Analysis on Wine Quality

1. Introduction

It has long been said that decent quality wines have this crazy ability to transport you to the time they were made. Of course, not all wines have this ability. There are many wine products that are made in such a way that will remove all but hints of their origin. For some wine drinkers, this is the real difference between a good and a great wine. Out of our shared interest in wine, we decide to explore how we are able to predict human wine taste based on fixed wine attributes.

In this paper, we utilize six machine learning and regression methods to model and predict wine quality. We also analyze the results and figure out the variables that play decisive roles in wine quality. In addition, we comment on the performance of the models and select the models we think that could be further polished and applied into real life scenarios. Our model can not only predict wine quality but also provide more insight into indicators on wine quality for wine sellers.

2. Data

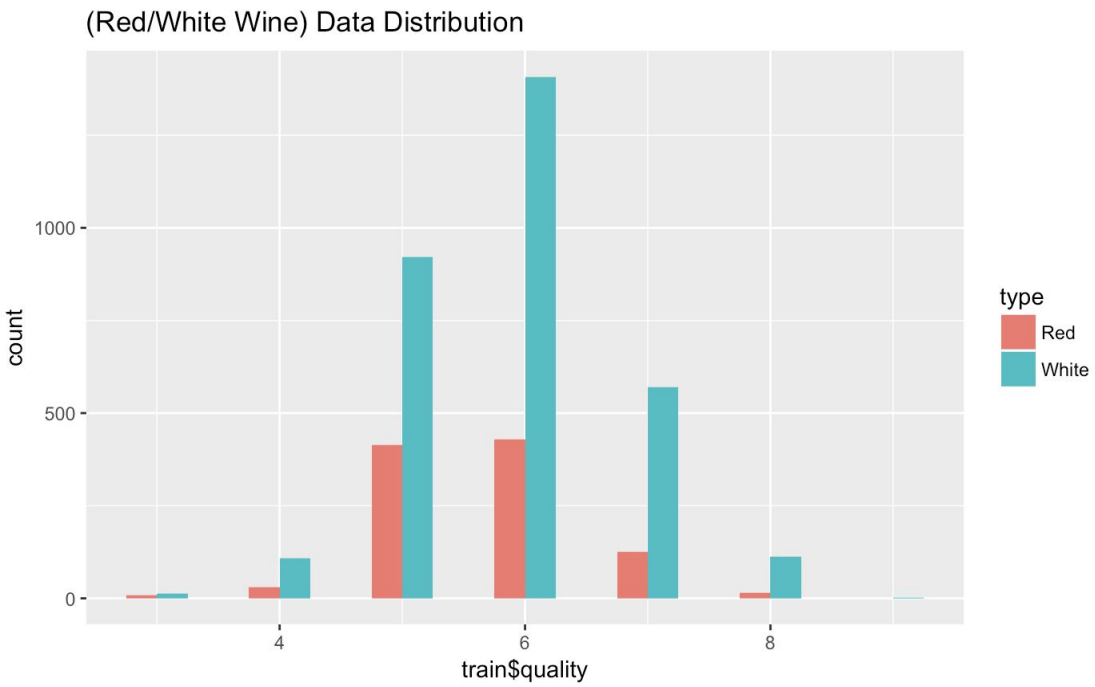
We obtain two datasets on the Portuguese "Vinho Verde" red wine and white wine from UCI Machine Learning Repository[1]. The red wine dataset contains 1599 observations and the white wine dataset contains 4898 observations. Both datasets contain 11 input features: Fixed Acidity, Volatile Acidity, Alcohol, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, PH, Citric Acid, Density, Chlorides, and Residual Sugar. The outcome of these two datasets is the quality, which are classes labeled from 3 to 9 where 3 is the lowest quality and 9 is the highest quality. We combine these two datasets for a quality prediction model, where we have 6497 observations on 11 features as indicated above plus a wine type indicator (red wine or white wine) to predict the quality of wine.

a. Initialize data (density plot)

For model training and selection purpose, we split our data into three parts: training set (3/5 data), testing set (1/5 data), and hidden set (1/5 data). We proceed to model training, testing and selection using the training and testing set. In the last stage, we validate our selected models utilizing the hidden set.

