Grading wines from physicochemical properties with proportional odds model

Maobin Guo

Summary

This report analyzes the impact of physicochemical properties[1] on wine grades by examining the association between the properties and the wine flavor scores. To explore the relationship, a proportional odds model was used in this article. The data used in this paper was created by Cortez et al. (1998). It includes two datasets related to red and white variants of the Portuguese "Vinho Verde" wine.[2]

Introduction

Establishing an objective method to evaluate wine quality has always been a hot issue for the industry. Such an objective evaluation is valuable because it can help winemakers produce better wines and help wine merchants set suitable prices for thousands of wines more wisely. Furthermore, consumers can also have a reliable approach to choosing the best wines. The model established in this article is a practical attempt to answer this research question. Specifically, this article will try to use the obtained model to answer the following interesting questions:

1. Which factors can most affect the quality of wines?
2. Which factors have the same effect on both red wine and white wine? Which factors only affect red wine or white wine?
3. In recent years, sulfur dioxide-free wine has been made popular because some vendors claim that sulfur-dioxide-free wines tasted better. Will the data and model in this article support this kind of business trend?

Data

Data Description

This analysis consists of two datasets: white wine containing 4898 observations and red wine containing 1599 observations. The two datasets share the same structure and attributes, and there are no missing fields. The two separate datasets were combined into a big dataset by adding a new column to indicate red wine or white wine. The predictor variables include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol; and the response variable is wine quality. All variables are continuous; however, some variables will be categorized for modeling. The detailed information of this transformation will be described in the following section.

Some variables seem to be associated with similar physicochemical properties, for example, fixed acidity and volatile acidity. This situation may raise concerns about the high correlation between the predictors. However, correlation analysis confirms that there are no high correlations between these variables. For example, the absolute correlation coefficient value shows correlation between fixed acidity and volatile acidity is as low as 0.21. The researchers may have already considered this problem when they chose measurement indicators.

[1] The physicochemical properties includes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

[2] For more details, please refer http://www.vinhoverde.pt/en/

Table 1: Quality Classification

quality	class	data size
quality <5	acceptable	2384
5 <= quality <7	good	2836
quality >= 7	premium	1277

Data Transformation

The type of response variable 'quality' was transformed from numerical to ordinal to apply the proportional odds model. To do so, the wines were classified into three groups: acceptable, good, and premium. The rule to classify the wine's grades is recorded in table 1. Another variable that was categorized is density because its distribution is too skewed, and it cannot be fixed by a transformation such as log operation. Wines whose density is higher than 1.00 (excluded) will be tagged as 'heavy', while wines whose density is lower than 1.00 (included) will be tagged as 'light.' Additionally, chlorides level values were log transformed because the distribution is significantly skewed.

Exploratory Data Analysis (EDA)

Generally, red wine and white wine are drunk in different situations and people grade them differently. Due to these dissimilarities, people tender to grade the two kinds of wines with distinct criteria. For example, the white wine is normally served at lower temperatures, making people less sensitive to its high acidity. Moreover, many people prefer high acidity when they drink white wine because the high acidity can complement white meat or seafood . For red wine, people tend to find a balance between acid levels and acid-producing red meat. Hence, many interaction effects on predictors are expected between red wine and white wine. There are five obvious interactions identified by exploratory data analysis and were confirmed in the model selection phase. The predictors that involve in interactions include: citric acid, volatile acidity, fixed acidity, pH, and sulphates. As expected, all four indicators related to acidity (citric acid, volatile acidity, fixed acidity, and pH) have interaction effects between white wine and red wine. In addition to acid-related factors, sulphates also play different roles in the quality of red wine and white wine. Figure 1 displays these five interaction effects.

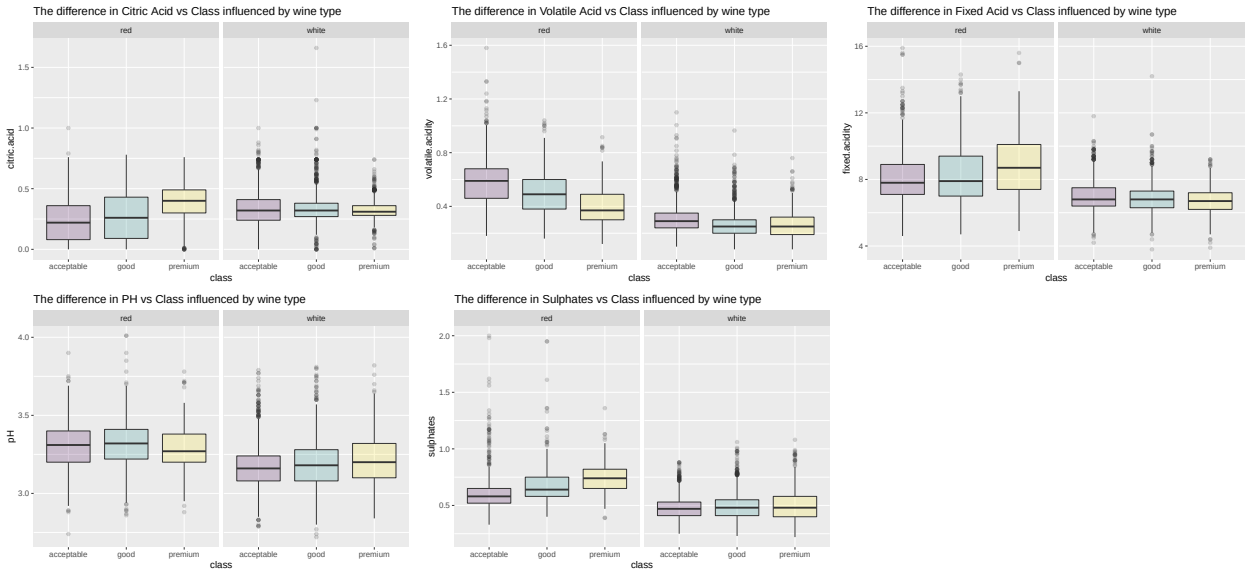

Figure 1: Interactions (Data Source: Cortez et al)

