Classification of Wine Qualities

Robert Dam

2021-12-10

Introduction

Expertise in winemaking leads to high quality wines. Winemakers have worked towards the tried and true methods, but why do those methods work? Are we able to differentiate high quality wines from low quality ones with more than just a taste test? In this study, we will examine several physicochemical properties of low quality wines compared to high quality wines. These physicochemcial properties include measures like pH, sulphates, density, residual sugar, and more. If we are able to accurately classify wines from physicochemical properties, we gain an additional perspective into what makes a high quality wine. This would provide a quantitative measure of wine quality. Such insight can provide additional guidance for winemakers to understand the effectiveness of their techniques. Additionally, one could target certain levels of these properties to create higher quality wines.

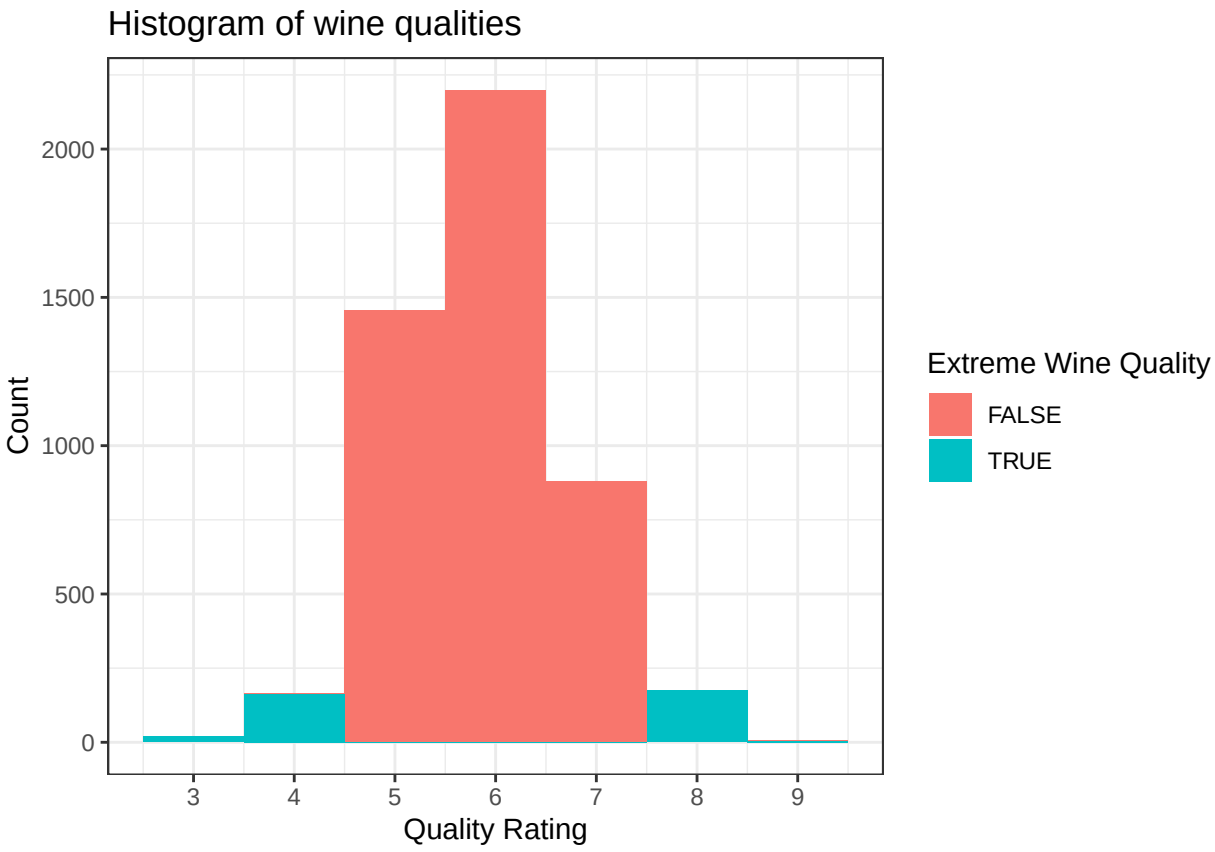
With this goal in mind, we have a couple questions of interest. First, we would like to know how effective this method of wine quality classification is. Can we accurately differentiate low and high quality wines with physicochemical properties? We would also like to know which physicochmecial properties are important in this classification task. Do all of these properties provide useful information or are there only some that matter? We will be using multiple methodologies to answer these questions. We will run multiple logistic regression and linear discriminant analysis to classify wines by quality. We will then select the one that performs better by the criterion of classification accuracy on a test set. To examine which variables are important we will use PCA and best subset selection for logistic regression.