Figure 1 displays the distribution of red wine and white wine from our training set. we can tell that the spread for the quality for both Red and White seems to exhibit two similar normal distributions except for the fact that white wine distribution exhibits a peak quality around quality rating of 6 while red wine exhibits a peak quality rating at around 5. Also, only white wine seems to have been rated with a quality of 9 from the given sample.

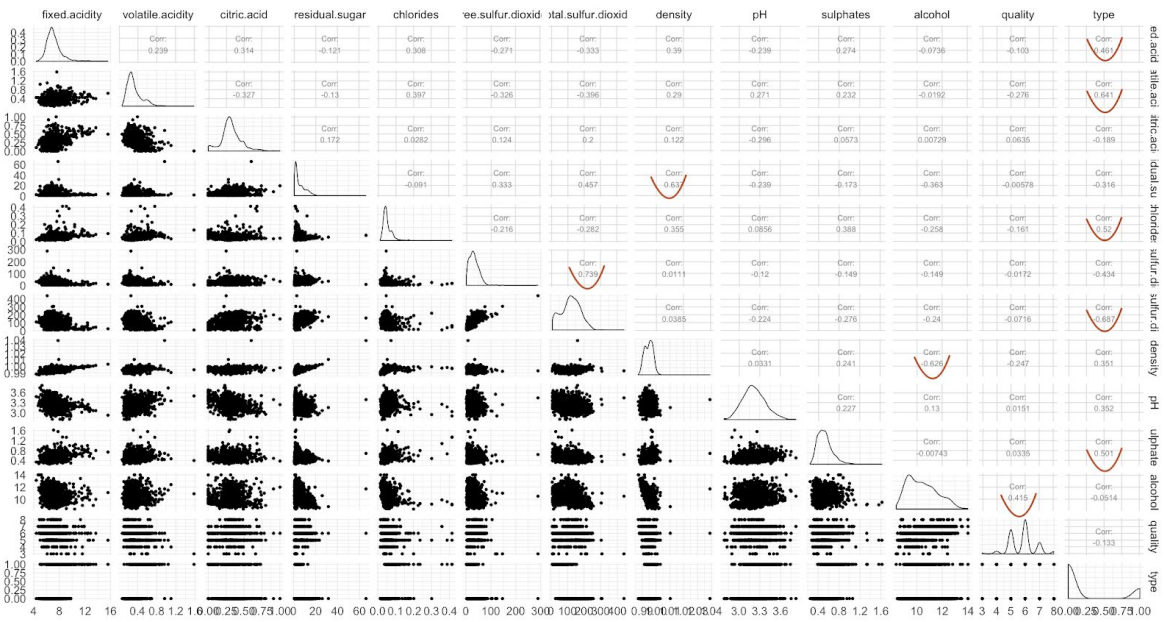
[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.



b. Collinearity

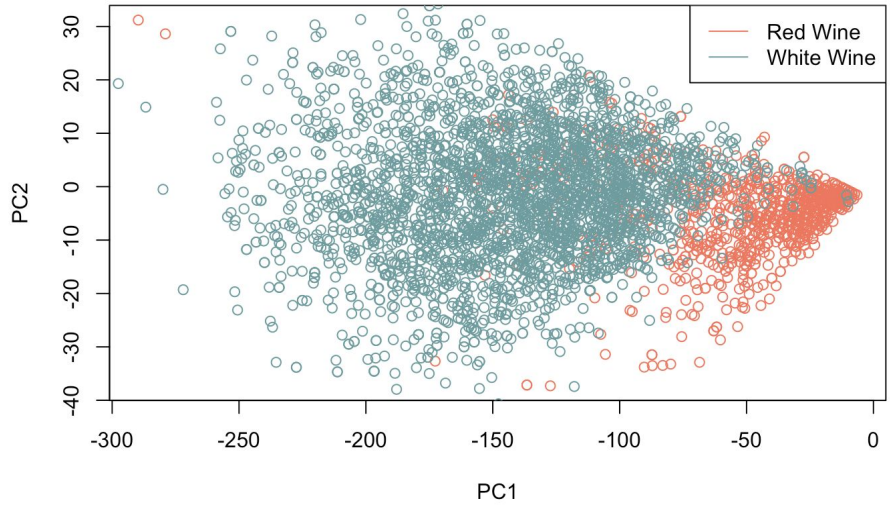
Although multicollinearity has little impact over the behavior of models and associated statistics, it becomes a problem if we want to learn and interpret the effects of how each feature contributes to quality estimation. Therefore, our collinearity analysis would suggest how we regard our predictive models.

