University of Glasgow

Assessing the Impact of Coffee Characteristics on Quality Classification on GLM

Hanwen Yuan, Zhujunyi Li, Chunyao Hou

School of Mathematics and Statistics,
University of Glasgow, UK

Contents

University of Glasgow

Introduction

1. Coffe quality assessment is a crucial aspect of the coffee industry, impacting both market value and consumer preferences. The study aims to explore the factors that influence whether a batch of coffee is classified as “Good” or “Poor”.

2. Our dataset includes more than 1000 coffee samples from different countries and records sensory attributes (aroma, flavor and acidity), production characteristics (harvest year, altitude) and defect counts. The score threshold (82.5 as “Good”, <82.5 as “Poor”).

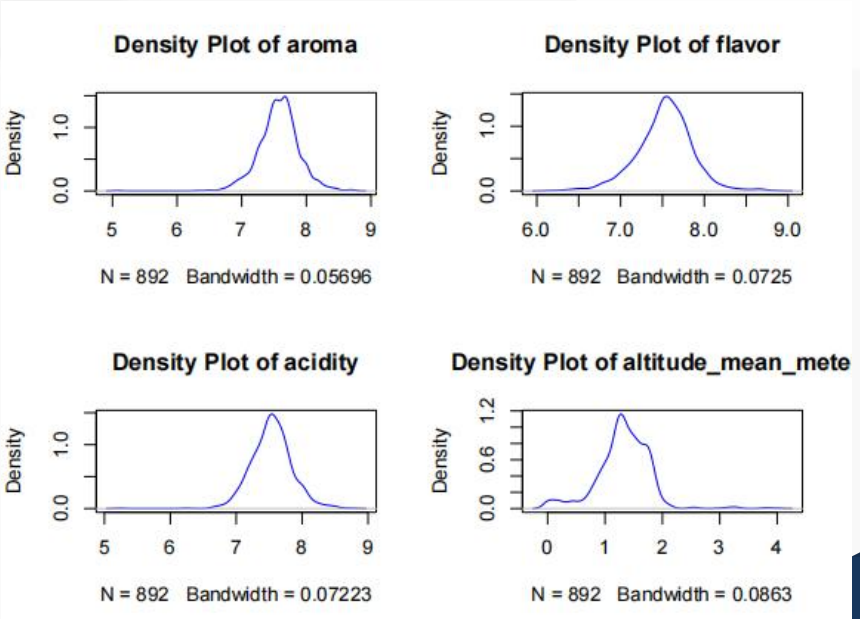
3. The primary goal of this study is to analyze the relationship between coffee quality and factors mentioned above then decide which factors has significant influence in coffee quality classfication.

University of Glasgow

EDA

1. Data preparing and Cleaning
2. Graphical summaries
3. Numerical summaries

EDA



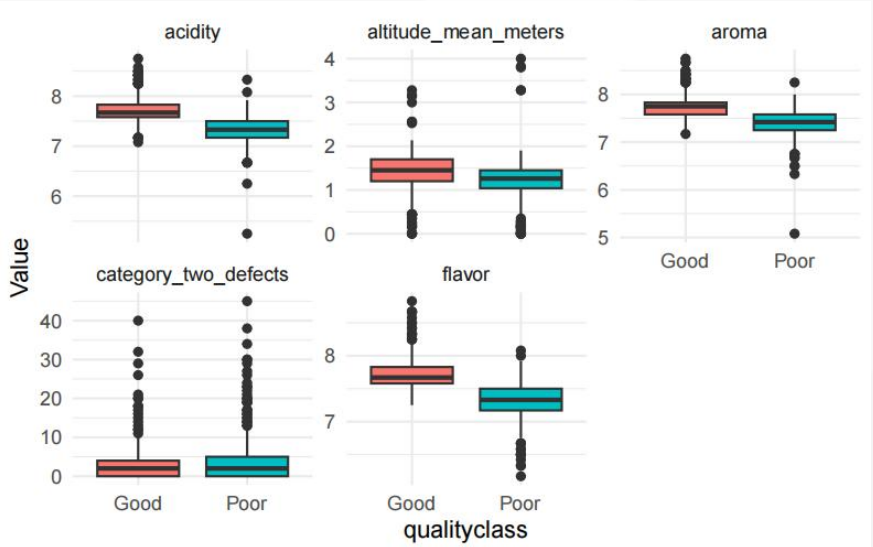
The first three plots show **near-Normal Distributions**, indicating a relatively symmetric evaluation across samples.
However, the distribution of altitude is right-skewed with a few higher altitude outlier. This justifies our decision to standardize the altitude variable to reduce scale effects and improve model interpretability.