Data Description

(Cortez et al., 2009)

Our data comes from UC Irvine's Machine Learning Repository. It contains nearly 4,900 samples of a white wine variant of the Portuguese vinho verde wine. Vinho verde is a wine that comes from the Minho region of Portugal and it accounts for about 15% of total Portuguese wine production. These samples come from lab tests conducted by a wine certification organization, the CVRVV, from May 2004 to February 2007.

The response variable is sensory data quality ratings on an integer scale of 0 (very bad) to 10 (excellent). Blind taste tests were conducted to rate each wine sample and the median score is taken as the quality rating. As we can see from the histogram of wine qualities, the data is unbalanced. Most of the wines in this data set have average/good ratings of 5, 6, or 7. Also, the wine qualities in the data set only range from 3 to 9. Our goal is more to classify between very high and very low quality wines rather than average quality ones so we will only consider wines of ratings 3, 4, 8 and 9. We will consider wines of quality 3 and 4 to be low quality and wines of quality 8 and 9 to be high quality. With this new subset, we have 363 total samples, with balanced classes: 183 are considered low quality and 180 are considered high quality. We futher split 70% of our data into a training set and 30% into a test set to calculate test prediction error.



Our possible predictor variables are physicochemical properties of each wine sample from lab tests. The table below contains each variable name and the units they are represented in.

Table 1: Units of Physicochemcial Properties in Data Set

Property	Units
Fixed acidity	g(tartaric acid)$/dm^3$
Volatile acidity	g(acetic acid)$/dm^3$
Citric acid	g/dm^3
Residual sugar	g/dm^3
Chlorides	g(sodium chloride)$/dm^3$
Free sulfur dioxide	mg/dm^3
Total sulfur dioxide	mg/dm^3
Density	g/cm^3
pH	pH
Sulphates	g(potassium suplhate)$/dm^3$
Alcohol	$vol\%$

Exploratory Candidate Models: Logistic Regression and LDA

To classify low quality wines from high quality ones we will consider two methodologies: logistic regression and linear discriminant analysis (LDA).

With no prior knowledge, we try an exploratory initial fit, fitting the model with every possible parameter. For logistic regression, this results in a quite strong total classification accuracy of 85.32% on the test set.

Within low quality wines this accuracy is 87.27% and within high quality wines this accuracy is 83.33%. For LDA, we have similar, but lower classification accuracies of 83.49% total, 85.45% within low quality wines, and 81.48% within high quality wines. This initial result suggests that either modelling approach can lead to some strong results that can bring insight into how these physicochemical properties influence wine quality. Initially, logistic regression performs better, but we need to do more investigation on which variables to select before settling on a final model.

However, there are problems with the naive full first-order models. We see from the table below that many of the logistic regression coefficients are not significant at level 0.05 which means they may not provide useful information for classifying wine qualities. As we can also see from the VIF table, despite mostly low pairwise correlations between our predictor variables, there is some multicollinearity present in the full model. Residual sugar, density, and alcohol all have very high VIF values. This means their coefficients will have higher variances, the model will generalize less reliably, and the model is harder to interpret.

