Group 3

GETTING DRUNK

A Graduate Student's Guide to Surviving April

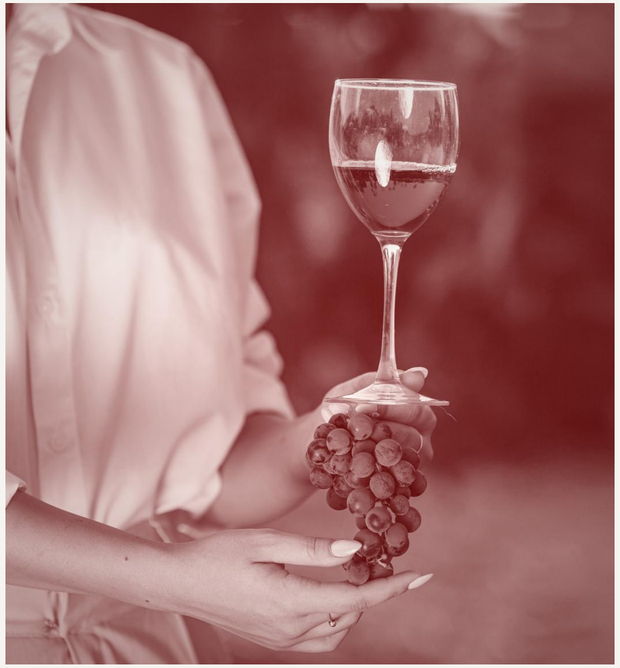
Chikai Chiang

Table Of Contents

01

Introduction

- The Problem
- The Solution
- Data Description



02

Methodology

- Logistic Regression
- Bayesian MCMC
- Newton Raphson
- SVM & MLP

03

Conclusion

- Comparison
- Conclusions
- Future Work

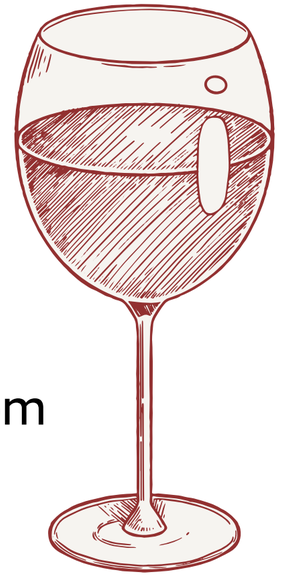
The Problem

- Graduate students don’t have much money and therefore need to choose good wines with limited financial resources.
- However, many grad students don’t know what a “good wine” is.
- Wine taste may be subjective, but there is growing interest in identifying objective qualities that contribute to wine quality or at least identifying qualities that experts believe make a good wine.

The Solution

Leverage the Vinho Verde Wine Quality Dataset from the UCI Machine Learning Repository to address the following questions:

1. What chemical properties are most indicative of a good* wine?
2. What chemical properties are predictive of a wine's color?
3. Determine whether different chemical properties are valued in different color wines.



* Within this dataset, wine quality scores are provided by expert assessors on a scale from 3 to 9, but for the purpose of model comparison, it was dichotomized at ≥ 7.

Covariates

Acids

- Fixed Acids remain in a wine when it's boiled
- Volatile Acids evaporate
- Citric Acid is a fixed acid, traditionally the acid of choice to boost wine acidity

Sugars

- Residual sugar is the sugar remaining in wine after fermentation, directly affects sweetness

Salts

- Chlorides are indicative of sodium chloride content in wines
- Sulphates are salts or esters of sulfuric acid, naturally occurring

Stabilizers

- Free SO2: SO2 still available to protect wine, ensures freshness and stability of wine
- Total SO2: Includes already reacted SO2, affects taste and relevant for legal limits

Alcohol

- Density: A measure of fermentation, wines grow less dense with fermentation
- Alcohol: Another measure of fermentation, directly measures the amount of alcohol present

Acidity

- pH: Affects taste, color, and stability. Lower pH wines are "sharper" and higher pH wines have softer notes.