01
Outliers

Density plots
02

Boxplot
03

We check the data then found that data of altitude_mean_meters has **two outliers** which were higher than the highest mountain 8848 meters so we **remove** these two row.



Reveal clear differences in the **median values** of aroma, flavor, and acidity between the "Good" and "Poor" quality classes.
Notably, "Good" coffee samples tend to have **higher median scores** in these sensory attributes, suggesting their strong influence on quality classification.

EDA

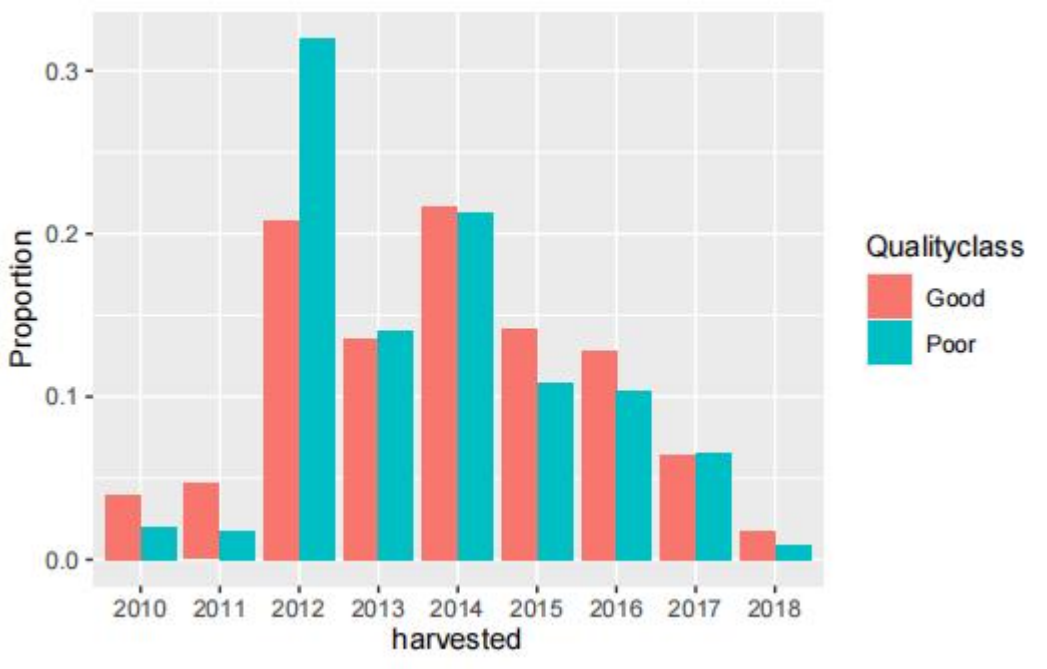


Table 1: Summary Statistics of continuous factors

Variable	Mean	Median	Std. Dev	Minimum	Maximum	IQR
aroma	7.57	7.58	0.32	5.08	8.75	0.33
flavor	7.53	7.58	0.33	6.17	8.83	0.42
acidity	7.54	7.50	0.31	5.25	8.75	0.42
category_two_defects	3.50	2.00	5.21	0.00	45.00	4.00
altitude_mean_meters	1.32	1.31	0.47	0.00	4.00	0.50

Table 2: Summary Statistics of harvested

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Count	27	29	235	123	192	112	104	58	12

About the harvested year the barplot shows no clear pattern in the distribution of quality classification over different years, indicating that harvest year might **not significantly affect** the classification, which aligns with the later formal analysis results.

