The Impact of Coffee Features on Quality Classification

Group_15

1 Data Introduction

This data come from the Coffee Quality Database (CQD). The database contains information from the CoffeevQuality Institute which is a non-profit organisation working internationally to improve the quality of coffee and the lives of the people who produce it. Each of the 5 datasets contain information on features of coffee and its production including an overall score of quality.

Table 1: Brief Explanation of the Data.

Variable	Explanation
country_of_origin	Country where the coffee bean originates from.
aroma	Aroma grade (ranging from 0-10)
flavor	Flavour grade (ranging from 0-10)
acidity	Acidity grade (ranging from 0-10)
category_two_defects	Count of category 2 type defects in the batch of coffee beans tested.
altitude_mean_meters	Mean altitude of the growers farm (in metres)
harvested	Year the batch was harvested
Qualityclass	Quality score for the batch (Good - >= 82.5, Poor - <82.5).

2 Data Cleaning

First clear the null lines in the data, and then delete all lines with the line name *Taiwan* since *Taiwan* is not a country.

```
#coffee.data <- read.csv("dataset15.csv")
coffee.data <- read.csv("C:/Users/hello/Downloads/dataset15.csv")
coffee <- na.omit(coffee.data)
coffee<- subset(coffee, country_of_origin != 'Taiwan')
```

Using Qualityclass as a classification variable, assign Qualityclass 0 and 1 in terms of it's original value poor and good.

```
coffee$Qualityclass <- as.integer(coffee$Qualityclass == "Good")
```

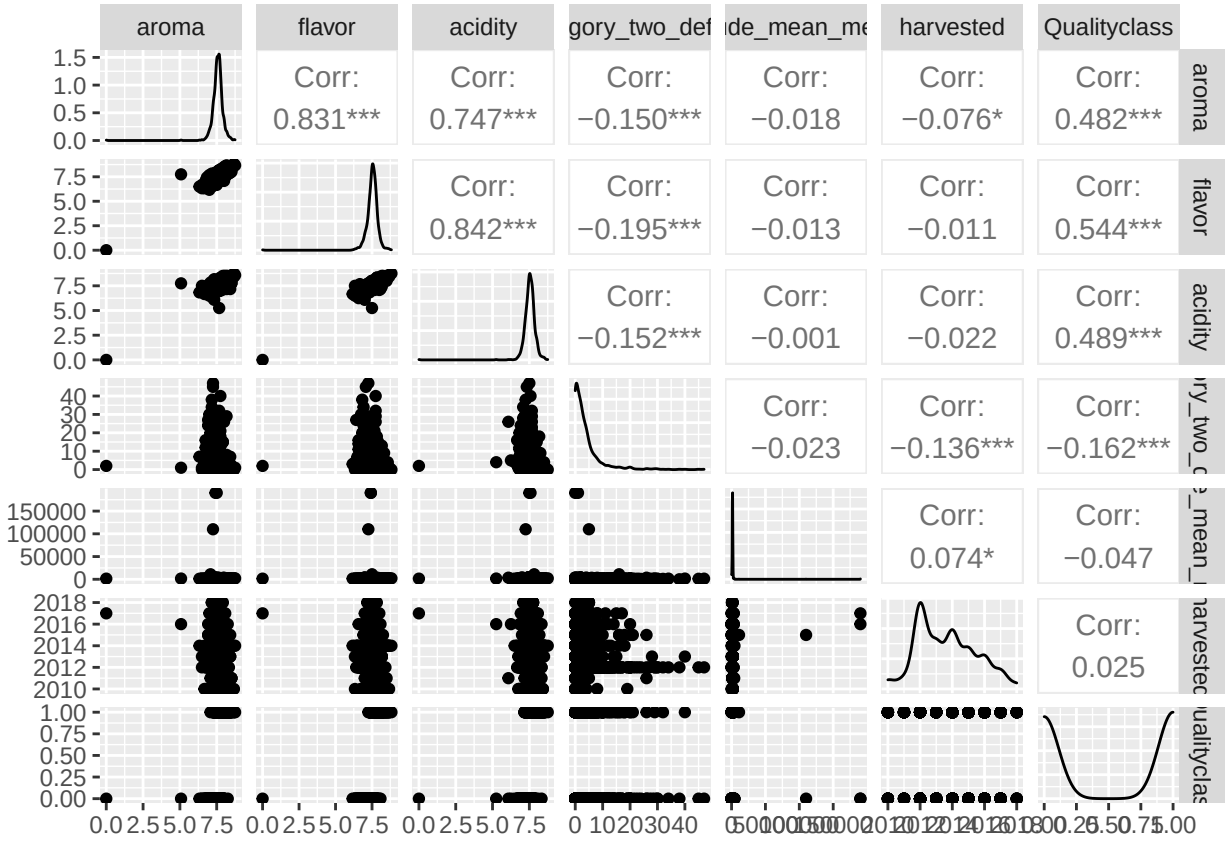
A subset with reasonable data distribution is selected from *coffee.data's* data set and stored in a new data set, *coffee*.

```
quantiles_aroma <- quantile(coffee.data$aroma,probs=c(0.25,0.75),na.rm=FALSE)
```