Model

Model selection

The model selection begins from a plain combination of all predictors and gradually adds interactions and transformations into the model. At first, the five interactions found in EDA phase were added, and these interactions actually decreased the AIC (Akaike Information Criterion) significantly. In addition to the decrease in AIC, the p-value of these interaction items also confirms that they are statistically significant. The residual sugar also shows some signs of interaction in the exploratory data analysis; however, the model shows that it is not statistically significant. Therefore, the interaction of residual sugar was discarded. After the exploration of interactions, the predictors' distributions were also checked. Many predictors' distributions are skewed; however, transforming them into normalized distribution did not improve the model's performance for most predictors except chlorides. Consequently, only the chloride was transformed. The other skewed predictors were preserved without any transformation.

Final model

$$\begin{aligned}\frac{Pr[Class_i \leqslant j|x_i]}{Pr[Class_i > j|x_i]} =& \beta_{0j} - \beta_1 Type_{i1} - \beta_2 PH_{i2} - \beta_3 Sulphates_{i3} - \beta_4 CitricAcid_{i4} - \\ & \beta_5 VolatileAcidity_{i5} - \beta_6 FixedAcidity_{i6} - \beta_7 ResidualSugar_{i7} - \\ & \beta_8 Log(Chlorides)_{i8} - \beta_9 FreeSulfurDioxide_{i9} - \beta_{10} TotalSulfurDioxide_{10} - \\ & \beta_{11} DensityBi_{i11} - \beta_{12} Alcohol_{i12} - \beta_{13} Type * PH_{i13} - \\ & \beta_{14} Type * Sulphates_{i14} - \beta_{15} Type * CitricAcid_{i15} - \beta_{16} Type * VolatileAcidity_{i16} - \\ & \beta_{17} Type * FixedAcidity_{i17} - \beta_{18} Type * ResidualSugar_{i18}, j = 1, 2\end{aligned}$$

Model Summary

The baseline values incorporated in the intercept are red wine, heavy density and values of zero for all continuous predictors.

According to the model summary (Appendix II), sulphates, citric acid, volatile acidity, fixed acidity, residual sugar, log(chlorides) (log term of chlorides), free sulfur dioxide, total sulfur dioxide, density, alcohol, and interactions of pH and type, sulphates and type, citric acid and type, volatile acidity and the type and fixed acidity and type are strongly associated (statistically significant) with the differences in grade at the 0.05 level. The two intercepts of grades, namely, between acceptable vs. good; and good vs. premium are also statistically significant. Here are some important factors:

1. For any fixed wine level and controlling other factors, if a wine's sulphates increases $1 g/dm^3$, the estimated odds of the wine to get a higher grade will increase by 1069.55 % (p<0.01). The likely range (95% CI) of this value is (520.88% , 2103.09%).
2. For any fixed wine level and controlling other factors, if a wine's citric acid increase $1 g/dm^3$, the estimated odds of the wine to get a higher grade will decrease by 78.89 % (p<0.01). The likely range (95% CI) of this value is (48.83%, 91.29%).
3. For any fixed wine level and controlling other factors, if a wine's volatile acidity increases $1 g/dm^3$, the estimated odds of the wine to get a higher grade will decrease by 96.93 % (p<0.01). The likely range (95% CI) of this value is (93.27%, 98.6%).
4. For any fixed wine level and controlling other factors, if a wine's alcohol increases $1 vol.\%$, the estimated odds of the wine to get a higher grade will increase by 153.79 % (p<0.01). The likely range (95% CI) of this value is (139.47% , 168.96%).
5. For any fixed wine level and keeping other variables constant, for 1 unit increase in PH of a white wine, its estimated odds to get a higher grade will increase by 127.87 %. The likely range (95% CI) of this value is (64.29% , 216.04%).

6. For any fixed wine level and keeping other variables constant, for $1g/dm^3$ increases in sulphates of a white wine, its estimated odds to get a higher grade will decrease by 63.87 %. The likely range (95% CI) of this value is (19.07%, 83.87%).

7. For any fixed wine level and keeping other variables constant, for $1g/dm^3$ increases in citric acid of a white wine, its estimated odds to get a higher grade will increase by 327.41 %. The likely range (95% CI) of this value is (55.68% , 1073.44%).

8. At the significance level of p<0.01, keeping other variables constant, for $1g/dm^3$ increases in volatile acidity of a white wine, its estimated odds to get a higher grade will decrease by 80.18 %. The likely range (95% CI) of this value is (45.94%, 92.73%).

Due to space limitations, the Confidence interval(CI) data appears in the appendix III.