Data Description (n=6497)

Variable name	Role	Type	Mean (SD)	Median (min, max)	Missing values
fixed acidity	Feature	Continuous	7.22 (1.30)	7.00 (3.80, 15.9)	no
volatile acidity	Feature	Continuous	0.34 (0.17)	0.29 (0.08, 1.58)	no
citric acid	Feature	Continuous	0.32 (0.15)	0.31 (0, 1.66)	no
residual sugar	Feature	Continuous	5.44 (4.76)	3 (0.6, 65.8)	no
chlorides	Feature	Continuous	0.06 (0.04)	0.05 (0.01, 0.61)	no
free sulfur dioxide	Feature	Continuous	30.5 (17.7)	29 (1, 289)	no
total sulfur dioxide	Feature	Continuous	116 (56.5)	118 (6, 440)	no
density	Feature	Continuous	1.00 (0.003)	1.00 (0.99, 1.04)	no
pH	Feature	Continuous	3.22 (0.16)	3.21 (2.72, 4.01)	no
sulphates	Feature	Continuous	0.53 (0.15)	0.51 (0.22, 2)	no
alcohol	Feature	Continuous	10.5 (1.19)	10.3 (8, 14.9)	no
quality	Outcome	Continuous	5.82 (0.87)	6 (3, 9)	no
color	Outcome	Binary (red/white)	Red 1,599 (24.6%) White 4,898 (75.4%)		no
binary quality	Outcome	Binary (≥7/<7)	≥7: 1,277 (19.7%) <7: 5,220 (80.3%)		no

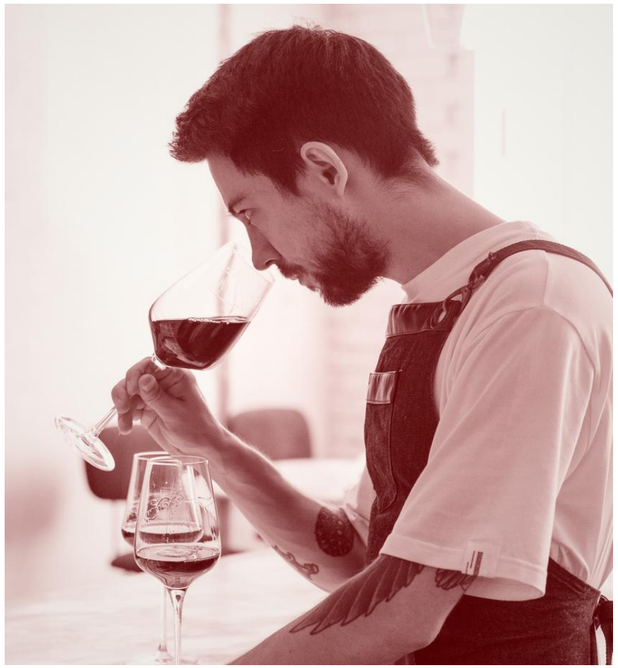
Note: n (%) was used for binary variables.

Data Exploration

Start off by splitting 80% of the data into a train set and 20% into a test set. Then, use the training set to build models with cross validation of k = 10 where possible, and then utilize the test set to ascertain model performance.

Table: Outcome Summary in Training and Test Subsets

	Wine Color		Wine Quality		Wine Quality
	Red	White	≥7	<7	
Training set	1,279 (24.6%)	3,918 (75.4%)	1,018 (19.6%)	4,179 (80.4%)	5.82 (0.88)
Test set	320 (24.6%)	980 (75.4%)	259 (19.9%)	1,041 (80.1%)	5.83 (0.87)