2.1 Correlation Test

Correlation test for eight variables:

```
coffee_nocountry <- coffee[,-which(names(coffee) == "country_of_origin")]
ggpairs(coffee_nocountry) +
  theme(plot.background = element_rect(
    fill = "transparent",
    colour = NA,
    size = 1))
```



From the results of the correlation test, only aroma, flavor, and acidity are highly correlated with Qualityclass, while the correlation between category_two_defects, altitude_mean_meters, harvested, and Qualityclass is very low.

2.2 Cleaning outliers

The IQR method is used to identify outliers in coffee. The IQR (interquartile range) is a measure of the spread of the middle 50% of the numeric variables in coffee, calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the dataset. According to the results of the correlation test, we decided to clean up the outliers only for the three variables aroma, flavor, and acidity.

```
quantiles_aroma <- quantile(coffee$aroma,probs=c(0.25,0.75),na.rm=FALSE)
IQR_aroma <- IQR(coffee$aroma)
Lower_aroma <- quantiles_aroma[1]-1.5*IQR_aroma
Upper_aroma <- quantiles_aroma[2]+1.5*IQR_aroma
coffee<- subset(coffee,coffee$aroma>Lower_aroma & coffee$aroma< Upper_aroma)

quantiles_flavor <- quantile(coffee$flavor,probs=c(0.25,0.75),na.rm=FALSE)
IQR_flavor <- IQR(coffee$flavor)
Lower_flavor <- quantiles_flavor[1]-1.5*IQR_flavor
Upper_flavor <- quantiles_flavor[2]+1.5*IQR_flavor
coffee<- subset(coffee,coffee$flavor>Lower_flavor & coffee$flavor< Upper_flavor)

quantiles_acidity <- quantile(coffee$acidity,probs=c(0.25,0.75),na.rm=FALSE)
IQR_acidity <- IQR(coffee$acidity)
Lower_acidity <- quantiles_acidity[1]-1.5*IQR_acidity
Upper_acidity <- quantiles_acidity[2]+1.5*IQR_acidity
coffee<- subset(coffee,coffee$acidity>Lower_acidity & coffee$acidity< Upper_acidity)


# Store the cleared data
write.csv(coffee,file="coffee_clean.csv",row.names=FALSE)
```

2.3 Correlation test between variables

Correlation test of data after cleaning:

```
coffee.clean <- read.csv("coffee_clean.csv")
cor(coffee[,2:7])
```

```
##                             aroma      flavor      acidity category_two_defects
## aroma                  1.00000000  0.67160282  0.538079249          -0.19698110
## flavor                 0.67160282  1.00000000  0.705622108          -0.23495703
## acidity                0.53807925  0.70562211  1.000000000          -0.16888150
## category_two_defects -0.19698110 -0.23495703 -0.168881503           1.00000000
## altitude_mean_meters -0.03490461 -0.02777236 -0.006467399          -0.02396095
## harvested            -0.03412132  0.04100676  0.047885262          -0.13277330
##                      altitude_mean_meters   harvested
## aroma                        -0.034904613 -0.03412132
## flavor                       -0.027772358  0.04100676
## acidity                      -0.006467399  0.04788526
## category_two_defects         -0.023960951 -0.13277330
## altitude_mean_meters          1.000000000  0.07402986
## harvested                     0.074029856  1.00000000
```

