**BANA 7047-002 - Final Project-Group 9**

**Professor Yan Yu**

**What Makes my Glass of Wine so Tasty?**



**Ashwita Saxena**

**Bharath Vattikuti**

**Mark McNall**

**Xi Ru**

## Introduction

What could be better than having a glass of wine after a long, tiring day? Though we have all had our experiences with tasting different types of white wines, have you ever wondered what makes a wine taste better than another? To answer this question, we are going to analyze different physicochemical constituents of a type of white wine and see which ones from those predict whether a wine tastes good or not.

## Motivation and Objective

The motivation for the project is to find a new dataset in a field untouched during classes, labs, and homeworks, and apply some of the techniques we have learned throughout the semester. Our objective is to identify the key factors leading to greater white wine quality and find answers to the question, “ What makes my glass of wine so tasty?” by creating predictive models. Using a wide range of models, including ordinal logistic regression, classification trees (CART), Random Forest, generalized additive models (GAM), neural networks and Gradient Boosting (GBM), we want to predict whether a particular wine sample would taste bad, okay or good. Each model will be evaluated based on their interpretability & predictability.

## Dataset Variables

We have chosen to use dataset from UCI Machine Learning Repository “Wine Quality Data Set”. This dataset contains samples of white wine called ‘vinho verde’, from North of Portugal. The white wine data in total contains 4898 observations (wines) and 12 variables. Among the 12 variables, 11 physicochemical constituents of the wine are our predictors, and one sensory wine quality measure is our response.

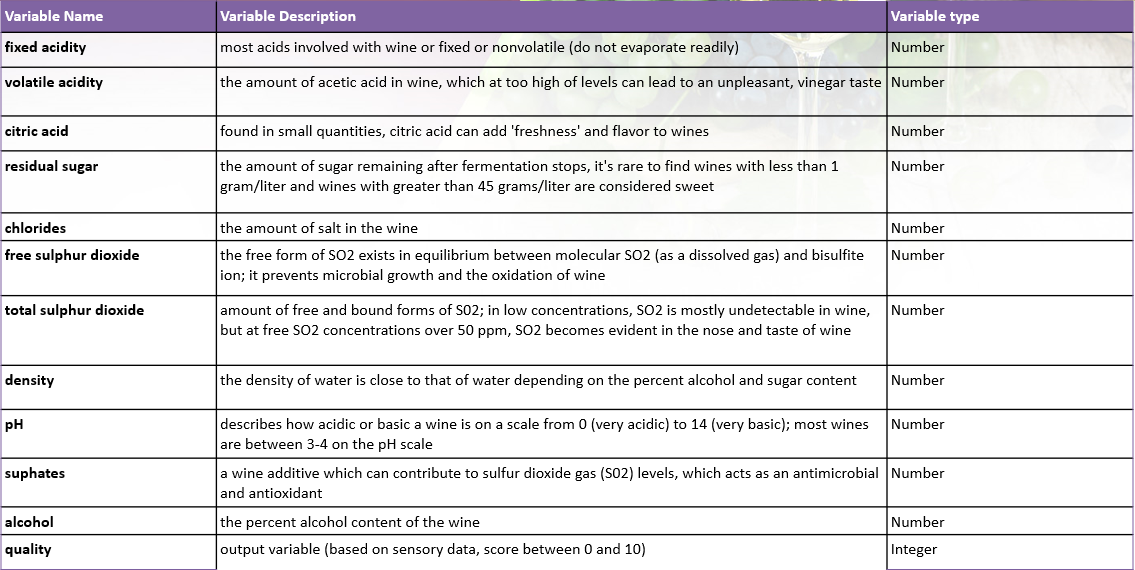


Table 1: Variable Directory

Since the aim of our model is to predict wine quality based on the physicochemical factors in the dataset by comparing our results with the sensory quality measure, we created a new ‘response’ variable which holds value 0 when the wine is bad (quality 0-4), 1 when wine is okay (quality 5-6), 2 when wine is good quality (quality 7-10). In this way we will be able to predict factors that decide whether a wine is good, okay or bad, making it into an ordinal classification modeling problem.

## Exploratory Data Analysis

We performed some basic exploratory data analysis to understand each predictor and the distribution of each predictor with respect to the response variable (wine quality). This analysis provides us some insight on identifying the key factors before our model building process. Our EDA was mainly focused on creating correlation matrix, density plot and boxplots for all variables.