The summary table shows aroma, flavor and acidity have **similar distributions**. The variable of defects and altitude have different distributions. And the data of harvested shows they concentrated in **2012-2016**.

University of Glasgow

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X_{aroma} + \beta_2 \cdot X_{flavor} + \beta_3 \cdot X_{acidity} + \beta_4 \cdot X_{defects} + \beta_5 \cdot X_{altitudes} + harvested$$

Formal Data Analysis

We considered the harvested of 2010 as the baseline. The function below:

$$harvested = \mathrm{I}_{2011}(x) + \mathrm{I}_{2012}(x) + \mathrm{I}_{2013}(x) + \mathrm{I}_{2014}(x) + \mathrm{I}_{2015}(x) + \mathrm{I}_{2016}(x) + \mathrm{I}_{2017}(x) + \mathrm{I}_{2018}(x)$$

$$\mathrm{I}_j(x) = \begin{cases} 1, & \text{if group of harvested } x \text{ is considered as } j, \\ 0, & \text{Otherwise} \end{cases}$$

α: Represents the baseline log-odds when all explanatory variables are set to zero.

$\{\beta_i\}, i = 1, \ldots, 5$, are the coefficients . Which means when X_i increase 1, the probability will change according to the β_i.

p=Prob (good) represent the probability of coffee quality being good.

Model Selection

Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(13)$	723.60
Pseudo-R^2 (Cragg-Uhler)	0.74
Pseudo-R^2 (McFadden)	0.59
AIC	540.86
BIC	607.97

	Est.	S.E.	z val.	p
(Intercept)	-122.73	9.14	-13.43	0.00
aroma	5.02	0.73	6.84	0.00
flavor	7.29	0.89	8.18	0.00
acidity	3.86	0.70	5.48	0.00
category_two_defects	0.03	0.03	1.00	0.32
altitude_mean_meters	0.59	0.24	2.42	0.02
harvested2011	-0.09	1.09	-0.08	0.93
harvested2012	-0.70	0.91	-0.77	0.44
harvested2013	-0.25	0.92	-0.27	0.78
harvested2014	0.06	0.92	0.07	0.94
harvested2015	-0.47	0.93	-0.51	0.61
harvested2016	0.38	0.96	0.40	0.69
harvested2017	0.15	0.96	0.16	0.87
harvested2018	1.59	1.26	1.27	0.20

Standard errors: MLE

The model 1 (**full model**) with 6 variables outcome showed the p-value of defects was **higher than 0.05**.

Model Selection

Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(12)$	722.58
Pseudo-R^2 (Cragg-Uhler)	0.74
Pseudo-R^2 (McFadden)	0.58
AIC	539.89
BIC	602.20

	Est.	S.E.	z val.	p
(Intercept)	-122.23	9.10	-13.44	0.00
aroma	5.02	0.73	6.84	0.00
flavor	7.25	0.89	8.11	0.00
acidity	3.83	0.70	5.44	0.00
altitude_mean_meters	0.61	0.24	2.49	0.01
harvested2011	-0.05	1.09	-0.05	0.96
harvested2012	-0.58	0.90	-0.64	0.52
harvested2013	-0.20	0.92	-0.22	0.83
harvested2014	0.10	0.92	0.11	0.91
harvested2015	-0.41	0.92	-0.44	0.66
harvested2016	0.45	0.95	0.47	0.64
harvested2017	0.22	0.96	0.23	0.81
harvested2018	1.67	1.25	1.33	0.18

Standard errors: MLE

The model 2 (**removed defects**) with 5 variables outcome showed the p-value of harvested was higher than 0.05. **AIC and BIC decreased.**

Model Selection

Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(4)$	708.08
Pseudo-R^2 (Cragg-Uhler)	0.73
Pseudo-R^2 (McFadden)	0.57
AIC	538.38
BIC	562.35

	Est.	S.E.	z val.	p
(Intercept)	-119.12	8.69	-13.71	0.00
aroma	4.66	0.69	6.74	0.00
flavor	7.04	0.86	8.17	0.00
acidity	4.00	0.69	5.81	0.00
altitude_mean_meters	0.46	0.23	2.00	0.05

Standard errors: MLE

The model 3 (**removed harvested**) with 4 variables outcome showed the p-value of all variables were **lower than 0.05**. **AIC and BIC decreased**.