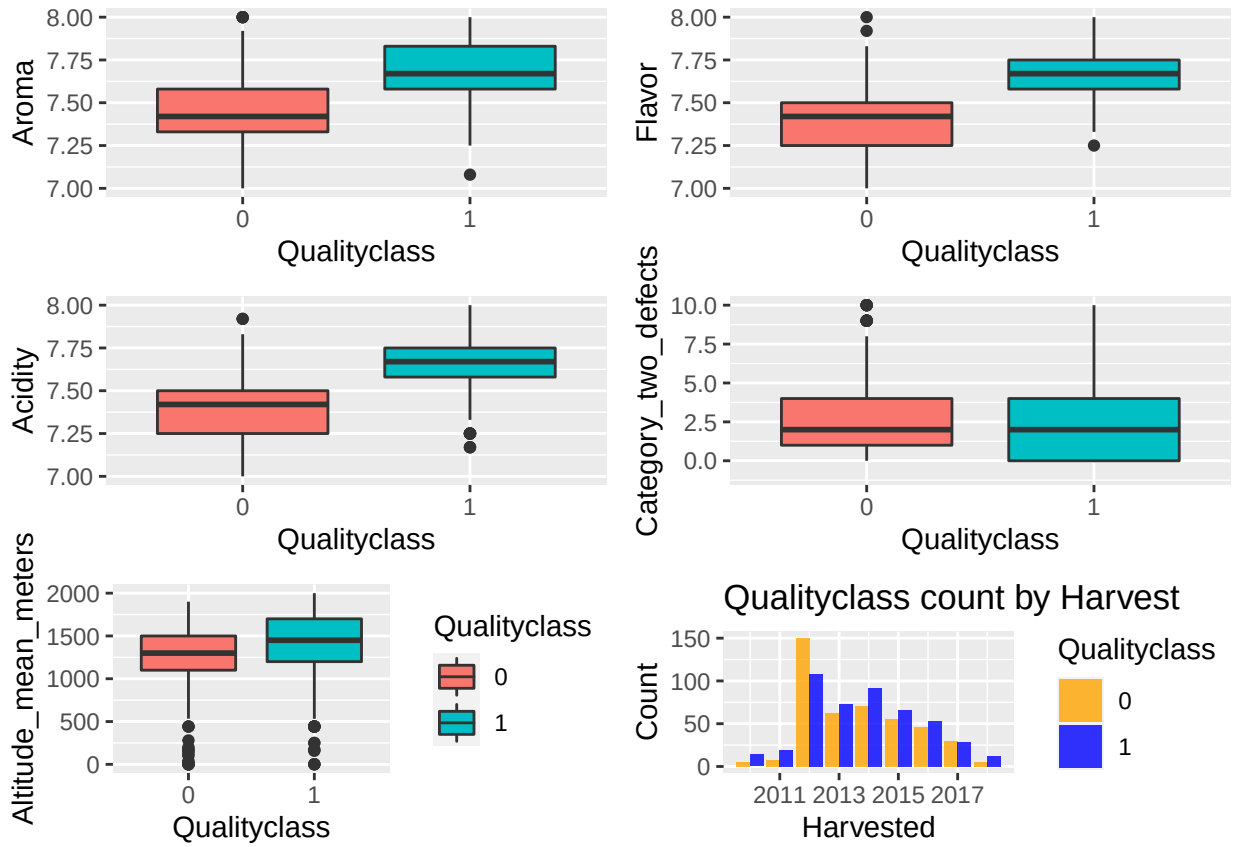
3 Data visualization

```
library(gridExtra)
coffee$Qualityclass <- factor(coffee$Qualityclass)
plot1 <- ggplot(data = coffee,aes(x = Qualityclass, y = aroma, fill = Qualityclass)) +
```

```
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Aroma") +
  ylim(7,8) +
  theme(legend.position = "none")

plot2 <- ggplot(data = coffee,aes(x = Qualityclass, y = flavor, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Flavor") +
  ylim(7,8) +
  theme(legend.position = "none")

plot3 <- ggplot(data = coffee,aes(x = Qualityclass, y = acidity, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Acidity") +
  ylim(7,8) +
  theme(legend.position = "none")

plot4 <- ggplot(data = coffee,aes(x = Qualityclass, y = category_two_defects, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Category_two_defects") +
  ylim(-1,10) +
  theme(legend.position = "none")

plot5 <- ggplot(data = coffee,aes(x = Qualityclass, y = altitude_mean_meters, fill = Qualityclass)) +
  geom_boxplot() +
  labs(x = "Qualityclass", y = "Altitude_mean_meters") +
  ylim(0,2000)
  theme(legend.position = "none")
```

```
## List of 1
##  $ legend.position: chr "none"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
## The number of different quality inspection results in each year
plot6 <- ggplot(coffee, aes(x=harvested, fill=as.factor(Qualityclass))) +
  geom_bar(position="dodge", alpha=0.8, stat="count") +
  scale_fill_manual(values=c("Orange", "Blue")) +
  labs(x="Harvested", y="Count", fill="Qualityclass") +
  ggtitle("Qualityclass count by Harvest")

## Merge six charts
## Arrange plot1 to plot6 in a grid of 2 rows and 3 columns
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, nrow=3, ncol=2)
```



