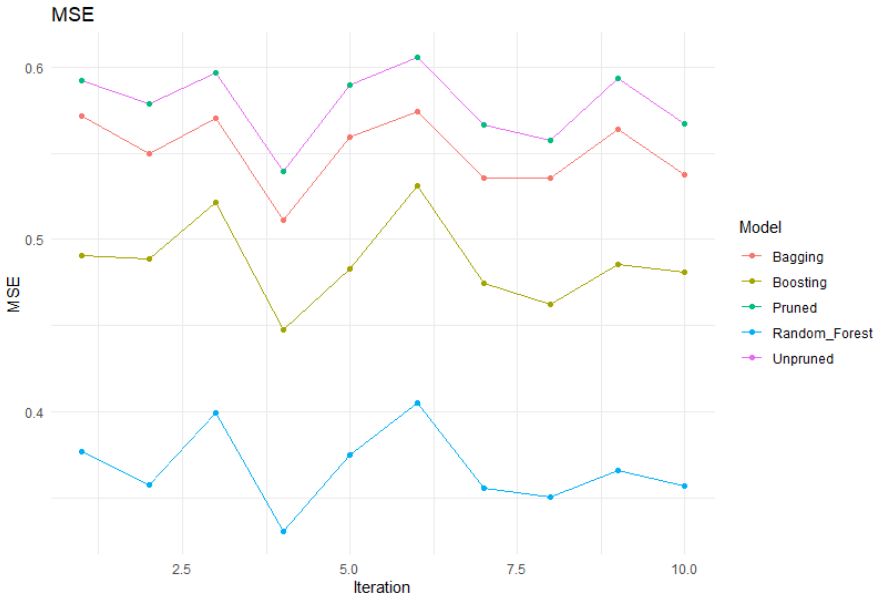
Mean of Test MSEs - Bagging, Random Forests, and Boosting:

Sample commands:

```
#Random Forest
rf_model <- randomForest(quality ~ ., data = train_data)
rf_pred <- predict(rf_model, test_data)
rf_mse <- mean((test_data$quality - rf_pred)^2)
# Boosting
boosting_model <- gbm(quality ~ ., data = train_data, distribution = "gaussian",
n.trees=          100, shrinkage = 0.1, interaction.depth = 3)
```

```
boosting_pred <- predict(boosting_model, test_data, n.trees = 100)
boosting_mse <- mean((test_data$quality - boosting_pred)^2)
# Bagging
bagging_model <- bagging(quality ~ ., data = train_data, nbagg = 100)
bagging_pred <- predict(bagging_model, test_data)
bagging_mse <- mean((test_data$quality - bagging_pred)^2)
```