Model Selection

Observations	892
Dependent variable	Qualityclass_dummy
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(3)$	704.14
Pseudo-R^2 (Cragg-Uhler)	0.73
Pseudo-R^2 (McFadden)	0.57
AIC	540.32
BIC	559.50

	Est.	S.E.	z val.	p
(Intercept)	-118.95	8.65	-13.76	0.00
aroma	4.81	0.69	6.97	0.00
flavor	6.89	0.85	8.12	0.00
acidity	4.06	0.68	5.95	0.00

Standard errors: MLE

The model 4 (**removed altitudes**) with 3 variables outcome showed the p-value of all variables were **lower than 0.05**. **AIC increased and BIC decreased**.

Model Selection

Table 5: Comparison for the 4 models

Model	AIC	BIC
model1	540.86	607.97
model2	539.89	602.20
model3	538.38	562.35
model4	540.32	559.50

We found that model 3 and model 4 have **closely AIC and BIC**. Then which one be the best?

Table 1: model-1 summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
1236.46	891	-256.43	540.86	607.97	512.86	878	892

Table 2: model-2 summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
1236.46	891	-256.94	539.89	602.2	513.89	879	892

Table 3: model-3 summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
1236.46	891	-264.19	538.38	562.35	528.38	887	892

Table 4: model-4 summary

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
1236.46	891	-266.16	540.32	559.5	532.32	888	892

Hypothesis testing using deviance:
M0=model4, M1=model 3;
D0=532.32, q=3; D1=528.38, p=4;
D0-D1=532.32-528.38=3.94;
Chi-square(p-q=4-3=1)=3.84; It has 3.94>3.84, then reject H0.

University of Glasgow

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X_{aroma} + \beta_2 \cdot X_{flavor} + \beta_3 \cdot X_{acidity} + \beta_4 \cdot X_{altitudes}$$