Correlation Matrix

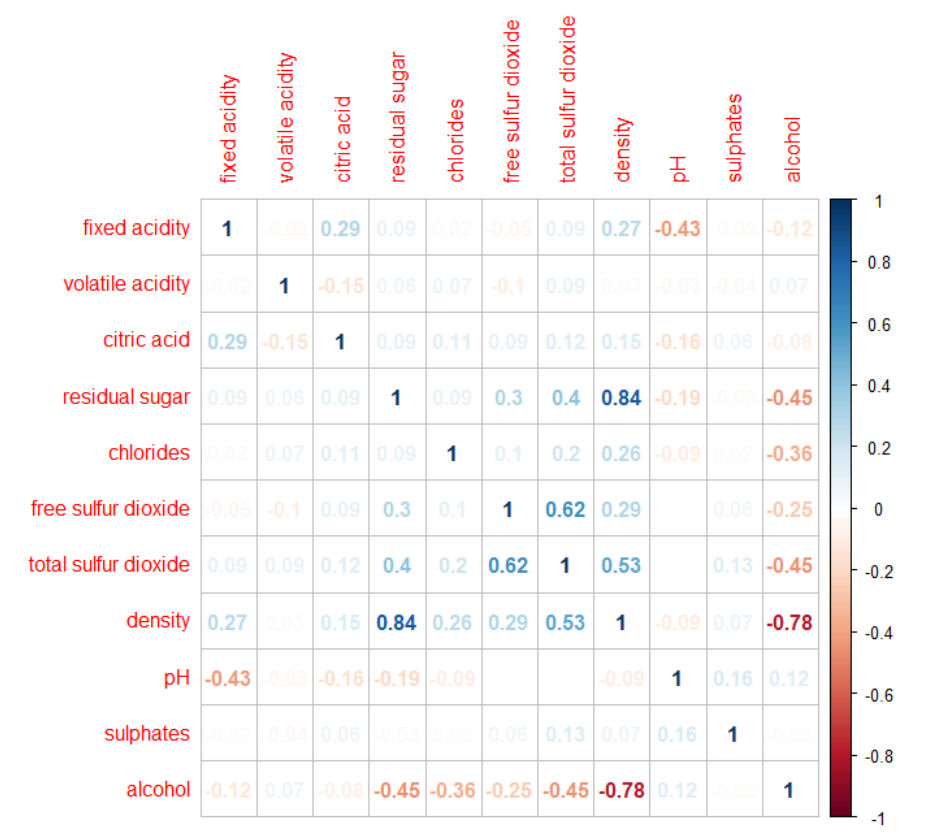


Fig 1: Correlation Matrix

We created a correlation matrix of all the predictors to see if any multicollinearity exists among the predictors. Based on our analysis, the density and residual sugar are highly collinear with a correlation coefficient of 84%. Also, density and alcohol content are highly negatively correlated with a correlation of -78%. This makes sense as density of a wine depends on its alcohol content and sugar content. We also confirmed the high correlation by looking at scatterplots of these variables. Hence, we decided to not include density in our modeling processes to avoid multicollinearity.