4 Standardized data

After data cleaning, we standardized the data:

```
## Creates a logical vector that determines which columns in the data box coffee contain
## numeric data. It uses the function sapply to determine whether the data type of each
## column is numeric.
numeric_cols <- sapply(coffee,is.numeric)
## The columns containing only numerical data in data box coffee are extracted and
## assigned to new data box numeric_coffee
numeric_coffee <- coffee[,numeric_cols]
## A new data box scaled_coffee is obtained by standardizing numerical data in numeric_coffee.
## The standardization process is to transform the data range of different variables into
## the same scale for easy comparison.
scaled_coffee <- as.data.frame(scale(numeric_coffee))
## Add the category_two_defects column from the original data box coffee to scaled_coffee.
scaled_coffee$category_two_defects <- coffee$category_two_defects
## Assign the numerical data in the standardized data frame scaled_coffee back to the
## corresponding column in the original data frame coffee, and overwrite the original data.
coffee[,numeric_cols] <- scaled_coffee
```

5 Modeling

5.1 Fit 15 generalized linear models

With Qualityclass as the response variable and aroma, flavor, and acidity as independent variables, 15 different generalized linear models were fitted, and Deviance, and AIC of each model were calculated.

```
model1 <- glm(Qualityclass~aroma+flavor+acidity,family=binomial(link="logit"), data=coffee)
model2 <- glm(Qualityclass~aroma,family=binomial(link="logit"),
              data=coffee)
model3 <- glm(Qualityclass~flavor,family=binomial(link="logit"),
              data=coffee)
model4 <- glm(Qualityclass~acidity,family=binomial(link="logit"),
              data=coffee)
model5 <- glm(Qualityclass~aroma+flavor,family=binomial(link="logit"),
              data=coffee)
model6 <- glm(Qualityclass~flavor+acidity,family=binomial(link="logit"),
              data=coffee)
model7 <- glm(Qualityclass~aroma+acidity,family=binomial(link="logit"),
              data=coffee)
model8 <- glm(Qualityclass~aroma*flavor+acidity,family=binomial(link="logit"),
              data=coffee)
model9 <- glm(Qualityclass~aroma+flavor*acidity,family=binomial(link="logit"),
              data=coffee)
model10 <- glm(Qualityclass~aroma*acidity+flavor,family=binomial(link="logit"),
              data=coffee)
model11 <- glm(Qualityclass~aroma*flavor+aroma*acidity,family=binomial(link="logit"),
              data=coffee)
model12 <- glm(Qualityclass~aroma*flavor+acidity*flavor,family=binomial(link="logit"),
              data=coffee)
model13 <- glm(Qualityclass~aroma*acidity+flavor*acidity,family=binomial(link="logit"),
              data=coffee)
model14 <- glm(Qualityclass~aroma*flavor+acidity*flavor+acidity*aroma,family=binomial(link="logit")
               ,data=coffee)
model15 <- glm(Qualityclass~aroma*flavor*acidity,family=binomial(link="logit"),
              data=coffee)
anova(model1,model2,model3,model4,model5,
      model6,model7,model8,model9,model10,
      model11,model12,model13,model14,model15)
```

```
## Analysis of Deviance Table
##
## Model  1: Qualityclass ~ aroma + flavor + acidity
## Model  2: Qualityclass ~ aroma
## Model  3: Qualityclass ~ flavor
## Model  4: Qualityclass ~ acidity
## Model  5: Qualityclass ~ aroma + flavor
## Model  6: Qualityclass ~ flavor + acidity
## Model  7: Qualityclass ~ aroma + acidity
## Model  8: Qualityclass ~ aroma * flavor + acidity
## Model  9: Qualityclass ~ aroma + flavor * acidity
## Model 10: Qualityclass ~ aroma * acidity + flavor
## Model 11: Qualityclass ~ aroma * flavor + aroma * acidity
```

```