Model Assessment

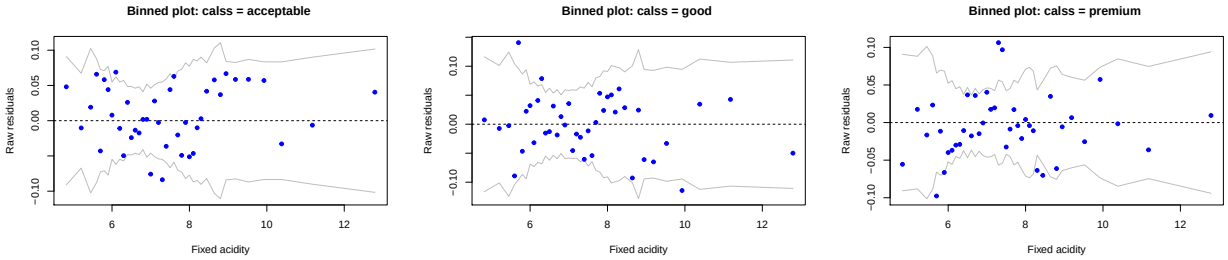

Figure 2: Residual binplot for fixed.acidity (Data Source: Cortez et al)

Table 2: Confusion Matrix (Data Source: Cortez et al)

	Sensitivity	Specificity
Class: acceptable	0.63	0.81
Class: good	0.66	0.55
Class: premium	0.30	0.94

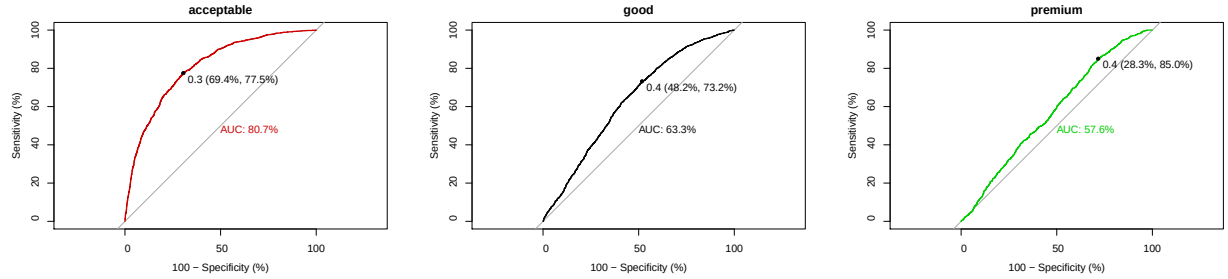

Figure 3: ROC Curve (Data Source: Cortez et al)

According to the model summary table, most of the predictors are observed to be significant at the conventional 0.05 significance level. From the binned raw residuals versus every continuous predictors (Appendix I), it can be seen that almost all the residual points randomly appear within two standard errors of the mean. Therefore, we believe that our model is sufficient to accurately estimate the population and answer the research questions of interest.

The confusion matrix shows that the sensitivity is satisfactory for classification of acceptable (0.63) and good (0.66) classes. However it is poor for the premium (0.3) class. The confusion matrix also suggests that while

the specificity is high for classification of acceptable (0.81) and premium (0.94) classes, however it is markedly lower for the good class(0.55).

The ROC curves indicate that the accuracy of classification for acceptable, good and premium are 80.69%, 63.32% and 57.63%, respectively, which means the model's outcome for the acceptable class is more realiable than that of the good and premium classes. In general, according to the ROC curves, the model's accuracy is acceptable.

Discussions

According to the model, alcohol content is the most important factor that affects both red wine and white wine. The high alcohol content could increase the odds of a particular wine to get a better quality classification. On the contrary, high chlorides and heavy density will lower the probability of a wine to get a better grade. Residual sugar, free sulfur dioxide, and total sulfur have only a slight impact on wine quality.

For red wines, the most important factors for improving their odds of a good rating are ranked in order of importance as follows: low volatile acidity, high sulphates, and low citric acid. While fixed acidity and PH value have much less influence on the red wine rating, especially the PH value. For white wines, the factors that will promote their ratings are ranked in order of importance as follows: low volatile acidity, high citric acid, low sulphates, and high pH. Low fixed acidity will also help white wines to get higher grade, but its importance is much less than the first four factors.

Finally, the two indicators of sulfur dioxide (free sulfur dioxide and total sulfur dioxide) in the model fail to show that free sulfur dioxide-free wine will taste better.

Conclusions

In summary, a proportional odds model was developed to assess the effects of various physicochemical properties on wine grades. Some interesting questions were answered on the basis of this model. These answers may provide some insights for both wine manufacture, merchants and consumers.

There are several limitations to this analysis, though. First, the sample size of red wine and white wine is imbalanced, and this problem may result in a biased model. Secondly, the classification accuracy for good and premium is not satisfactory. These problems may reduce the effectiveness of our model and conclusions. In order to obtain more effective results, some sophisticated machine learning methods, such as neural networks, support vector machines, and random forest, can be used in the future studies.

Reference

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (1998) Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47**, 547–553.