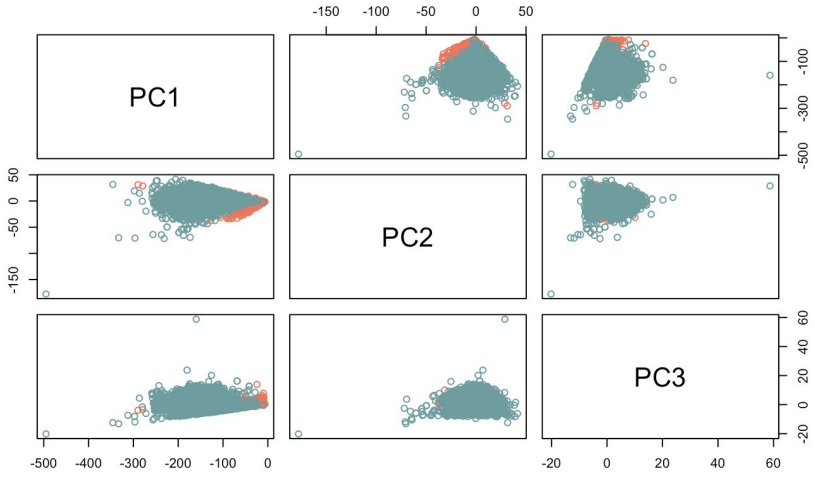
Figure 2 shows a scatterplot matrix of correlations between features. Red circle suggests that the correlation between the two corresponding variables is greater than 0.40. In general, we can tell that type is highly correlated with sulfur and chloride; alcohol is negatively correlated with density; quality is positively correlated with alcohol; alcohol is negatively correlated with density; density is positively correlated with residual sugar; total sulfur dioxide is positively correlated with free sulfur dioxide. These high correlations suggest that regression models can be used for prediction while coefficients are not meaningful.



c. PCA

Figure 3 and Figure 4 display the principal component projections on wine type. Although PC1 and PC2 account for most of the variance in the data, about half of the data points are still overlapped. We also plot the projected data on all paired combinations of first three PC's, in hope that we can find a hyperplane that separates the points linearly, because they explain 99.9% of the variance. Still, the projection on PC1 and PC2 is the best result we have. These plots may suggest that red wine data and white wine data are not linearly separable.





3. Methodology

Our goal is to model and predict wine quality by utilizing machine learning and regression methods including ordered logistic, random forest, k-Nearest neighbor (kNN), extreme gradient boosting (XGBoost), linear regression, and multi-class support vector machine (multi-class SVM). In the section below, we will report the accuracy of each methods and analyze on its performance. We will also validate the model on our hidden set of data.

a. Evaluation

To better evaluate the result, we will assess our models from three aspects. Firstly, we calculate the percentage when the prediction is exactly the same as the actual data (Absolute Accuracy). This is the most straightforward way to evaluate the power of each method. Secondly, we calculate the percentage with a margin of 1 (Marginal Accuracy). In this case, we believe that categorizing a wine of quality 7 as 6 or 8 is still acceptable, so we allow this margin of error when evaluating accuracy. Thirdly, we calculate the Marginal Squared Error (MSE) for each method. The reason we use MSE here is because we want to evaluate how bad the predictions are and magnify the effects of the ones that are more off.

