Red Wine Quality

Applied Data Science Project

Mary Damilola Aiyetigbo

1 Introduction

For this project, I'm analyzing wine dataset gotten from the UCI Machine Learning repository. Quality evaluation is often part of the wine certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices). Wine certification is generally assessed by physicochemical features and sensory tests [1]. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values[2]. I want to focus on applying machine learning model to analyse the effect of the wine features in determining the quality of wine. The primary goal is to explore the dataset using different models and comparing the results to determine which model performs best in analysing the wine data. Supervised learning approach will be used for this analysis because the dataset contains a labeled output variable ”Quality”. The response variable is a numeric data, hence I will be using regression models to predict how the values of each input features affects the quality of a wine. Also, because quality is a score between 0 and 10, I felt that I might be making a lot of assumptions if I decide to use classification model to determine the cutoff point to categorize this variable into ”low” and ”high” quality

The objectives of this project are as follows

1. To experiment with different regression models to determine which model fits best
2. To answer this question; How does the values of the physicochemical features in wine affect the quality of a wine?

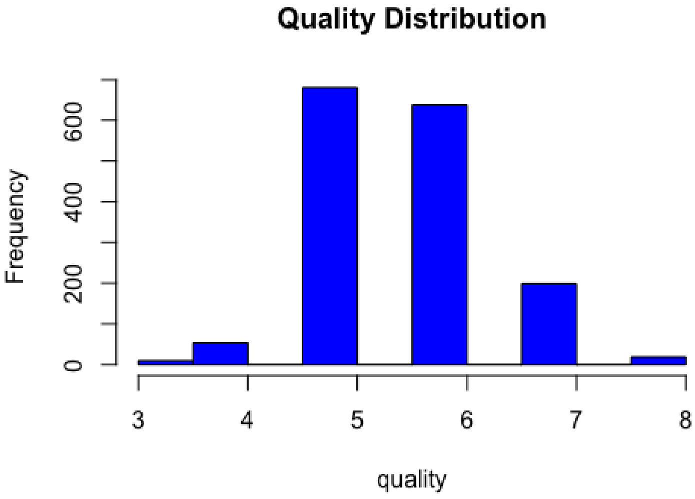
2 DataSet

The dataset used for the project is the red wine quality dataset gotten from `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`. The repository contains dataset for both white wine and Red wine which was collected in October 2009. I will be using the Red Wine Dataset for the purpose of this project. The dataset contains 1599 observations and 12 features. Each observation in this dataset is given a “quality” score between 0 and 10. The dataset contains 11 input variables (based on physicochemical tests) listed below to determine the quality of a wine:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulfates
11. Alcohol

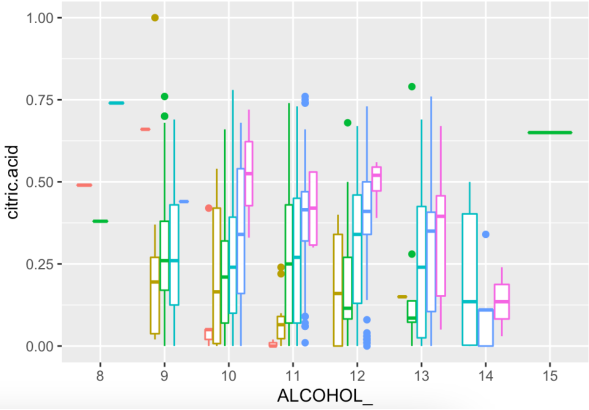
2.1 Data Visualisation

I used the histogram in the figure below to visualise the distribution of the quality variable.