Appendix I (Binplots for all continuous predictors)

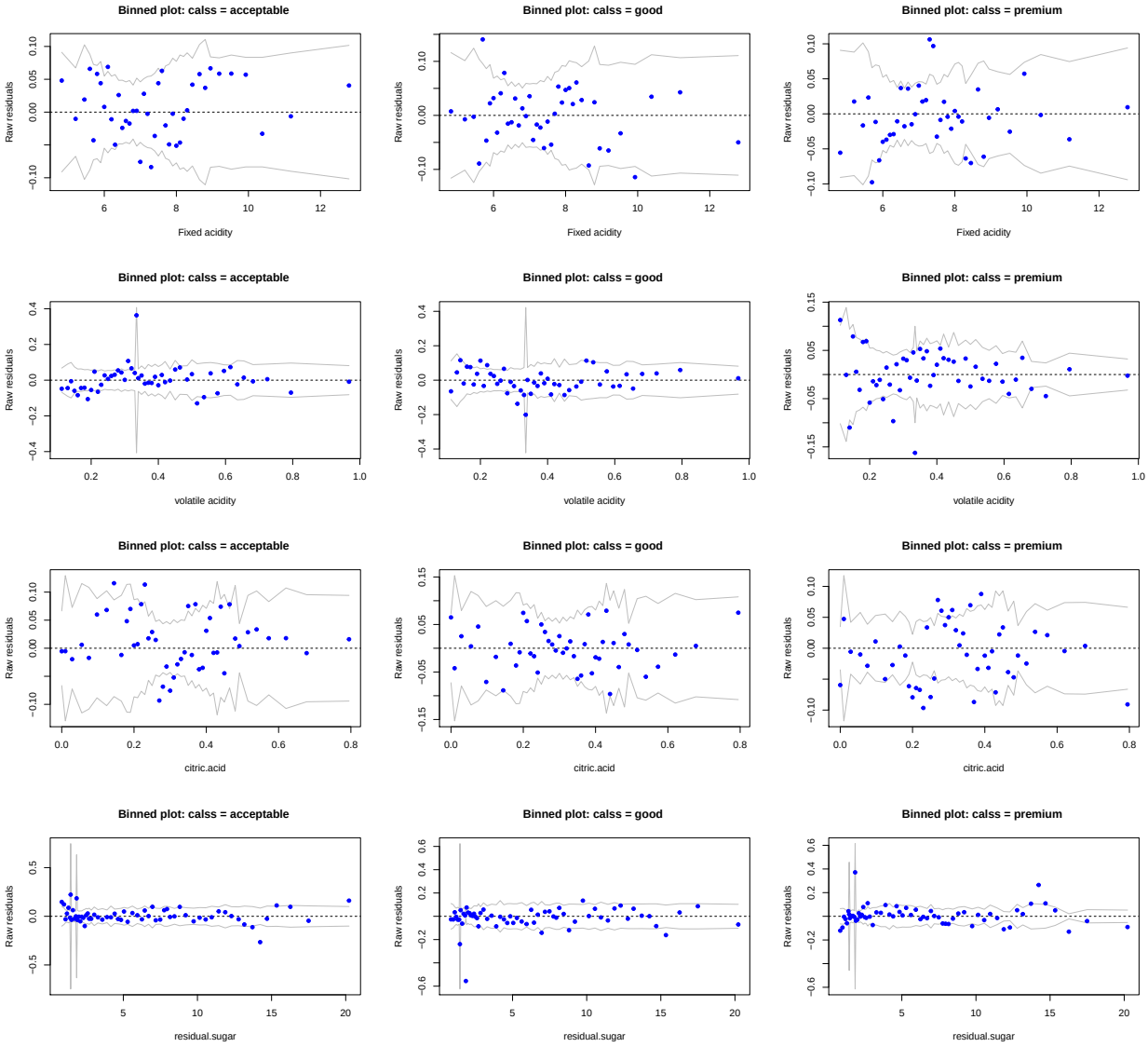

Figure 4: Residual binplots (Data Source: Cortez et al)