01

Introduction

- The Problem
- The Solution
- Data Description

02

Methodology

- Logistic Regression
- Bayesian MCMC
- Newton Raphson
- SVM & MLP

03

Conclusion

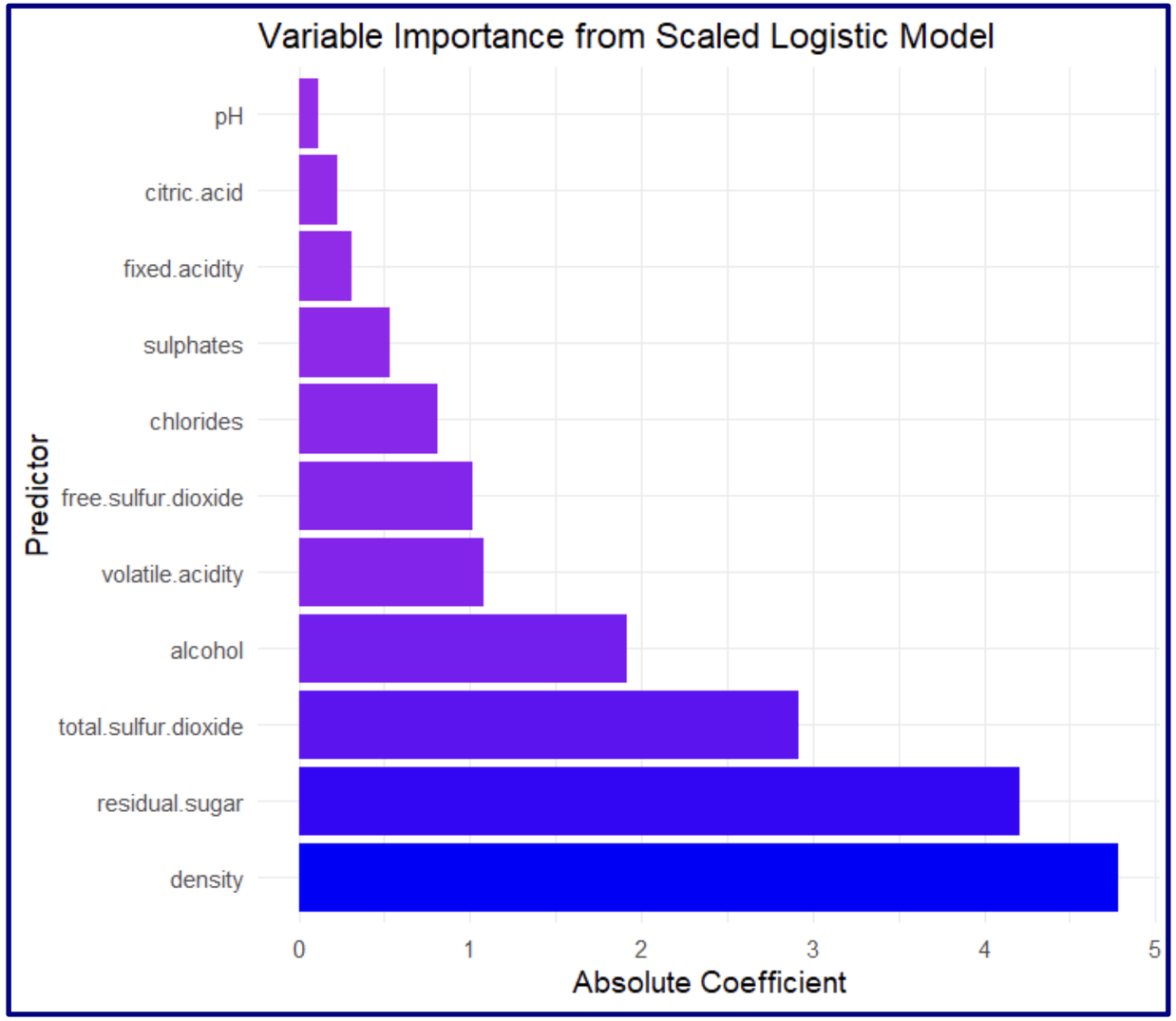
- Comparison
- Conclusions
- Future Work