## Model 12: Qualityclass ~ aroma * flavor + acidity * flavor
## Model 13: Qualityclass ~ aroma * acidity + flavor * acidity
## Model 14: Qualityclass ~ aroma * flavor + acidity * flavor + acidity *
##     aroma
## Model 15: Qualityclass ~ aroma * flavor * acidity
##    Resid. Df Resid. Dev Df Deviance
## 1        895     560.73
## 2        897     850.57 -2 -289.848
## 3        897     689.10  0  161.470
## 4        897     833.77  0 -144.661
## 5        896     612.86  1  220.901
## 6        896     623.63  0  -10.769
## 7        896     668.01  0  -44.377
## 8        894     556.92  2  111.091
## 9        894     559.31  0   -2.387
## 10       894     559.35  0   -0.039
## 11       893     556.16  1    3.189
## 12       893     555.58  0    0.577
## 13       893     558.13  0   -2.554
## 14       892     555.07  1    3.063
## 15       891     554.98  1    0.089
```

```
summary(model1)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity, family = binomial(link = "logit"),
##     data = coffee)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8724  -0.3618   0.0297   0.4818   2.2729
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.006953   0.106623   0.065    0.948
## aroma       1.207428   0.166860   7.236 4.62e-13 ***
## flavor      1.967400   0.221293   8.890  < 2e-16 ***
## acidity     1.152111   0.170101   6.773 1.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1245.07  on 898  degrees of freedom
## Residual deviance:  560.73  on 895  degrees of freedom
## AIC: 568.73
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model8)
```

```
##
```

```
## Call:
## glm(formula = Qualityclass ~ aroma * flavor + acidity, family = binomial(link = "logit"),
##     data = coffee)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7017  -0.3253   0.0585   0.4796   2.2915
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.01235    0.10801   0.114    0.909
## aroma         1.27183    0.17133   7.423 1.14e-13 ***
## flavor        2.04471    0.22836   8.954  < 2e-16 ***
## acidity       1.13575    0.16896   6.722 1.79e-11 ***
## aroma:flavor -0.55495    0.27687  -2.004    0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1245.07  on 898  degrees of freedom
## Residual deviance:  556.92  on 894  degrees of freedom
## AIC: 566.92
##
## Number of Fisher Scoring iterations: 7
```

```
AIC(model1)
```

```
## [1] 568.7271
```

```
AIC(model2)
```

```
## [1] 854.5747
```

```
AIC(model3)
```

```
## [1] 693.1046
```

```
AIC(model4)
```

```
## [1] 837.7657
```

```
AIC(model5)
```

```
## [1] 618.8648
```

```
AIC(model6)
```

```
## [1] 629.634
```

```
AIC(model7)
```

```
## [1] 674.0113
```

```
AIC(model8)
```

```
## [1] 566.9203
```

```
AIC(model9)
```

```
## [1] 569.3077
```

```
AIC(model10)
```

```
## [1] 569.3465
```

```
AIC(model11)
```

```
## [1] 568.1577
```

```
AIC(model12)
```

```
## [1] 567.5803
```

```
AIC(model13)
```

```
## [1] 570.1348
```

```
AIC(model14)
```

```
## [1] 569.0716
```

```
AIC(model15)
```

```
## [1] 570.9826
```

```
Modelname <- c("aroma+flavor+acidity","aroma","flavor","acidity",
               "aroma+flavor","flavor+acidity","aroma+acidity","aroma*flavor+acidity",
               "aroma+flavor*acidity","aroma*acidity+flavor",
               "aroma*flavor+aroma*acidity","aroma*flavor+acidity*flavor",
               "aroma*acidity+flavor*acidity","aroma*flavor+acidity*flavor+acidity*aroma",
               "aroma*flavor*acidity")
AIC <- c(model1$aic,model2$aic,model3$aic,model4$aic,
         model5$aic,model6$aic,model7$aic,model8$aic,
         model9$aic,model10$aic,model11$aic,model12$aic,
         model13$aic,model14$aic,model15$aic)
Deviance <- c(model1$deviance,model2$deviance,model3$deviance,
```

```
              model4$deviance,model5$deviance,model6$deviance,
              model7$deviance,model8$deviance,model9$deviance,
              model10$deviance,model11$deviance,model12$deviance,
              model13$deviance,model14$deviance,model15$deviance)
table_values <- data.frame(Modelname, Deviance, AIC)
table_values <- as.data.frame(table_values)
table_values %>%
  transmute(Modelname=Modelname, Deviance=Deviance, AIC=AIC) %>%
  kable(caption = '\\label{tab:summary} Deviance and AIC of each model',
booktabs = TRUE, linesep = "", digits = 2) %>%
  kable_styling(font_size = 16, latex_options = "hold_position")
```

