Zhuo Zhou, Shiqi Yu, Yueran Yang

STSCI 5740 Final Project

20 November 2021

Statistical Learning for Wine Quality Prediction

1. Introduction

Vinho Verde is a unique product of the Minho region of Portugal. It has a moderate alcohol content and is particularly appreciated for its freshness. The dataset contains various features for the attributes of wine, and the output variable is ordinal data(quality). Therefore, it can be used by public research for regression and classification tests. We have two types of wine in the dataset which are red wine and white wine. In this report, we are going to fit multiple regression and classification methods in this multi-class dataset and compare them.

2. Analysis

2.1 Data Exploration

The wine quality dataset contains 12 input variables with 6497 entries. The first column is a categorical variable with two types of wine: red and white. All of the other input variables are numerical attributes of wine based on physicochemical tests. The output variable “quality” is based on sensory data scored between 0 and 10. The whole dataset is complete with no missing values.

2.2 Explanatory Visualization

Below(Figure 2.1 and Figure 2.2) are the histogram and boxplot of each variable, showing how they are distributed in the dataset. This is helpful for finding outliers and unrealistic values in the dataset, as well as how balanced the distribution of different classes is. There seem to be few outliers in each boxplot; however, we assumed that every data point in this dataset is meaningful due to the lack of observations(only 6497 observations), and we do not remove them.