Logistic regression

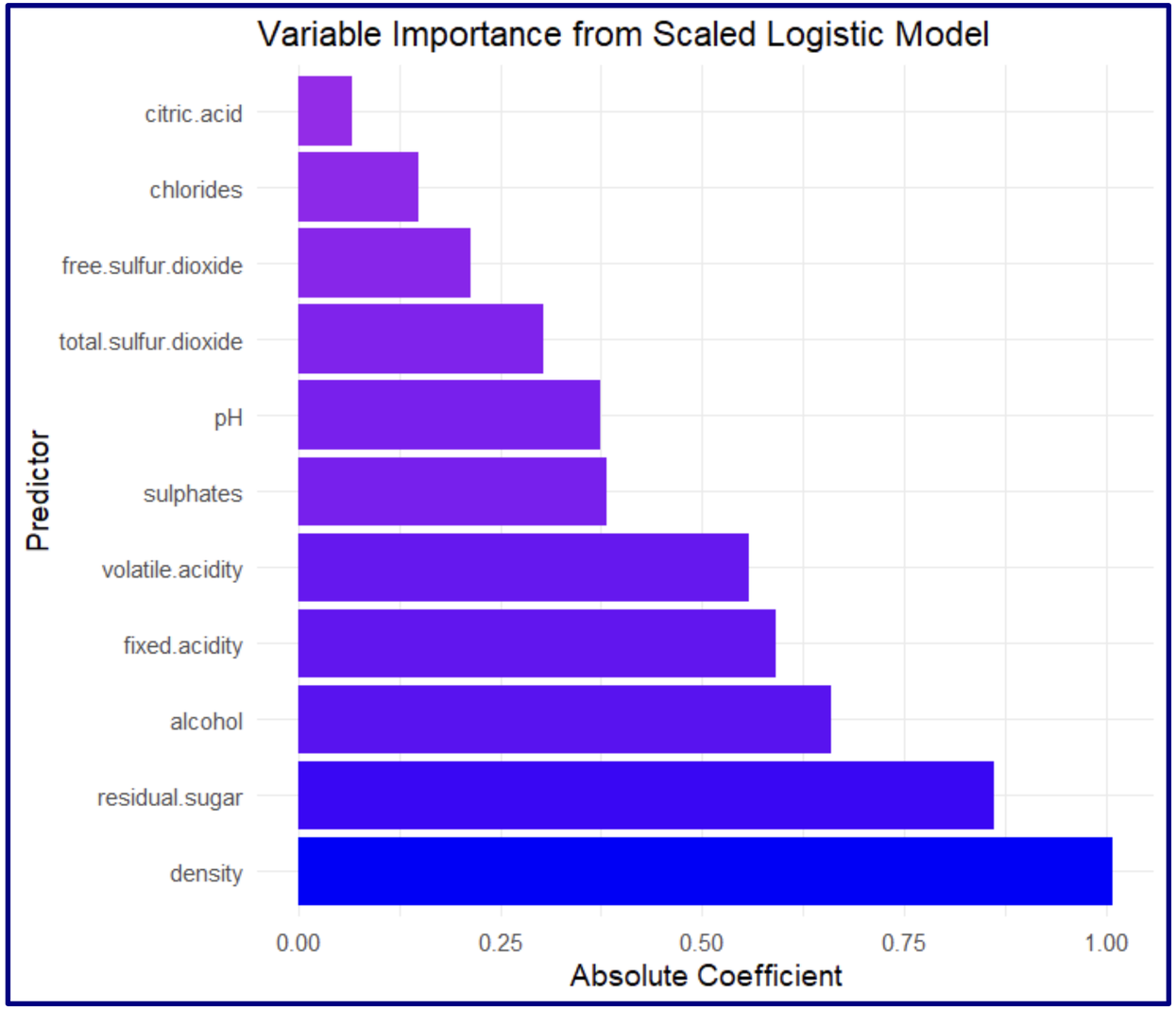
Start off simple – fit a simple logistic regression model with all the covariates!

Outcome: Wine Color

Prediction	Reference red	Reference white
red	317	1
white	3	979

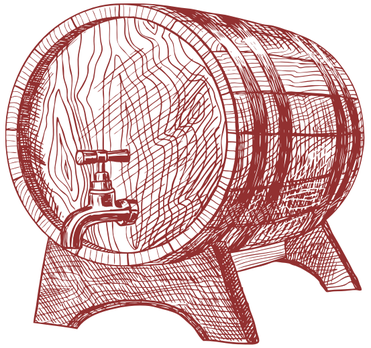


Outcome: Wine Quality



Prediction	Reference < 7	Reference ≥ 7
< 7	988	196
≥ 7	53	63

Bayesian Approach?