Table 2: Deviance and AIC of each model

Modelname	Deviance	AIC
aroma+flavor+acidity	560.73	568.73
aroma	850.57	854.57
flavor	689.10	693.10
acidity	833.77	837.77
aroma+flavor	612.86	618.86
flavor+acidity	623.63	629.63
aroma+acidity	668.01	674.01
aroma*flavor+acidity	556.92	566.92
aroma+flavor*acidity	559.31	569.31
aroma*acidity+flavor	559.35	569.35
aroma*flavor+aroma*acidity	556.16	568.16
aroma*flavor+acidity*flavor	555.58	567.58
aroma*acidity+flavor*acidity	558.13	570.13
aroma*flavor+acidity*flavor+acidity*aroma	555.07	569.07
aroma*flavor*acidity	554.98	570.98

5.2 Multicollinearity test of model

VIF (Variance inflation factor) is a statistical index used to detect whether there is collinearity between independent variables. The higher the VIF, the stronger the collinearity between the independent variables, which may lead to instability and inaccuracy of the model.

```
vif(model1)
```

```
##    aroma   flavor  acidity
## 1.008356 1.029070 1.026516
```

```
vif(model5)
```

```
##    aroma   flavor
## 1.001863 1.001863
```

```
vif(model6)
```

```
##   flavor  acidity
## 1.015289 1.015289
```

```
vif(model7)
```

```
##    aroma  acidity
## 1.007338 1.007338
```

```
vif(model8)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##        aroma       flavor      acidity aroma:flavor
##     1.080629     1.123385     1.025315     1.179135
```

```
vif(model9)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##          aroma         flavor        acidity flavor:acidity
##       1.008685       1.046731       1.112711       1.098556
```

```
vif(model10)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##         aroma       acidity        flavor aroma:acidity
##      1.010254      1.047653      1.032025      1.022463
```

```
vif(model11)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##         aroma        flavor       acidity  aroma:flavor aroma:acidity
##      1.071227      1.135566      1.069947      1.220518      1.081700
```

```
vif(model12)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##          aroma         flavor        acidity   aroma:flavor flavor:acidity
##       1.110921       1.142509       1.142270       1.216605       1.127068
```

```
vif(model13)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##          aroma        acidity         flavor  aroma:acidity acidity:flavor
##       1.020543       1.115246       1.056513       1.040014       1.108063
```

```
vif(model14)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##          aroma         flavor        acidity   aroma:flavor flavor:acidity
##       1.104590       1.155531       1.153137       1.301887       1.165130
##  aroma:acidity
##       1.153612
```