Density plots

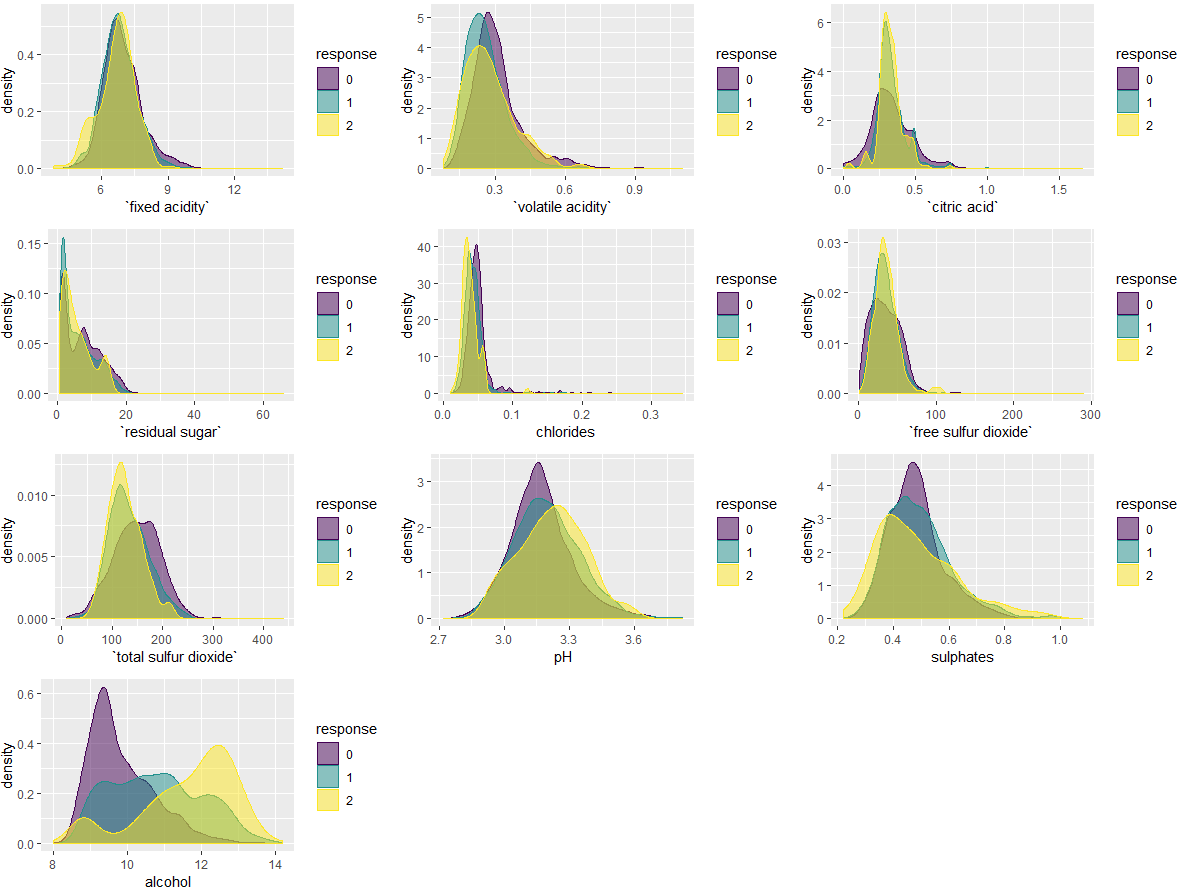


Fig 2: Density plots for all variables wrt response

We then looked at density plots of each variable with respect to good, bad and okay wine. The results look like the above figure. We can intuitively say that alcohol content should be one of the significant factors in predicting wine quality. Also, other factors such as total sulphur dioxide, pH content, sulphates, volatile acidity and chlorides should predict the quality of wine.

Boxplots

We also looked at outliers via boxplots and noticed that very few variables have outliers. In order not to lose on important information, we decided to keep the outliers in the dataset. In addition, we observed that most of predictors have small variance, expect alcohol. This could once again confirm our conclusion from previous EDA, because there is significant difference among different wine qualities in terms of alcohol content. The bad and okay wines have more outliers as compared to good wine.

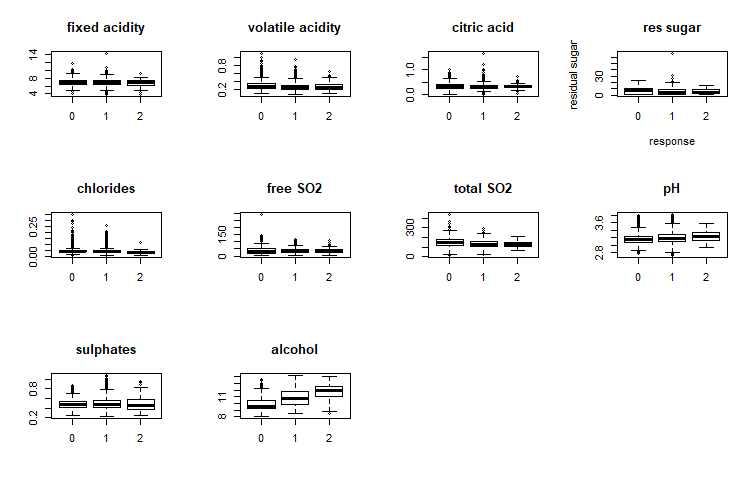


Fig 3: Boxplots for all variables w.r.t. response

## **Methods and Models**

In this project, we performed six data mining models mentioned above and compared the misclassification rate of all the models. The first step of our model building process was variable selection, which selects the important predictors that help predict wine quality in the final models. After building each model, model comparison will be ran and we will present the best models in terms of both interpretability and predictability. Before we build our predictive model, we split our dataset into 80% training and 20% testing for bad, okay and good wines respectively. By taking 80% each of good, bad and okay wines for training dataset and 20% each of good, okay and bad wines for the testing dataset, we were able to maintain the original proportion of the ordinal response. The following are the detailed description of each model and analysis the results from each model.

### Ordinal Logistic Regression Model (POLR)

Since our response is ordinal variable, we chose ordinal logistic regression model for this dataset to predict white wine quality. The below Table 1 is summary table for the OLRM and identify the key predictors based on odds ratio. Most of variables had significant p-value, while fixed acidity, citric acid, total sulfur dioxide and pH had larger p-value. The sulphates had the highest odds ratio of 3.7091, which had the same interpretation as binary logistic model and means there are 3.7091 unit of changes in odds for every unit increases in sulphates, all else constant. Table 2 shows the confusion matrix and misclassification rate of this model. The in-sample MR for this ordinal logistic model is 0.2348, while the out-of-sample MR is slightly larger and that is 0.2398.