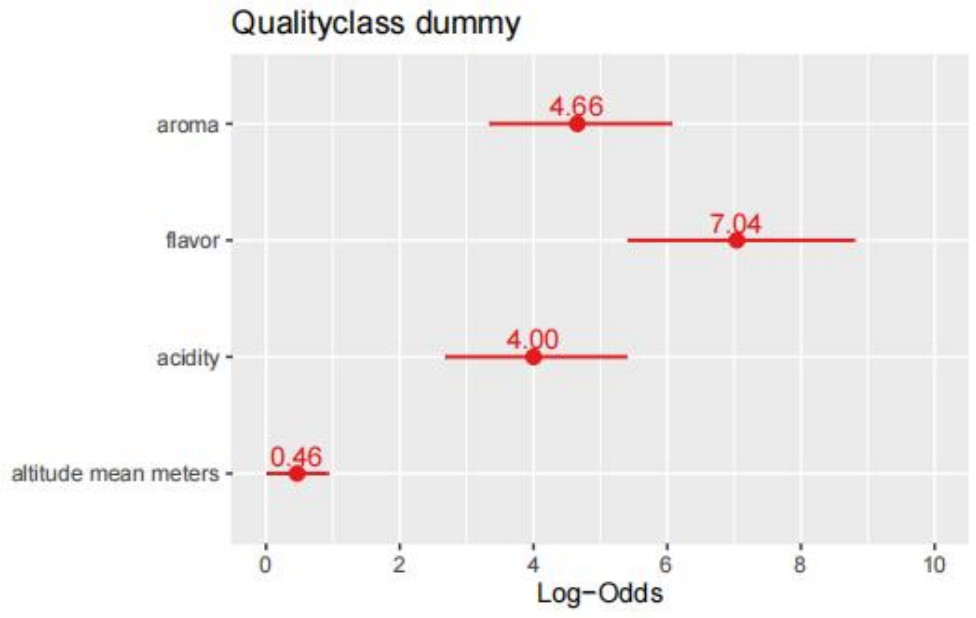
Result

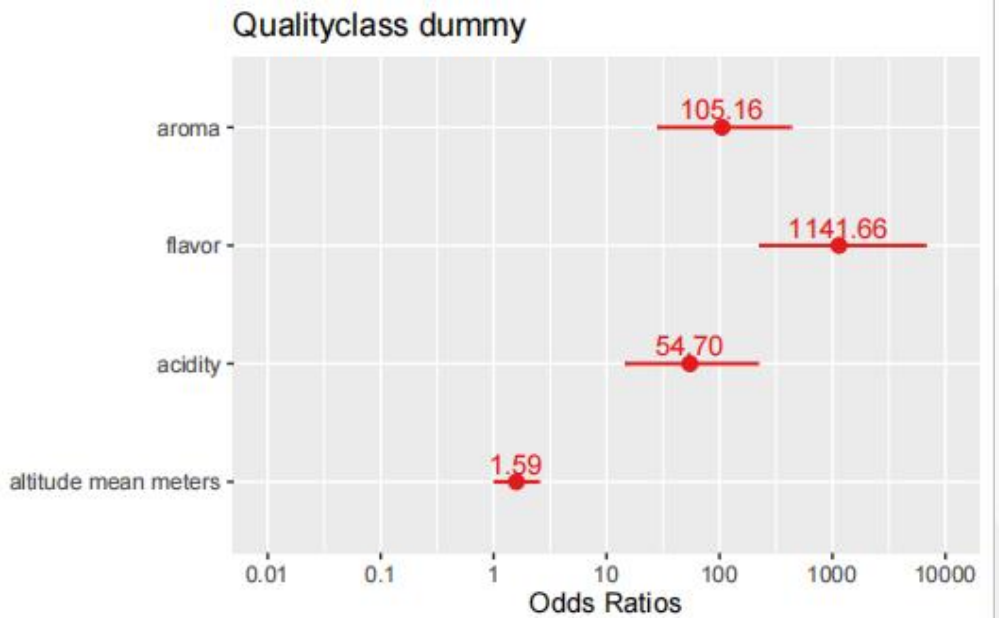
Table 7: Confidence Interval of model-3

	2.5 %	97.5 %
(Intercept)	-137.1043034	-103.0008636
aroma	3.3414665	6.0512855
flavor	5.4098353	8.7926752
acidity	2.6776888	5.3801746
altitude_mean_meters	0.0062384	0.9181653

Finally we choose the model 3 for the best model in this research. And its CI showed not include 0, that was significant for coffee.

