Predicting Red Wine Quality

by: Eric Dahlberg, Jonathan Hill, and Alex Reichard

Abstract

For our project we will be looking at a kaggle dataset in order to determine red wine quality. It uses the following variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality. The first 11 objective variables will be used as the input for our model to determine the subject score of wine quality on a scale from 1-10. In order to accomplish this, we will do a basic regression analysis using the tensor flow package for python. Our hypothesis is that these inputs have a major correlation to how they are perceived in a quality test. We will test this by splitting our data to see how accurately we can predict these quality scores. We will also be taking a confidence interval. The data set we will be using is a Portuguese wine data set.

1 Introduction

What makes a wine of high quality? Researchers have tried to answer this question through various studies, often categorizing wine as ”good” or ”bad”. We will build upon this existing data to examine which factors have the greatest impact on the quality score of wine. This study is significant because it can not only help consumers determine which wines are of high quality, but it can also assist winery owners in producing the best possible wine. Knowing the key characteristics that contribute to high-quality wine can be especially valuable for entrepreneurs who are looking to enter the industry without incurring large costs from trial and error. Our unique contribution to this research is to evaluate the accuracy of various predictive models in determining wine quality. This is important because it will show which model is most effective for this particular data set. In general, our experiment aims to use empirical data to assign an objective score for quality. This is a valuable ability because it allows us to quantitatively assess objective values.

2 Past Work

An article published in ”Past Data Science” [3] used the same dataset as us to predict wine quality from the features, and compared the performance of five different quantization models: decision trees, random forest, adaboost, gradient boosting, and xgboost. Decision trees are the most basic of these models, in which the data is compared in a tree-like structure and the leaves represent the categorical outcomes. The other four models enhance decision trees to potentially improve the accuracy of the predictions. In this study, the researchers divided the categorical wine quality ratings from 1 to 10 into ”good” quality (scores 7 to 10) and ”bad” quality (scores 1 to 6) to create a binary classification problem. The findings showed that the xgboost and random forest models performed the best, with weighted accuracy around 92%. [3]

3 Data Visualization

The dataset we used has 1599 observations of different red variants of a Portuguese wine. Each observation has 12 variables–fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. For our project, we will focus on quality and its relationship with the other variables to observe which ones have the biggest impact on the quality score given to each observation. To familiarize ourselves with the dataset we first looked at the quality distribution (Figure 1) to see which scores appear the most.