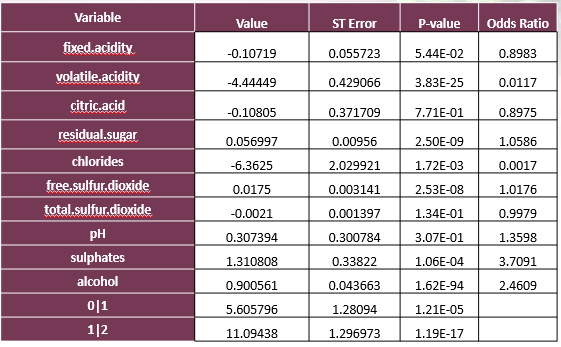


Table 1: Summary Table for OLRM

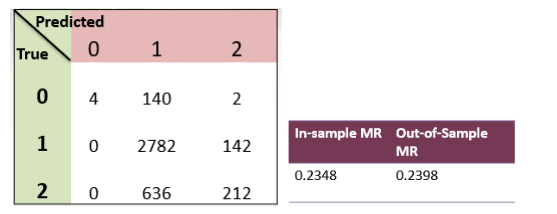


Table 2: Confusion Matrix & Misclassification Rate for OLRM

### Classification Tree

The classification tree is also a good choice for modeling this wine data in terms of interpretability. Figure 4 displays the final tree model we created and allows us to predict the quality of any new wines with those same predictors. There are total 10 terminal nodes and 9 splits. Each terminal node contains the number of observations that fall into this specific category(bad, okay, good). The confusion matrix in Table 3 illustrated that there is no prediction for bad wine (category 0) using classification tree. The misclassification rate of training data is 0.2111 and that of testing data is 0.2449, which are slightly smaller than the MR from ordinal logistic model.

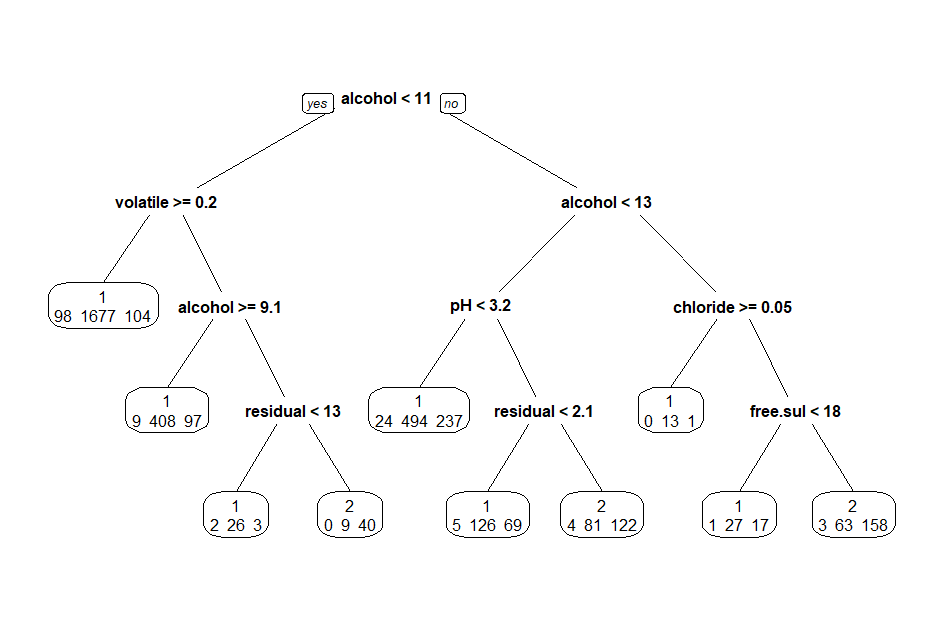


Fig 4: Classification Tree

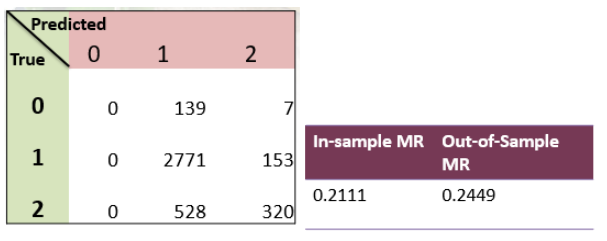


Table 3: Confusion Matrix & Misclassification Rate for CART

### Random Forest

The third model we developed was random forest, which is an ensemble learning method for classification and offers greater accuracy by constructing a large number of classification trees by randomly selecting variables from the original dataset and bootstrapping samples. We can also understand the importance of independent variables after running this model. The most important variable that predicts quality (Figure 5) is alcohol content, whereas fixed acidity and citric acid are the least important based on mean decrease gini criterion.

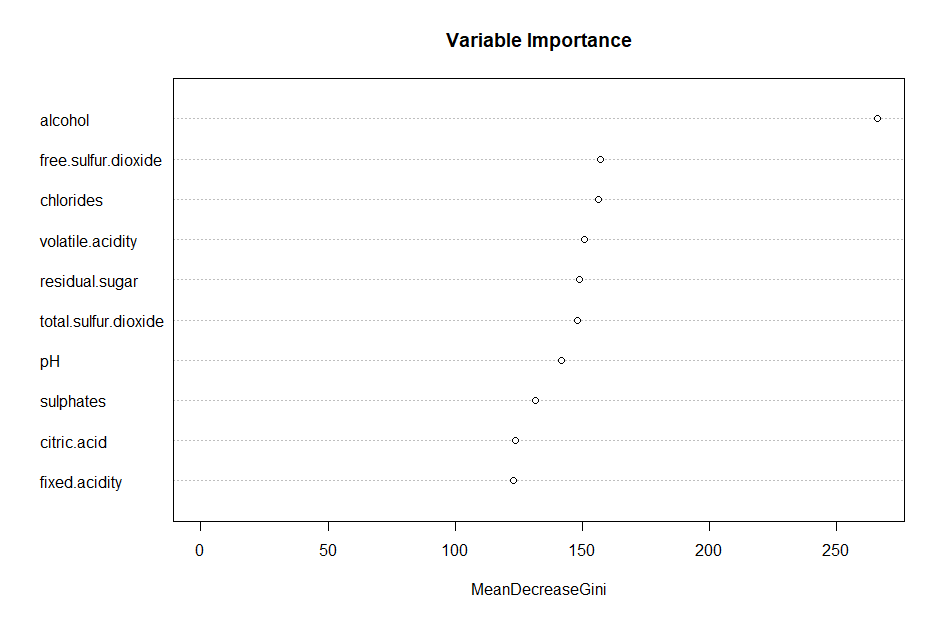


Fig 5: Variable Importance for Random Forest

Figure 6 was generated to determine the optimal number of trees we should build in our forest. Based on the results, the number of trees in our random forest should be less than 50. In addition, the in-sample misclassification rate for random forest is much lower and equals to 0.1542. In the same way, the out-of-sample misclassification rate is 0.1449 , which shows that testing data had slightly better performance than training data.

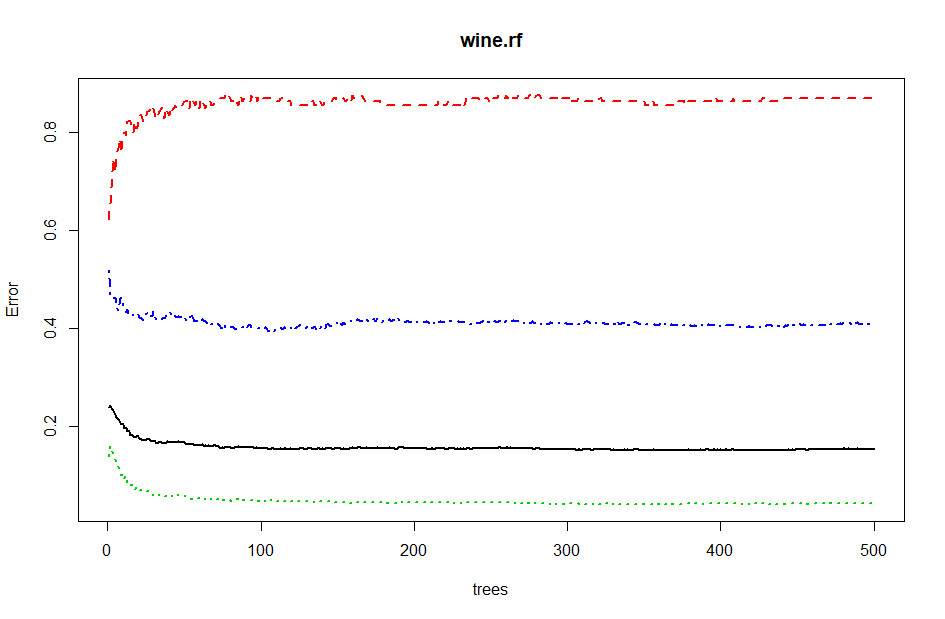


Fig 6: Error vs Number of Trees

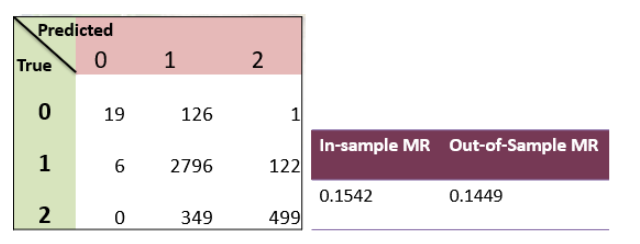
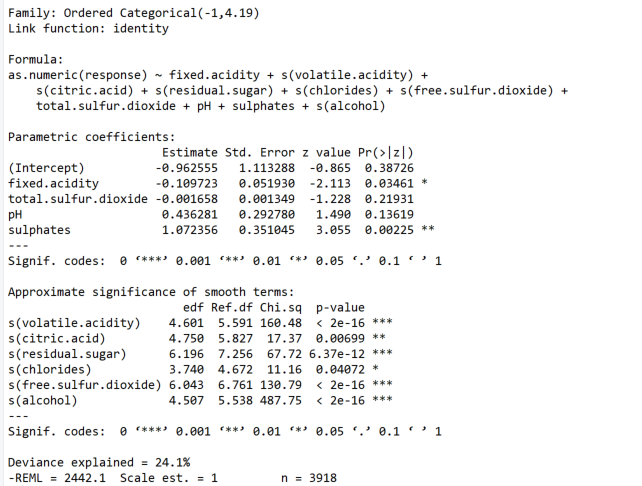


Table 4: Confusion Matrix & Misclassification Rate for Random Forest

### Generalized Additive Model (GAM)

To better fit non-linear predictors (if any), we use Generalized additive model (GAM) by using smooth functions on some of the predictors. After a series of trials, we selected six out of ten predictors to be added to smooth functions, as their effective degree of freedom was greater than 1 and p-values were significant. The final GAM model is “response ~ fixed.acidity + s(volatile.acidity) + s(citric.acid) + s(residual.sugar) + s(chlorides) + s(free.sulfur.dioxide) + total.sulfur.dioxide + pH + sulphates + s(alcohol)”. The GAM plot also confirmed the non-linearity of those smooth terms. This model predicted a lot bad wines, while the other models only predicted a few. The in-sample misclassification rate is 0.2282 and the out-of-sample MR is 0.2306.