Log-odds/odds

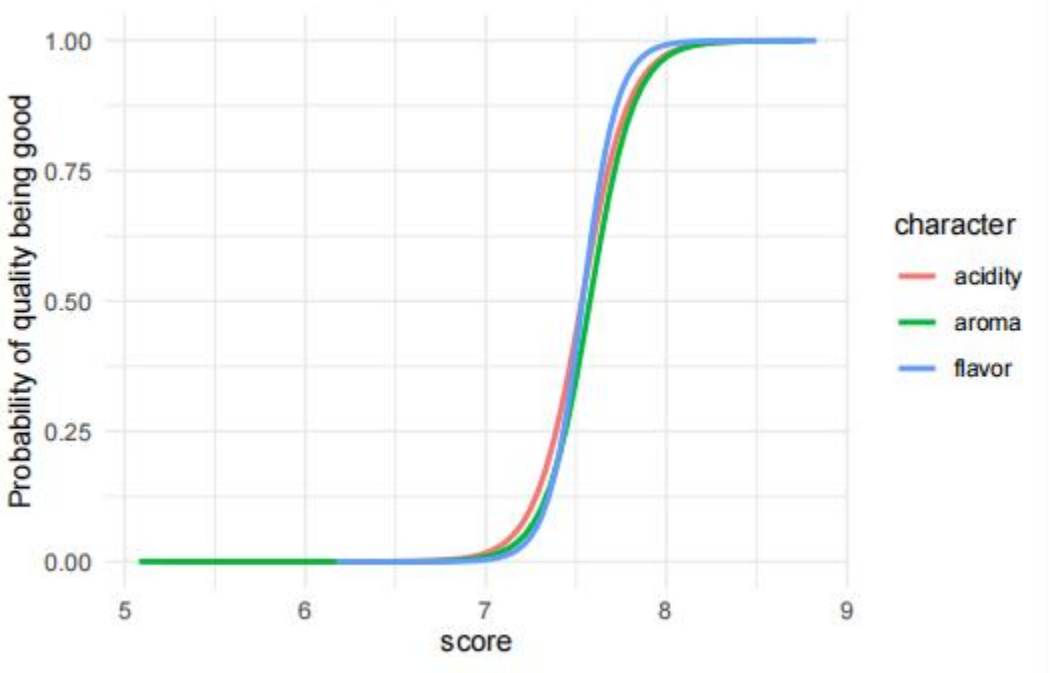


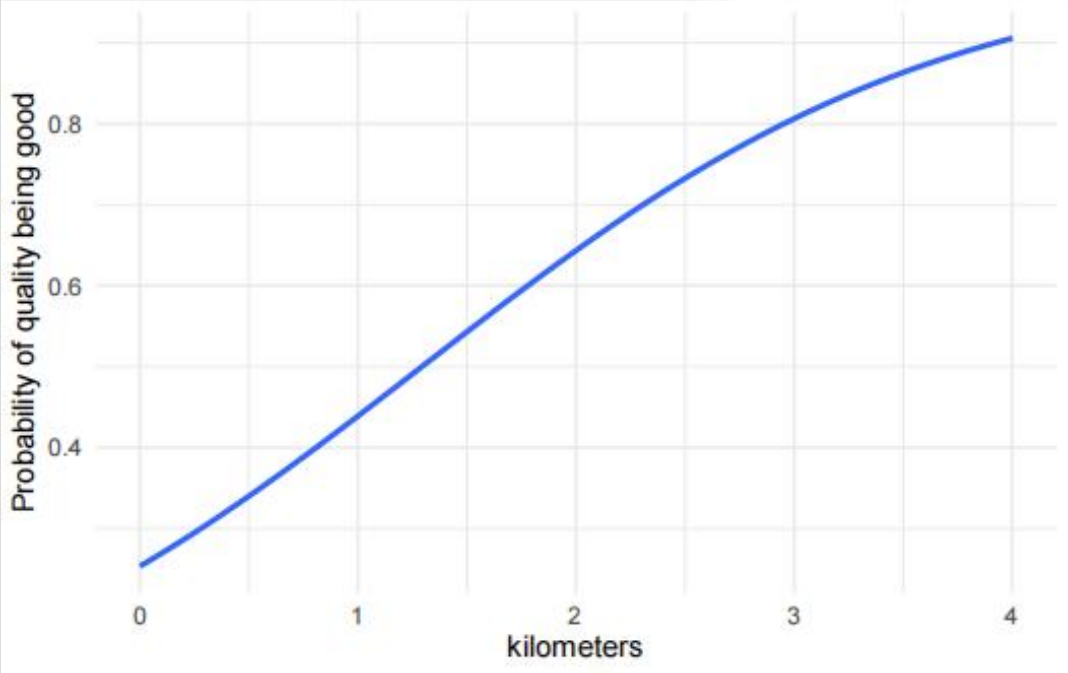


All four variables retained had a **significant positive effect** on coffee quality classification. Especially the coefficients for flavour was **7.04** that an increase in score per unit significantly increases the log odds of a coffee being classified as 'Good'. In addition, acidity and altitude_mean_meters also show positive effects, although the magnitude is relatively small.