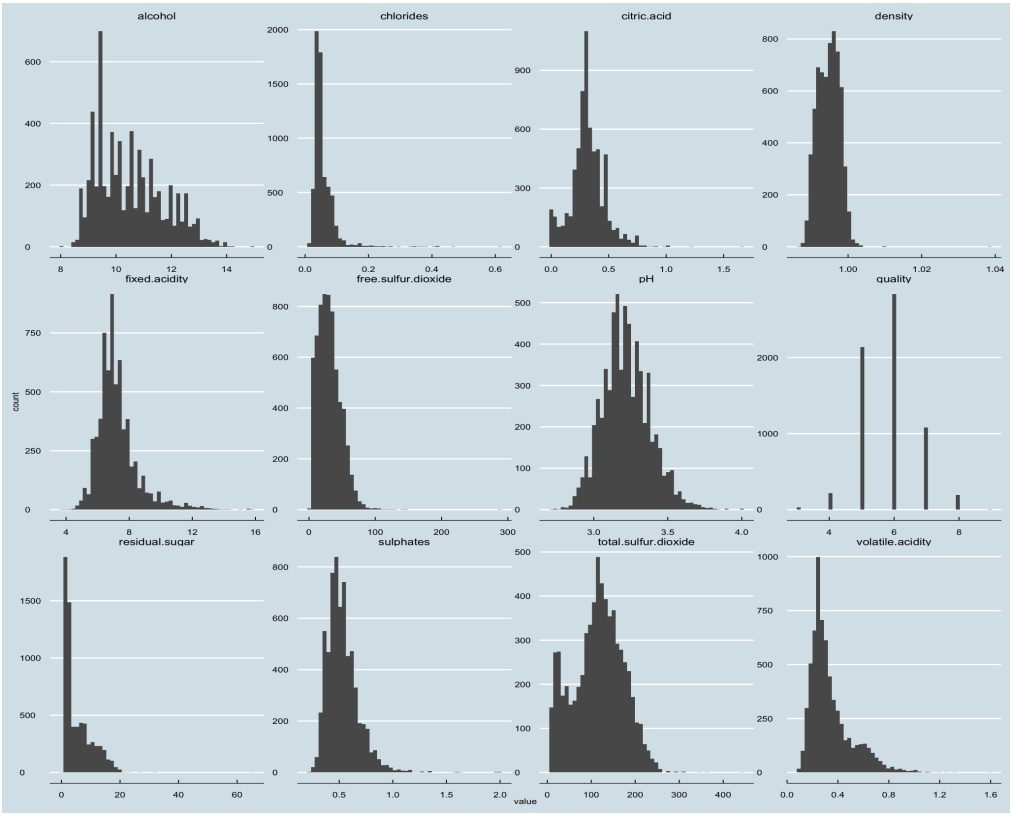

Figure 2.1

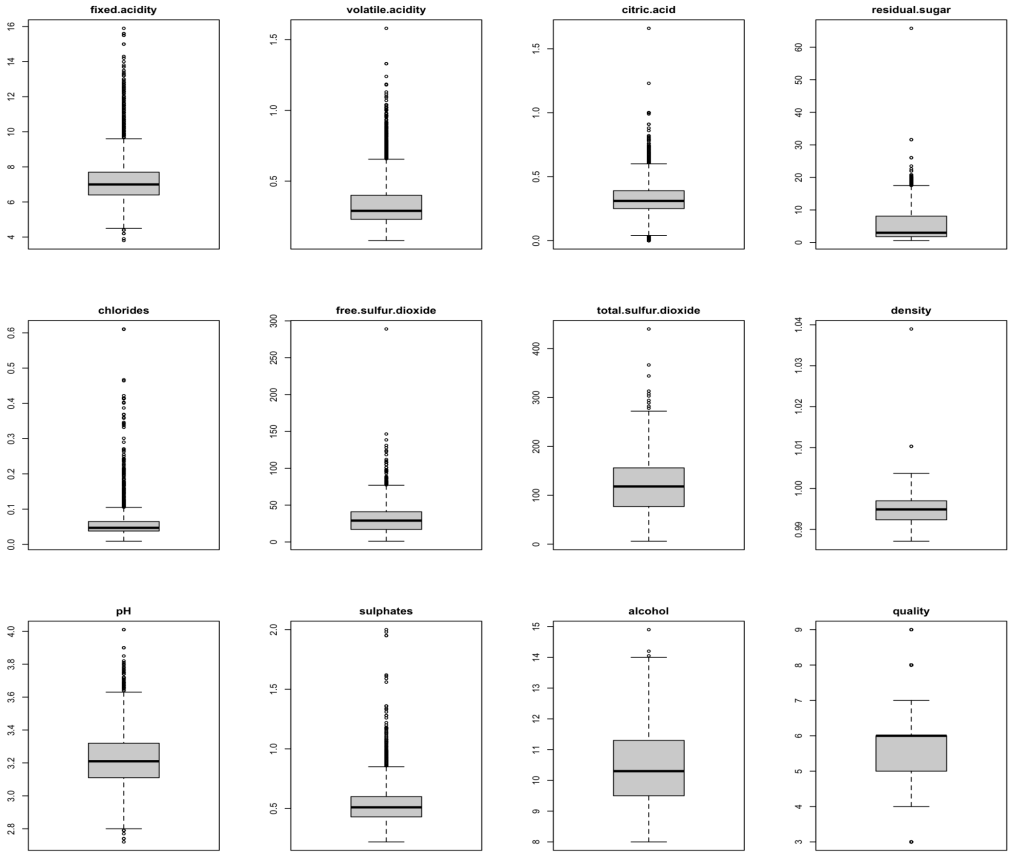

Figure 2.2

Figure 2.3 shows the correlation matrix among different variables. Based on the correlation plot, we get the following observations:

1. Total sulfur dioxide is positively correlated with free sulfur dioxide
2. Fixed acidity is positively correlated with density and citric acid
3. Citric acid is negatively correlated with pH and volatile acidity.
4. Most of the features do not have an explicit linear relationship with the target variable. We expect a linear model might not perform well.

