Model Selction for the Quality Classification of Vinho Verde Wine

Alex Rintamaa

2024-05-12

Abstract

The analysis of Wine quality shows how different models can predict results at different percentages and the importance of model selection for classification. The data in question is Wine ratings data from Kaggle. For the purposes of this study, we will be using six different models to classify the Wine quality as either good or bad. The six models that will be used are Logistic, K-Nearest Neighbors, Gradient Boosting Method, Support Vector Machine, Random Forrest, and Neural Network. This report will show the methods to which each model was used to predict the Wine quality, and how Random Forrest is the best model for classification of this data. This report is good for anyone looking to understand more about how different models can be used for classification

Introduction

The report entails the use of a data set that was downloaded from Kaggle. The dataset is entirely about the measurement of wines, specifically, the red variants of the Portuguese "Vinho Verde" wine. Wine quality the response variable will be measured from the predictors: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The predictors in this dataset are all different chemicals present in the wine. The purpose of this report is to show how these predictors are selected by each model to classify the wine as either good or bad.

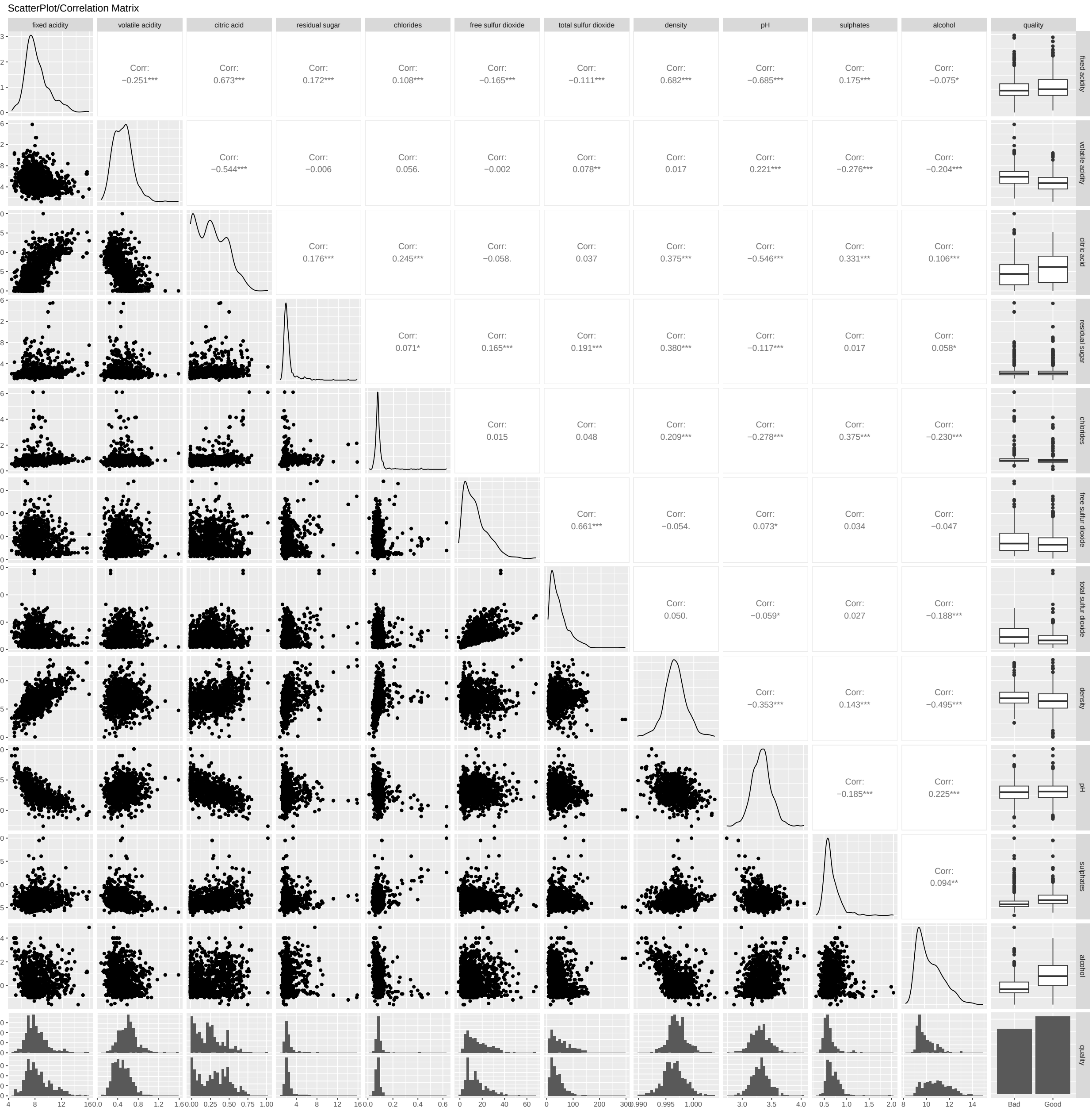
Methods

Data Preprocessing

The wine data originally picked out for the project, has 13 columns. 11 of these columns will be used for the purpose of being predictor variables, 1 of the columns will be our response variable, and the last column listed as ID, will be removed. Getting the data setup for usage throughout the report required a small amount of data cleaning. The first step would be removing the ID column as it offers nothing beneficial to the report. The second step was changing the response variable quality into a classification of either good or bad. In order to do this a cut function was used in which any quality rating of 0 through 5 is classified as bad, and any quality rating of 6 through 10 would be classified as good. For the initial process of exploring the data a ggpairs() function was used to create a correlation matrix. The predictor variables, each received a correlation number and the singular response variable recieved side-by-side box plots showing the different between good and bad with each predictor.

EDA

Figure 1



Upon inspection of Figure 1, the scatterplot/correlation matrix displays some quite interesting results. For the purposes of this report the most important column will be the quality column as this is the response variable, and relationship between each of the predictors is important. When looking at the column quality with the row fixed acidity, it can be seen that the for both good and bad classifications, the median fixed acidity number is very close, however, when looking at the proportion that fall within the 50% of data, it can be seen that more of the good classifications fall closer to the median. This would initially suggest that fixed acidity does not have much of an effect on the classification of wine quality. The next row within the the quality column to be looked at would be volatile acidity level. Upon inspection of this side-by-side boxplot, the median acidity level for wines classified as bad, is a bit higher than the median acidity level for wines classified as good. Also, it can be seen that the "bad" wines have both more outliers, along with a couple wines that have very high volatile acidity levels. This would suggest that, volatile acidity level has some sort of effect on the classification of wines, although it is likely a rather