Red Wine Review

Using Logistic Regression, LDA and Nonparameteric model approach

MSDS6372 – Project 2

November 2018

Vitaly Briker | Kelly Carter | An Nguyen

Table of Contents

[Introduction 3](#_Toc531545869)

[Data Description 3](#_Toc531545870)

[Exploratory Data Analysis 3](#_Toc531545871)

[Exploratory Analysis 3](#_Toc531545872)

[Objective 1 – Logistic Regression Model 4](#_Toc531545873)

[Problem Statement and Approach 4](#_Toc531545874)

[Assumptions 5](#_Toc531545875)

[Model Fit 5](#_Toc531545876)

[Parameter Interpretation 5](#_Toc531545877)

[Conclusion 5](#_Toc531545878)

[Objective 2 – Logistic Regression 2 6](#_Toc531545879)

[Summary 6](#_Toc531545880)

[Model Fit 6](#_Toc531545881)

[Parameter Interpretation 6](#_Toc531545882)

[Conclusion 6](#_Toc531545883)

[Objective 2 – Linear Discriminant Analysis or Quadratic Discriminant Analysis 6](#_Toc531545884)

[Assumptions 6](#_Toc531545885)

[Model Fit 6](#_Toc531545886)

[Conclusion 6](#_Toc531545887)

[Objective 2 – Nonparametric approach 7](#_Toc531545888)

[Problem Statement and Approach 7](#_Toc531545889)

[Assumptions 7](#_Toc531545890)

[Wine with quality rate of 5 and less called “poor” wine and wine 6 and more “good” wine. 7](#_Toc531545891)

[Split observation into train set as 2/3 of data and 1/3 is as test set. 7](#_Toc531545892)

[Each wine is assumed to be independent from the others. 7](#_Toc531545893)

[Model Fit 7](#_Toc531545894)

[Conclusion 7](#_Toc531545896)

[**Variable Importance** 8](#_Toc531545897)

# Introduction

In this project, we attempted to classify red wines that are related to the Portugese “Vinho Verde” variants. Using a data set provided by the UCI machine learning repository, <https://archive.ics.uci.edu/ml/datasets/wine+quality>. The data set included fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality as variables. Using the quality variable, based on sensory data score between 0 and 10, we categorized the wines into a ‘poor’ category or a ‘fine’ category; 5 or less in quality was desiganted as ‘poor’ and greater than 5 is ‘fine’. For the purpose of this project our objectives are as follows:

1. Build a logistic regression model using the provided data.
2. Building on the regression model above add complexity to the model.
3. Create a competing model using LDA or QDA.
4. Use a nonparametric model approach in a third model.

## Data Description

Source: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

* Contains 1599 records

For the purpose of this project, we randomly split the data set into a training set and a test set 50/50:

* train: wine\_train.csv
* test: wine\_test.csv

## Exploratory Data Analysis

The following analysis is a prediction of wine quality based on 11 physiochemical test measures and grape type. The data used includes 6497 wines which are all variants of the Portuguese "Vinho Verde". Each wine has been given a quality rating on a scale from one to ten (worst to best); this variable will be changed to a binary response of poor (1-5) or fine (6-10) for the purpose of this project. All physiochemical measures are represented as continuous variables and are as follows: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. There is an additional categorical variable for grape type where 0 represents red grapes and 1 represents white grapes. The objective of this analysis is to examine the relationships between the predictor variables and wine quality in the hope to identify important measures that can predict a wine’s quality so that the subjective assessment of wine quality and pricing can become more controlled. The ultimate goal is to develop an objective model using some or all of these variables to predict wine quality.

## Exploratory Analysis

This analysis begins by first looking at histograms (APPENDIX EDA A) of the variables to assess for any deviations in normality or potential outliers in the data. Of utmost concern is the response variable as it a discrete variable based on an ordinal ranking. Looking at this plot, it can be seen that the full range is [3, 9] where the majority of wine rankings fall between the 5 to 7 range and that the data do not appear to deviate far from a normal distribution. Add in chart for binary response and description.