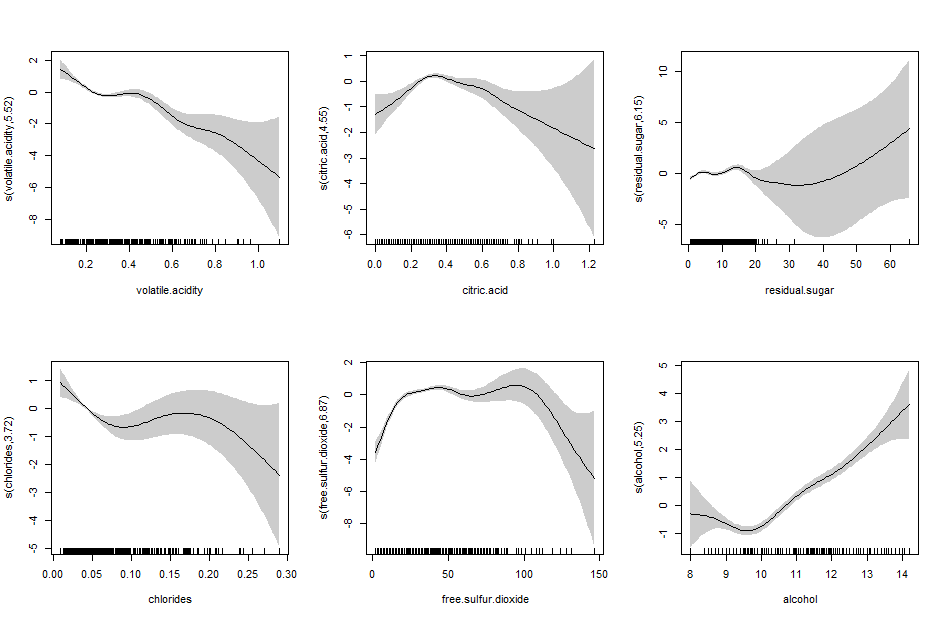


Fig 7: Smoothing terms for GAM

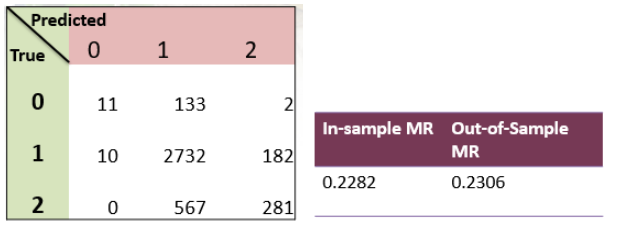
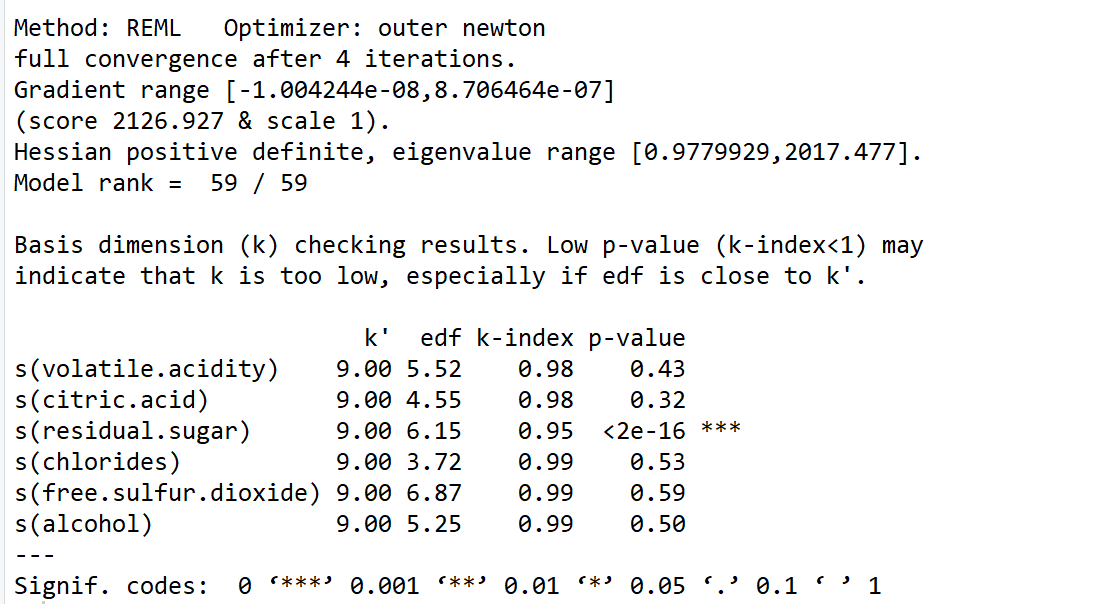