small effect. The next row for analysis would be the citric acid predictor. Upon inspection of these boxplots, the level of citric acid appears to have a slight correlation with the classification of wine, as a result of the median citric level being a bit higher for the wines classified as good. For the predictor residual sugar level, it can seen that for this type of wine, the sugar level is relatively low. Upon initial analysis, it appears that sugar level has no correlation with the classification of wines. Looking at chloride levels as they relate to the quality of wine, it can be seen that this specific wine has a pretty consistent low level of chloride. This would suggest that chloride has little to now effect on the difference in classification of wines. The relationship that will be looked at is free sulfur dioxide and wine quality. In looking at the side-by-side boxplots it can be seen that there is not much difference between the good and bad wines for free sulfur dioxide, this would suggest little to correlation between free sulfur dioxide and wine classification. In looking at how total sulfur dioxide relates to the quality of wine, it can be seen that wines classified as bad then to have a higher amount of total sulfur dioxide, whereas wines classified as good tend to have a lower level, albeit there are some outliers that have very high levels. This would suggest that there is some correlation between free sulfur dioxide amounts and classification of wines. When looking at the density of drinks and there relationship with wine quality, it can be seen that that the wines classified as good are slightly lower than the wines classified as bad, however, more importantly there are much more outsiders on the lower end of density. This suggests that there is a slight relationship between classification and the density of the wine. In looking at the pH levels, there appears to be no relationship with good or bad classification. The boxplots appear almost exactly the same. The relationship between sulphates and wine quality show an appeared relationship. In looking at the boxplots, it can be seen that the sulphate levels of wines classified as good are higher than the sulphate levels of wines classified as bad. This suggests that the higher the sulphate level, the better quality the drink is. The last relationship to be looked at is alcohol amount within the wine and the quality of the wine. In looking at the boxplots, it can be seen that for wines that are rated to be better quality the alcohol is much higher. This set of side-by-side boxplots shows the biggest difference between the classification of good and bad wines. This suggest that the amount of alcohol within the wine has the biggest relationship with the wine quality and is the most important predictor of wine quality.

After exploring all of the predictors relationship with the response, it can be summarized to the most important predictors of wine quality are: volatile acidity, citric acid, total sulfur dioxide, sulphates, and alcohol.

Next the predictors need to be tested for multicollinearity. Multicollinearity is the relationship predictors have with each, for this analysis, looking at the correlation numbers within the upper half of the Scatterplot/Correlation coefficient will be important. In looking at Figure 1, there appears to be only a few predictor relationships that are at risk for multicollinearity. The relationships with the highest risk of multicollinearity are those with a correlation higher than 0.5 or less than -0.5. Looking at the figure these relationships are citric acid with fixed acidity, citric acid with volatile acidity, total sulfur dioxide with free sulfur dioxide, density with fixed acidity, pH with fixed acidity, and pH with citric acid.