The odds ratios for flavor and aroma are about **1141** and **105**, respectively, indicating that for every one unit increase in flavour and aroma, the chances of a coffee being rated as 'Good' are **hundreds or even thousands of times higher**. The odds ratios for acidity and altitude were **54.7** and **1.59** respectively, indicating that their positive effects on quality are also not negligible, but are slightly milder than those of flavour.

Probability Trends

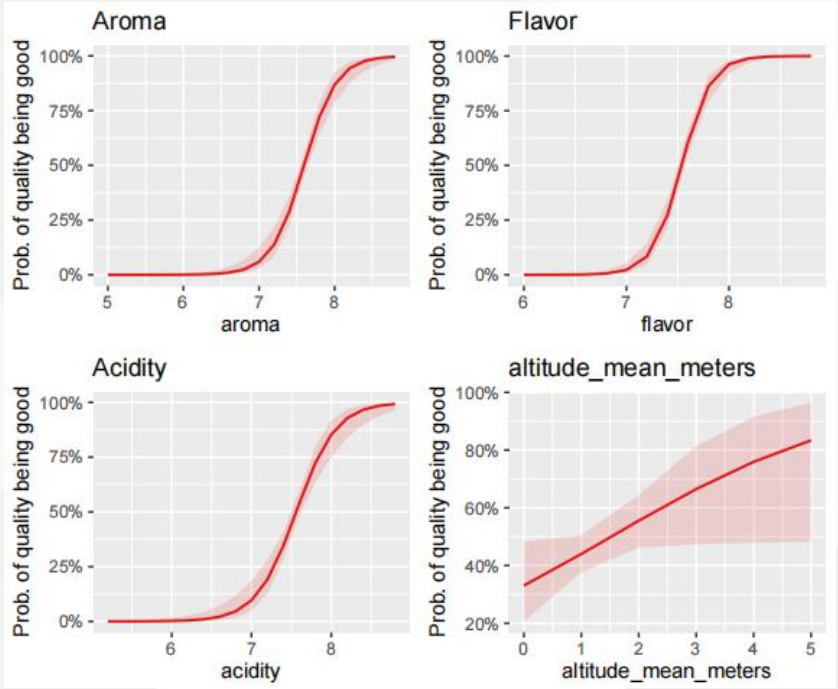




As shown in figure(left), all three sensory attributes positively impact the probability of coffee being classified as 'Good'. Notably, **flavor has the steepest curve**, indicating that improvements in flavor score have the most substantial effect on quality classification, especially within the **6.5 to 8** score range.

Figure(right) illustrated that higher altitudes are associated with a higher probability of being classified as 'Good'. . The positive trend becomes more pronounced between 1km and 3km.

Conclusion



In conclusion, after fitting and comparing **four logistic regression models**, we identified that aroma, flavor, acidity, and altitude_mean_meters are significant predictors of coffee qualityand **flavor** score have the **most substantial effect** on quality classification. Through stepwise elimination, the final model (**Model 3**) was chosen based on its relatively **lowest AIC** and sufficient explanatory power.

Limited

1. **Limited Variable Scope**: This study focused only on a subset of available variables, including sensory scores (aroma, flavor, acidity), average altitude, defect counts, and harvest year. However, other potentially influential factors such as **coffee variety, processing methods,** or **climatic conditions** were not included, which may lead to different results.

2. Another potential limitation is **multicollinearity** among the sensory attributes (aroma, flavor, acidity), which could affect the stability of coefficient estimates.

3. Additionally, this study did not account for **potential interaction effects** between variables. For example, the combined influence of aroma and flavor might have a synergistic effect on coffee quality, which could be explored in future models by including interaction terms.

University of Glasgow

Thanks for Listening!

Group 11 members:
Hanwen Yuan, Zhujunyi Li, Chunyao Hou, Wei Li, Congle Wang