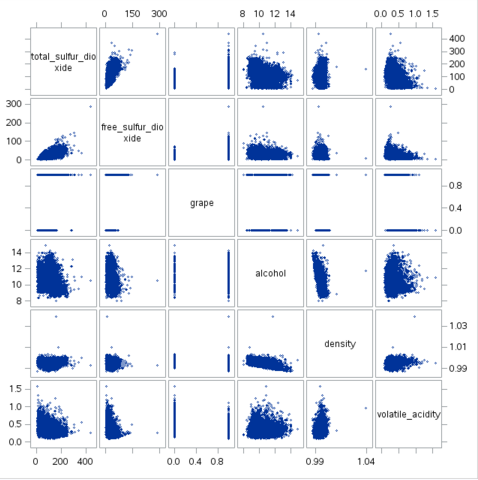
Many of the explanatory variables appear to have large outliers which means transformations might be needed with model fitting. Fixed acidity, volatile acidity, and citric acidity appear to have a relatively normal and symmetric distribution but have large outliers; chlorides, free sulfur dioxide, and total sulfur dioxide have similar characteristics but are not as symmetric.

Residual sugar is skewed right which could indicate the need for a logarithmic transformation. Density and alcohol have irregular shapes but do not appear to have outliers. pH and sulfates show signs of symmetry but both have a right skew; neither appear to have outliers.

Now that the individual variables have been examined, the relationships between the predictor variables themselves and response variable need to be examined. This can be done by examining scatterplots and correlations in order to determine data patterns and assure there is not multicollinearity within the model.

In Appendix EDA B, Pearson correlations between all variables are shown. Any correlation larger than 0.4 has been highlighted as this is indicative of a high correlation; as it can be seen there are many correlations which are higher than this threshold. This needs to be considered when the model is built as any underlying linear patterns between the predictor variable will cause redundancy within the model and potentially inflate model performance.

The largest of these correlations are between total sulfur dioxide and free sulfur dioxide at 0.72093, total sulfur dioxide and grape at 0.70036, alcohol and density at -0.68675, and grape and volatile acidity at -0.65304 (this likely okay to include since one of the variables is categorical and a linear relationship is very unlikely); in fact the all response variable combinations with high correlation include at least one of the previously mentioned variables. It is likely than any combination of these variables within a model will cause fit issues including inflation.

Additionally, the scatterplot between these variables show a linear trend between total sulfur dioxide and free sulfur dioxide and positive fanning patterns between both total sulfur dioxide and volatile acidity and free sulfur dioxide and volatile acidity. All of this evidence shows that including both total sulfur dioxide and free sulfur dioxide in the model will cause inflation. Of note, a scatter plot of all variables was not included due to SAS limitations.

When looking for any high correlation between the non-transformed response variable and a predictor variable the only high correlation seen is between alcohol and quality at 0.44432. There are also a couple moderate correlations between density and quality at -0.30586 and volatile acidity and quality at -0.26570.

Pricipal component analysis of the data indicated that there is a slight deviation between fine and poor wines indicating that further investigation is warranted. PCA plot between principal compent 1 vs principal compent 2, scree plot and eigenvectors can be found in Appendix EDA C.

# Objective 1 – Logistic Regression Model

## Problem Statement and Approach

By using various prediction modeling methods (logistic. LDA/QDA, random forest) we seek to predict if a wine is ‘fine’ or ‘poor’ based on its’ reported attributes. We first used basic logistic regression on the data provided to us from the kaggle database without any transformation or selection process. We then transformed the variables that were not normally distributed and applied a forward selection method to the logistic regression. The two models’ prediction efficiency is then compared to a quadratic discriminant analysis and random forest analysis approach.

## Assumptions

The response variable, wine quality, must be binary. In order to meet this assumption all wines with a wine quality of 5 or less is assigned a ‘poor’ outcome, more than 5 quality wines is assigned as ‘fine.’ Secondly, each observation is assummed to independent from one another.

Kelly can you take care of this? I am not 100% on the assumptions for logistic regression

## Model Fit

The Hosmer-Lemshow test concluded that the model had a good fit (p = 0.9932). There was no noticebly large outliers or leverage points, diagnostic table can be found in the appendix. The AIC model fit statistics is 825.998. The highest correct prediction percentage is at 75% with the cutoff set at 50%. At this setting the sensitivity is 75.1% and specificity is 74.9%. SAS output can be found in the appendix.

## Parameter Interpretation

Log(odds) = 118.5 + 0.1(fixed acidity) -3.1(volitile acidity) – 1.9(citirc acid) + 0.2(residual sugar) + 0.008(free sulfur dioxide) – 0.01(total sulfur dioxide) – 123.4(density) – 1.3(pH) + 4.9(sulphates) + 0.7(alcohol) – 8.6(chlorides)

# Objective 2 – Logistic Regression 2

## Summary

While running the logistic regression model using the provided data it was noticed that some of the variables were not normally distributed and were very skewed, namely residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, alcohol and fixed acidity. A log transformation was applied to all the named variables except for the fixed acidity as it had 0’s.

## Model Fit

After applying the Forward selection method to the log transformed and non-transformed variables seven variables were selected by the algorithm. The Hosmer-Lemshow test concluded that the model had a good fit (p = 0.7738). There was no noticebly large outliers or leverage points, diagnostic table can be found in the appendix. The AIC model fit statistics is 825.142. The highest correct prediction percentage is at 75.5% with the cutoff set at 50%. At this setting the sensitivity is 76.7% and specificity is 74.1%. SAS output can be found in the appendix

## Parameter Interpretation

Log(odds) = -11.7 – 1.1(log(chlorides)) – 0.3(log(total sulfur dioxide)) + 3.6(log(sulphates) + 9.0(log(alcohol)) – 3.2(volatile acidity) – 2.2(citric acid) – 2.0(pH)

## Conclusion

After log transformation of the varables and application of the selection method the sensitivy of the model improved slightly. The more “complicated” log regression model has a slightly lower AIC value and is 1% more sensitive with a smal drop, < 1%, in specificity. Even so the changes were neligible making the changes unimportant to our prediction output.

# Objective 2 – Linear Discriminant Analysis or Quadratic Discriminant Analysis

## Assumptions

From the EDA above variables that are not normally distrubuted have been log transformed, except for the fixed acidity variable.

Each wine is assumed to be independent from the others.

The Homogeneity of the covariances is violotated, p value less than 0.0001 indicating a rejection of the null hypothesis that there is homogeneity between the covariances. Therefore, Quadratic Discriminant Analysis will be used. See appendix for SAS output.

## Model Fit

Using all the available variables the QDA’s error rate for detecting fine wines were 26.5% and for detecting poor quality wines are 28.4% giving us a total of 27.47%, a little more than 25% acceptable cuttoff. This is assuming that there is a 50:50 split between fine and poor wines.

## Conclusion

Compared to our logistic regression models’ error rate the logistic regression did a nonimally better job. To obtain the error rate we used the classification table applied to the training data set. We chose to set the probablity level at 0.5000 because it gave the highest correct classification percentage, 75.5%. The error rate for detecting fine quality wine is 100 – 76.7 = 23.3% (100 – sensitivity) and the error rate for detecting poor quality wines is 100 – 74.1 = 25.9% (100 – specificity).

# Objective 2 – Nonparametric approach

## Problem Statement and Approach

By using non-parametric model approach detect the quality of wine “poor” or “fine” using random forest approach.