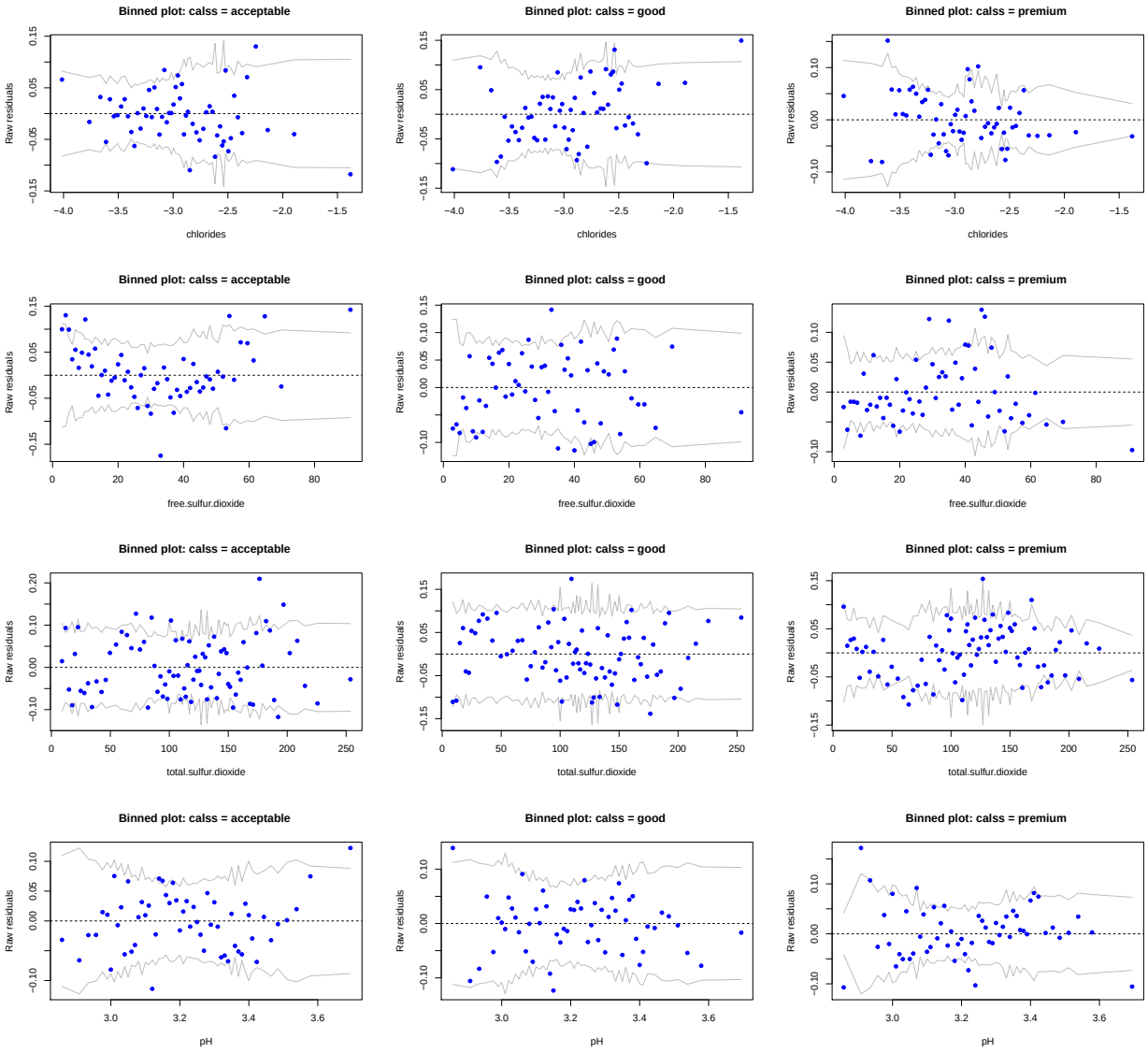

Figure 5: Residual binplots (Data Source: Cortez et al)

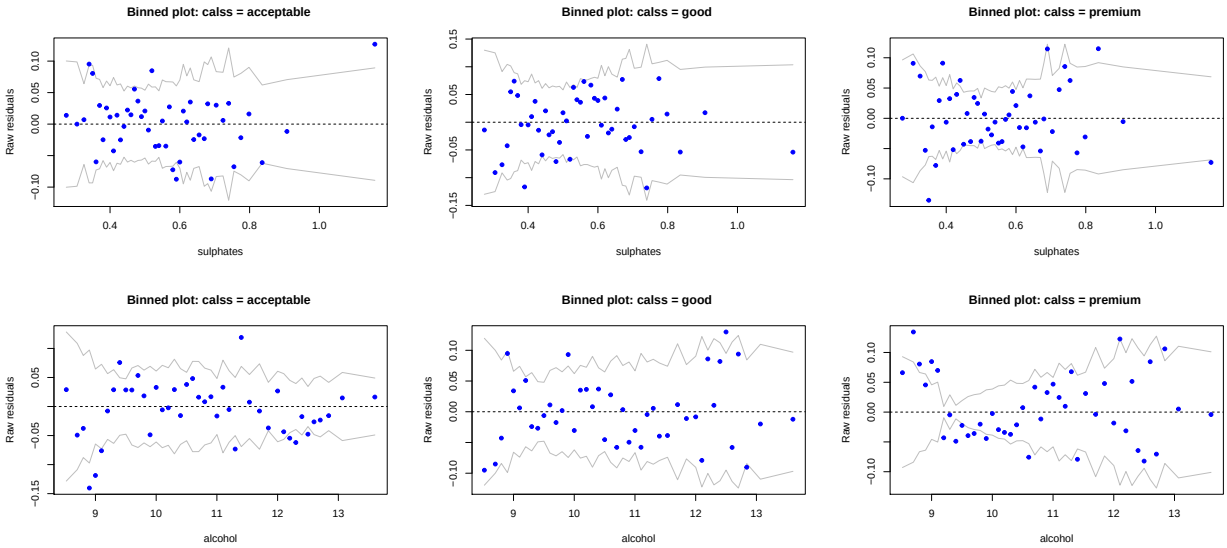

Figure 6: Residual binplots (Data Source: Cortez et al)

Appendix II (Model Summary)

Table 3: Model Summary

	Value	Std. Error	t value	p value
typewhite	0.01	0.44	0.01	0.99
pH	-0.08	0.16	-0.48	0.63
sulphates	2.46	0.32	7.61	0.00
citric.acid	-1.56	0.45	-3.44	0.00
volatile.acidity	-3.48	0.40	-8.70	0.00
fixed.acidity	0.21	0.04	5.36	0.00
residual.sugar	0.07	0.01	10.17	0.00
log(chlorides)	-0.39	0.09	-4.27	0.00
free.sulfur.dioxide	0.01	0.00	6.45	0.00
total.sulfur.dioxide	-0.01	0.00	-5.61	0.00
density_binlight	0.43	0.19	2.32	0.02
alcohol	0.93	0.03	31.45	0.00
typewhite:pH	0.82	0.17	4.93	0.00
typewhite:sulphates	-1.02	0.41	-2.47	0.01
typewhite:citric.acid	1.45	0.52	2.82	0.00
typewhite:volatile.acidity	-1.62	0.51	-3.16	0.00
typewhite:fixed.acidity	-0.28	0.05	-5.42	0.00
acceptable|good	11.91	0.49	24.32	0.00
good|premium	14.55	0.50	29.18	0.00

Appendix III (Confidence Interval)