Table 5: Confusion Matrix & Misclassification Rate for GAM

After we built our GAM model, we also performed GAM checking and examined the model assumptions. All the smoothing terms had basic dimension (k) of 9, while residual sugar had much lower p-value than others. Figure 8 includes the four residual plots to check the model assumption, such as deviance residual vs theoretical quantiles, residuals vs. linear predictor, histogram of residuals, response vs. fitted values. Those plots proved that GAM model didn’t violate the model assumptions.



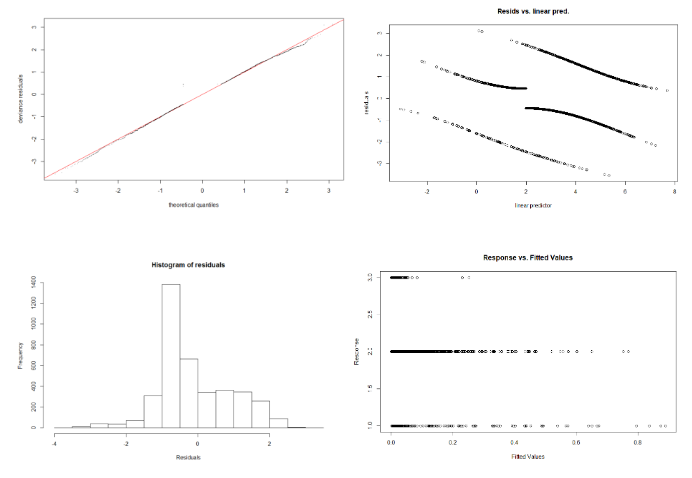


Fig 8: GAM checking

#### Neural Network

Neural network is also conducted in this project to obtain more accurate predictions. However, because it is a black-box, it is difficult to interpret the results. Figure 9 suggest us the optimal number of hidden layers should be used to gain better accuracy. Figure 10 is graphic representation of our neural network model. The confusion matrix suggests that none of bad wines were predicted and the in-sample MR is 0.2335. The out-of-sample MR is 0.2327, which is slightly higher than what we expected.