b. Method and Analysis

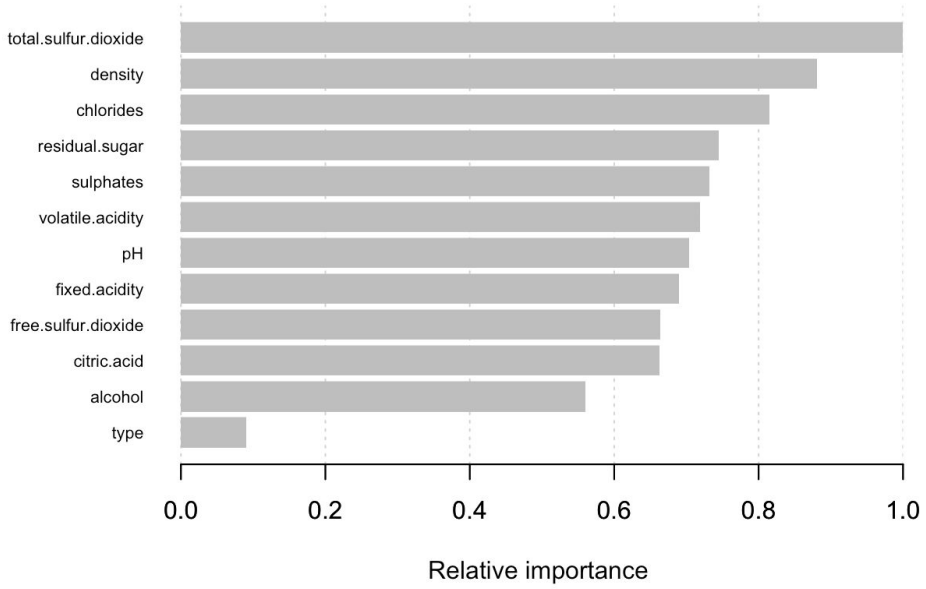
Model	Train			Test		
	Absolute Accuracy	Marginal Accuracy	MSE	Absolute Accuracy	Marginal Accuracy	MSE
Multi-class SVM	0.57	0.96	0.57	0.55	0.95	0.64
XGBoost	0.98	0.99	0.20	0.63	0.95	0.70
Ordered Logistic	0.53	0.95	0.63	0.54	0.94	0.68
Linear Regression	0.52	0.95	0.63	0.43	0.93	0.67
Random Forest	0.93	0.99	0.07	0.65	0.90	0.41
k-Nearest Neighbors	0.50	0.91	4.55	0.46	0.91	4.84

i. Random Forest

Considering the biggest benefit of the tree-based algorithms is that it is very intuitive and can be easily interpreted (by humans) as rules, i.e., wine-makers may get useful suggestions in producing the wine, we apply random forest algorithm firstly. In addition, the tree-based algorithms do not require any assumptions such as linearity and normalization, so it might perform well. To avoid the misleading accuracy problem that imbalanced data might bring, we view our outcome variable as continuous and then build the trees based on the regression function rather than classification function. In parameter estimation, we vary mtry from 1 to 12 and find the minimum test error at 6. As a consequence, the model performs well with the highest absolute accuracy at 0.65 and lowest MSE at 0.41 among all the models, and its marginal accuracy is also acceptable. However, the overly perfect fancy values of the train set indicates there might be overfitting problem.

ii. XGBoost

Being aware that gradient boosting is an advanced tree methods often times pushes classification accuracy forward, we utilize XGBoost to improve the prediction accuracy of classification. Given that we have sufficient amount of observations, we choose the max.depth restriction (from 1 to 12) base on their prediction accuracy on testing set. It turns out that max.depth=12 is selected for our purpose, where XGBoost model returns high marginal accuracy at 0.95 as well as exact accuracy at 0.63 on testing data. However, the fitting on the training set with a marginal accuracy at 0.99 and exact accuracy at 0.98 suggests that XGBoost is prone to overfitting when estimating exact accuracy. This result suggests that our model relies too much on training set information that is not well generalized. Therefore, our model may suffer from the risk of high variance in predicting exact accuracy. Therefore the XGBoost model seems reliable in predicting marginal wine quality, while potentially suffering from the risk of high variance in predicting exact wine quality. Figure X displays the relative importance of each feature, suggesting that the effects of all features except type indicator is evenly distributed while type is seemingly not as helpful in this XGBoost model.