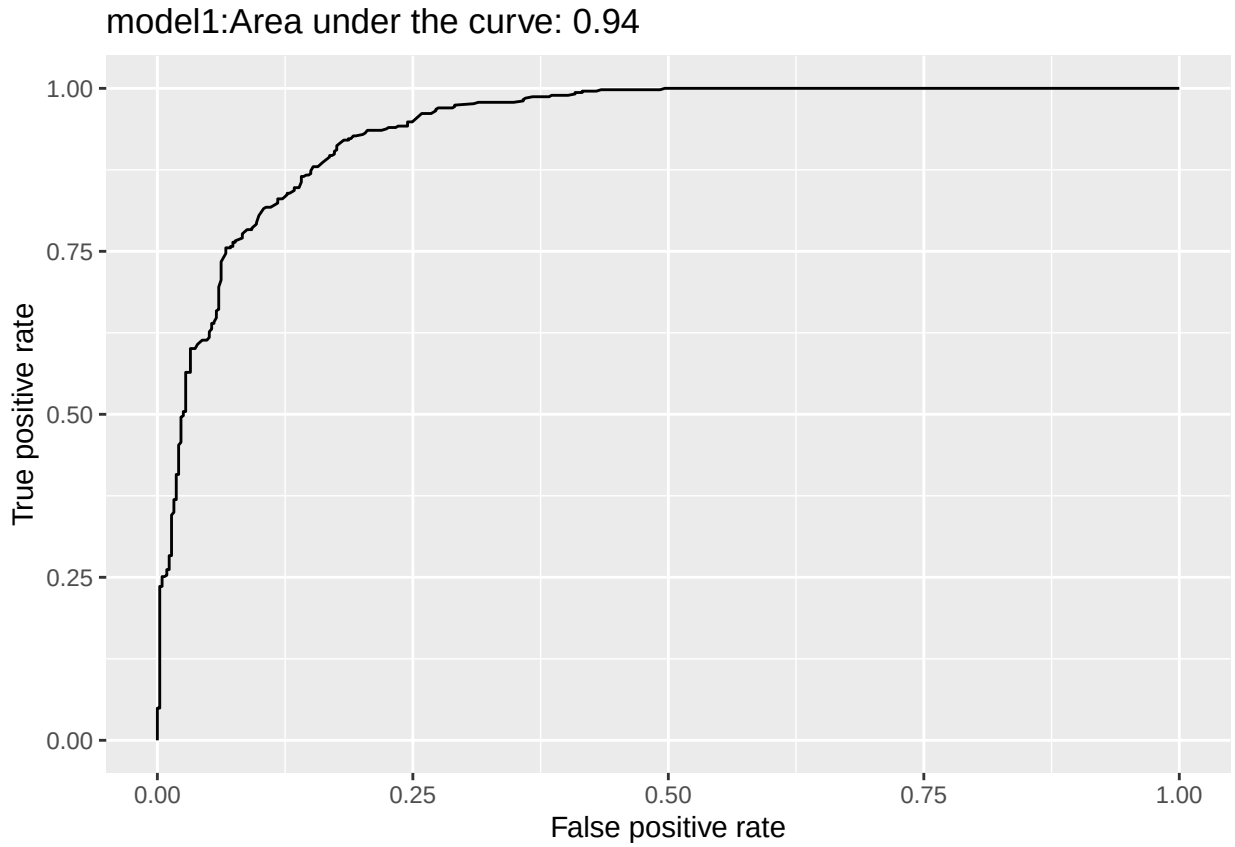
```
vif(model15)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

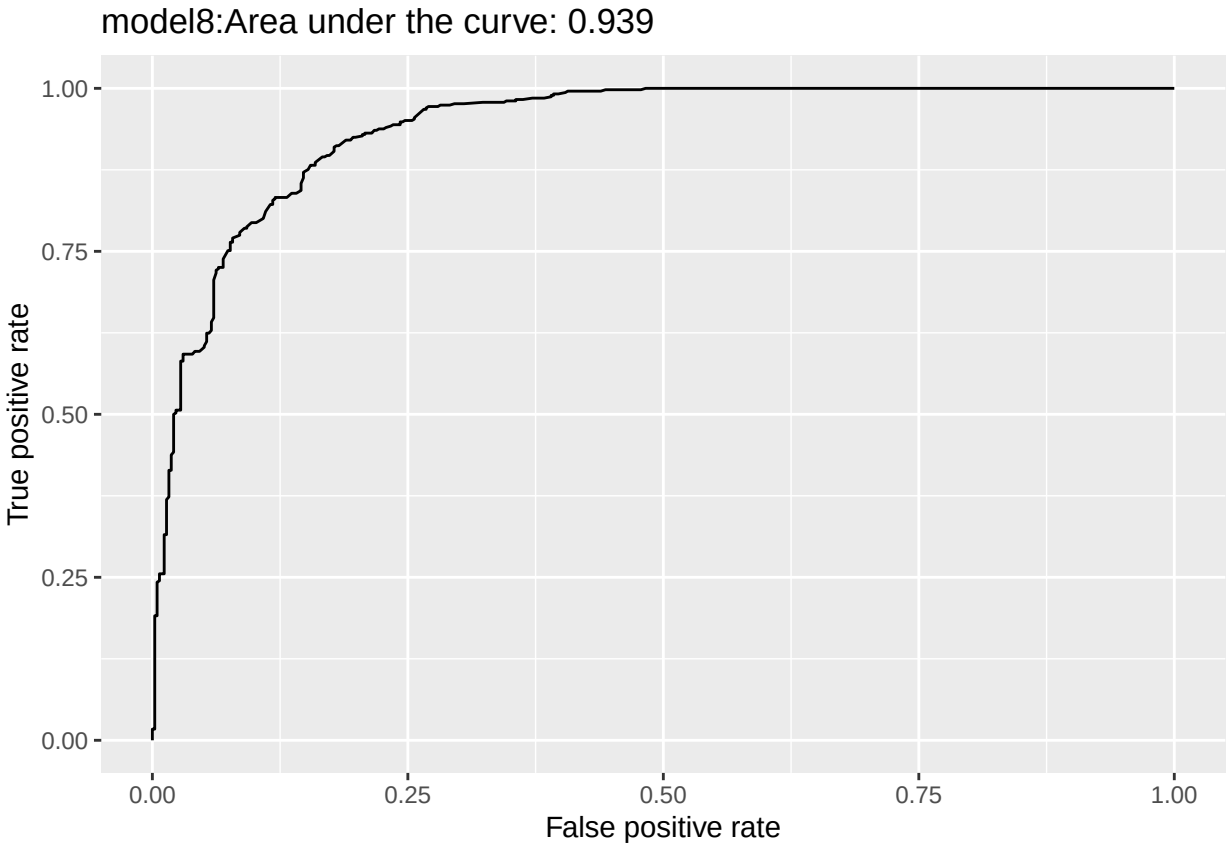
```
##                aroma              flavor             acidity
##             1.121892            1.152746            1.163232
##         aroma:flavor        aroma:acidity       flavor:acidity
##             1.261895            1.364864            1.484557
## aroma:flavor:acidity
##             1.696149
```

Comprehensive comparison of Deviance, AIC and VIF of each model, we think model1 and model8 are the two models with the best performance. The two Receiver Operating Characteristic (ROC) curves below show the performance of model1 and model8. The areas under the two ROC curves (AUC) are 0.94 and 0.939, which indicate a good classification performance.