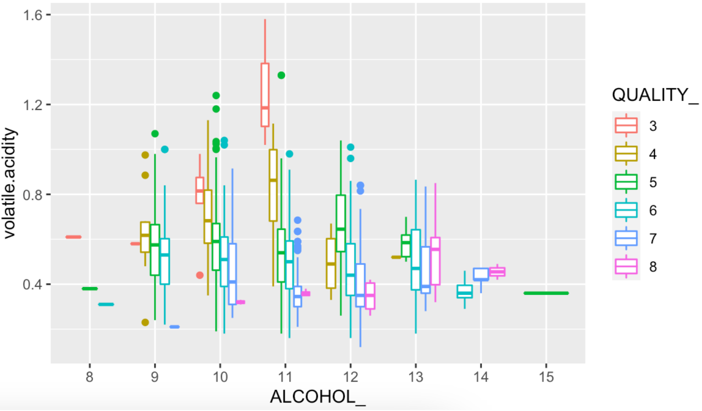


From the histogram above, the relationship shows a normal distribution and it was observed that a large amount of the observation have quality of between 5 and 6 which makes up of about 80% of the dataset.

I also used plots below to understand relationships between quality and some predictors. The first image shows that for wines with an alcohol level below 14, there is an increase in quality as the level of citric acid increases. While for the second image, quality of wine increases with low volatile acidity level and increase in alcohol level.



(a) Citric Acid + Alcohol



(b) Volatile Acidity + Alcohol

3 Model Selection and Validation

Regression model was used for the data analysis of this project because the purpose was to determine the quality of a wine and how it is affected by the physiochemical input features that make up the dataset. Three regression models were used for this project which are; Linear Regression, Random Forest and Support vector Machine (SVM). The dataset was split into training and test set with 70% of the data being used for training while 30% was used for test data. Mean Square Error (MSE) was used as a metric to evaluate the performance of each regression model and how well each model fits the data.

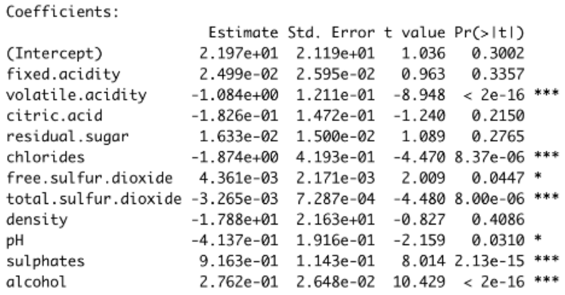
Model 1: I used Multi-Linear regression to build an optimal model for prediction of red wine quality and K-fold cross validation was used for the model assessment. After testing with 5 and 10 values of K, 5-fold cross validation was used because there no significant difference between the result of 5-fold and 10-fold cross validation. The image (a) below shows the correlation of each predictor with the response variable. Alcohol, sulphates, free sulphur dioxide and fixed acidity have positive relationship with quality while the others have negative relationship with quality. It also shows that alcohol, sulphates, free sulphur dioxide, total sulphur dioxide, chlorides and volatile acidity are statistically significant with quality. The 5-fold cross validation gave an MSE of 0.4217

Model 2: Support Vector Machine (SVM) model was used to predict quality in wine on the training data. 10-Fold cross validation and radial kernel was used for SVM using the function "tune" where the parameters with cost=1 and gamma=0.5 gave the best performance model. MSE gotten from the prediction on the test data for SVM model was 0.4346

Model 3: Random Forest regression tree was used as the third model for this project. This helps to create a random sample of multiple regression decision trees and merging them to obtain a more stable and accurate prediction. I decided to use the square root of the number of predictor (m=sqrt(P)) so as to achieve higher model accuracy by reducing the number of predictors used at each split. Random Forest gave an error rate of 0.3603 on the test data.