Model Selection and Training

For the overall analysis of the wine quality data, six different models were run, using the Caret packages train function for each model. Within the train function the response variable (quality) was called with all of the predictors being used in the training. The six different are Logistic, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) Radial, Gradient Boosting Machine (GBM), Random Forrest, and Neural Network.

Logistic Regression

Logistic Regression was selected as it is a very common simplistic method for binary classification. Binary Classification envolves the answer being either a 1 or a 0, or in this instance Good or Bad. For this method a 5-fold Cross Validation was used so that model can be compared with the other models in the analysis. A 5-fold Cross Validation splits the data into 5 equal but separate sections and compares the results with each other.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor was selected as a non-parametric but common option for classification models. K-Nearest Neighbors algorithm involves looking at the K Number of neighbors closest to the plotted point or projection, and giving the same classification of those projections to the current projection. This method is very good in a classification setting as the assumption that projections close to each other are the same. A 5-fold Cross Validation was used to split the data into 5 equal but separate sections and makes comparison with other models easier.

Support Vector Machine (SVM) Radial

Support Vector Machine was used for it's ability to handle nonlinear relationships. As seen in figure 1 above, there are quite a few predictor relationships which are non-linear and require a more complex model. For this model the hyperparameter radial was selected because of the nonlinear nature of the predictors. This data was Cross-Validated using a 5 fold as its processing time is much quicker in this case than a repeated cross validation approach despite it being more consistent.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine was selected for its learning capability, which uses weak learners to help make a stronger prediction model. Gradient Boosting Machine is seen as a better model for working with nonlinear relationships between predictors and the response. Gradient Boosting Machine is also seen as a more effective model type when it comes to working with data that is of multiple types (i.e Categorical, Numerical, etc.). A five-fold cross validation was used for its simplicity in comparing to other models.

Random Forrest

Random Forrest model was created for the purpose of creating a model based on decision trees. Random Forrest are known for its robustness to overfitting and feature importance estimation. This is important for the understanding the underlying relationships between predictors as from figure 1 they are nonlinear. A 5-fold cross validation approach is used here for its simplicity and speed in comparing models.

Neural Network

A forward propagation neural network was employed comprising a singular layer of neurons with nonlinear activation functions. The neural network function used comes from the nnet() function because of its ability to work with the caret package. The neural network makes only one layer and uses different rates and methods of decay to get a more accurate prediction. Neural Networks are very useful for classification because of there ability to balance model complexity and generalization.

Every model will be assessed on Accuracy. By using the same cross-validation method for each of the models, a table of each model is able to be easily created. Each model's hyperparameters were tuned to achieve optimal performance.