```
coffee.pr <- predict(model1, type="response")
score <- prediction(coffee.pr,coffee$Qualityclass)
perf <- performance(score,"tpr","fpr")
auc <- performance(score,"auc")
perfd <- data.frame(x= perf@x.values[1][[1]], y=perf@y.values[1][[1]])
p1 <- ggplot(perfd, aes(x= x, y=y)) + geom_line() +
  xlab("False positive rate") + ylab("True positive rate") +
  ggtitle(paste("model1:Area under the curve:", round(auc@y.values[[1]], 3)))
p1
```




```
coffee.pr <- predict(model8, type="response")
score <- prediction(coffee.pr,coffee$Qualityclass)
perf <- performance(score,"tpr","fpr")
auc <- performance(score,"auc")
perfd <- data.frame(x= perf@x.values[1][[1]], y=perf@y.values[1][[1]])
p1 <- ggplot(perfd, aes(x= x, y=y)) + geom_line() +
  xlab("False positive rate") + ylab("True positive rate") +
  ggtitle(paste("model8:Area under the curve:", round(auc@y.values[[1]], 3)))
p1
```



6 Conclusion

According to the two models we finally select,

$$model1 : \log\left(\frac{p}{1-p}\right) = 0.006953 + 1.207428 \cdot \text{aroma} + 1.967400 \cdot \text{flavor} + 1.152111 \cdot \text{acidity}$$

$$model8 : \log\left(\frac{p}{1-p}\right) = 0.01235 + 1.27183 \cdot \text{aroma} + 2.04471 \cdot \text{flavor} + 1.13574 \cdot \text{acidity} - 0.55495 \cdot \text{aroma} \cdot \text{flavor}$$

we believe that the quality of coffee is positively correlated with aroma, acidity and flavor, while the three features of category_two_defects, altitude_mean_meters and harvested have no significant impact on the quality of coffee.