Table 2: VIF scores and coefficient p-values for predictor variables in full model

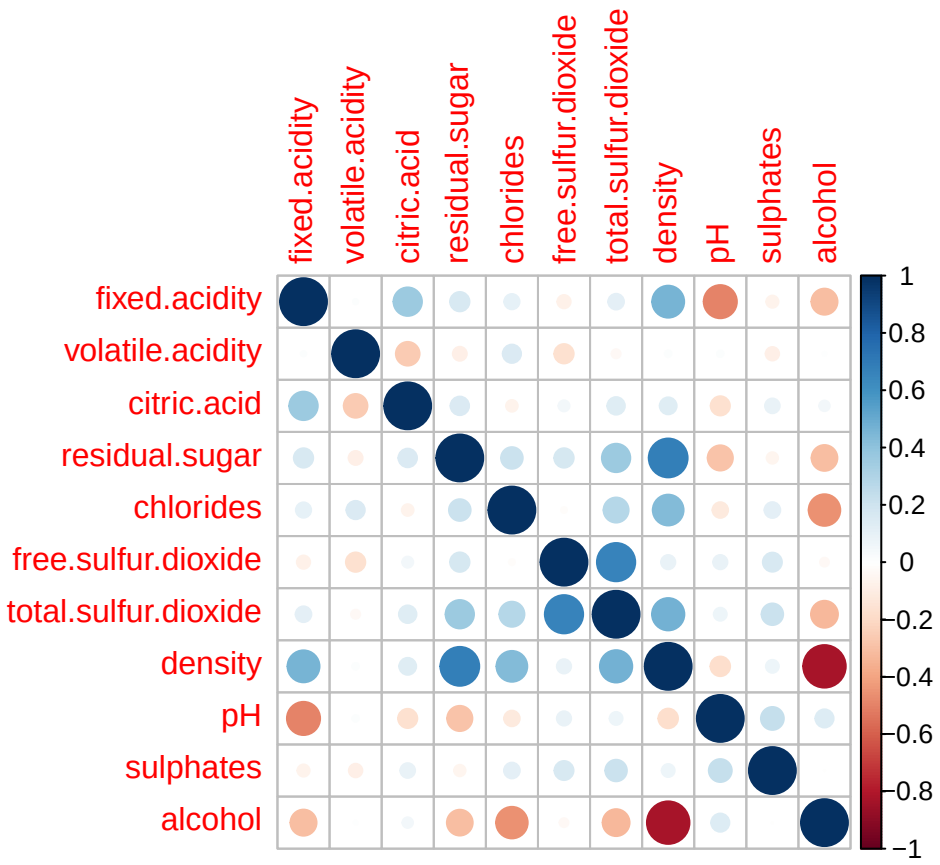
Variable	VIF	p-value of coefficient
Fixed acidity	4.20	0.13
Volatile acidity	1.57	2.81e-05 (*)
Citric acid	1.32	0.69
Residual sugar	15.33	0.0005 (*)
Chlorides	1.39	0.51
Free sulfur dioxide	2.48	0.32
Total sulfur dioxide	3.38	0.84
Density	47.15	0.03 (*)
pH	3.43	0.007 (*)
Sulphates	1.44	0.80
Alcohol	14.92	0.52

Variable Selection: Best Subset Selection

So to avoid the issues with the full first-order model and gain an understanding of crucial variables we want to find a subset of important variables. To gain insight into this problem we conduct principle component analysis (PCA) and best subset selection procedures.

a) Correlation plot

We first examine the correlation structure between our predictor variables using correlation plots.



We first create a Correlogram using the "corrplot" package to make it easier to visualize and compare the correlation coefficients.

The blue circles represent positive correlation coefficients, while the red circles represent negative correlation coefficients. The darker and larger circles represent correlation coefficients that are relatively large in absolute value.

We observe that alcohol and density have the largest correlation coefficient (in absolute value) of -0.82. Furthermore, residual sugar and density have the second largest correlation coefficient (in absolute value) of 0.69. Besides these pairs of variables, we observe that most of our variables are not highly correlated with each other.

However, there are a few variables that raise concerns about multicollinearity. The most obvious one is density. We observe from the Correlogram that density appears to have a relatively moderate to high degree of correlation with 5 other variables. We also observe that alcohol and total sulfur dioxide each have a moderate to high level of correlation with 4 or 5 other variables.

b) Best subset selection

We can also perform an exhaustive best subset selection procedure for logistic regression to select the set of variables that perform best by the Bayesian Information Criterion (BIC). Since we only have eleven predictor variables under consideration an exhaustive search is computationally feasible. Best subset selection will choose the subset of variables that leads to the lowest BIC logistic regression model. We can then compare the results of this procedure to the correlation and PCA analyses above to check for consistency.

