Red Wine Quality Project

Varantorn Weruwanarak (873300483)

Siddhesh Padwal (873305527)

Master of Computer Engineering
Ritchie School of Engineering and Computer Science
University of Denver

***Abstract* – In this paper, we attempt to find which chemical properties influence the quality of red wines. The first attempt we use correlation to analyze between chemical and quality. After that, we applied linear regression model to predict the outcome of a test set data. However, because of there are a lot of variables are responsible for the quality of the wine and result from correlation maybe incorrect. So, for Second model we use random forest to classify and extract the feature that influence the wine quality respectively. And we create a random forest model to predict the outcome of test data. Finally, we can see how relative between the result from linear regression model and random forest model and which one has more accuracy for test data.**

I. INTRODUCTION

How we can determine which red wine has better quality. It is difficult to determine an empirical relation between the subjective quality of a wine and its chemical composition. Wine makers want to know what they can do to their processes to optimize the quality of their wine. While attempts have been made to build classifiers for wine from chemical data, not all algorithms have been tested.

In addition to testing new algorithms and variations of algorithms for prediction purposes, we are also interested in analyzing the data itself. For example, do some ingredients have a stronger impact on the perceived wine quality that others? Should winemakers be focusing more on certain ingredients than others?

So, based on knowledge from Probability and Statistics for Data Science I, we applied correlation and linear regression model also random forest to analyze which chemical properties influence the quality of red wines and which model will give better accuracy for predict.

II. Dataset

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Input variables (chemical properties)
(based on physicochemical tests):

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile
2 - volatile acidity: the amount of acetic acid in wine
3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
4 - residual sugar: the amount of sugar remaining after fermentation stops
5 – chlorides: the amount of salt in the wine
6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 and bisulfite ion
7 - total sulfur dioxide: amount of free and bound forms of S02
8 – density: the density of water is close to that of water depending on the percent alcohol and sugar content
9 – pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels
11 – alcohol: the percent alcohol content of the wine

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

III. Analyze dataset with correlation

First of all, we analyze for any variables and plot the distribution of each of the variable to see how there are relationship between properties.

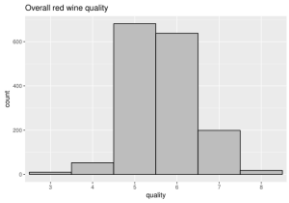

Figure 1 Overall red wine quality

From graph above we can see that most of quality of wine are in score 5 and 6 respectively.

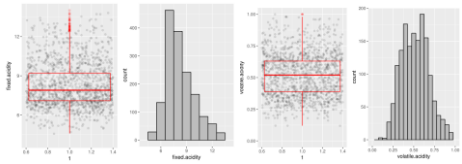

Figure 2 Fixed Acidity (left) Volatile Acidity (right)

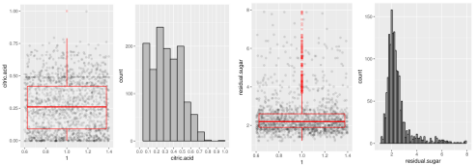

Figure 3 Citric Acid (left) Residual Sugar (right)

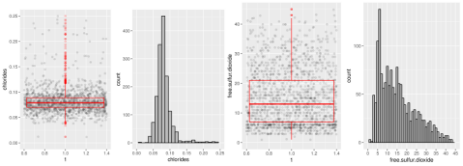

Figure 4 Chlorides (left) Free Sulfur Dioxide (right)

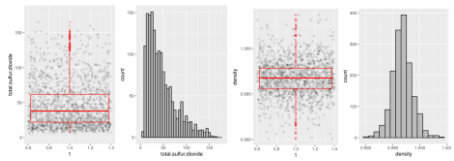

Figure 5 Total Sulfur Dioxide (left) Density (right)