Table 4: Confidence Interval

	2.5 %	97.5 %
typewhite	-0.86	0.87
pH	-0.38	0.23
sulphates	1.83	3.09
citric.acid	-2.44	-0.67
volatile.acidity	-4.27	-2.70
fixed.acidity	0.14	0.29
residual.sugar	0.06	0.08
log(chlorides)	-0.57	-0.21
free.sulfur.dioxide	0.01	0.02
total.sulfur.dioxide	-0.01	-0.00
density_binlight	0.07	0.79
alcohol	0.87	0.99
typewhite:pH	0.50	1.15
typewhite:sulphates	-1.82	-0.21
typewhite:citric.acid	0.44	2.46
typewhite:volatile.acidity	-2.62	-0.62
typewhite:fixed.acidity	-0.39	-0.18

Appendix IV (Multiclass ROC)

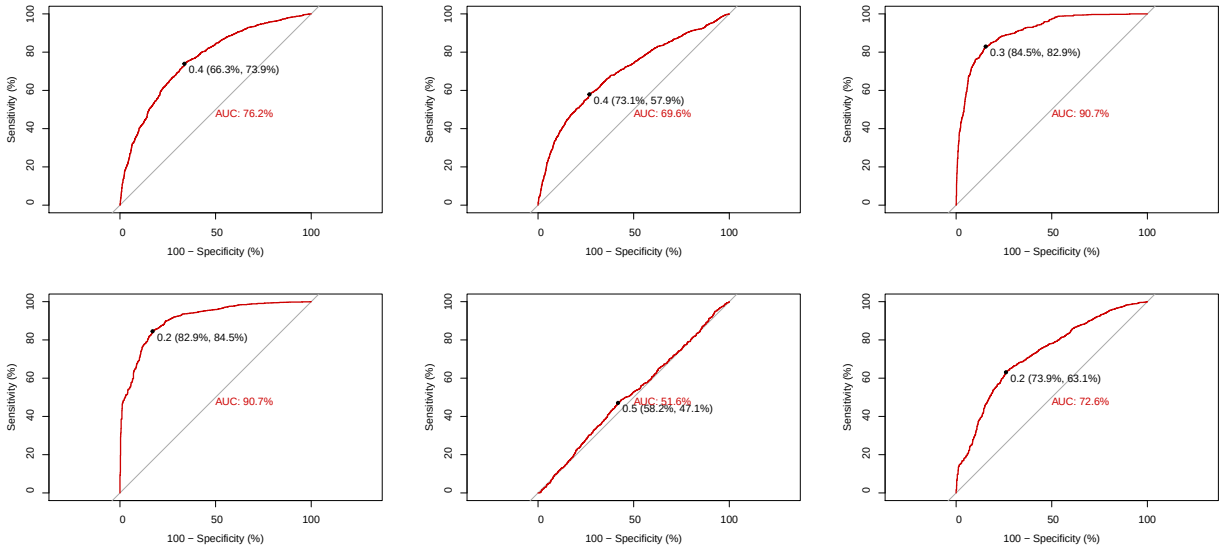

Figure 7: Parewise ROC

Appendix V (Code)

```
#
# Author : Maobin Guo
# File :  EDA

library(ggplot2)
library(MASS)
library(car)
#library(corr)
library(sjPlot)

rm(list = ls())

wine <- read.csv("../Data/wine.csv")

table(wine$quality)


wine$class <- as.factor(wine$class)


str(wine)
summary(wine)


# Check response

hist(wine$quality)

# Create cented pridctor variable
for (cn in colnames(wine[,1:11])){
    print(paste(cn, "_c", sep = ""))
    wine[, paste(cn, "_c", sep = "")] = wine[, cn] - mean(wine[, cn])
}




ggplot(wine,aes(y=quality, fill=type)) +
                geom_boxplot() +
                theme_classic() +
                theme(legend.position="none") +
                facet_wrap( ~ type,ncol=4)

par(mfrow=c(1, 2))
hist(wine[wine$type == "white", "quality"])
hist(wine[wine$type == "red", "quality"])
par(mfrow=c(1,1))

## split quality to quantile

table(wine$quality)
```

```
# premium     :  7,8,9con
# Good        :  5,6
# Acceptable :  3,4

# Bing quality to class
wine[wine$quality >=7, "class"] = "premium"
wine[wine$quality >= 5 & wine$quality < 7, "class"] = "good"
wine[wine$quality < 5, "class"] = "acceptable"
table(wine$class)

wine$class <- ordered(wine$class, levels=c("acceptable","good","premium"))


# Bin density to class
wine[wine$density <=1, "density_bin"] = "light"
wine[wine$density >1 , "density_bin"] = "weight"
wine$density_bin <- ordered(wine$density_bin, levels=c("light","weight"))


############  Corrlation

num_predictors = data[, c(1:11)]

dcor <- cor(num_predictors)

corrplot(dcor, method="square")


############  Interaction

# Interaction:  Type VS total.sulfur.dioxide
ggplot(wine,aes(x=class, y= total.sulfur.dioxide, fill=class  )) +
        geom_boxplot(alpha=0.2) +  theme(legend.position="none") + facet_wrap( ~ type)



#Interaction
ggplot(wine,aes(x=class, y= volatile.acidity, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)
#Interaction
ggplot(wine,aes(x=class, y= citric.acid, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)

#Interaction
ggplot(wine,aes(x=class, y= residual.sugar, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine,aes(x=class, y= chlorides, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)

#Interaction ?
ggplot(wine,aes(x=class, y= free.sulfur.dioxide, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)
```

```
#Interaction
ggplot(wine,aes(x=class, y= total.sulfur.dioxide, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine,aes(x=class, y= density, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)

#Interaction
ggplot(wine,aes(x=class, y= pH, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)

#Interaction
ggplot(wine,aes(x=class, y= sulphates, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine,aes(x=class, y= alcohol, fill=class)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)

#=======================================

ggplot(wine, aes(y= quality, x=alcohol)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y= quality, x=sulphates, color=sulphates)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

#
ggplot(wine, aes(y=quality, x= fixed.acidity, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine, aes(y=quality, x= volatile.acidity, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= citric.acid, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= residual.sugar, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= chlorides, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= free.sulfur.dioxide, color=quality)) +
```

```
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= total.sulfur.dioxide, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)

ggplot(wine, aes(y=quality, x= density, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine, aes(y=quality, x= pH, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine, aes(y=quality, x= alcohol, color=quality)) +
        geom_smooth(method = "lm", col = "red3") +
        geom_point(alpha = .5, colour = "blue4") +  theme(legend.position="none") + facet_wrap( ~ type)


ggplot(wine,aes(x=density_bin, y= alcohol, fill=density_bin)) +
        geom_boxplot(alpha=0.2)  +  theme(legend.position="none") + facet_wrap( ~ type)


# Model

m1 <- polr(class ~  fixed.acidity + volatile.acidity +
                    citric.acid + residual.sugar +
                    chlorides + free.sulfur.dioxide +
                    total.sulfur.dioxide + density +
                    pH + sulphates + alcohol,
                    data=wine)

m2 <- polr(class ~  fixed.acidity + volatile.acidity +
             citric.acid + residual.sugar +