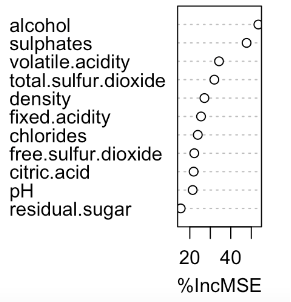
3.1 Validation

After running the three models, I used mean square error as a metric to evaluate my model prediction performance. I got an MSE of 0.04217 for linear regression while support vector machine gave and MSE of 0.4346 and random forest had an MSE of 0.3603 which is the lowest. As we expected, random forest gave the best model fit. I guess this is because it has higher bias during training due to the reduced number of predictors at split and this makes it to produce low variance on test data.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2.197e+01  2.119e+01   1.036   0.3002    
fixed.acidity        2.499e-02  2.595e-02   0.963   0.3357    
volatile.acidity    -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid         -1.826e-01  1.472e-01  -1.240   0.2150    
residual.sugar       1.633e-02  1.500e-02   1.089   0.2765    
chlorides           -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide  4.361e-03  2.171e-03   2.009   0.0447 *  
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density             -1.788e+01  2.163e+01  -0.827   0.4086    
pH                  -4.137e-01  1.916e-01  -2.159   0.0310 *  
sulphates            9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol              2.762e-01  2.648e-02  10.429  < 2e-16 ***
```

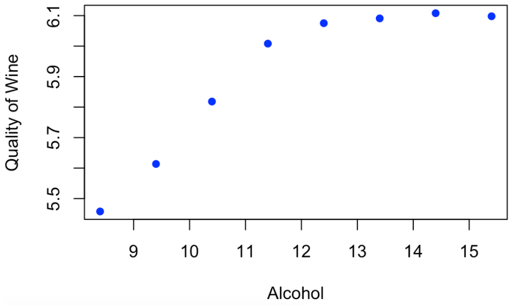
(a) Coefficient of Variables



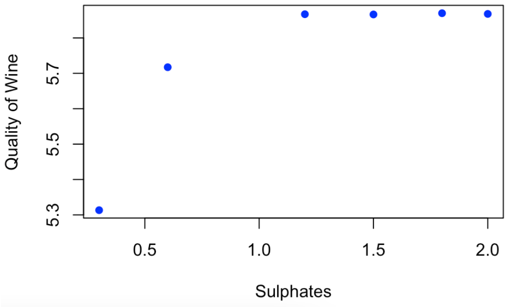
(b) Important Features

4 Model Interpretation

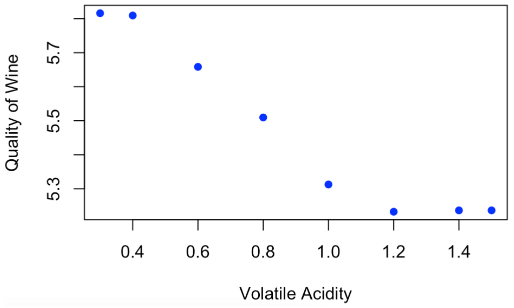
For the purpose of the project, random forest was the best model because it had the lowest test error rate and fits the data better. I also checked the top predictors that are important to the model. It showed that alcohol, sulphates, total sulphur dioxide and volatile acidity are the top predictors which is similar to the significant predictors gotten with linear regression. I also made predictions with some cases to check how increase in the values of these features affects the quality of wine. Figures (a) and (b) below shows respectively that increase in alcohol and sulphate level of wine increases the quality of the wine. While figures (c) and (d) shows that increase in Volatile acidity and Chloride decreases the quality level. This is expected because amount of chlorine for example is an indicator of saltiness and the right proportion is required while volatile acidity is the presence of acetic acid bacteria which a high proportion can cause the wine to have unpleasant taste and smell.



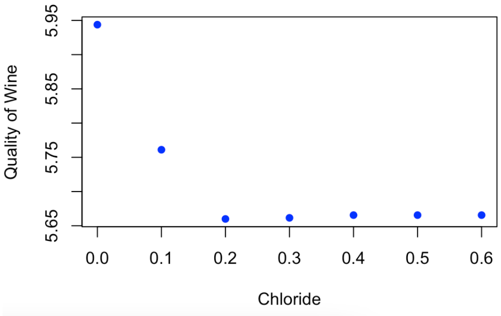
(a) Prediction for Alcohol



(b) Prediction for Sulphate



(c) Prediction for Volatile Acidity



(d) Prediction for Chloride

By analyzing the physicochemical features of red wines, I was able to create a model that can help industry producers, distributors, and businesses to predict the quality of red wine products and have a better understanding of how each feature will impact the red wine quality. Random Forest model outperformed other models used and I was able to identify four important features (volatile acidity, total sulfur dioxide, sulphates, and alcohol) and their effect on quality of wine.

References

[1] S. Ebeler. *Flavor Chemistry - Thirty Years of Progress, chapter Linking flavour chemistry to sensory analysis of win.* Kluwer Academic Publishers, 1999.

[2] 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, 1905.