Well-studied Dataset

Want to take advantage of wealth of previous studies

Posterior Predictive Distributions

Allow for probabilistic statements about results (naturally fascinating as biostats students)

Rationale

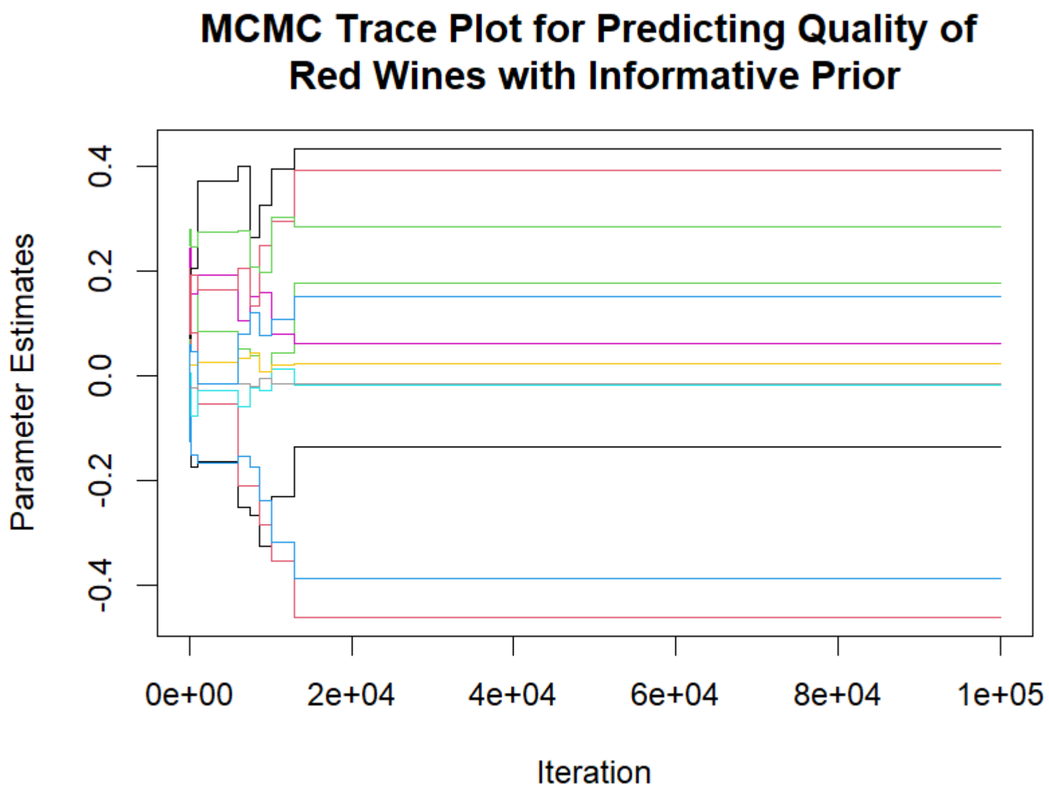
Naturally, start off with choice of prior:

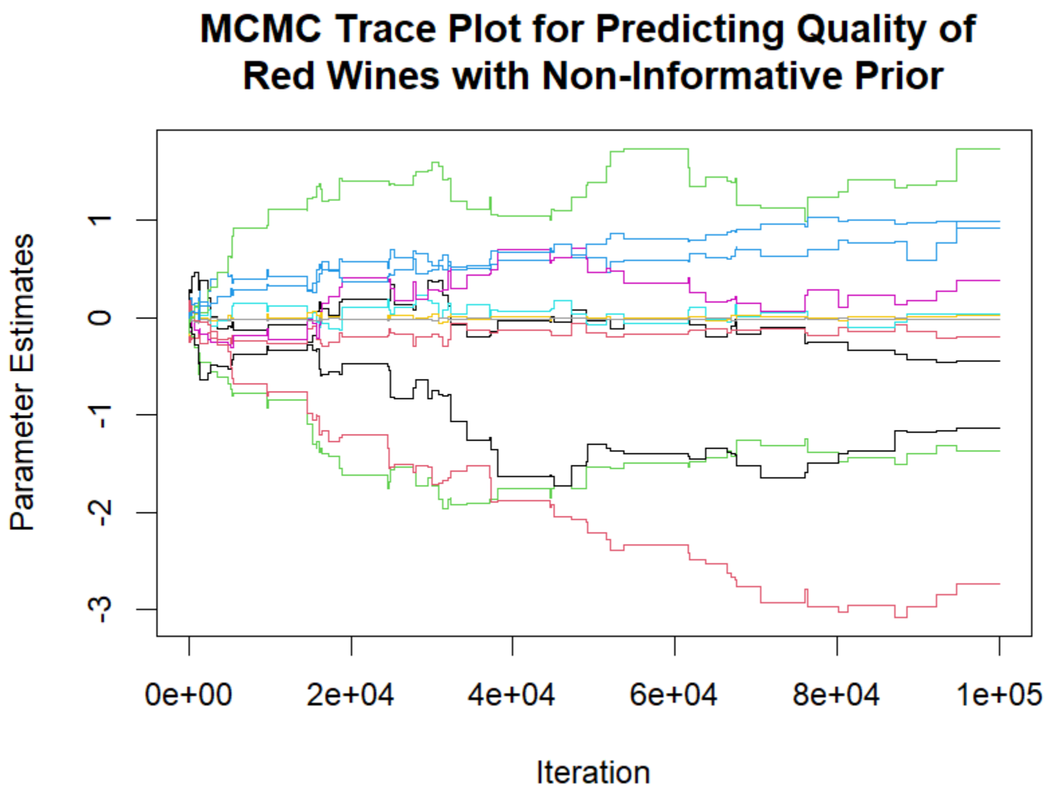
- Agyemang et al. fit several models for logistic regression and provided parameter and variance estimates → Have lots of data, so natural decision to use Normal prior

Problems?

- Different prior distributions?
- Pool different estimates?
- Power priors?
- Is this prior even good?

Trace Plots for Predicting Quality of Red Wines





Binomial Regression
with Newton Raphson

Goals

Want to fit binomial regression

Quality ranges from 3 to 9, subtract 3 such that our outcomes follows:
Y = (0, … , 6)

Link Function

Picking the Proper Link

A natural choice is the canonical link, for the binomial distribution, this is the logit

Binomial Distribution w/ Logit Link

$$\text{logit}(p_i) = \beta_0 + x_{1,i}\beta_1 + \cdots + x_{11,i}\beta_{11}$$

Log Likelihood Under the Canonical Link

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \mathbf{x}_i^\top \boldsymbol{\beta} - 6\log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))$$

Fitting via Newton Raphson

1. Fit an initial linear model with transformed response (closed form solution)

$$\log \frac{\tilde{Y}_i}{6 - \tilde{Y}_i} = \beta_0 + x_{1,i}\beta_1 + \cdots + x_{11,i}\beta_{11} + \varepsilon_i$$

2. First and second derivatives are as follows:

$$\partial_\beta \ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(Y_i - \frac{n \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right) \mathbf{x}_i^\top = \sum_{i=1}^{m} (Y_i - n p_i)\, \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{S},$$

