writeup

STA 210 - Summer 2022

Team 2 - Alicia Gong, Ashley Chen, Abdel Shehata, Claire Tam

6-16-2022

Introduction and Data

Introduction About 234 million hectoliters of wine were consumed in 2020, worldwide, with the US making up approximetly 14% of that consumption (Karlsson 2020). Since Wine composition and wine quality varies widely, it raises the question: what makes a good wine?

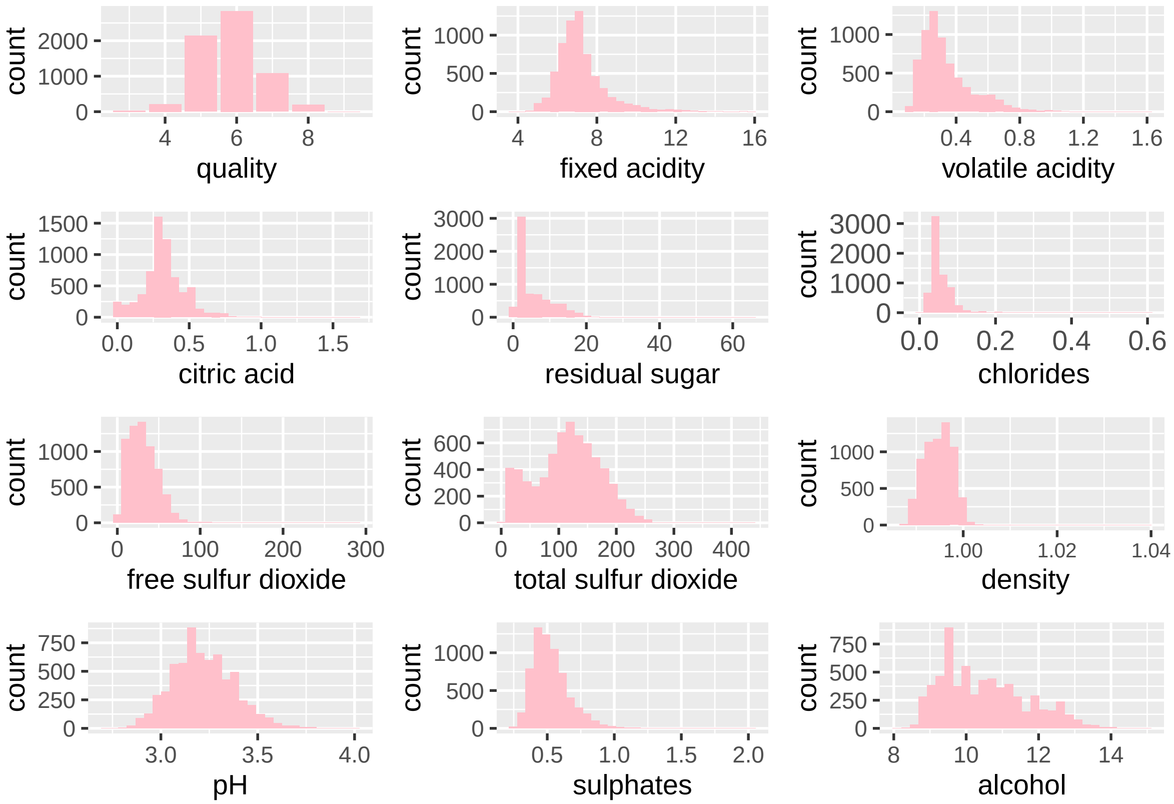
To answer that question, we will analyze the wine quality dataset from Vinho Verde vinyard in Portogal, and more importantly try to narrow down our question to make it possible for it to be supported by evidence. Below is the introduction to our research:

Project Goal: To identify variables that are important in explaining variation in the response. "Vinho Verde" is the kind of wine exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal. The Vinho Verde wine has its own unique system of production and is only produced from the indigenous grape varieties of the region. Vinho Verde region is one of the largest and oldest wine regions in the world, and is home to thousands of producers, generating a wealth of economic activity and jobs, and strongly contributes to the development of Minho province and the country. The Vinho Verde wine also enjoys high reputation worldwide. It is recurrently awarded in national and international competitions. The goal of this dataset is to model wine quality based on physicochemical tests. We believe that this dataset can also be used to analyze the relationship between different chemical compositions and the ratings of wine quality. We believe that this dataset can also be used to analyze what chemical factors are attributable to the final rating of Vinho Verde wine. Our research may shed light on future research and development directions for improving the quality of Vinho Verde wine, which may also contribute to the competitiveness of Portuguese wine industry.

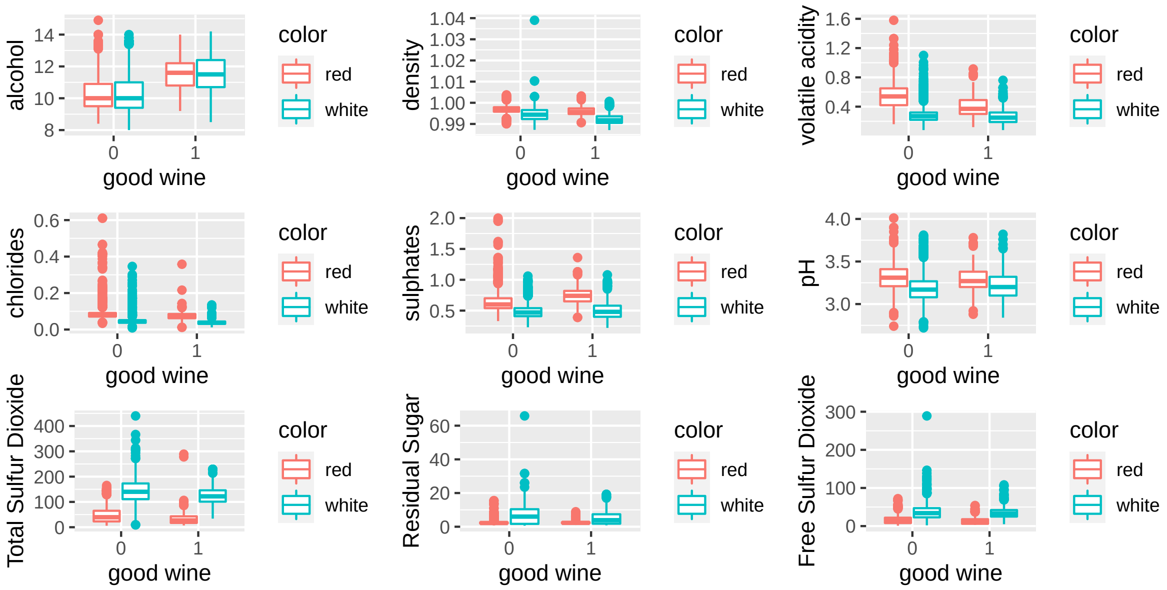
Our goal is to produce a classification model that best explains how different chemical compositions of the Portuguese "Vinho Verde" wine affects the variation of the wine quality.

Data Introduction The Wine Quality dataset was collected from Vinho Verde wine Samples, from the North of Portugal. The data was originally donated in 2009 by Professor Cortez. The specific mechanism of the collection of the data was lab work done on different wines to measure their chemical attributites (like acidity etc.). The quality of the wine however was obtained through the average rating of three wine experts. The dataset is divided into two: Red wine and White wine. Red Wine has 1599 observations, and white wine has 4898 observations (each observation being a specific wine). Information about the wine include but are not limited to:PH,Density,Acidity, and alcohol content.

Each observation is a specific wine from the Vinho Verde region. Thus, there might be a little uncertainity in collecting the exact numbers for each numbers. However, since the Vinho Verde region is a vast region spreading 15500 hectareas of vineyards in far-north Portugal, this uncertainity shouldn't be significant in our analysis or project. Thus, we will assume that the datas are independent and random



Exploratory Data Analysis



Methodology

Best AIC Model We choose best AIC model. What is that? Why? Reason and justification.

Data split: we split the data into 25% testing set and 75% training set. ::: {.cell}

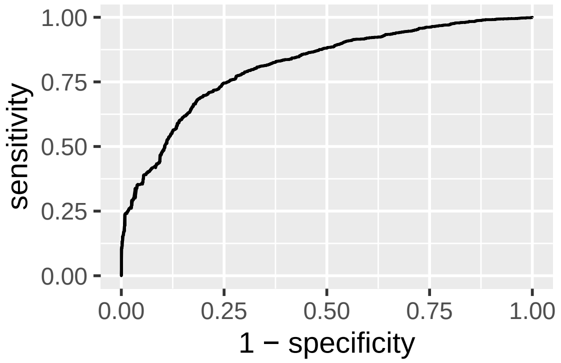
:::

Model output

term	estimate	std.error	statistic	p.value
(Intercept)	373.853	75.499	4.952	0.000
c_fixed.acidity	0.467	0.074	6.268	0.000
c_volatile.acidity	-3.712	0.434	-8.557	0.000
c_residual.sugar	0.223	0.030	7.360	0.000
c_chlorides	-8.086	3.052	-2.650	0.008
c_free.sulfur.dioxide	0.012	0.004	3.504	0.000
c_total.sulfur.dioxide	-0.005	0.002	-2.937	0.003
c_density	-394.148	76.494	-5.153	0.000
c_pH	2.345	0.416	5.636	0.000
c_sulphates	2.298	0.333	6.907	0.000
c_alcohol	0.505	0.093	5.453	0.000
colorwhite	-0.716	0.282	-2.541	0.011

ROC curve prediction

```
New names:
* `` -> ...1
```

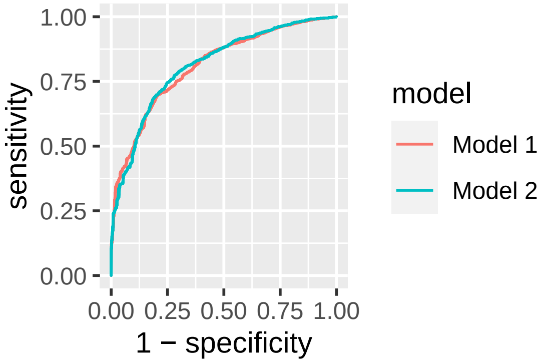


Model Evaluation

Logistic Model We compare the best AIC model with a logistic model with interactive terms.

term	estimate	std.error	statistic	p.value
(Intercept)	-6.040	75.975	-0.079	0.937
c_fixed.acidity	0.117	0.052	2.238	0.025
c_volatile.acidity	-4.007	0.443	-9.048	0.000
c_residual.sugar	0.170	0.026	6.631	0.000
c_chlorides	-11.626	3.227	-3.603	0.000
c_total.sulfur.dioxide	-0.010	0.004	-2.806	0.005
c_density	-6.545	75.897	-0.086	0.931
c_sulphates	2.770	0.611	4.533	0.000
c_alcohol	0.965	0.123	7.819	0.000
color_white	297.197	79.950	3.717	0.000
c_total.sulfur.dioxide_x_color_white	0.011	0.004	2.833	0.005
c_density_x_color_white	-295.272	79.490	-3.715	0.000
c_sulphates_x_color_white	-0.750	0.719	-1.044	0.297
c_alcohol_x_color_white	-0.407	0.145	-2.813	0.005

Model Selection Compare two ROC curves. ::: {.cell} ::: {.cell-output-display}

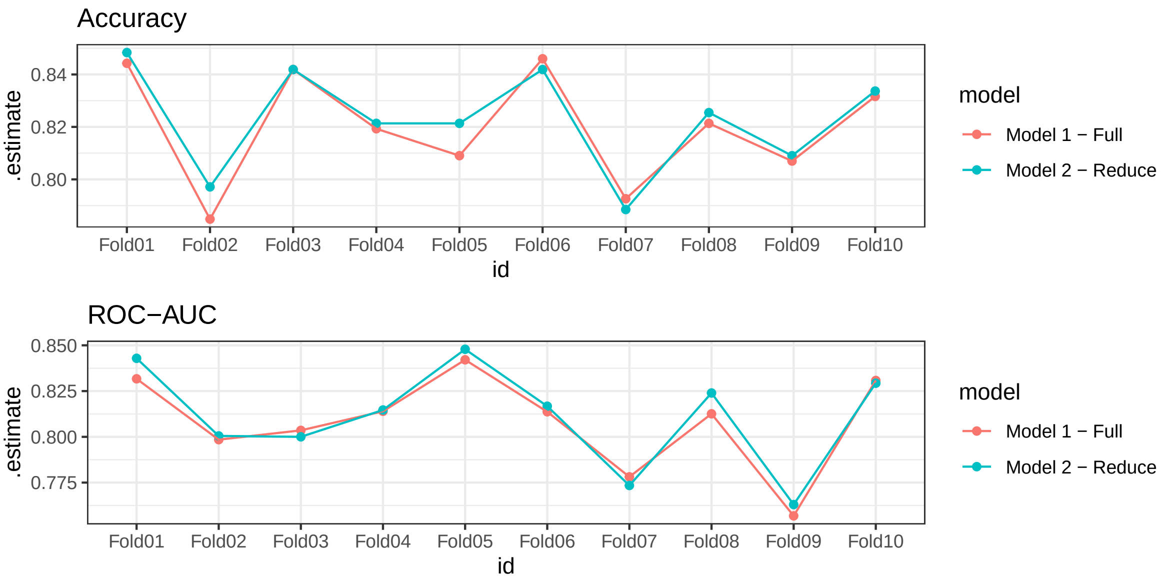


::: ::: The two ROC curves are very close, suggesting that the two models' performances are similar.

model	.metric	.estimator	.estimate
Logistic Model	roc_auc	binary	0.811
Best AIC Model	roc_auc	binary	0.814

Cross Validation To perform the cross validation, we split the training data into 10 folds. ::: {.cell}

:::



The two models have similar accuracy and the roc-auc for 10 folds, so due to the principles of parsimonious, we prefer the reduce one.

Check Conditions Not every variables satisfy the linearity condition. (This can be mentioned in the limitation part of the final report.) The independence is satisfied. The Vinho Verde region is a vast region spreading 15500 hectareas of vineyards in far-north Portugal, and the data are not collected across time.

Results

- model interpretation
- full model
- best AIC model
- AIC BIC ROC
- model selection: do not include the interactive terms due to parsimonious. because the ROC does not improve much.

Discussion & Conclusion

- what chemical components contribute to wine quality
- future research suggestions
- suggestions to wine valley?

Reference

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009.