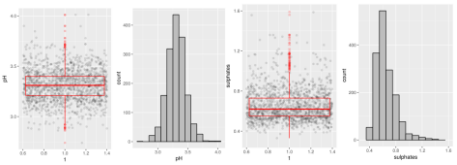

Figure 6 pH (left) Sulphates (right)

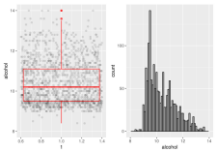

Figure 7 Alcohol

After we visualize graph from plot of distribution we can see that there are mix of skew and normal distribution. For example, fixed acidity, residual sugar, chlorides, total sulfur dioxide, sulphates and alcohol have skew distribution but volatile acidity, citric acid and pH are similarly with normal distribution. So, we couldn't find any relation for each chemical propertie.

We applied correlation between each chemical and quality and we obtained result as table below.

Correlation with quality	Result value
Fixed Acidity	0.1240516
Volatile Acidity	-0.3905578
Citric Acid	0.2263725
Residual Sugar	0.01373164
Chlorides	-0.1289066
Free Sulfur Dioxide	-0.05065606
Total Sulfur Dioxide	-0.1851003
Density	-0.1749192
pH	-0.05773139
Sulphates	0.2513971
Alcohol	0.4761663

Figure 8 Result for each correlation between feature and quality

In figure 8 shown the result after we applied correlation, we have four strong correletion value that could have signification with quality of wine are Alcohol, Sulphates, Ctiric Acid and Volatile Acidity respectively. However, for Volatile Acidity, the result give a negative value that maybe if increase volatile acidity, wine quality will degrade. In the other hand, for Alcohol, Sulpahtes and Citric Acid, the result give a positive value so we will get a better quality if have strong these chemicals. For visualize these result we can plot graph as below.

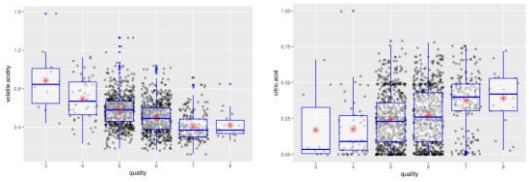

Figure 9 Volatile Acidity vs quality (left) Citric Acid vs quality (right)

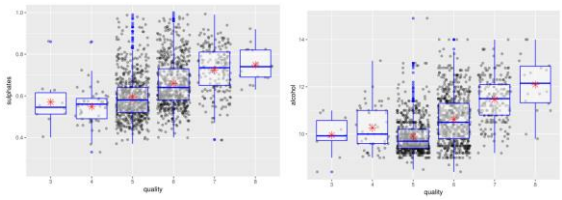

Figure 10 Sulphates vs quality (left) Alcohol vs quality (right)

In summary, as we can visual analyze from graph between each four chemical and quality. They are relevant with correlation result, that wine quality will degrade if volatile acidity increases and better quality if strong citric acid, sulpahtes and alcohol. However, we still can see some fluctuate value in boxplot, for example in alcohol, overall is wine quality increase if alcohol increase but in quality in score 4 and 5 should continue but from boxplot in score 5 is less that 4 and continue increase.

So, we still not guarantee that quality will degrade if volatile acidity increases and better quality if strong citric acid, sulpahtes and alcohol. We need to applied linear fitted model to see a slope.

IV. Analyze dataset with linear model

We attempt to know which feature from four above that has the most effect with quality of wine respectively. So, we applied linear model and give a result as below.

```
Call:
lm(formula = as.numeric(quality) ~ alcohol + sulphates + volatile.acidity +
    citric.acid, data = dat)

Coefficients:
     (Intercept)           alcohol         sulphates  volatile.acidity       citric.acid
         2.64592           0.30908           0.69552          -1.26506          -0.07913
```

(Intercept)	**2.64592**
Alcohol	0.30908
Sulphates	0.69552
Volatile Acidity	-1.26506
Citric Acid	-0.07913

Figure 11 Result from applied linear model

So, from linear modelling result above we can see that alcohol and sulphates have positive slope so high alcohol and sulphate content seems to produce better wines. In contrast, citric acid has weakly correlated plays a part in improving the wine quality. But volatile acidity has a major role to degrade a wine quality.

In addition, we applied linear modelling from alcohol and sulphates to verified with wine quality that strong alcohol and sulphates will increase wine quality.

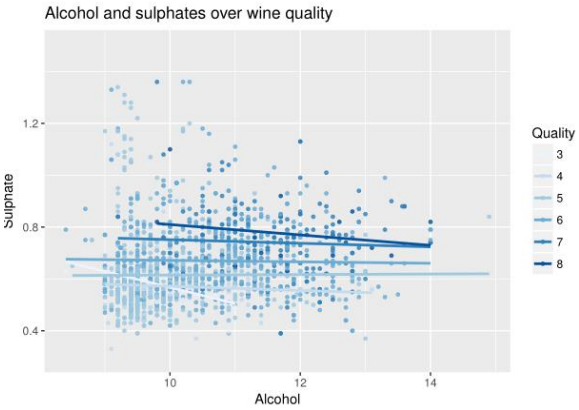

Figure 12 Alcohol and sulphates over wine quality

In this plot, we see that the best quality wine have high values for both Alcohol percentage and Sulphate concentration implying that High alcohol contents and high sulphates concentrations together seem to produce better wines.

V. Analyze dataset with random forest

We still attempt to know which one has the most effect between alcohol and sulphates. So, we applied random forest which base on builds multiple trees and using bootstrap and then takes the average of all the trees to arrive at the final model. This works by reducing the amount of correlation between trees, and thus helping reduce the variance of the final tree which feature has the most effect to quality of wine.

We build 500 random decision in random forest and give a result as below.

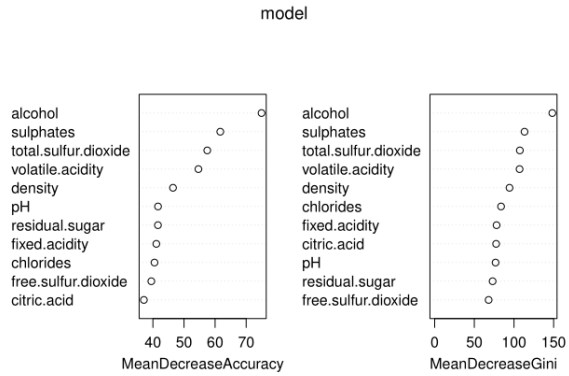

Figure 13 Result from random forest

From random forest model we can see that alcohol has more influence with wine quality than sulphates which the result relates with correlation and linear model too. However, from random forest we can see that total sulfur dioxide

and volatile acidity have influence with wine quality which we could not see there has strong correlation between total sulfur dioxide in and quality more than citric acid.

So, we applied predict model by separate dataset into two sets are train data and test data with probability 0.8 and see the result which one will give the better accuracy.

VI. Build a linear model and predict with test data. Calculate accuracy.

After we separate dataset into two groups with probability 0.8 and build a linear model with train data with all chemical properties. We obtain a predict value and calculate with expected value also find accuracy which given the result as below.

Number of Dataset	317
Number of match data	5
Accuracy	1.577287%

Figure 14 Accuracy for linear model

Because of dataset is non-linear but we try build a linear model so it gives a less accuracy with 1.57%. We can plot residuals and fitted value to guarantee that we have regression with relationship non-linear model as below.