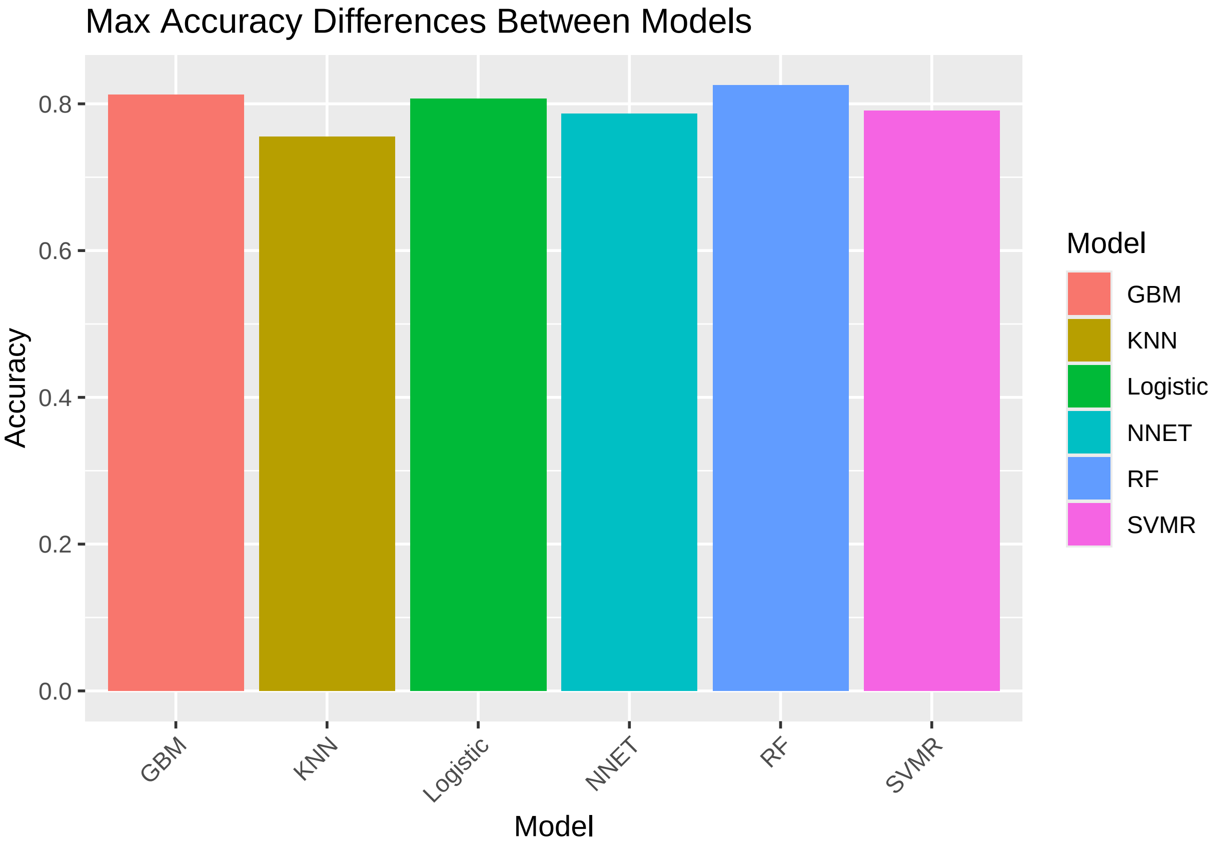
Results

Table 1

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Logistic	0.7192982	0.7379913	0.7467249	0.7524247	0.7510917	0.8070175	0
KNN	0.6842105	0.6929825	0.7368421	0.7225944	0.7434783	0.7554585	0
SVMR	0.7368421	0.7587719	0.7685590	0.7681184	0.7860262	0.7903930	0
GBM	0.7641921	0.7763158	0.7903930	0.7882747	0.7982456	0.8122271	0
RF	0.7554585	0.7903930	0.7982456	0.7961656	0.8114035	0.8253275	0
NNET	0.7412281	0.7500000	0.7631579	0.7628537	0.7729258	0.7869565	0

Based on the data within table 1, it can be seen that the Logistic model has a minimum accuracy rate of 0.719, a median accuracy at 0.747, a mean at 0.752, and a max at 0.810. For the KNN model the minimum accuracy is at 0.684, the median accuracy at 0.737, the mean accuracy at 0.723, and a maximum accuracy at 0.755. For the SVM Radial model, the minimum accuracy is at 0.737, the median accuracy is at 0.769, the mean accuracy is at 0.768, and the maximum accuracy at 0.790. For the GBM model, the minimum accuracy is 0.764, with the median accuracy being 0.790, the mean accuracy at 0.788, and a maximum accuracy at 0.812. For the Random Forrest model, the minimum accuracy is at 0.755, with a median accuracy at 0.798, the mean accuracy at 0.796, and a maximum accuracy at 0.825. Lastly, for neural network, the minimum accuracy is 0.741, with a median accuracy at 0.763, a mean accuracy at 0.796, and a max accuracy at 0.787. From these different results, it can be seen that for minimum, median, and mean, and maximum, the Random Forrest model has the highest accuracy percentage. The gradient boosting machine is the overall second best model for predicting accuracy. For the Neural Network, SVM Radial and Logistic the accuracy is almost the same for minimum accuracy, median accuracy, mean accuracy, and maximum accuracy. The worst performing model in this instance is the KNN as it is the lowest in every single category.

Figure 2



Upon looking at figure 2, which is a plot of the Maximum Accuracy percentages it is clear to the see the results discussed above that the Random Forrest model outperfoms the Gradient Boosting Method. Logistic, Neural Network and SVM Radial all having the same values can be seen in the figure where all three models appear to have a maximum value right above 0.75.

Discussion

After looking at the results from the above tables and plots, it becomes clear that the best model for predicting the quality classification of wines is the Random Forrest. It is important to note that the important accuracy percentage is the Max. accuracy, as for any prediction done would be using the best model from each of the models. The random forrest using a bunch of decision trees paired with is ability to handle overfitting, had the best formula for predicting the classification of the Portuguese "Vinho Verde" wine as either good or bad.

Limitations

The rest of the models within this analysis all showed to semi-accurate at predicting the rates, with every model predicting the correct classification over the 75% of the time. It should be noted that a prediction accuracy of 86% is not super good, and this is only the best model out of the six models presented in this analysis

References

Yasser H., (2017). Wine Quality Dataset. Retrieved from Kaggle: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset