Project2

Hiba Hajali

20/03/2022

1 Introduction

Coffee is one of the most popular beverages worldwide and has a vast market. Research on coffee quality can help coffee farmers understand the quality of the coffee they grow to make more accurate market planning. The researchers obtained data containing features of coffee and its production from the Coffee Quality Institute, a coffee research institute. They used this data to analyze the impact of these coffee features (such as acidity) on coffee quality scores.

In the following sections, the researchers will use the Generalized Linear Model to model the Qualityclass variables, obtain the optimal model by comparison, and analyze each variable to determine its impact on coffee quality.

2 Explanatory Analysis

The numbers of the missing values in each column. From the output we find that NA values is concentrate on variable of 'harvested' and 'altitude mean meters'.

```
##    country_of_origin                aroma               flavor
##                    0                    0                    0
##              acidity category_two_defects altitude_mean_meters
##                    0                    0                  162
##            harvested         Qualityclass
##                   55                    0
```

The data after we remove the missing values, 858 observations were obtained. And data set includes 8 variables in total. To answer the research question, the Quality class serves as response variables, the country_of_origin, aroma, flavor, acidity, category_two_defects, altitute_mean_meters and harvested serves as explanatory variables.

Since there are only two kinds of results, good and bad, the response variable is dichotomous and follows binomial distribution, which is suitable for binary logistic regression in GLM, not multivariate logistic regression.

```
## Rows: 858
## Columns: 8
## $ country_of_origin    <chr> "Guatemala", "China", "Colombia", "Guatemala", "C~
## $ aroma                <dbl> 7.92, 7.67, 7.75, 7.83, 7.67, 8.17, 7.83, 7.67, 7~
## $ flavor               <dbl> 7.67, 7.67, 7.50, 7.67, 7.42, 8.00, 7.50, 7.75, 7~
## $ acidity              <dbl> 7.75, 7.67, 7.50, 7.33, 7.33, 7.17, 7.42, 7.67, 7~
```

```
## $ category_two_defects <int> 3, 3, 0, 1, 5, 0, 2, 1, 4, 0, 10, 0, 4, 4, 2, 4, ~
## $ altitude_mean_meters <dbl> 1650.00, 1600.00, 1750.00, 1310.64, 1600.00, 1750~
## $ harvested            <int> 2015, 2015, 2013, 2013, 2011, 2014, 2013, 2015, 2~
## $ Qualityclass         <chr> "Good", "Good", "Good", "Poor", "Poor", "Good", "~
```

The number of unique values in country of origin:

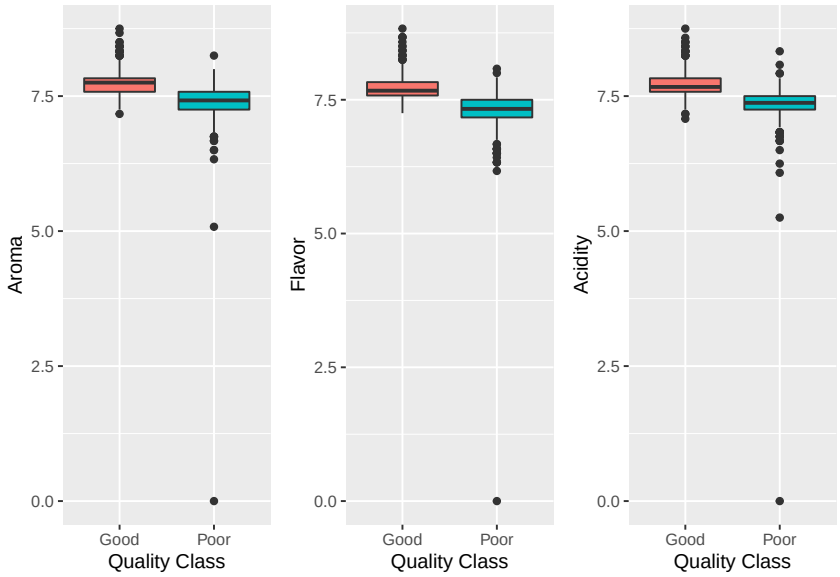
```
## [1] 34
```

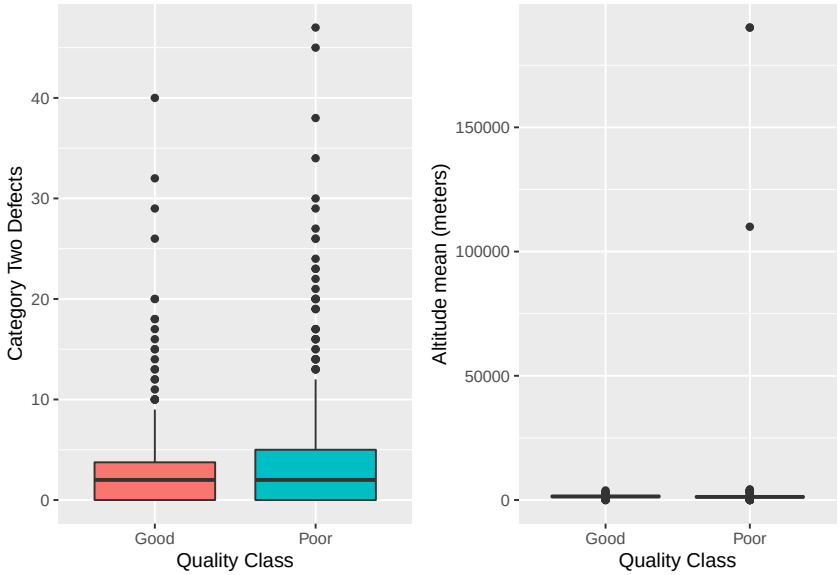
The number of unique values in harvest year:

```
## [1] 9
```

Correlation analysis diagram shows that there are strong positive correlations between flavor,aroma and acidity, so multicollinearity should be paid attention to in the subsequent modeling process. The correlation between other variables is weak.

Box plots showing the distribution of the quantitative variables.





The box plot shows that three variables, aroma,flavor and acidity, have more impact on good coffee quality than bad coffee quality. However, Category Tow Defects and Altitude mean (meters) have almost the same influence on coffee quality as bad, and there are some abnormal values.

2.1 Bar Charts:

The Percentages:

Table 1 showing the percentage of the quality classes for each country. Colombia has the most proportion of good quality coffee, while Mexico has the most bad coffee, Ethiopia only has good coffee.

Table 1: The Proportion of Quality Classs in Different Country

country_of_origin	Good	Poor
Brazil	47.3% (35)	52.7% (39)
Burundi	0.0% (0)	100.0% (1)
China	70.0% (7)	30.0% (3)
Colombia	78.8% (89)	21.2% (24)
Costa Rica	58.8% (20)	41.2% (14)
Cote d?Ivoire	0.0% (0)	100.0% (1)
Ecuador	50.0% (1)	50.0% (1)
El Salvador	78.6% (11)	21.4% (3)
Ethiopia	100.0% (19)	0.0% (0)
Guatemala	48.4% (62)	51.6% (66)
Haiti	16.7% (1)	83.3% (5)
Honduras	22.2% (8)	77.8% (28)
India	50.0% (5)	50.0% (5)
Indonesia	58.8% (10)	41.2% (7)
Kenya	93.3% (14)	6.7% (1)
Laos	33.3% (1)	66.7% (2)
Malawi	10.0% (1)	90.0% (9)
Mauritius	0.0% (0)	100.0% (1)
Mexico	25.9% (48)	74.1% (137)
Myanmar	0.0% (0)	100.0% (8)
Nicaragua	18.8% (3)	81.2% (13)
Panama	75.0% (3)	25.0% (1)
Papua New Guinea	100.0% (1)	0.0% (0)
Peru	0.0% (0)	100.0% (1)
Philippines	40.0% (2)	60.0% (3)
Rwanda	100.0% (1)	0.0% (0)
Taiwan	38.5% (20)	61.5% (32)
Tanzania, United Republic Of	41.4% (12)	58.6% (17)
Thailand	68.8% (11)	31.2% (5)
Uganda	73.1% (19)	26.9% (7)
United States	70.0% (7)	30.0% (3)
United States (Hawaii)	100.0% (2)	0.0% (0)
United States (Puerto Rico)	50.0% (1)	50.0% (1)
Vietnam	66.7% (4)	33.3% (2)

Table 2 indicates the percentage of the quality classes for each harvest year. In 2012, 2016 and 2017 there was a larger proportion of bad coffee, and in all other years there was a larger proportion of good coffee.

Table 2: The Proportion of Quality Classs in Different Harvested Year

harvested	Good	Poor
2010	64.0% (16)	36.0% (9)
2011	72.0% (18)	28.0% (7)
2012	40.0% (96)	60.0% (144)
2013	51.7% (61)	48.3% (57)
2014	50.3% (94)	49.7% (93)
2015	54.6% (59)	45.4% (49)
2016	47.8% (43)	52.2% (47)
2017	44.4% (24)	55.6% (30)
2018	63.6% (7)	36.4% (4)

3 Formal Anaylsis

3.1 Multicollinearity Checking.

Before start to built model, the multicollinearity examine between our potential variables. Here, kappa function helps to evaluate the multicollinearity. kappa value for the explanatory variables of this research is 21.31, that is lower than 100. In general, it considered that the multicollinearity is weak.

```
## [1] 21.31623
```

3.2 Models comparison and Selection:

The strategy to establish model is the top-down approach. Starting with a model contains all variables, then remove the variables one by one.

```
levels(data$Qualityclass)
```

```
## [1] "Good" "Poor"
```

```
#check the baseline and contributor for our response variables in models.
#the first returned result is our baseline and the second returned is the contributor.
```

For model_1 to model_5, the positive outcome is Poor quality coffee. It means that coffee with poor quality has been treated as positive outcomes in our model.

Table 3: The Result of Model comparison

Formula								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + altitude_mean_meters + harvested								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + harvested								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects								
Qualityclass ~ aroma + flavor + acidity + category_two_defects								
Qualityclass ~ aroma + flavor + acidity								
Rank	Df.res	AIC	AICc	BIC	McFadden	Cox.and.Snell	Nagelkerke	p.value
40	818	508.0	512.2	702.9	0.642	0.589	0.785	0
39	819	506.4	510.4	696.6	0.641	0.589	0.785	0
38	820	506.5	510.3	691.9	0.640	0.588	0.784	0
5	853	529.5	529.6	558.0	0.565	0.543	0.724	0
4	854	527.7	527.7	551.4	0.565	0.543	0.724	0

Table 3 summaries the results of model comparison and provides fit information. AIC, corrected AIC, BIC, p-value and pseudo R-square value are included, in which, pseudo R-square is calculate by the method of McFadden, Cox and Snell and Nagelkerke.
AIC, BIC and p values act as criteria to evaluate the goodness-of-fit and the complexity of models. Pseudo R-square is not suit for our model comparison, because the predictor variables we apply are different. It is meaningful only compared models on the same data and same response variables.

For **Table 3**, the difference between AIC and AICc is approximetly 30, that value is quite similar. While, the value of BIC indicates huge deviation. Thus, we prefer choosing model_5 as final model that show smaller BIC value. It suggests a better balance of fitting and complexity. The p value indicates the results are strongly significant.

The formula of final model is given by,

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Aroma} + \beta_2 \cdot \text{Flavor} + \beta_3 \cdot \text{Acidity}$$

where $p = Prob(Poor)$ and $1 - p = Prob(Good)$. According to the property of our response variable (QualityClass), which also mentioned at **Exploratory Analysis**. The outcomes of QualityClass follows binomial distribution, thus, we apply logit link function, also called binary logistic regression model. This link function describe the relationship between predictor variables and response variables by log-scaled odds ratio.

3.3 Information of Final Model.

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2730  -0.4162   0.0028   0.3572   4.0021
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 117.1787     8.5874  13.645  < 2e-16 ***
## aroma        -4.3149     0.6961  -6.199 5.70e-10 ***
## flavor       -7.3938     0.8854  -8.350  < 2e-16 ***
## acidity      -3.8050     0.7036  -5.408 6.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1188.88  on 857  degrees of freedom
## Residual deviance:  517.67  on 854  degrees of freedom
## AIC: 525.67
##
## Number of Fisher Scoring iterations: 7
```

The summary of model_5 indicates aroma, flavor and acidity highly relevant with Quality Class. It significance code is 3 stars that means the p value close to 0.

It is notice that the of coefficients are negative for all predictors. According the relationship between the *log*(odds ratio) and the probability, the coefficient means the increase score of aroma, flavor and acidity will cause drops the probability of positive events, poor quality coffee, occurs.

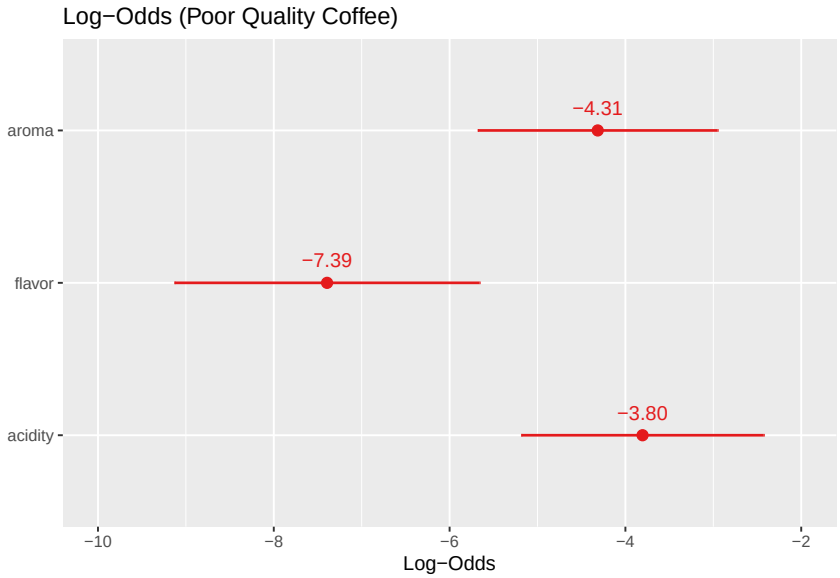


Figure 1: The Estimates of Coefficients for each predictor variables

Figure 1 could well describe the influence against each predictor. The absolute value of coefficients reflect the effect level, a higher absolute value present the predictor could change the outcome more. As for Figure 1, seems Flavor score could change the log-scaled odds more significant than others.

The confidence Interval (CI) checking **(Table 4)** indicates the estimates of parameters β base on the final model of this research. It informs the estimates log-scaled range of coefficients for each variable.
However, it seems all of 95% confidence interval is wide conbinated the result of **Figure 1** , that might suggest the variance within estimates coefficients is huge, and might effect the prediction ability of model_5. Then, we will evaluate the prediction of Model_5.

3.4 Confidence Intervals of Log-Odds:

Table 4: Confidence Intervals for log odds in Model 5

	2.5 %	97.5 %
(Intercept)	101.235732	134.953218
aroma	-5.720601	-2.988415
flavor	-9.197355	-5.721335
acidity	-5.211874	-2.449520