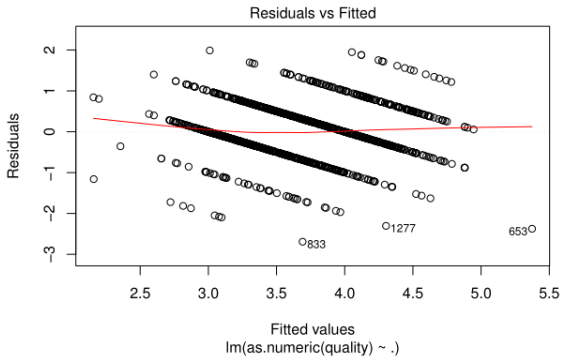

Figure 15 Residuals vs Fitted for quality and all chemicals

We can see that our plot from residuals and fitted value for quality and all chemicals properties did not come along with red line which it was not linear model. And we should not use linear model to predict for wine quality data.

VII. Build a regression model and predict with test data. Calculate accuracy.

We use the same train data that we use in linear model. We build with 700 random decision in random forest. We obtain a predict value and calculate with expected value also find accuracy which given the result as below.

Number of Dataset	317
Number of match data	227
Accuracy	71.60883%

Figure 16 Accuracy for 700 random decision in random forest

From this result we can see that random forest give a better predict than linear model because we build a different decision tree and it better with non-linear model.

However, we can improve random forest by increase number of tress which we adjust to 1000 random decision in random forest and given a result as below.

Number of Dataset	317
Number of match data	229
Accuracy	72.23975%

Figure 17 Accuracy for 700 random decision in random forest

From this result, number of match data between predict and expect has increase by two which give more a little accuracy.

In addition, after we plot the result from 1000 random decision in random forest, we can see that volatile acidity has more significant than total sulfur dioxide which similar with correlation and linear model from above.

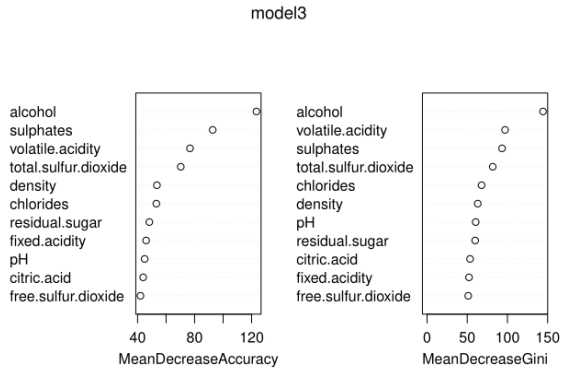

Figure 18 random forest with 1000 decision trees

VIII. Summary

In summary, we attempt to find which chemical properties has the most influence the quality of red wines. First of all, we plot a distribution for each feature but we have no clue because the result gives skew and normal distribution without any relation. So, we applied correlation between chemical properties and wine quality which given four strong correlation are Alcohol, Sulphates, Volatile Acidity and Citric Acid. In addition, Alcohol, Sulphates and Citric Acid has positive correlation which mean if increase these chemicals will give better wine quality. In contrast, wine quality will degrade if Volatile Acidity increase. And after we tried to use boxplot to guarantee this assumption we found that some value has fluctuate, so we need to use linear model to analyze. Result form linear model given a positive slope only Alcohol and Sulphates, so we can conclude that these two chemicals are influence with wine quality. However, we still did not know which one has more significant, so we applied random forest to determined the result. After we applied random forest with 500 random decision trees we found that alcohol has the most influence with wine quality and sulphates, total sulfur dioxide and volatile acidity respectively. Unfortunately, we did not conclude total sulfur dioxide from correlation and linear model. So, we still wonder which model should be better between linear model and random forest. We separate a dataset into two groups are train and test data and create model. From linear model, the result give accuracy with 1~2% but random forest gives 71~72% and will increase accuracy if we add more number of decision trees.

Conclusion, the random forest give the best predict model and Alcohol is the most influence with wine quality. Sulphates has the less significant than Alcohol for effect with wine quality. However, if we increase chemicals both of them the wine quality will better.

IX. Reference

[1] Dataset

https://onlinecourses.science.psu.edu/stat857/node/223

[2] Random forest

https://www.r-bloggers.com/predicting-wine-quality-using-random-forests/