             chlorides + free.sulfur.dioxide +
             total.sulfur.dioxide + density +
             pH + sulphates + alcohol + type,
             data=wine)

anova(m1, m2, test = "Chisq")
summary(m2)
confint(m2)


m3 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                    fixed.acidity ) + residual.sugar + chlorides +
                    free.sulfur.dioxide +  total.sulfur.dioxide + density +
                    alcohol, data=wine)
anova(m2, m3, test = "Chisq")
```

```
m31 <- polr(class ~  type * ( pH + sulphates + citric.acid +
                          volatile.acidity  +  fixed.acidity + residual.sugar) +
                          chlorides + free.sulfur.dioxide +
                          total.sulfur.dioxide + density_bin + alcohol,
                     data=wine)
anova(m3, m31, test = "Chisq")
summary(m31)



m311 <- polr(class ~  type * ( pH + sulphates + citric.acid +
                          volatile.acidity  +  fixed.acidity + residual.sugar) +
                          chlorides + free.sulfur.dioxide +
                          total.sulfur.dioxide + density_bin + alcohol
                          ,
                     data=wine)
summary(m31)
anova(m31, m311, test = "Chisq")


#=========================================================================


m32 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                    fixed.acidity ) + residual.sugar + log(chlorides) +
                    free.sulfur.dioxide +  total.sulfur.dioxide +
                    density_bin + alcohol, data=wine)
anova(m31, m32, test = "Chisq")
summary(m32)


m33 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                    fixed.acidity ) + residual.sugar + log(chlorides) +
                    log(free.sulfur.dioxide) +  total.sulfur.dioxide +
                    density_bin + alcohol, data=wine)
anova(m32, m33, test = "Chisq")
summary(m33)


m34 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                    fixed.acidity ) + log(residual.sugar) + log(chlorides) +
                    log(free.sulfur.dioxide) +  total.sulfur.dioxide +
                    density_bin + alcohol, data=wine)
anova(m33, m34, test = "Chisq")
summary(m34)


m35 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                    fixed.acidity ) + log(residual.sugar) + log(chlorides) +
                    log(free.sulfur.dioxide) +  log(total.sulfur.dioxide) +
                    density_bin + alcohol, data=wine)
anova(m34, m35, test = "Chisq")
summary(m35)
```

```
m36 <- polr(class ~  type * ( pH + log(sulphates) + citric.acid + volatile.acidity +
                    fixed.acidity ) + log(residual.sugar) + log(chlorides) +
                    log(free.sulfur.dioxide) +  log(total.sulfur.dioxide) +
                    density_bin + alcohol, data=wine)
anova(m35, m36, test = "Chisq")
summary(m36)



m5 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                            fixed.acidity + residual.sugar + chlorides) +
                            free.sulfur.dioxide + total.sulfur.dioxide +
                            density + alcohol, data=wine)
anova(m3, m5, test = "Chisq")


m10 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity) +
                   fixed.acidity + residual.sugar + chlorides +
                   free.sulfur.dioxide +  total.sulfur.dioxide + density +
                   alcohol, data=wine)
anova(m3, m10, test = "Chisq")



#=====================================================


m6 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity
                            + fixed.acidity+ residual.sugar) +
                    + chlorides +
                   free.sulfur.dioxide +  total.sulfur.dioxide + density +
                   alcohol, data=wine)
anova(m5, m6, test = "Chisq")

m7 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity
                            + fixed.acidity + chlorides) +
                   + residual.sugar  +
                   free.sulfur.dioxide +  total.sulfur.dioxide + density +
                   alcohol, data=wine)
anova(m3, m7, test = "Chisq")

m8 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity
                            + fixed.acidity + chlorides) +
                   + residual.sugar  +
                   free.sulfur.dioxide +  total.sulfur.dioxide + density +
                   alcohol, data=wine)
anova(m3, m8, test = "Chisq")



m4 <- polr(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                   fixed.acidity + residual.sugar + chlorides +
```

```
                  free.sulfur.dioxide +  total.sulfur.dioxide + density +
                  alcohol),  data=wine)

anova(m5, m4, test = "Chisq")



##############################
# Auto matic

m0 <- polr(class ~ 1,  data=wine)

Model_stepwise_aic <- step(m0,
                           scope = m4,
                           direction = "both",
                           trace = 0)
summary(Model_stepwise_aic)

##############################
final_model <- m31


##################
# Append P-value
(ctable <- coef(summary(final_model)))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