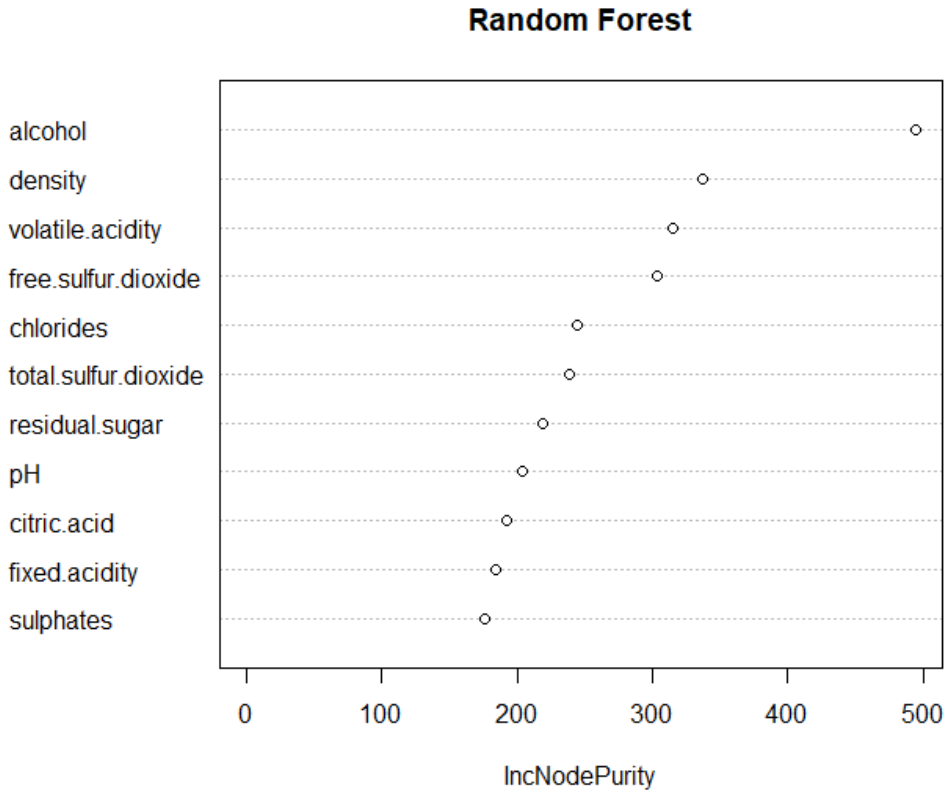
Iteration	Unpruned	Pruned	Bagging	Random_Forest	Boosting
1	0.5924379	0.5924379	0.5717439	0.3765573	0.4908016
2	0.5784968	0.5784968	0.5496868	0.3576226	0.4887689
3	0.5967286	0.5967286	0.5701301	0.3993006	0.5217563
4	0.5392724	0.5392724	0.5111641	0.3308368	0.4475778
5	0.5895229	0.5895229	0.5592165	0.3750538	0.4831882
6	0.6060111	0.6060111	0.5744114	0.4049574	0.5308124
7	0.5662966	0.5662966	0.5354291	0.3553774	0.4747784
8	0.5574197	0.5574197	0.5356604	0.3504955	0.4623301
9	0.5932771	0.5932771	0.5636655	0.3658692	0.4852577
10	0.5671186	0.5671186	0.5373648	0.3570568	0.4809127



The Mean of Test MSEs concludes that the random forest method produced the lowest MSE values compared to the unpruned, pruned, bagging, and boosting methods. Based on this result, we implemented the random forest method to determine the most significant predictors.

Random Forest:

	IncNodePurity
fixed.acidity	184.2600
volatile.acidity	315.3986
citric.acid	192.4062
residual.sugar	219.0714
chlorides	244.4757
free.sulfur.dioxide	302.9807
total.sulfur.dioxide	238.2083
density	337.4709
pH	203.5677
sulphates	175.9536
alcohol	494.5513



Utilizing the variable importance plot, we gather the same results as the pruned tree: alcohol, denisty, volatile acidity, and free sulfur dioxide are most significant in determining wine quality.

4. Conclusion (Erica, Daisy)

The linear regression model and the regression tree models both concluded that alcohol, volatile acidity, density, and free sulfur dioxide are the most significant predictors in determining the quality of white wine. The linear regression model revealed the significant predictors using p-values while the regression tree model determined the most signifcant predictors using recursive splits based on the impact of each predictor on the quality. The regression tree model was more effective in determining which predictors were significant because we were able to utilize a pruned tree that showed the best subtree with the best splits, as well as a variable importance plot that gave us a visual of the most significant predictors using node purity.

Although it was simple to note which predictors are important to determining wine quality, it was difficult to analyze which predictors led to a higher wine quality rating. Based off of the results from the regression tree model it seems that there are combinations that would lead to a higher wine quality such as higher levels of alcohol and higher levels of free sulfer dioxide. Another possibility that needs to be noted is that since the wine quality was rated off of sensory each observation had different preferances to wine quality. Therefore the wine quality results might not accurately represent wine quality since it is trying to predict wine quality off of chemical predictors rather than predictors that would be easier to analyze the quality of wine such as sweet, dry, or bitter wine.

Source code

References: (Gabriel)

1. Modeling wine preferences by data mining from physicochemical properties
2. Dataset: University of California, Irvine Machine Learnign Repository Center for Machine Learning and Intelligent Systems