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](#_Toc530320891)

[Data Description 3](#_Toc530320892)

[Exploratory Analysis 3](#_Toc530320893)

[Objective 1 – Logistic Regression Model 3](#_Toc530320894)

[Problem Statement and Approach 3](#_Toc530320895)

[Assumptions 3](#_Toc530320896)

[Model Fit 3](#_Toc530320897)

[Parameter Interpretation 3](#_Toc530320898)

[Conclusion 3](#_Toc530320899)

[Objective 2 – Logistic Regression 2 4](#_Toc530320900)

[Summary 4](#_Toc530320901)

[Model Fit 4](#_Toc530320902)

[Parameter Interpretation 4](#_Toc530320903)

[Conclusion 5](#_Toc530320904)

[Objective 2 – LDA 5](#_Toc530320905)

[Problem Statement and Approach 5](#_Toc530320906)

[Assumptions 5](#_Toc530320907)

[Model Fit 6](#_Toc530320908)

[Parameter Interpretation 6](#_Toc530320909)

[Conclusion 6](#_Toc530320910)

[Objective 2 – Nonparametric approach 6](#_Toc530320911)

[Problem Statement and Approach 6](#_Toc530320912)

[Assumptions 6](#_Toc530320913)

[Model Fit 6](#_Toc530320914)

[Parameter Interpretation 6](#_Toc530320915)

[Conclusion 6](#_Toc530320916)

# 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 Analysis

Kelly

# Objective 1 – Logistic Regression Model

## Problem Statement and Approach

Kelly

## Assumptions

Kelly

## Model Fit

Kelly

## Parameter Interpretation

Kelly

## Conclusion

Kelly

# 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. 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

# 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

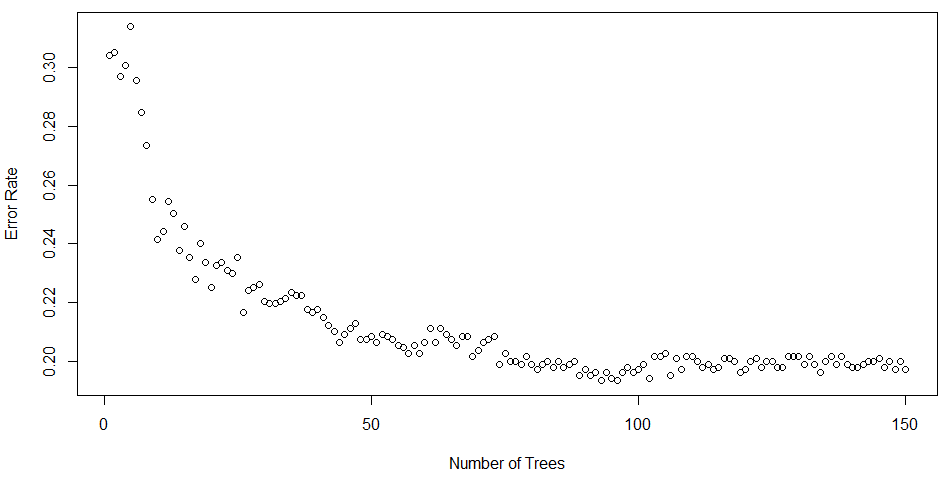
## Wine with quality rate of 5 and less called “poor” wine and wine 6 and more “good” wine.

## 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.



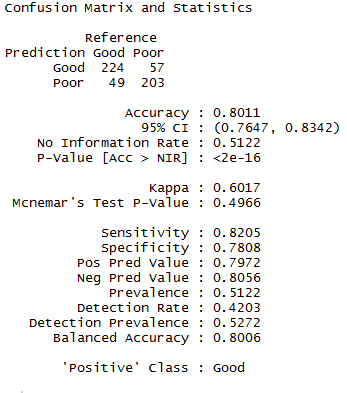
## 

## 

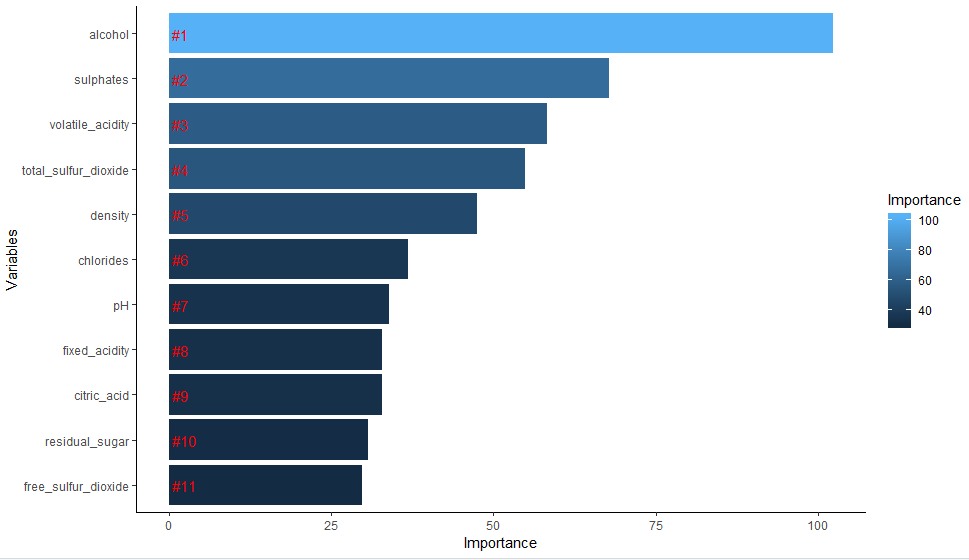
## Conclusion

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.