From this procedure, we arrive at a logistic regression model with five variables: fixed acidity, volatile acidity, residual sugar, density, and pH. Comparing the resulting variables with our observations above, we see that

we do not include the potential problem variables of alcohol, and total sulfur dioxide. We do include density, but in the absence of those other correlated variables this should not be too much of an issue.

Fitting Logistic Regression and LDA with Chosen Variables

a) Interpretation of results

We will now fit logistic regression and LDA with the variables chosen by the best subset selection procedure: fixed acidity, volatile acidity, residual sugar, density, and pH.

From our chosen logistic regression model, we have two variables with negative coefficients: volatile acidity and density. Both are strongly significant testing against the null hypothesis of no effect, with z-values of -4.58 and -7.86 respectively. This suggests that holding all other variables constant, decreases in volatile acidity or density lead to increases in the probability of a wine being high quality.

The other three variables: fixed acidity, residual sugar, and pH have positive coefficients high significance levels. Their z-values are 2.61, 7.21, and 4.25 respectively. We can interpret these coefficients as holding all other variables constant, increasing these variables marginally will lead to increases in the probability of a wine being high quality. Do note that this result only applies to ranges where these variables were observed and predictions should not be extrapolated beyond that range.

If look at the linear discriminant analysis results on the same variables, we see that the group means of each variable for low quality versus high quality wines mostly align with the logistic regression results. In every case except for fixed acidity the group means match up with their logistic regression coefficient signs (i.e. higher group mean for high quality wines versus low quality wines match with positive coefficients and vice versa). This means that the controlling effect that our variables have on each other do not reverse the general trend of each group for each variable except for with fixed acidity.

LDA has similar performance to logistic regression and it provides us with another interpretation. There is approximately a separating hyper-plane between high and low quality wines based on our chosen variables. This is a relatively simple interpretation that says that wines with higher residual sugar, and pH, but lower fixed acidity, volatile acidity, and density within a given range will tend to be higher quality wines.

b) Comparing logistic regression and LDA by accuracy

Examining the classification accuracy table, we see that both models perform similarly, but logistic regression performs better across the board. We can also see that this logistic regression model has a small improvement of accuracy over the full model, but it also avoids the issues of overfitting and multicollinearity present in the full model. This model will likely have less variability and generalize better so we will take it as our final model.

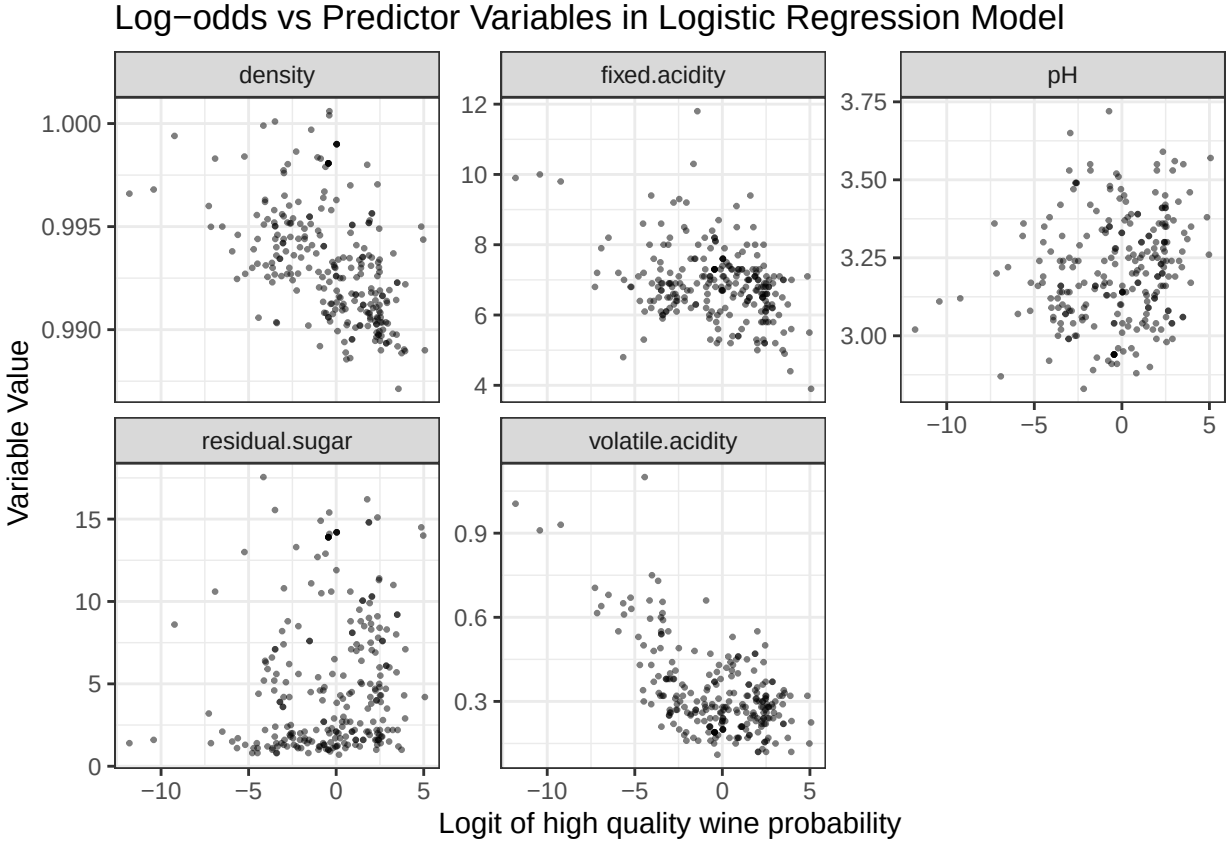
Table 3: Classification accuracy (on test set) of logistic regression and LDA with chosen variables