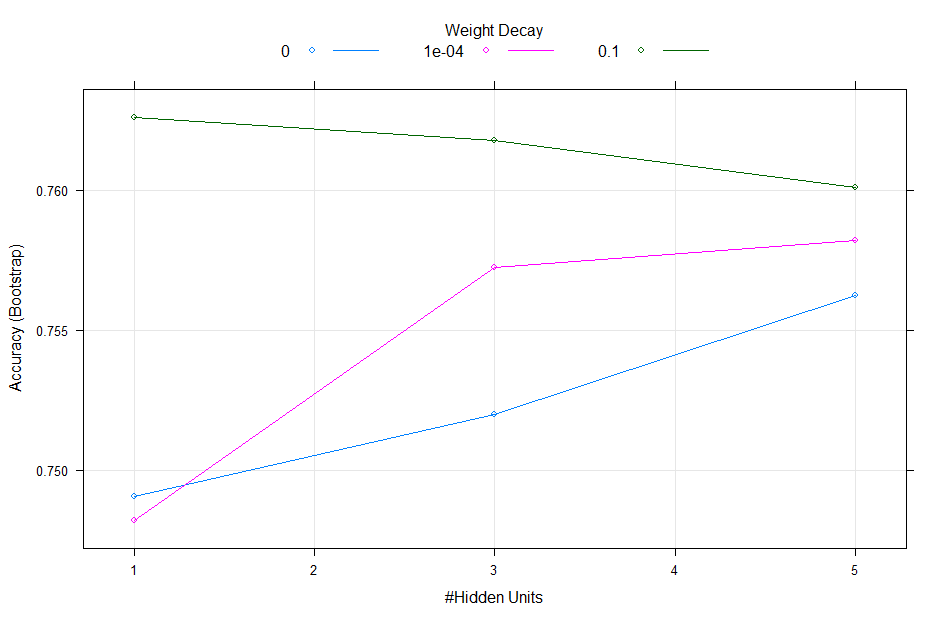


Fig 9: Accuracy vs. Number of Hidden Units

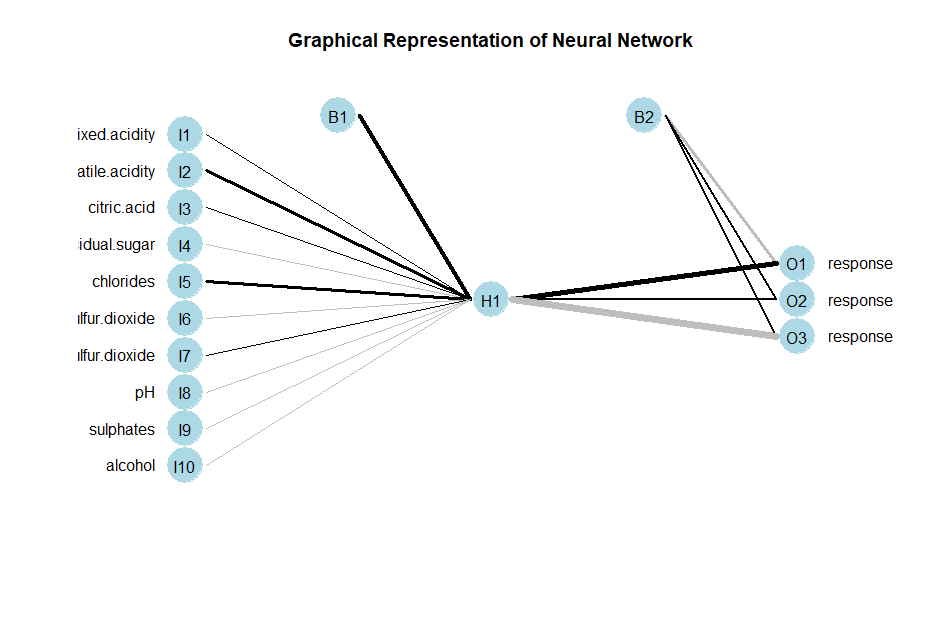


Fig 10: Graphic Representation of Neural Network

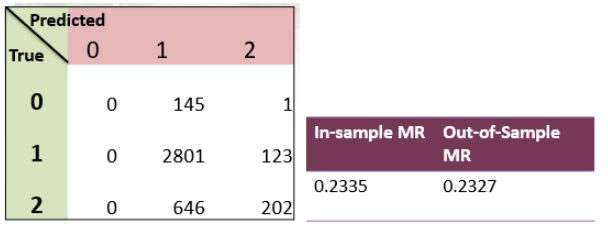


Table 6: Confusion Matrix & Misclassification Rate for GAM

### Gradient Boosting Machine (GBM)

GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful “committee” that are often hard to beat with other algorithms. This model will give us accurate predictions as compared to other tree models. Based on the Figure 11, we can conclude that the variable with highest relative influence on quality of wine is alcohol and chlorides had lowest relative influences in GAM. The MR for this model can be seen in Table 7 below.

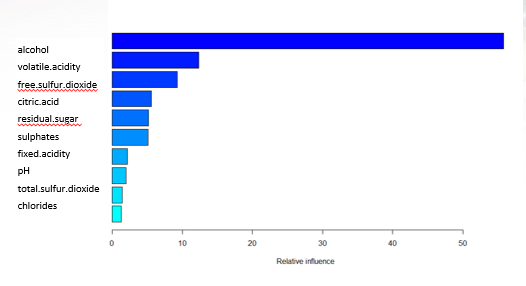


Fig 11: Relative Influences for GAM

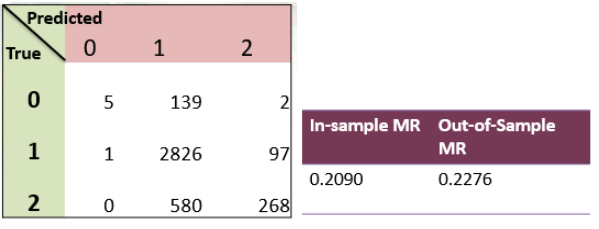


Table 7: Confusion Matrix & Misclassification Rate for GBM

## Model Comparison

After we build our six predictive model for white wine quality, it is important to compare the models and find out the best model to use. We assess our model performance based on the misclassification rate. The model with lower MR has better chance to be selected. Although the MR of five models are very similar, random forest was quite different and had the lowest MR for both in-sample and out-of-sample data. Another key observation was that there was only slight difference between in-sample and out-of-sample MR for all the models. This shows that the model performance is relatively good.

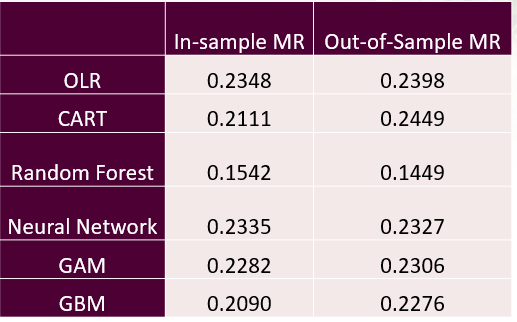


Table 12: Model Comparison in terms of MR

## Conclusion

In conclusion, we select random forest as our final model for predicting white wine quality. We observed that alcohol was the most important predictor in predicting wine quality, hence, it is most important for the winemakers to control the alcohol content in order to produce good quality wines.

**References**

* 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.

* <https://archive.ics.uci.edu/ml/datasets/wine+quality>
* <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>