3.5 The Predictive Probability Plot with marginal effect:

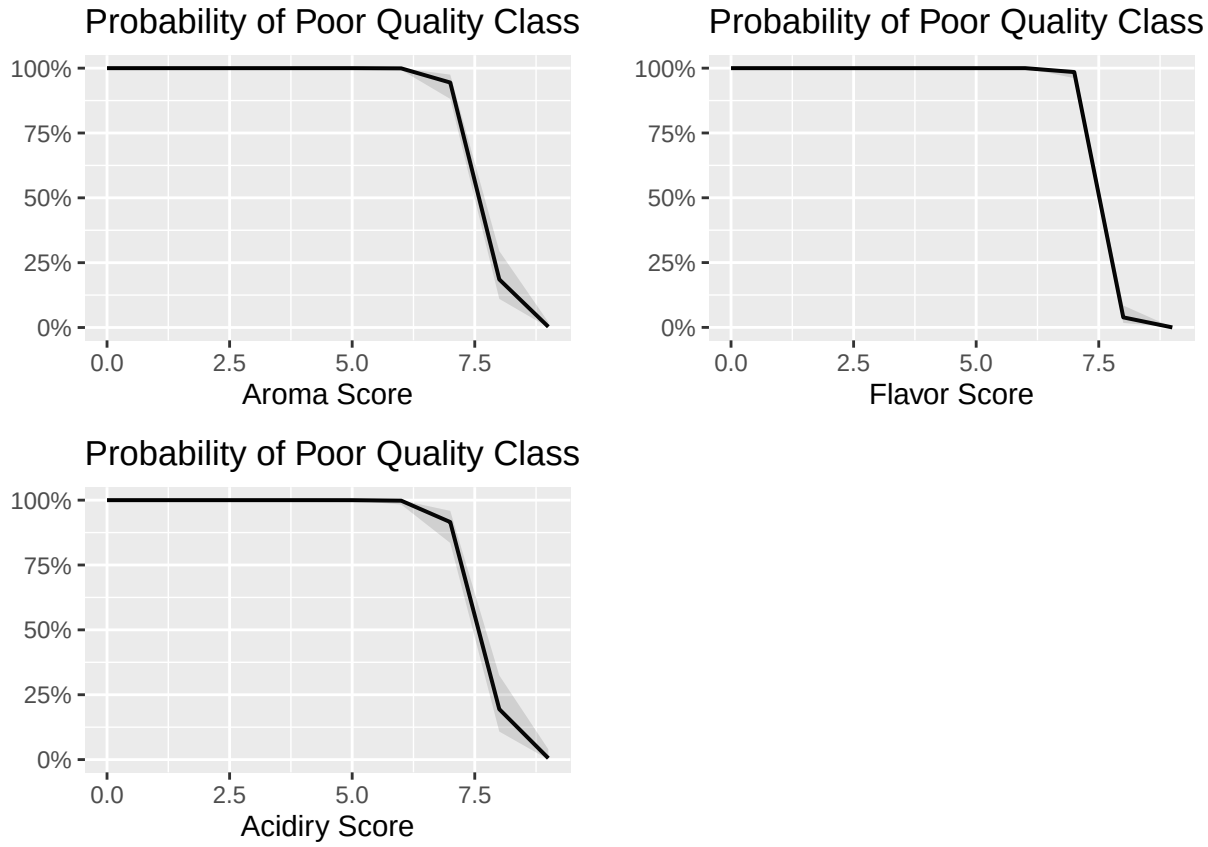


Figure 2: The Predictive Probability of Poor Quality Coffee against Score

In Model_5 we use logit link function, which related with the probability. We examine the predicted probability changes with predictor by Model_5. From the predictive probability plot, the probability of poor quality rapid decrease when the score higher 6 on aroma, flavor and acidity. In which, the decrease trend of flavor most significantly, its prediction line steeper then other predictors.

4 Extend Analysis – Prediction Assesment.

Extend analysis will apply two machine leaning techniques and evaluate the prediction ability by accuracy, sensitivity and specificity.

The data set divided to train data and test data at ratio 8 to 2. and fit the model_5 at train data to obtain the parameters. The predictive data were generate by fitted model_5 in train_data. to establish confusion matrix with predictiev data and actual data.

4.1 Confusion Matrix.

Table 5: Accuracy of Prediction.

	Value
Accuracy	0.8694639
Kappa	0.7399662
AccuracyLower	0.8338570
AccuracyUpper	0.8998668
AccuracyNull	0.5244755
AccuracyPValue	0.0000000
McnemarPValue	0.0003085

Table 6: The Resule of Sensitivity and Specificity of Prediction.

	Value
Sensitivity	0.8133333
Specificity	0.9313725
Pos Pred Value	0.9289340
Neg Pred Value	0.8189655
Precision	0.9289340
Recall	0.8133333
F1	0.8672986
Prevalence	0.5244755
Detection Rate	0.4265734
Detection Prevalence	0.4592075
Balanced Accuracy	0.8723529

Table 7: Confuse table.

	Actual Good	Actual Bad
0	190	42
1	14	183

Confusion matrix could summaries the prediction result and solving classified questions. The result will indicates the 2 types of errors with are false negatives (FN) and false positives (FP). Applying in this research, Confusion matrix classify the event by 4 type:

1.True positive (TP): the model predict coffee quality is poor, and the actual observation is poor.
2.False positive (FP): the model predict coffee quality is poor, and the actual observation is good.
3.True negative (TN):the model predict coffee quality is good, and the actual observation is good.
4. False negative (FN):the model predict coffee quality is good, and the actual observation is poor.

The result of confusion matrix indicate highest value on specificity. However, accuracy and sensitivity might relative low. The prediction summary indicates the most occurred error type is FP, which means we wrongly classify the coffee of prediction as poor quality, but the actual class is good.

The lower value of Accuracy is contribute by FP. The more of FP error occurs, suggest the count of TP will decrease. The TP is one parameters to calculate model accuracy. Thus, it can be considered that the more error event of FP occur, it will decrease the model accuracy. For the model sensitivity, TP error will present a direct influence. It is defined by the calculation formula of sensitivity.

4.2 ROC Curve

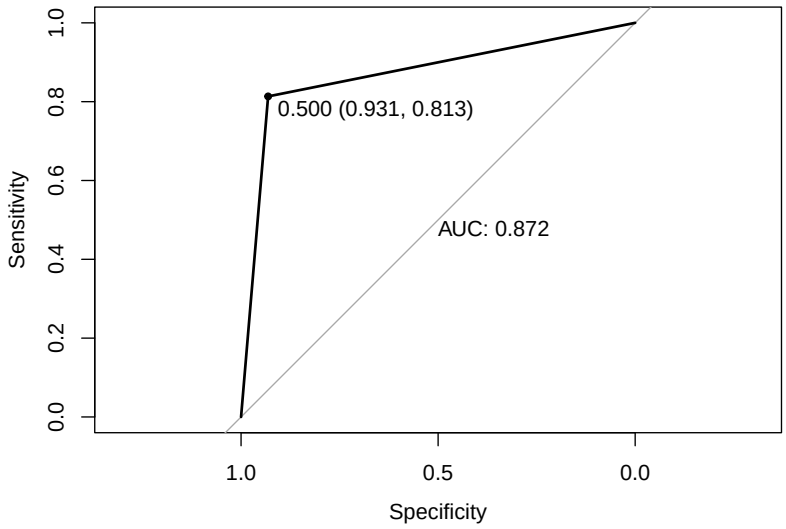


Figure 3: ROC cureve for model predicton

Receiver Operating Characteristic curve(ROC curve) is a method used to indicates the determination ability. It is usually apply to the binary system, which suit for the final model. When Model_5 apply to ROC curve, its indicates similar result with confusion matrix, that is high specificity and related lower sensitivity. But, it needs to notice that the AUC (Area Under the Curve) is over 0.8. In general, it is considers the model have predictive meaning when AUC value over 0.5.

According to the prediction assessment, a possible strategy that could improve the coffee quality model is decrease FP. occurs.

5 Conclusion

Aroma,flavor and acidity affect the quality of coffee. With the increase of aroma, the probability of good coffee quality increases. Similarly, as flavor and acidity increase, the probability of good coffee quality increases. The variation of flavor has the greatest impact on the quality of coffee, while the variation of acidity has the least impact on the quality of coffee.

All but three of the variables in the final model had very little effect on the quality of the coffee.

Further study: Evaluate the predictive power of the selected model by comparing the values of accuracy, sensitivity, specificity and AUC with the confusion matrix and ROC.