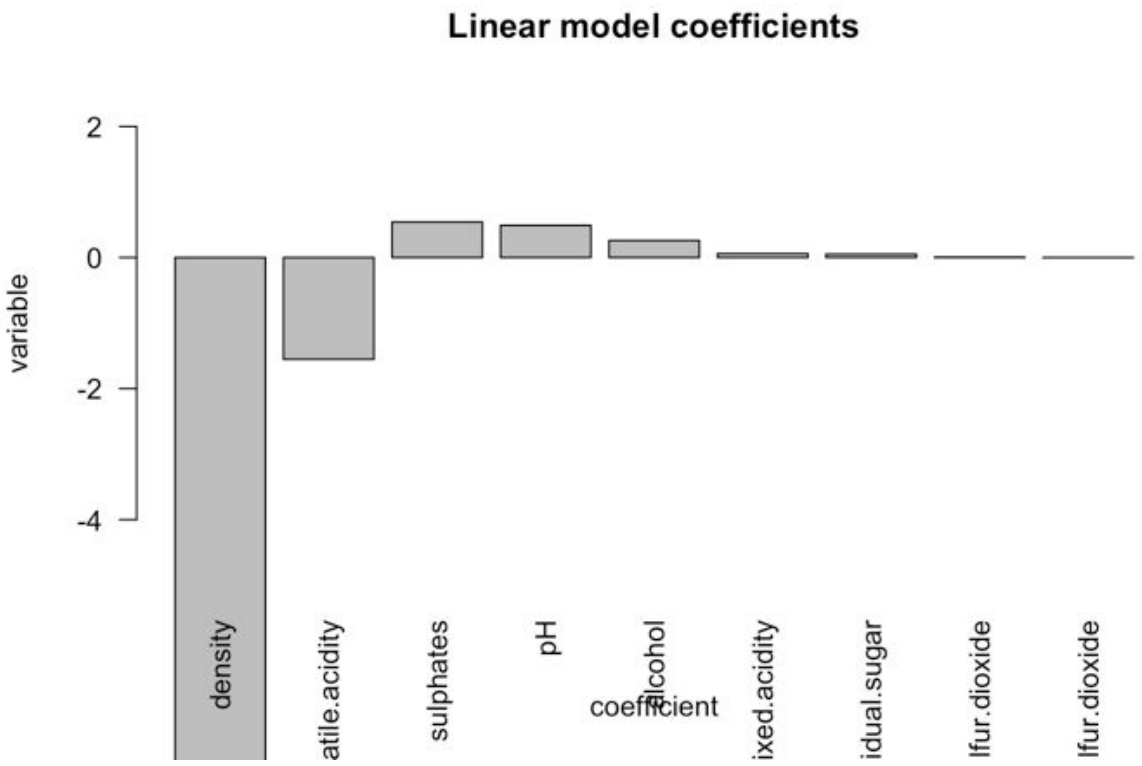


iii. Multi-class SVM

Considering that the data may not be linearly separable, we think SVM would be a good model to try because it not only performs well on binary outcome classification but also is able to deal with non-linear separable data with RBF kernel. To make use of its strong classifying ability, we implement a multi-class SVM with the help of the e1071 package in R. We build 1 SVM classifier for every 2 different outcomes. Because our outcome variable ranges from 3 to 9, we have 21 binary classifiers in total. Then we apply all these classifiers to the testing data and obtain 21 classification results for each observation. Finally, a majority voting is conducted to select final results.