Model	Total Accuracy	Within Low Quality Accuracy	Within High Quality Accuracy
Logistic regression	86.24%	85.45%	87.04%
LDA	82.59%	80.00%	85.19%

Final Model Diagnostics

We now examine how well our selected logistic regression model follows the model assumptions.

First, logistic regression assumes that the predictors are linear with respect to the log-odds of the outcome variable. From the logit vs predictor variables scatter plots we can see that most of the chosen variables are quite linear with the logit of high quality wines. The only real concern may be that residual sugar tends to cluster along the low values, but other than that the linearity assumption seems sound. This also means that no transformations of variables are needed.



We also need to check the degree of multicollinearity in the fitted model. From the VIF table, we see now that most of the chosen variables have pretty low VIF scores. Only residual sugar and density have scores over 2, but they are still under 10 which indicates that there is not too much multicollinearity. This result is much better than in the full model fit case so it seems like this assumption holds reasonably.

This chosen model follows the assumptions reasonably, needs no further transformations, and has a pretty strong overall classification accuracy.

Table 4: VIF scores for predictor variables in chosen logistic model

Variable	VIF
Fixed acidity	1.89
Volatile acidity	1.29
Residual sugar	4.69
Density	5.32
pH	1.96

Conclusion

Our logistic regression is quite effective when it comes to wine quality classification since we can fairly accurately differentiate between low and high quality wines given their physicochemical properties. The total classification accuracy of our final logistic regression model was 86.24% in the test set. Within the low quality wines in the test set, the accuracy of our model was 85.45%, while it was 87.04% accurate within the high quality wines in the test set.

Furthermore, we found that not all of the physicochemical properties provide useful information for this classification task. Through principal component analysis, we found that it would not be useful to include alcohol, total sulfur dioxide, and density together in a model due to multicollinearity. Our best subset selection procedure using Bayesian Information Criterion (BIC) points to similar results, but suggests that we should include density in the model. This should not be a problem so long as we exclude the other two highly correlated variables.

We have some cautions with applying these results though. Our classification is between the more extreme quality wines. For average quality wines the data is much more mixed and classification may not be as fruitful. So we recommend to use this model more to distinguish between exceptionally high or low quality wines, but it will not necessarily perform as well to distinguish average quality wines from extreme quality wines.