ctable



###### Diagnostics
#looking at the residuals may not be that meaningful here because we only have two binary variables

pred <- fitted(final_model)

wine$pred <- predict(final_model)

head(pred[sample(nrow(pred)),])


rawresid1 <- (as.numeric(wine$class) == 1) -  pred[,1]
rawresid2 <- (as.numeric(wine$class) <= 2) -  rowSums(pred[,1:2])
rawresid3 <- (as.numeric(wine$class) <= 3) -  rowSums(pred[,1:3])

tapply(rawresid1, wine$pH, mean)
tapply(rawresid2, wine$pH, mean)
tapply(rawresid3, wine$pH, mean)



###### Diagnostics
```

```
library(nnet)
mlm <- multinom(class ~  type * ( pH + sulphates + citric.acid + volatile.acidity +
                                  fixed.acidity ) + residual.sugar + chlorides +
                        free.sulfur.dioxide +  total.sulfur.dioxide + density +
                        alcohol, data=wine)


M1 <- logLik(m3)
M2 <- logLik(mlm)
(G <- -2*(M1[1] - M2[1]))
pchisq(G,3,lower.tail = FALSE)


###### Predictions



#predicted probabilities for cases in the model
predprobs <- fitted(final_model)
#look at first five rows just to see what results
predprobs[1:5,]



###### Diagnostics
####diagnostics comparing average raw residuals across bins based on predictor values
#for viewcat = 1:  create a raw residual using only the first column of the predicted probabilities
rawresid1 <- (wine$class == "acceptable") -  predprobs[,1]

#for viewcat = 2:  create a raw residual using only the second column of the predicted probabilities
rawresid2 <- (wine$class == "good") -  predprobs[,2]

rawresid3 <- (wine$class == "premium") -  predprobs[,3]



##can do binned plots for continuous variables
#make a 2 by 2 graphical display
par(mfcol = c(3,1))
binnedplot(wine$fixed.acidity, rawresid1, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned pl
binnedplot(wine$fixed.acidity, rawresid2, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned pl
binnedplot(wine$fixed.acidity, rawresid3, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned pl


par(mfcol = c(3,1))
binnedplot(log(wine$pH), rawresid1, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned plot: vi
binnedplot(log(wine$pH), rawresid2, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned plot: vi
binnedplot(log(wine$pH), rawresid3, xlab = "Prenumbers", ylab = "Raw residuals", main = "Binned plot: vi

#repeat for the other continuous predictors.....  all looks okay!

## Accuracy
```

```
pred_classes <- predict(final_model)
Conf_mat <- confusionMatrix(as.factor(pred_classes),as.factor(sesame$viewcat))
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[,c("Sensitivity","Specificity")]


## Individual ROC curves for the different levels
#here we basically treat each level as a standalone level
par(mfcol = c(2,2))
roc((wine$class=="acceptable"),predprobs[,1],plot=T,print.thres="best",legacy.axes=T,print.auc =T,
    col="red3",percent=T,main="Group 1")
roc((wine$class=="good"),predprobs[,2],plot=T,print.thres="best",legacy.axes=T,print.auc =T,
    col="gray3",percent=T,main="Group 2")
roc((wine$class=="premium"),predprobs[,2],plot=T,print.thres="best",legacy.axes=T,print.auc =T,
    col="green3",percent=T,main="Group 3")

#we can also combine them into a single plot
#


## Multi-class ROC curve (average of all pairwise comparisons)
par(mfcol = c(3,4))
multiclass.roc(wine$class,predprobs,plot=T,print.thres="best",legacy.axes=T,print.auc =T,col="red3",per
#multiclass ROCs can be hard to interpret, so don't get too hung up on them



## Interpretating the results
#Using the coefficients of `viewenc` in final model,
#For a child who is encouraged to watch Sesame Street,
#the odds of watching Sesame Street once or twice a week versus (level 2)
#watching it rarely (level 1) are 18.5 times higher (95% CI: 6.3 to 55.0)
#than the corresponding odds for a child not encouraged to watch Sesame Street.

#For a child who is encouraged to watch Sesame Street,
#the odds of watching Sesame Street three to five times a week versus
#watching it rarely are 12.8 times higher (95% CI: 4.3 to 38.0)
#than the corresponding odds for a child not encouraged to watch Sesame Street.

#For a child who is encouraged to watch Sesame Street,
#the odds of watching Sesame Street more than five times a week versus
#watching it rarely are 10.4 times higher (95% CI: 3.4 to 31.4)
#than the corresponding odds for a child not encouraged to watch Sesame Street.



################################################################################
################################################################################
### Play around with the data some more and see if you can do better!!! ###
################################################################################
################################################################################
```

```
#1. Is all continues variable need centered ?


#3. if log is good for AIC but not for accuracy/recall , is this transformation still meaningful ?

#4. transformation: center & log => together ?

#5. how to check outlier ?

#6. The density is better now, so is categorize is a way to
#   treat skew distribution.
```