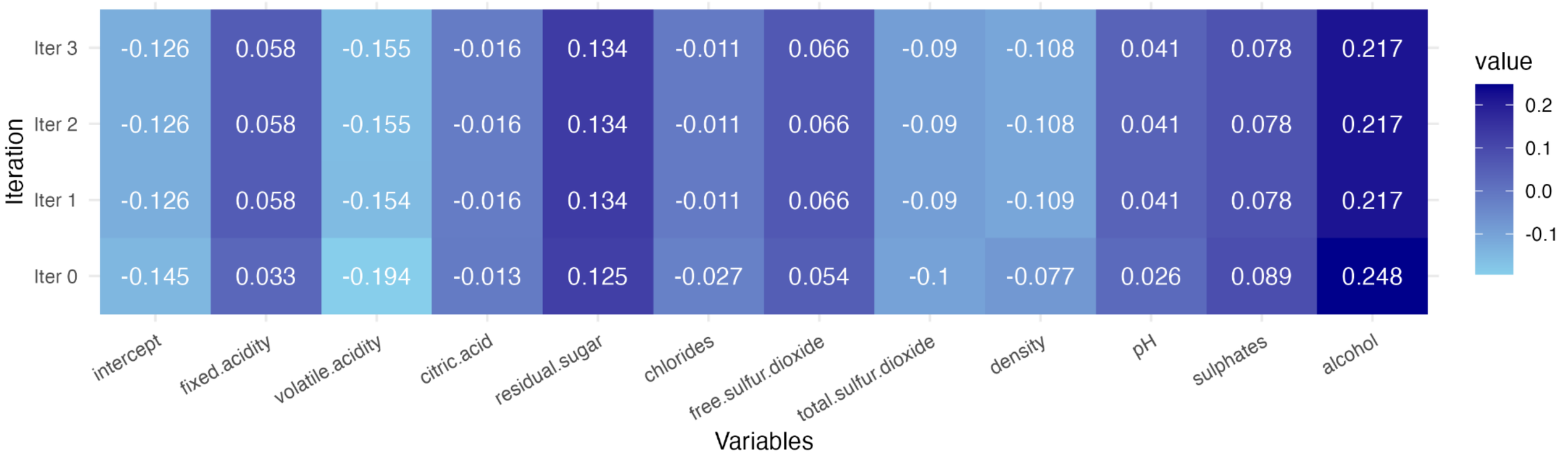
$$\partial_\beta^2 \ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} -n \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))^2} \mathbf{x}_i \mathbf{x}_i^\top = -\mathbf{X}^\top \mathbf{V} \mathbf{X},$$

3. NR update with tolerance 1e-10 then follows:

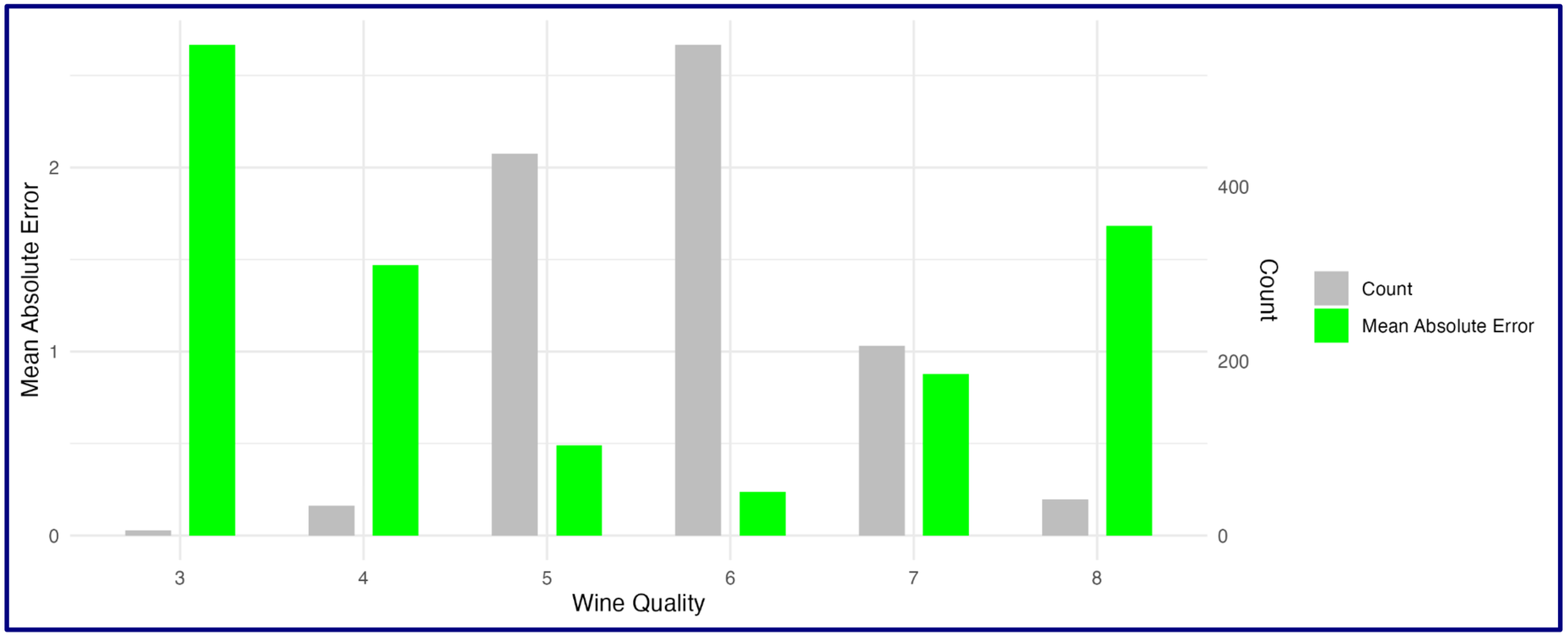
$$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} + (\mathbf{X}^\top \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{S} \big|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}}$$