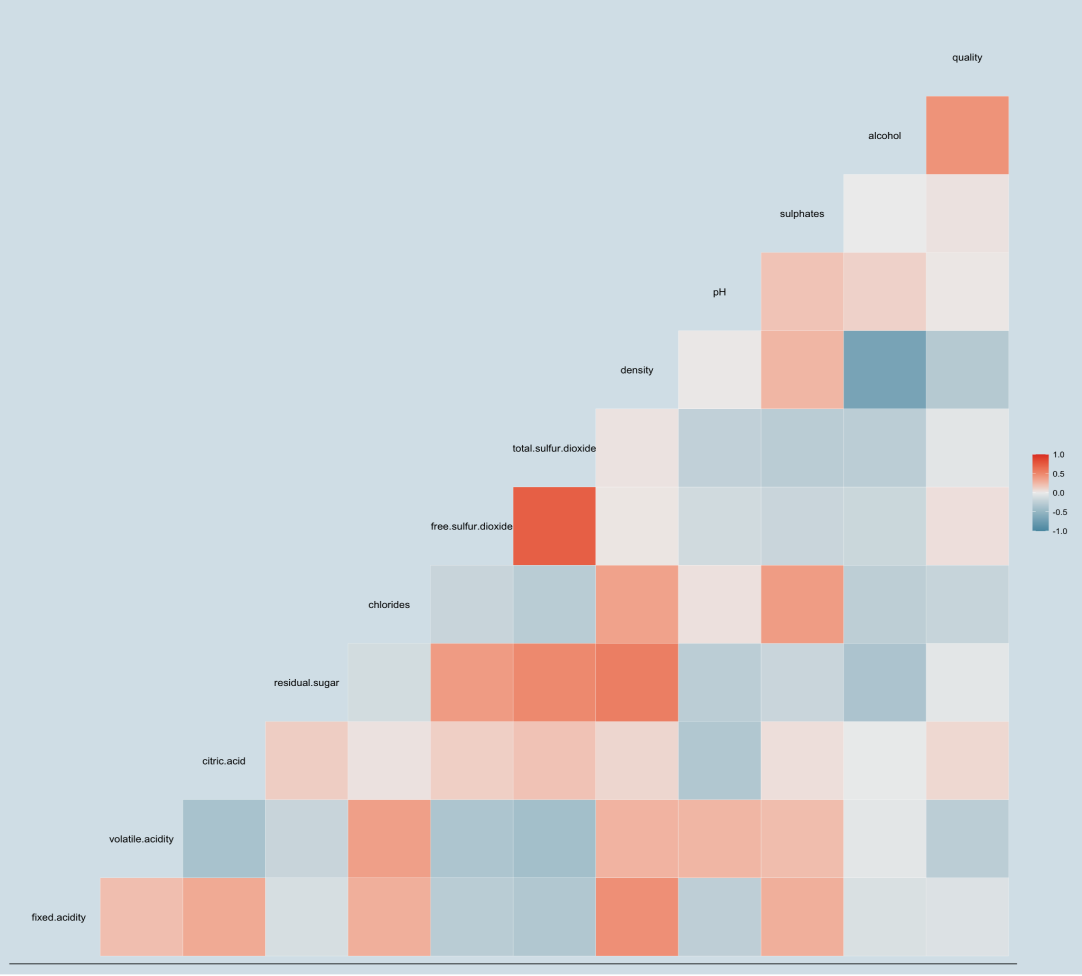

Figure 2.3

Clearly, we have collinearity among variables. We applied techniques such as regularizations to overcome them in the later sections.

Figure 2.4 shows some basic numerical statistics for each input variable. We can easily find that the ranges of different variables vary significantly, so we applied normalization on the

whole dataset to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

We split the dataset based on the types of wine. We will fit models on red wine and white wine separately.

```
    type           fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
Length:6497        Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
Class :character   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
Mode  :character   Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700
                   Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
                   3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
                   Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100
free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates         alcohol
Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50
Median : 29.00      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100   Median :10.30
Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49
3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30
Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90
```

Figure 2.4

3. Methodology

3.2 Regression Model Fit

3.2.1 Best Subset Selection on White Wine

Our first goal is to create a regression model to predict White Wine quality. To do so, we choose the Best Subset Selection. There are a total of 11 variables in our dataset. By performing linear regression on each model size, the Adjusted R-squared, Mallow's Cp, and BIC criteria are calculated on the white wine training data. To further ensure our model selection, we compute the prediction error in each model size. Again, the model with 8 predictors results in the smallest prediction error(about 58.61%). Therefore, our final model for White Wine becomes:
quality =5.64+0.087 fixed.acidity -0.31volatile.acidity + 0.35residual.sugar +0.067free.sulfur.dioxide -0.40density +0.10pH + 0.086sulphates + 0.27alcohol. To see how our model behaves, we create a confusion matrix and the accuracy of this model is about 50.99%.

3.2.1 Best Subset Selection on Red Wine

By applying the above steps on Red Wine training data, Adjusted R-squared tells us that the best model is the one with all the 11 predictor variables, Mallow's Cp tells us that the best model is the one with all the 9 predictor variables, and BIC tells us to stick with 6 variables. The model with 7 predictors results in the smallest prediction error(about 42.24%) and our final model for Red Wine becomes:
quality=5.62-0.15volatile.acidity-0.066chlorides+0.10free.sulfur.dioxide-0.23total.sulfur.dioxide -0.087pH+0.15sulphates+0.33alcohol. Accuracy of this model is about 58.46%.