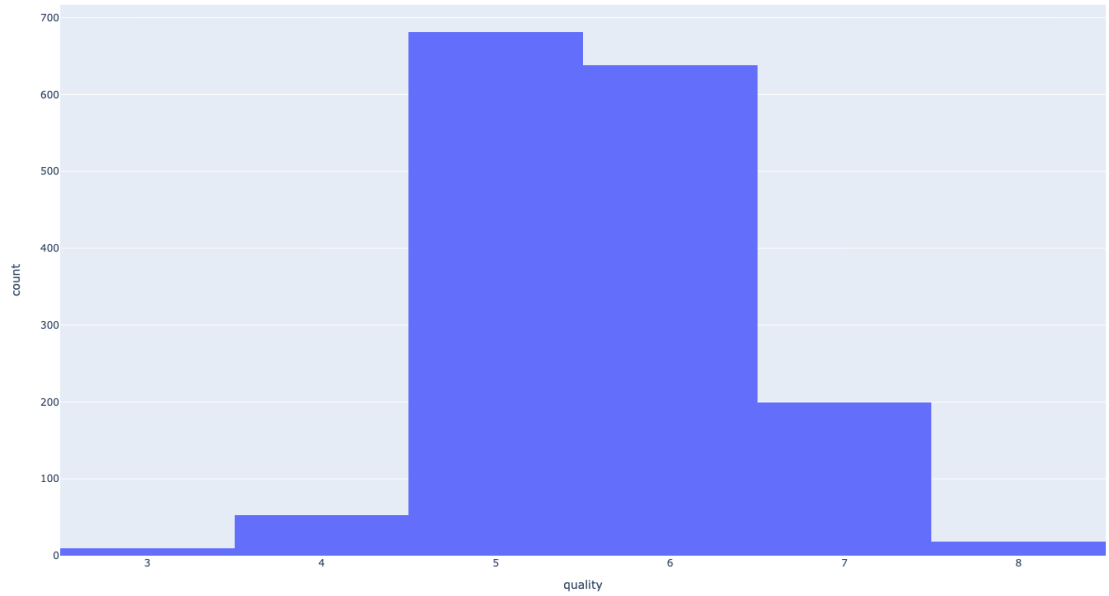

Figure 1: Count of Red Wine Bottle Quality

Quality is is measure on a scale from 1 to 10, with 10 being the highest quality. As you see in the figure above, a quality score of 5 is the most common, followed by a score 6.
Next, to get a sense of which variables could have a significant effect on quality, we created heatmap to show the most important variables when it came to producing good quality wine.

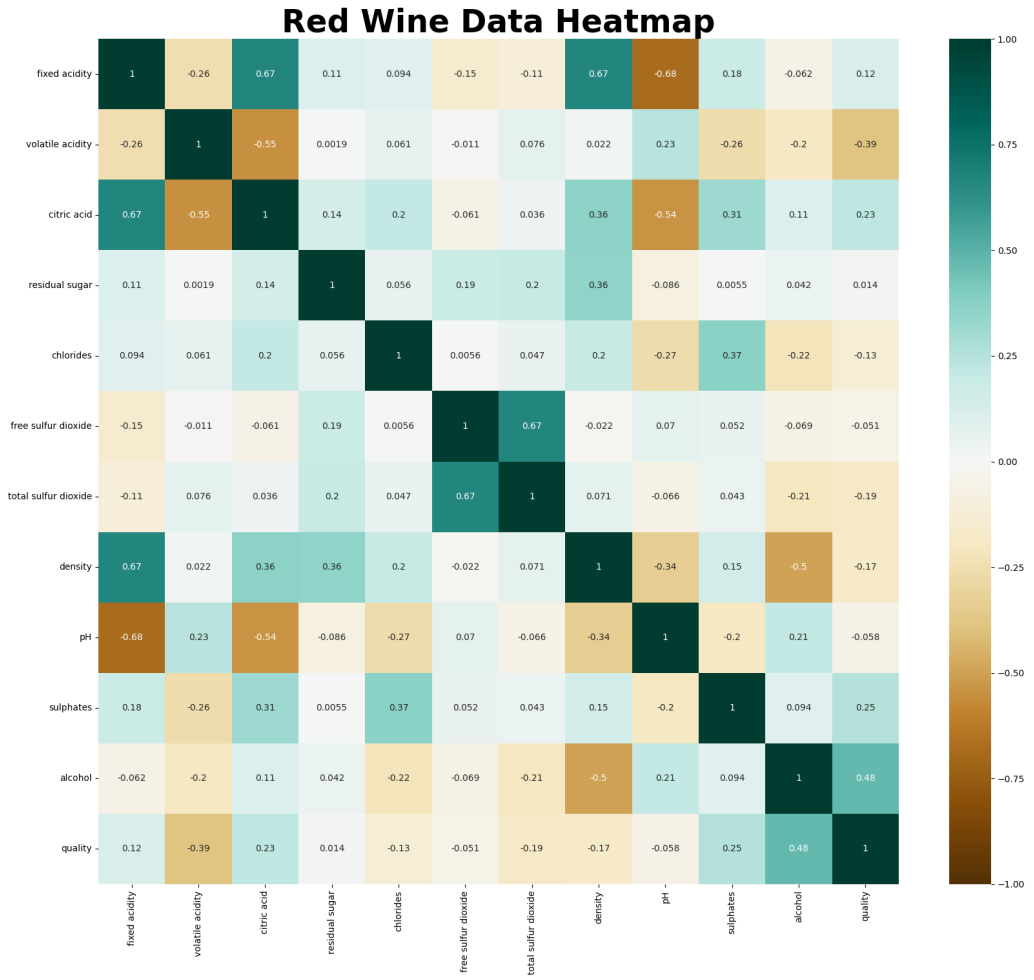

Figure 2: Heatmap of Read Wine Traits

We see from our heatmap (Figure 2), volatile acidity, alcohol, citric acid and sulphates all appear to have significant relationships with quality compared to the other variables. To explore these relationships in more depth, we made graphs of each one to get a better understanding of our data and the influence from each variable on quality. These Results are in figures 3 and 4.