In conclusion, the physicochemical properties that allowed us to create a fairly strong classification model were fixed acidity, volatile acidity, residual sugar, density, and pH. These results may be of interest to winemakers who might look for ways to create better quality wine for their consumers. Generally, our results support the idea that certain physicochemical properties should be focused on while trying to produce wine.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
log_reg_cm <- function(test, logreg) {
  predicted <- ifelse(predict(logreg,type="response",test) > 0.5, 1, 0)
  return(table(predicted, test$good))
}
library(ggplot2)
library(tidyverse)

#Read data
wine <- read.delim("winequality-white.csv", header=T, sep = ";")
wine$good <- factor(ifelse(wine$quality>=7,1,0))

#Only considering extreme wine qualities
wine_extr <- wine[wine$quality < 5 | wine$quality > 7,]

train_test_split <- function(seed, train_split, data) {
  set.seed(seed)
  train_index <- sort(sample(1:nrow(data),as.integer(nrow(data)*train_split),replace=F))
  train <- data[train_index,]
  test <- data[-train_index,]
  train_test <- list(train=train, test=test)
  return(train_test)
}

train_test_data <- train_test_split(39882, 0.7, wine_extr)

wine_extr_train <- train_test_data$train
wine_extr_test <- train_test_data$test
quality_labs <- 3:9
ggplot(data=wine, aes(x=quality)) + geom_histogram(bins=7, aes(fill=quality>7|quality<5)) +
  labs(title="Histogram of wine qualities", x="Quality Rating", y="Count") +
  scale_x_continuous(labels = quality_labs, breaks = quality_labs) +
  guides(fill=guide_legend(title="Extreme Wine Quality")) + theme_bw()
#Exploratory plots
hist(wine$fixed.acidity)
hist(wine$volatile.acidity)
hist(wine$citric.acid)
hist(wine$residual.sugar)
hist(wine$chlorides)
hist(wine$free.sulfur.dioxide)
hist(wine$total.sulfur.dioxide)
hist(wine$density)
hist(wine$pH)
hist(wine$sulphates)
hist(wine$alcohol)
hist(wine$quality)

#Exploratory Logistic Reg
exp_log_reg <- glm(good ~ . - quality, data = wine_extr_train, family = binomial)
summary(exp_log_reg)
car::vif(exp_log_reg)
```

```
logit_cm <- log_reg_cm(wine_extr_test, exp_log_reg)
logit_cm

#Exploratory LDA
library(MASS)
exp_lda_fit <- lda(good ~ . - quality, data = wine_extr_train)
exp_lda_fit
exp_lda_cm <- table(predicted = predict(exp_lda_fit,wine_extr_test)$class,actual = wine_extr_test$good)
exp_lda_cm
wine_features <- wine_extr[-c(12,13)]
corrplot::corrplot(cor(wine_features))
#Best subset selection for logistic regression
wine_for_bestglm <- wine_extr_train
wine_for_bestglm$quality <- NULL
names(wine_for_bestglm)[names(wine_for_bestglm) == "good"] <-  "y"
logit_bestglm <- bestglm::bestglm(wine_for_bestglm, family=binomial,IC="BIC",method="exhaustive")
summary(logit_bestglm$BestModel)

bestmodel_cm <- log_reg_cm(wine_extr_test, logit_bestglm$BestModel)
bestmodel_cm
#Exploratory LDA
lda_fit <- MASS::lda(good ~ fixed.acidity + volatile.acidity + residual.sugar + density + pH, data = wi
lda_fit
lda_cm <- table(predicted = predict(lda_fit,wine_extr_test)$class,actual = wine_extr_test$good)
lda_cm
#Checking linearity with logit assumption
logit <- log(predict(logit_bestglm$BestModel, type = "response")/(1-predict(logit_bestglm$BestModel, ty
predictors <- colnames(wine_extr)
wine_for_plot <- wine_extr_train[,c(1,2,4,8,9)] %>%
  mutate(logit = logit) %>%
  gather(key = "predictors", value = "predictor_val", -logit)
ggplot(wine_for_plot, aes(logit,predictor_val)) + geom_point(size = 0.5, alpha = 0.5) +
  facet_wrap(~predictors, scales = "free_y") + labs(title="Log-odds vs Predictor Variables in Logistic

#VIF of best subset model
#car::vif(logit_bestglm$BestModel)
```

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.