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:

```
##    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:

```
## 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", "~
```

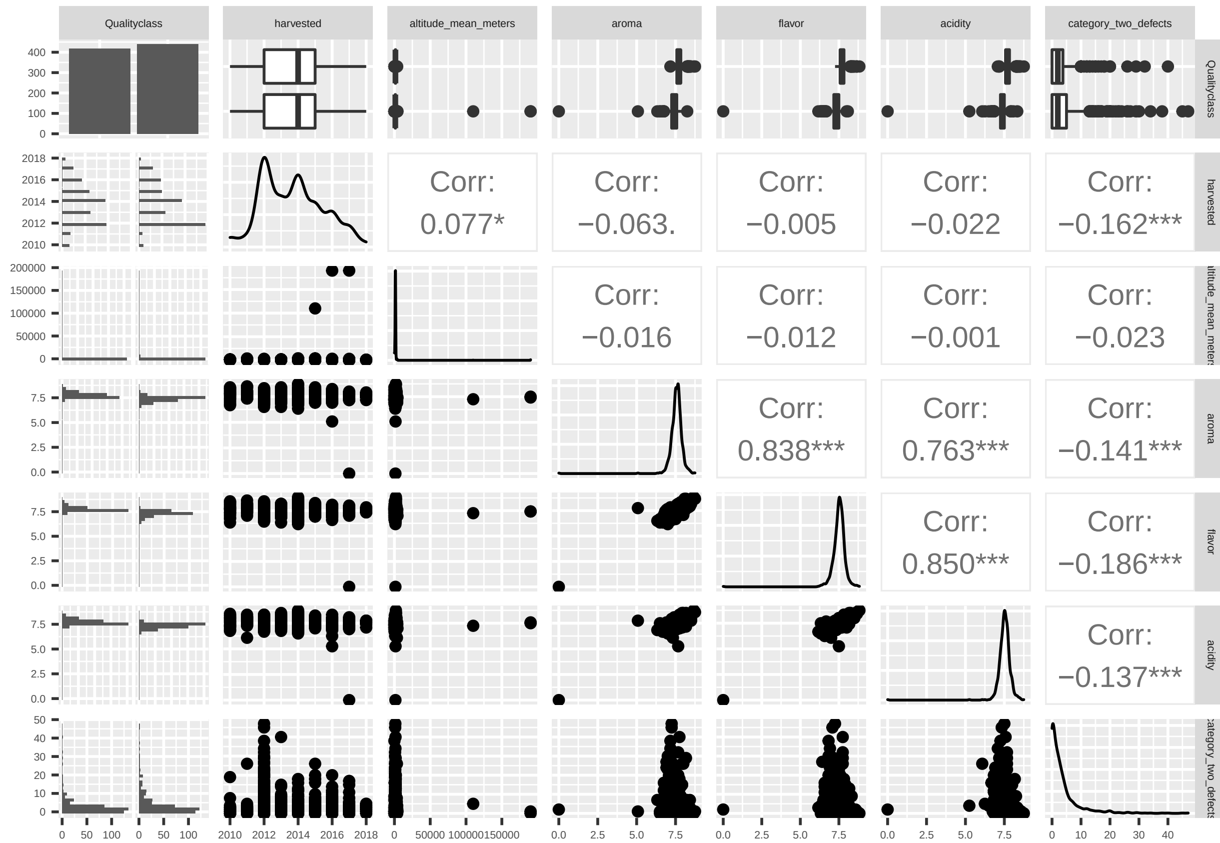
The number of unique values in country of origin:

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

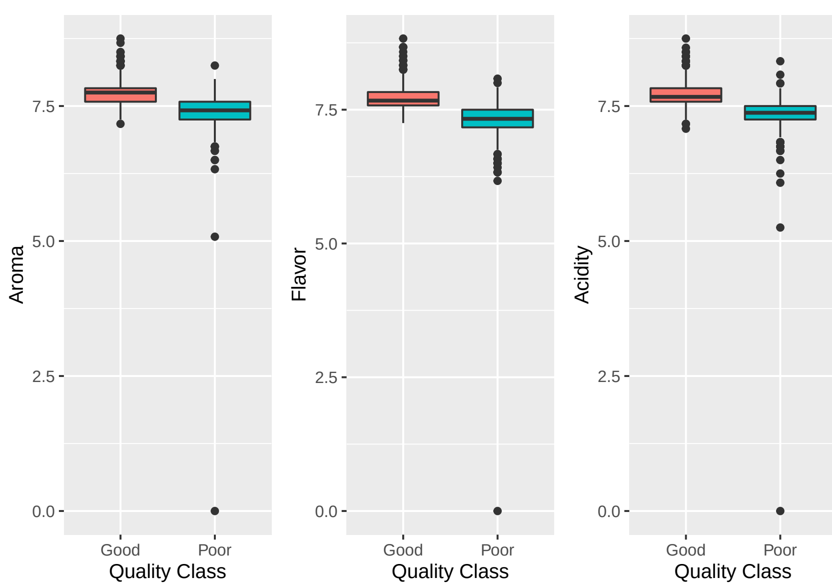
The number of unique values in harvest year:

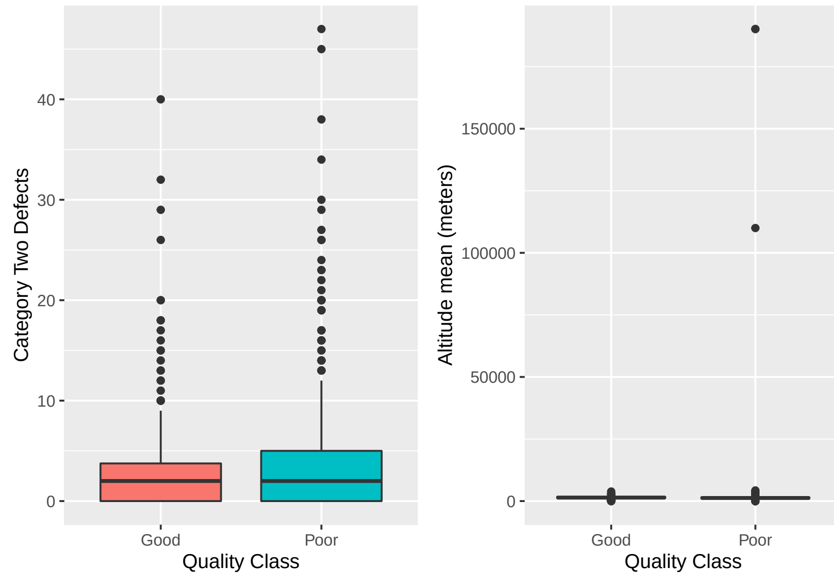
```
## [1] 9
```

The correlation between the quantitative variables:

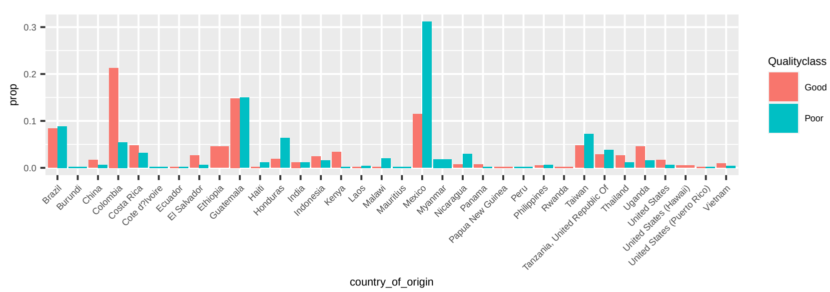


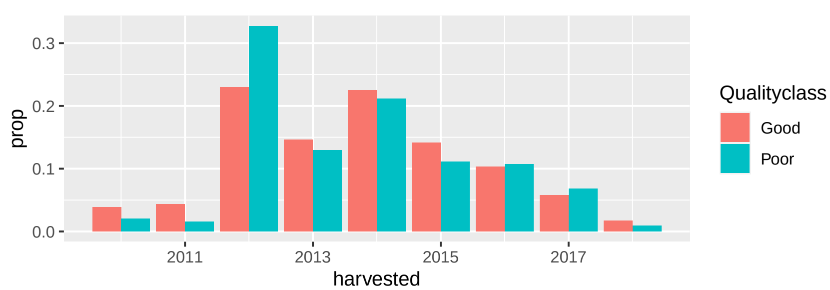
Box plots showing the distribution of the quantitative variables

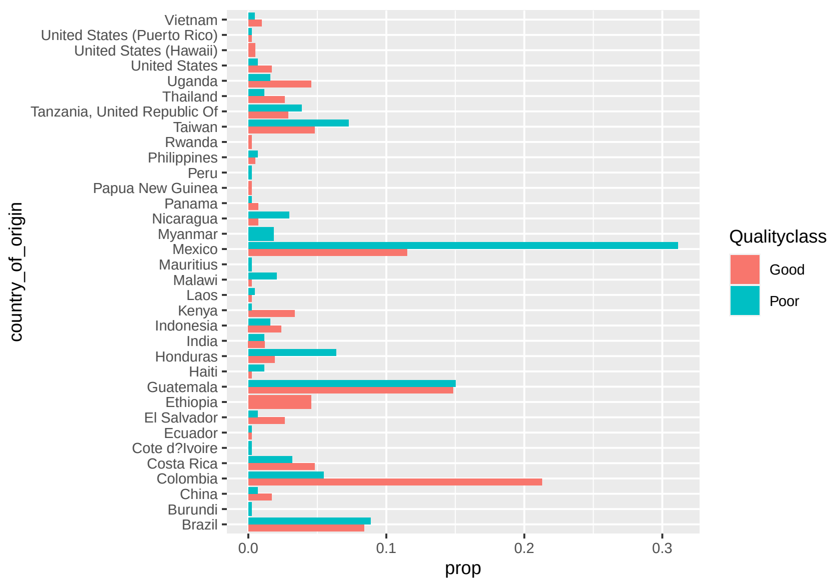


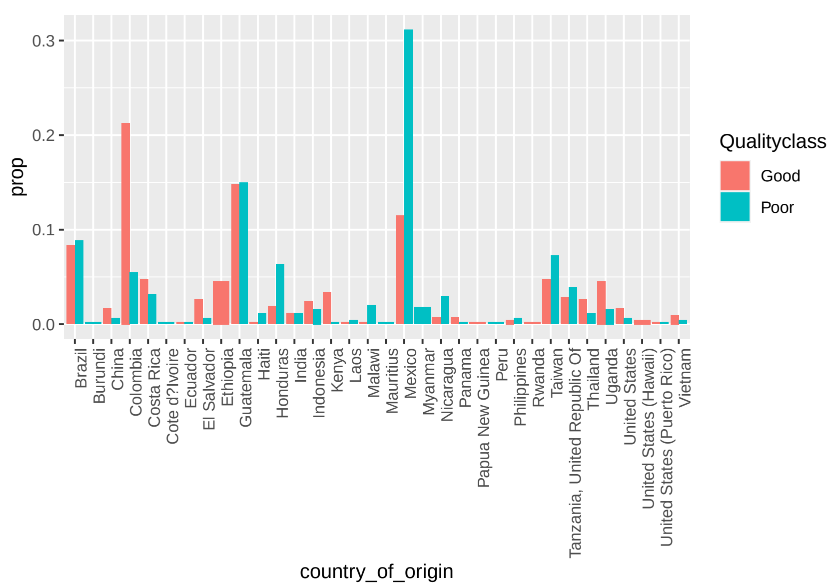


2.1 bar charts:









The percentages:

Table showing the percentage of the quality classes for each country

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 showing the percentage of the quality classes for each harvest year:

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 Analsis

Model 1:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Country} + \beta_2 \cdot \text{Aroma} + \beta_3 \cdot \text{Flavor} + \beta_4 \cdot \text{Acidity} + \beta_5 \cdot \text{Category Two Defects} + \beta_6 \cdot \text{Harvested} + \beta_7 \cdot \text{Altitude}$$

Model 2:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Country} + \beta_2 \cdot \text{Aroma} + \beta_3 \cdot \text{Flavor} + \beta_4 \cdot \text{Acidity} + \beta_5 \cdot \text{Category Two Defects} + \beta_6 \cdot \text{Harvested}$$

Model 3:

$$ln\left(\frac{p_{poor}}{1-p_{poor}}\right) = \alpha + \beta_1 \cdot \text{Country of origin} + \beta_2 \cdot \text{aroma} + \beta_3 \cdot \text{flavor} + \beta_4 \cdot \text{acidity} + \beta_5 \cdot \text{category two defects}$$

Model 4:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Aroma} + \beta_2 \cdot \text{Flavor} + \beta_3 \cdot \text{Acidity} + \beta_4 \cdot \text{Category Two Defects}$$

Model 5:

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

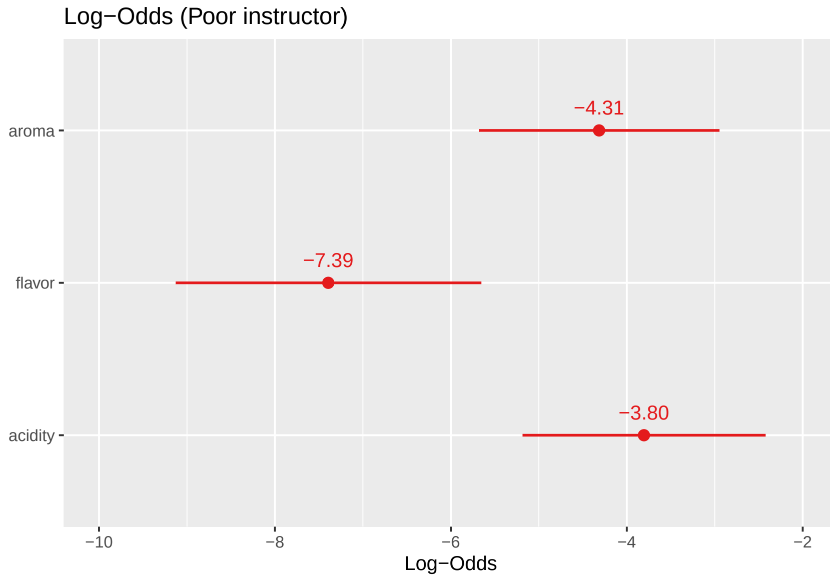
```
## [1] "Good" "Poor"
```

3.1 Models comparison and Selection:

Table 3: The Result of Model comparison

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

3.2 log Odds:

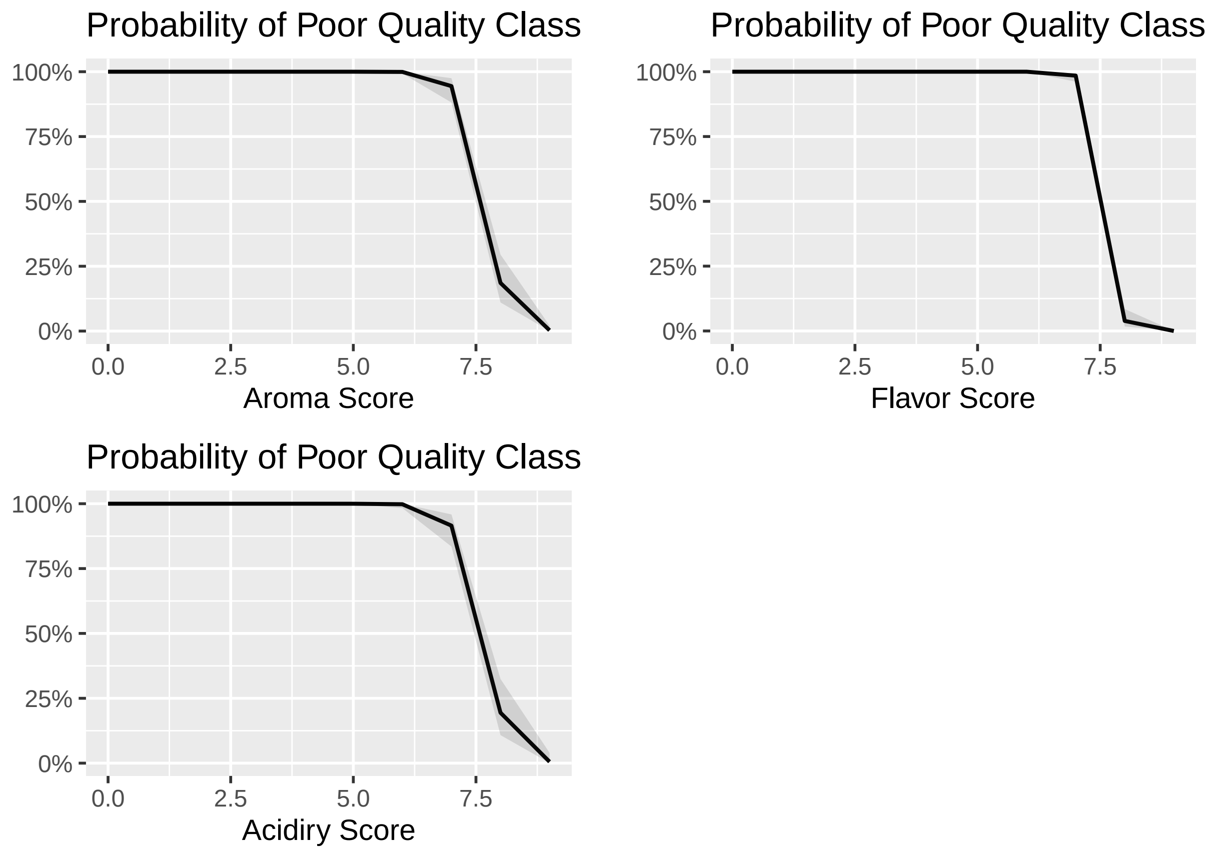


3.3 Confidence Intervals:

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

3.4 The Probability Plot:



4 Extend Analysis – Prediction Assesment.

4.1 Confusion Matrix

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity, family = binomial(link = "logit"),
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3509  -0.3845   0.0034   0.3167   3.7376
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 125.0857    10.3106  12.132  < 2e-16 ***
## aroma        -5.1905     0.8475  -6.125 9.08e-10 ***
## flavor       -6.9357     1.0111  -6.860 6.91e-12 ***
## acidity      -4.4311     0.8479  -5.226 1.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 950.29  on 685  degrees of freedom
## Residual deviance: 379.62  on 682  degrees of freedom
## AIC: 387.62
##
## Number of Fisher Scoring iterations: 8
```

```
## [1] "0" "1"
```

Table 5: Accuracy of Prediction.

	Value
Accuracy	0.8313953
Kappa	0.6627907
AccuracyLower	0.7669156
AccuracyUpper	0.8840801
AccuracyNull	0.5000000
AccuracyPValue	0.0000000
McnemarPValue	0.0093296

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

	Value
Sensitivity	0.7441860
Specificity	0.9186047
Pos Pred Value	0.9014085
Neg Pred Value	0.7821782
Precision	0.9014085
Recall	0.7441860
F1	0.8152866
Prevalence	0.5000000
Detection Rate	0.3720930
Detection Prevalence	0.4127907
Balanced Accuracy	0.8313953

Table 7: Confuse table.

	Actual Good	Actual Bad
0	79	22
1	7	64

4.2 ROC Curve

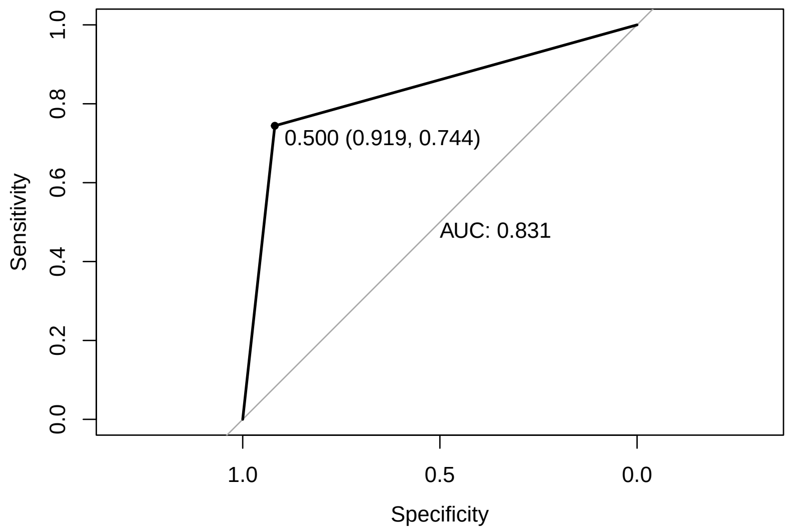


Figure 1: ROC cureve for model predicton

5 Conclusion