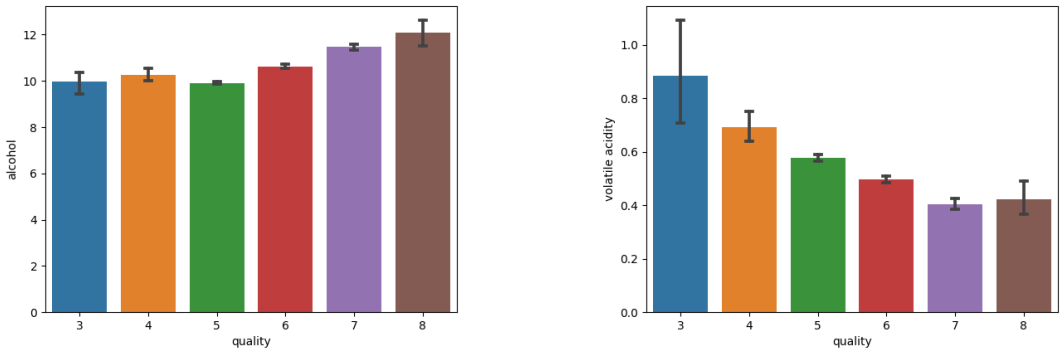

Figure 3: Alcohol and Volatile Acidity vs Quality

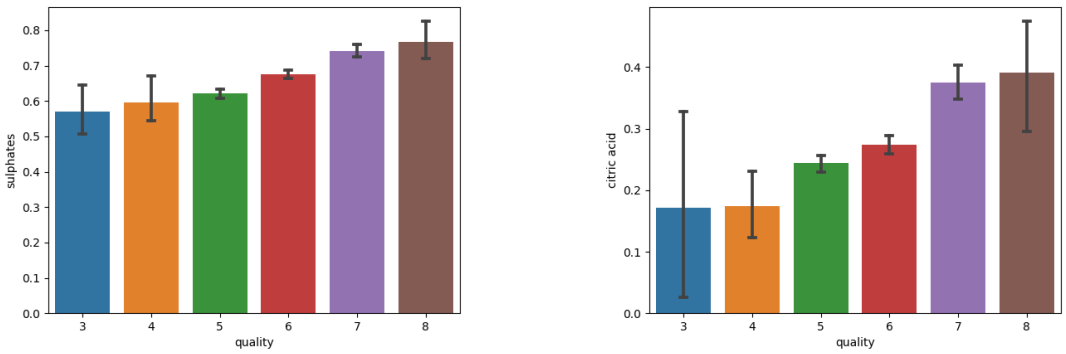

Figure 4: Sulphates and Citric Acid vs Quality

4 Wine quality Data Project

In our work, we aim to extend the existing research on wine quality prediction by attempting to classify the data into specific quality ratings, rather than simply distinguishing between ”good” and ”bad” wines. The split of wine quality at a score of 7, as used in the ”Predicting Wine Quality with Several Classification Techniques” article, appears arbitrary and does not provide a detailed comparison of the relative quality of different wines. To address this, we plan to develop a machine learning system that takes in the 11 features mentioned above and outputs a specific quality rating. To optimize the accuracy of our model, we will compare a number of different techniques and evaluate their performance using the mean absolute standard error score. This score is based off of the equation:

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{A_{t-1}} \tag{1}$$

The mean absolute standard error is a measure of the mean absolute deviation of a set of values from their mean. It is calculated as the mean absolute deviation of the values from their mean, divided by the standard deviation of the values. The lowest output determines the best model that, on average, produces values that are the closest to the real quality scores.