The absolute accuracy is not as high as the results from random forest, but the marginal accuracy is at a satisfactory level. We believe that there is no sign of overfitting because the accuracy given by training data and testing data are at a similar level. However, the model loses its interpretability because we use a non-linear kernel and a bagged result. Overall, SVM can be a credible model for our dataset.

iv. Linear regression



We try to fit the data in a linear regression for the sake of prediction because our outcome variable is not simply binary but can be viewed as continuous in the sense of integers. The model performs surprisingly well. Marginal accuracy in both training and testing sets is higher than 95%, although the absolute accuracy in the test set is far below SVM. We plot the significant coefficients by their magnitudes and try to figure out the impact of each variable on the result. It shows that density has a much greater negative coefficient comparing to all other variables. It is followed by volatile acidity, sulphate and pH. The coefficients may not be very trustworthy due to the collinearity in our data, so we would not make conclusion on which variables are making more impact based on this. The prediction results are still considered to be reliable nonetheless.

v. Ordered logistic

Ordered logistic regression is worth trying out because we have an ordered classification problem. It shows a strong predicting power and the marginal accuracy comes out at 95%. Although we can interpret the odd ratio, we decide not to do so because of one special assumption associated with the ordinal logistic regression, the proportional odds assumption. The proportional odds assumption assumes that each variable is posing same effect on the odd ratio regardless of the order they are at. We are not sure if this assumption is fulfilled in our case. If some variables are serving as the threshold setting features, then it is very likely that the proportional odds assumption is violated as well. To further explain on this, for example, if any wine with a density lower than 10 is guaranteed to receive a quality classification of 6 or above, but it is no longer a significant factor for further quality determination, then it is very likely that density does not have the same effect on the outcome in every order. Hence, ordered logistic model will not be chosen for this dataset.

vi. KNN

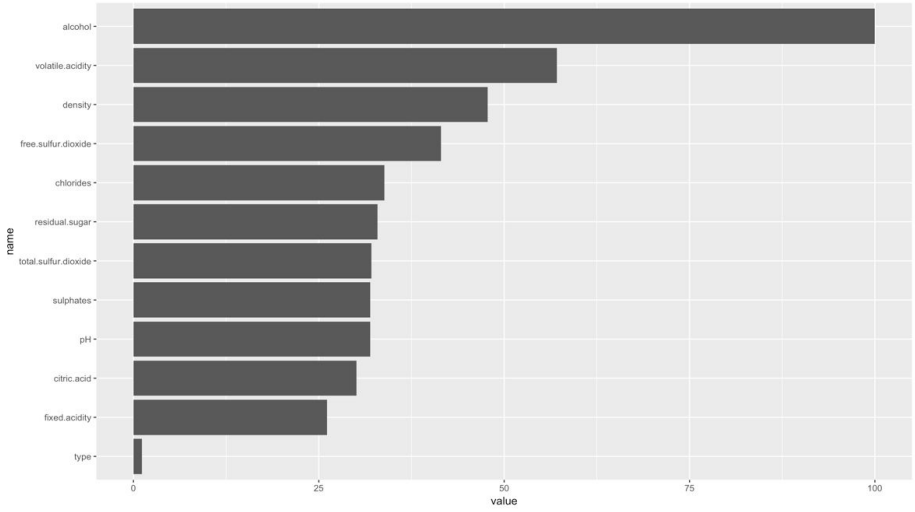
Our sufficient training set at low dimension (12 features) suggests that KNN can be a reasonably simple and interpretable method to use. The distance of every feature is normalized. We choose the function input k between 1 and 63 that returns the highest training accuracy. It turns out that k=38 nearest neighborhood (in Euclidean distance) is chosen for our model. It follows that we predict the quality of each new observation by averaging the quality of 38 nearest neighbor points around the new data. As expected, the result is pretty good, where the marginal accuracy of testing set is 0.91 while the exact accuracy of testing is 0.46. The prediction accuracies on training set and testing set are similar, with no obvious effect of overfitting. Our potential concern may be that our chosen k is large such that our model with such high bias may under-fit new data.

4. Interpretation

a. Variable importance

Here we plot the "Global variable importance", which is the mean decrease of accuracy in Gini index over all out-of-bag cross-validated predictions in Random Forest model. GINI importance measures the average gain of purity by splits of a given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes.

Alcohol is the most important variables here indicating it has a huge impact on the human wine taste and even slight changes will largely affect the wine quality. While it is interesting that the importance of the type of wines are quite low suggesting that the rules to measure the quality of red wine and white wine are essentially same.



b. MSE comparison

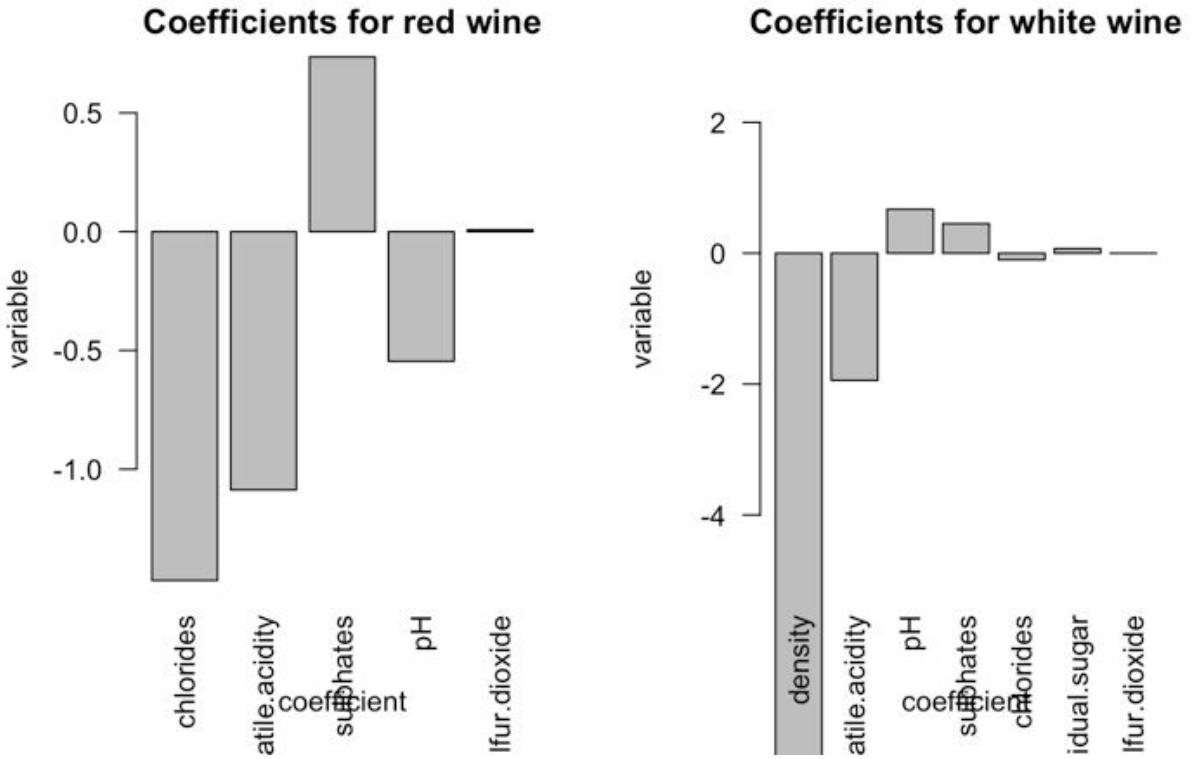
MSE values here enable us to view the model from a new angle: instead of considering how accurate the predictions are, we can now think about how off those wrongly classified results are. If we look at the methods that work well here, which are multi-class SVM, random forest, XGBoost and linear regression, we may notice that a higher accuracy does not grant a smaller MSE. XGBoost and random forest are the models that differ in MSE the most. In XGBoost, about 63% of the predictions are accurate and an additional 32% of the data contributes to MSE by 1. In random forest, about 65% of the predictions are accurate and an additional 25% of the data contributes to MSE by 1. About 5% of the data in XGBoost and 10% of the data in random forest

are contributing to MSE by 4 or more, However, the MSE of XGBoost is twice as much as the MSE of random forest. It implies that although XGBoost gives a higher marginal accuracy, it also gives more extreme error than random forest does when we are considering the wrongly classified observations.

c. Interpretation comparison

The only interpretable results we have are the one from random forest and the one from linear regression, but it is difficult to make any individual conclusion due to collinearity. Therefore, we try to draw some conclusion by comparing them. Both models pick out density, volatile acidity and some sulfur-related variables. Random forest also assigns alcohol a greater importance. It indicates that alcohol, density, volatile acidity and sulfur level may have a relatively higher possibility in affecting the quality of wine.

d. Red/White comparison



We run linear regression and random forest on red wine dataset and white wine dataset separately and want to see if there is any similarity or distinction. Again, random forest shows that alcohol, sulfur-related compounds, volatile acidity and density are the top ones in play. Although the order of variable importance in each dataset changes slightly in general, alcohol is still the most important one. When we look at the results from linear regression, however, the only and the biggest difference is that chloride has the largest effect in magnitude on the red wine dataset. Mapping back to the variable importance graph from random forest, this is indeed the next most important variable other than the ones we mention above. It may imply that chloride level is the biggest difference between the judgment of red wine and white wine.

e. Model validation

Model	Hidden data		
	Absolute Accuracy	Marginal Accuracy	MSE
Multi-class SVM	0.57	0.96	0.58
XGBoost	0.68	0.96	0.45
Linear Regression	0.56	0.96	0.59
Random Forest	0.70	0.91	0.35

To see if it is true the four models we believe have better performance among all the algorithms we apply, we utilize our hidden set of data to validate them. Table 2 shows the results which are consistent with the previous results in the model training process.

5. Conclusion and next step

To sum up, each model has its advantages and its limitation. Random forest gives us highest absolute accuracy, lowest MSE and also preserves the model interpretability. Its marginal accuracy is not as high as XGBoost and multi-class SVM. XGBoost does well on both absolute accuracy and marginal accuracy, but it has the highest MSE among the four models and we cannot make further inference on the variable importance based on its results. Multi-class SVM has a high marginal accuracy and a moderate level of MSE, while its absolute accuracy is far lower than the two models above and the coefficients are not interpretable. Results from linear regression are impressive considering it is the simplest model, although its interpretation is susceptible to collinearity.

a. Real Life Interpretation

Variable	Interpretation
Alcohol	Alcohol level
Volatile acidity	May lead to wine fault
Density	Clarity
Free sulfur dioxides	Preservative
Chlorides	Terroir of the grapes
Residual sugar	Sweetness
Total sulfur dioxides	Preservative
Sulphates	Preservative

pH	Strength of acidity
Citric acid	Acidification of wine when it is not acidic enough
Fixed acidity	Major type of acid in wine
Type	Whether it is red wine or white wine

The variables in the interpretation table are listed according to the variable importance given out by random forest. Based on the conclusion we reach in the previous sections, we think that the alcohol level, clarity, amount of preservatives and the freshness of the taste may be deciding the quality of the wine. As for the difference between red wine and white wine, it seems that red wine lovers are more serious about the terroir and origin of the grapes.

b. Real Life Application

We believe that the results of our study can serve as an alternative indicator on wine quality for the wine sellers. Instead of spending huge money on hiring the professionals, the wine sellers can get the evaluations that are very similar to the professionals' opinions using our models with certain input data given. Knowing which factors lead to a wine with higher quality also benefits the wineries. Other than following the traditional process to make a good wine, they can try to further improve the taste by controlling the level of each chemicals in their wine.