## Assumptions

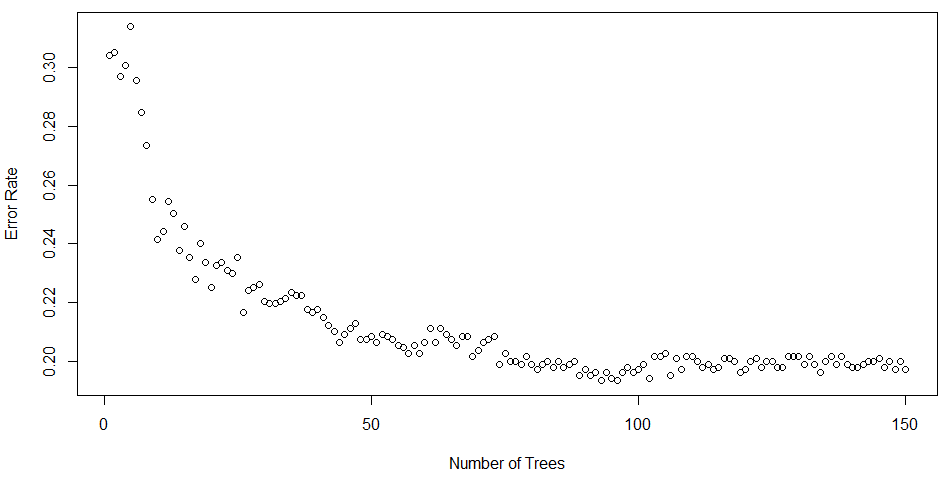
## Wine with quality rate of 5 and less called “poor” wine and wine 6 and more “good” wine.

## Split observation into train set as 2/3 of data and 1/3 is as test set.

## Each wine is assumed to be independent from the others.

## Model Fit

Initially we used all available variables for model creation and detecting error rate but finally detected that using 5 variables is enough to get best error rate. The best amount of trees based on error rate graph is around 100.

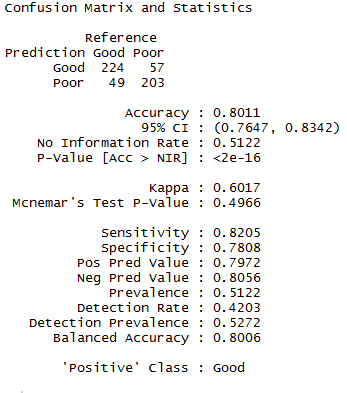


## 

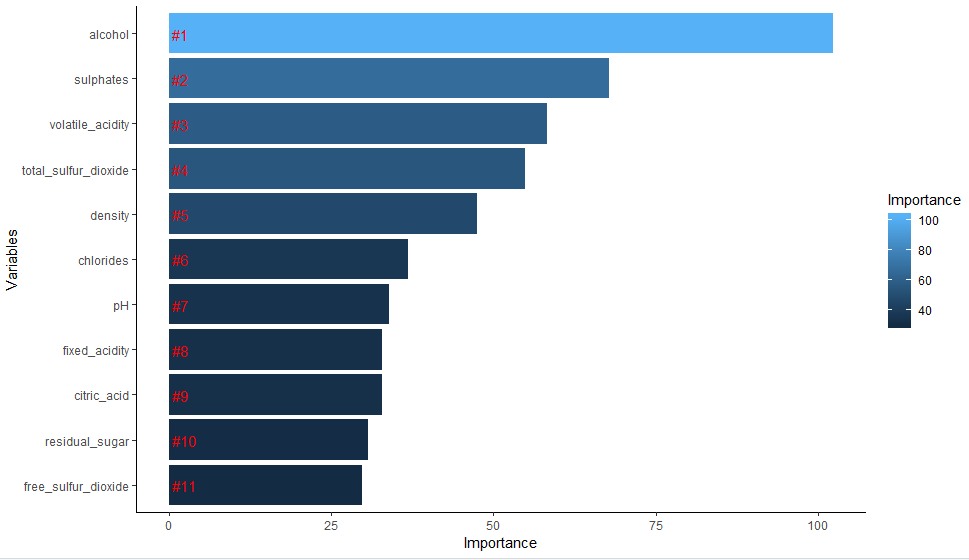
## Conclusion

Utilizing Random forest to obtain the error rate for detecting “fine” quality wine is 100 – 81.3 = 18.7% (100 – sensitivity) and the error rate for detecting poor quality wines is 100 – 78.08 = 22% (100 – specificity).

**Confusion matrix result:**



### **Variable Importance**



The overall accuracy of our model is pretty good at around 80% overall. However, we could clearly see that it is much better in predicting bad wines than good ones.