There are two types of learning models we will be testing in our experiments: regression and quantization. Regression models aim to create a line of best fit through the multi-dimensional space defined by our features, with outputs that are continuous and range from 1 to 10 on the quality spectrum. On the other hand, quantization models require outputs to be whole numbers that fall into specific categories.We will be comparing the performance of these two types of models to understand which is most effective for predicting quality in our dataset. We will be using three regression models:

1. Linear Regression: The best fit line is linear, determined by a linear equation.

2. Polynomial regression: The best fit line is polynomial, determined by a polynomial equation.

3. Deep learning regression: Uses layers of a neural network to feed our features through layers of the trained model instead of an equation.

We will also have 2 categorical data models:

1. Decision tree: The features run through a decision tree that can result in ten outcomes, each being a different wine quality score.

2. Random forest: The features run through many decision trees, and the quality category is determined by the most frequent outcome from the trees.

Although linear regression can be a very good tool, it may not be perfect for the categorical data, which is the benefit of the decision trees.

5 Prediction algorithms

Before analyzing our data, we standardized all of the feature values using the z-score method. This transformed the values to a scale ranging from -1 to 1, ensuring that the importance of each feature was not influenced by its magnitude. Next, we separated the data into feature values (X) and target values (Y), where the 11 features comprised the X data and the quality ratings were in the Y data. To properly evaluate the success of our model training, we split the data into training and test sets. We allocated 80% of the data for training and 20% for testing, and we randomized the data to eliminate any potential patterns in the distribution. This allowed us to assess the model's performance on unseen data.

We implemented linear and polynomial regression models using the sklearn package, utilizing the support vector regression technique. This method creates a hyperplane that minimizes the distance from all data points, allowing for an accurate estimate of the quality at any combination of feature values. The linear regression model returned a mean absolute standard error of .51, indicating that it was, on average, .45 units away from the true wine quality rating on a scale of 1 to 10. The polynomial model returned a mean absolute standard error of .77. These results indicate that the linear model is more suitable for our data. We also implemented a deep learning regression model, which consisted of one inner layer with 11 nodes and a ReLU activation function. We ran 10 epochs on a batch size of 100 and obtained a mean absolute standard error of .51, similar to the linear model. However, the accuracy of the model stopped improving after around 8 epochs. Additionally, the deep learning model required several hours to run, while the support vector regression model took almost no time to complete.

We also implemented decision tree and random forest models using the sklearn package. On average, the decision tree model returned a mean absolute error score of .45. The random forest model performed better, with a mean absolute standard error of .37. It is worth noting that while standardizing the data had an impact on the performance of the regression models, it did not affect the decision tree and random forest models. This suggests that these models are not sensitive to the scale of the data and do not require standardization.

6 Comparing linear SVR and Random Forest

In our study, the quantified random forest model outperformed the other algorithms. Among the regression models, the linear support vector model was the most effective, given its simplicity. The mean absolute standard error scores for the random forest and linear support vector models were .37 and .51, respectively. In comparison, the mean absolute standard deviation of our data was .69, which would be the result if we simply assigned every wine the average quality rating from our training data. These results demonstrate the usefulness of the random forest and linear support vector models, as they outperform the most basic technique of assigning the average rating to all wines. In the following section, we will compare the outcomes of these two models in more detail.

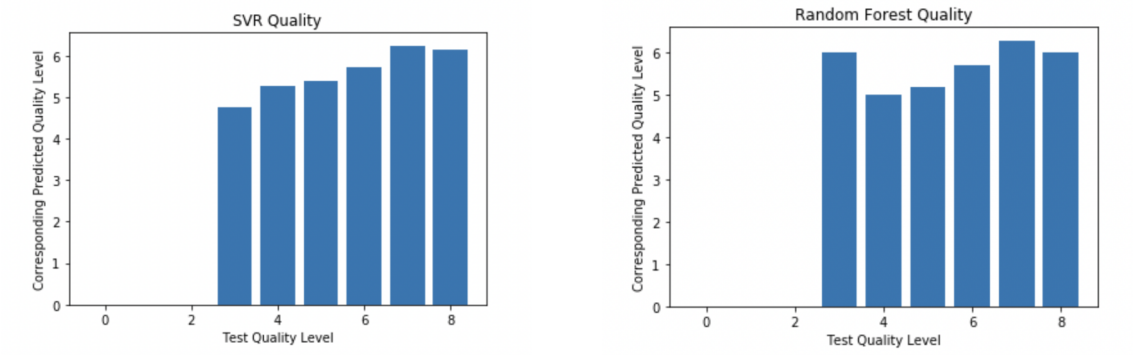

Figure 5:

Figure 5 illustrates the relationship between the average predicted quality ratings and the actual quality ratings of the test data. Both the linear support vector regression and random forest models show a positive correlation between the two, meaning that the predicted quality ratings increase as the actual quality ratings increase. The linear support vector regression model appears to have a more linear relationship between the predicted and actual quality ratings, which is a desirable trait. However, the predicted values are consistently higher than the actual values, leading to an upward skew in the data. For example, wines with an actual quality rating of 3 have a mean predicted quality rating of 5, and this pattern continues for higher quality ratings. On the other hand, the random forest model also exhibits an upward skew, but to a lesser extent. The box plot for this model is less linear than that of the linear support vector regression model. Overall, the random forest model appears to be a better fit for the data due to its lower degree of skewness.

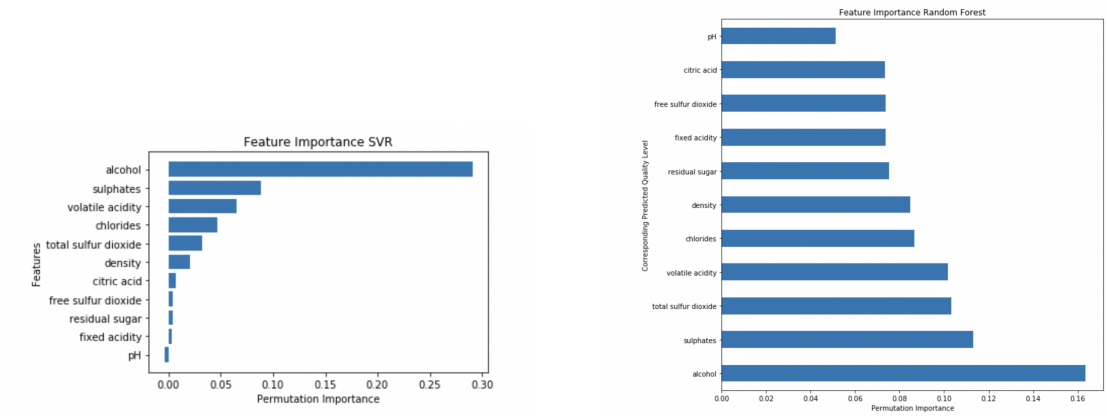

Figure 6:

Figure 6 shows the feature importance of the linear support vector regression and random forest models. Because the data has been standardized, the permutation importance measures indicate the reliance of the algorithms on each feature. In the random forest model (Figure 6b), the algorithm utilizes a number of features to a similar extent. In contrast, the linear support vector regression model (Figure 6a) heavily relies on alcohol content, with only a few other highly correlated features contributing to the prediction. This observation suggests that the random forest model may be more effective due to its use of all features, rather than just the most highly correlated ones. To test this hypothesis, we removed the features with lower correlation from the dataset and found that the accuracy of all models decreased. Overall, these

results suggest that having a larger number of features, even those with low correlation, may improve the performance of machine learning models.

7 Conclusion

As discussed in the above sections, the random forest algorithm is the most accurate model with a mean absolute value of .37. The next best, Decision tree model returned a mean absolute error score of .45. Deep learning regression of .51 mean absolute value score, linear regression a .51 mean absolute value score, and polynomial regression with a .77 mean absolute value score. Most of these performed better than the standard deviation, and it was clear that they were viable models to determine objective data.

References

[1] Vishal Kumar. (September 2017). Prediction of quality of Wine. Retrieved winequality-red.csv https://www.kaggle.com/code/vishalyo990/prediction-of-quality-of-wine/data.

[2] " Wine Quality Data Set." UCI Machine Learning Repository: Wine Quality Data Set, https://archive.ics.uci.edu/ml/datasets/wine+quality.

[3] Predicting Wine Quality with Several Classification Techniques. https://towardsdatascience.com/predicting-wine-quality-with-several-classification-techniques-179038ea6434.