3.2.3 Lasso Regression on White Wine

In this section, we fit a Lasso Regression model on White Wine training data and specify $\alpha = 1$. To determine what value is used for λ, we perform a k-fold cross validation with k = 10 and identify the best $\lambda = 0.006034$ which produces the lowest test MSE. The best model under Lasso Regression becomes:

quality=5.69-0.31volatile.acidity+0.20residual.sugar-0.022chlorides+0.074free.sulfur.dioxide-0.021total.sulfur.dioxide-0.17density+0.051pH+0.064sulphates+0.36alcohol. To see how our model behaves, we create a confusion matrix and the accuracy of this model is about 50.71%.

3.2.4 Lasso Regression on Red Wine

By applying the same steps to Red Wine training data, we identify the best $\lambda = 0.00422$ which produces the lowest test MSE. The best model under Lasso Regression becomes:

quality=5.61+0.0058fixed.acidity-0.16volatile.acidity-0.011critic.acid-0.0388residual.sugar-0.059chlorides+0.078free.sulfur.dioxide-0.20total.sulfur.dioxide-0.081pH+0.14sulphates+0.33alcohol. Accuracy of this model is about 59.08%.

3.2.5 Ridge Regression on White Wine

In this section, we fit a Ridge Regression model on White Wine training data and specify $\alpha = 0$. To determine what value is used for λ, we perform a k-fold cross validation with k = 10 and identify the best $\lambda = 0.0397$ which produces the lowest test MSE. The best model under Lasso Regression becomes:

quality=3.70+0.013fixed.acidity-0.30volatile.acidity-0.012critic.acid+0.21residual.sugar-0.038chlorides+0.086free.sulfur.dioxide-0.0045total.sulfur.dioxide-0.19density-0.060pH+0.071sulphates+0.33alcohol. Accuracy of this model is about 50.92%.

3.2.6 Ridge Regression on Red Wine

By applying the same steps to Red Wine training data, we identify the best $\lambda = 0.0381$ which produces the lowest test MSE. The best model under Lasso Regression becomes:

quality=3.61+0.023fixed.acidity-0.16volatile.acidity-0.015critic.acid-0.034residual.sugar-0.057chlorides+0.083free.sulfur.dioxide-0.20total.sulfur.dioxide-0.025density-0.071pH+0.14sulphates+0.31alcohol. Accuracy of this model is about 58.66%.

3.3 Classification Model Fit

3.3.1 Support Vector Machine

According to a research paper written by the Department of Information Systems/R&D Centre Algoritmi, University of Minho and Viticulture Commission of the Vinho Verde Region (CVRVV), we reproduce the same tuning method of the SVM model from it as a benchmark.

We create a grid for two parameters in the SVM model:

$$\gamma \in \{2^0, 2^{0.05}, 2^{0.10}, \dots, 2^2\} \qquad C \in \{2, 3\}$$

Then we search for the best combinations using 5-fold cross validation with the number of repeat equals 5. The best parameters for red wine are $\gamma = 1$ and $C = 2$. The best parameters for white wine are $\gamma = 1.03$ and $C = 2$. We try different values of tolerance in the model fit, but the accuracy only has minor changes.

Finally, we used the optimized parameters and got 66.71% accuracy for white wine with tolerance = 0.25, and 64.93% accuracy for red wine with tolerance = 0.5.

3.3.2 Random Forest

We fit our dataset using random forests in this section. First, we use 10-fold cross validation to find the best value of the number of variables randomly sampled as candidates at each split. We create a grid for this mtry parameter with a range as follows.

$$mtry \in \{2, 4, 6, 8, 10, 12\}$$

We apply grid search for both red wine and white wine. The optimal value of mtry for red wine is 4, and for white wine is 2. Finally, we used the optimized parameters and get 68.14% accuracy for white wine, and 67.64% accuracy for red wine. Random forest so far has the best performance.

3.3.3 Logistic Regression

Logistic regression on multi-class classification problems is to learn a logistic regression function for each class, and when predicting, the model selects the class whose expression has the highest value (probability). We tried logistic regression with both backward feature selection and forward feature selection on white and red. For each iteration, we keep the model with the lowest AIC. We stop the iteration when the AIC does not decrease for that iteration. After applying forward and backward selection on the data, we found on this dataset, the backward and forwards leads to the same model. For red wines, the final model is *quality ~ fixed.acidity + volatile.acidity + chlorides + citric.acid + residual.sugar + total.sulfur.dioxide + pH + sulphates + alcohol + density*. For white wines, the final model is *quality ~ alcohol + volatile.acidity + free.sulfur.dioxide + residual.sugar + fixed.acidity + chlorides + pH + density + sulphates*. The feature names in blue are the differences between models for white wines and red wines, indicating that for different kinds of wines, the standard to score quality contain different aspects.

By training these two final models and applying it on the test data, finally, the test accuracy is 59.92% for red wines and 53.37% for white wines.

3.3.4 K-Nearest Neighbors (KNN)

Before we apply KNN to this problem, we use 5-fold cross validation to decide on the best k for both white wines and red wines. The best k is 1 for white wines and 15 for red wines, with validation accuracy as 58.24% and 60.27%. Eventually, the test accuracy for white wines and red wines are 63.51% and 58.66% respectively.

3.3.5 QDA

QDA needs to calculate the prior variance for each class on training data. Since there are only 5 samples labeled as 9 in the whole dataset, in order to make QDA work, we combined label 8 and 9, label 3 and 4 for QDA specifically. The training accuracy and test accuracy for white wines are 50.16% and 47.92% respectively. For red wines, they are 60.53% and 57.62%. The gap between the training error and test error may indicate that the model can not learn the data well.

3.3.6 LDA

Compared to QDA, LDA uses the variance of the entire training data. Thus, for LDA, we use the original labels. However, the number of samples with label 9 or 3 is very scarce in the dataset. We can expect that the prediction for test data with these 2 labels cannot be very accurate, which would decrease the total accuracy. Eventually, the training and test accuracy are 53.37% and 52.07% for white wines, and 61.16% and 60.75% for red wines. Different from the results of QDA, there is no huge gap between test and training accuracy, which indicates that the model does not overfit the data as QDA does. LDA and QDA both show unsatisfactory performance on this classification problem.

3.3.7 Multilayer Perceptron (MLP)

A multilayer perceptron is a kind of neural network algorithm that only has one hidden layer. Here we use the *mlp* function in the *fdm2id* package, which is designed for classification problems. MLP is a very flexible model when the number of units in the hidden layer is very large. When tuning the model, we tried different hidden units from 2 to 14, and found that when the value exceeds 10, the results show obvious signs of overfitting, training accuracy increasing and test accuracy decreasing. As long as one hidden layer with 10 units is already too flexible for this dataset, we do not further consider the more complicated Neural Networks models. Eventually, we set the number of hidden units to 10 for white wines and to 6 for red wines. And

we set the decay rate in backpropagation to 0.005 for white wines and 0.001 for red wines. The activation function is softmax for both white and red.

The test accuracy for white and red wine are 54.87% and 60.33% respectively.

4. Model Evaluation and Comparison

Methods		Red Wines		White Wines	
		Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
Regression	Best Subset	0.5955	0.5846	0.5209	0.5099
	Lasso	0.5938	0.5908	0.5232	0.5071
	Ridge	0.5955	0.5866	0.5206	0.5092
Classification	Random Forest	/	0.6764	/	0.6814
	SVM	/	0.6493	/	0.6671
	Logistic Regression	0.6098	0.5992	0.5462	0.5337
	KNN	/	0.5866	/	0.6351
	QDA	0.6053	0.5762	0.5016	0.4792
	LDA	0.6116	0.6075	0.5336	0.5207
	MLP	0.6625	0.6033	0.5716	0.5487

5. Conclusion

In sum, we got the highest test accuracy (67.64% and 68.14%) for both Red Wines and White Wines by using Random Forest. Overall, regression methods work worse than classification methods. Due to the un-proportion of observations in each level of wine quality and the lack of data points, we are not able to reach a very good accuracy. There are much fewer observations in quality 3, 4, 8, and 9 than the others. If more observations from these four quality levels are found, better results will be achieved. Or, we may try other more complex models in the future.

Reference

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.