The Fitted Model

Newton Raphson converged in 3 iterations, with iteration 0 being the initial estimate from the specified linear model.



Performance on the Test Set



- Imbalanced wine quality scores
- More accurate for common wine qualities

- Overall MAE: 0.518

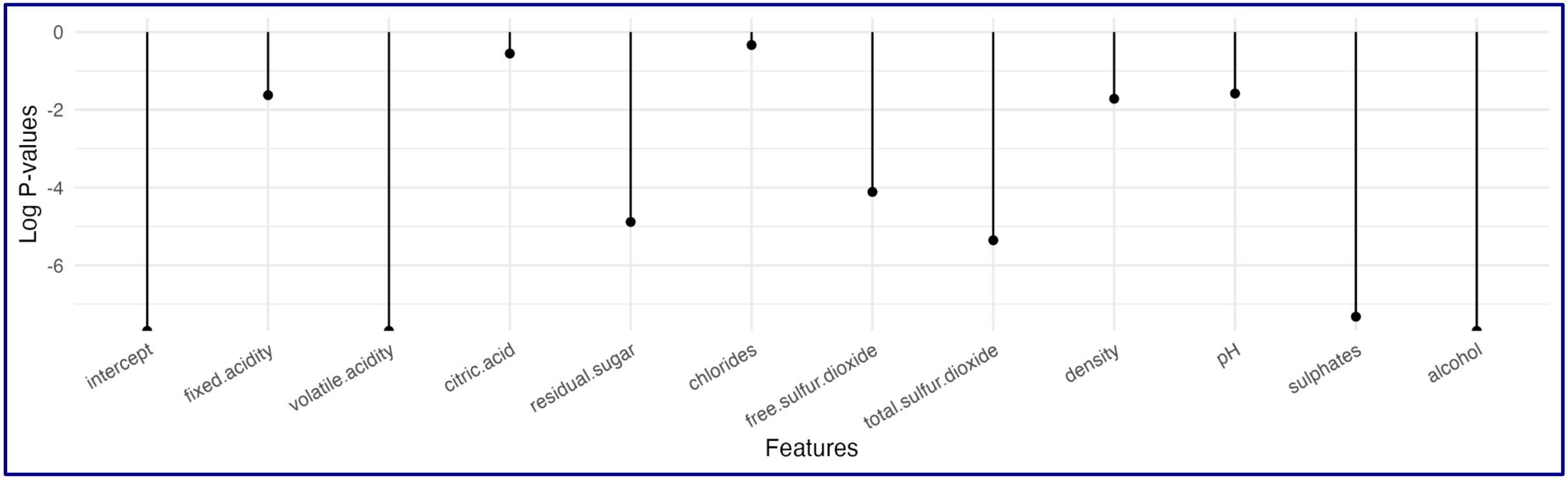
Testing Significant Features

Perform the Wald test for each feature:

- H_0: the ith feature can be removed.
- H_a: the ith feature cannot be removed.

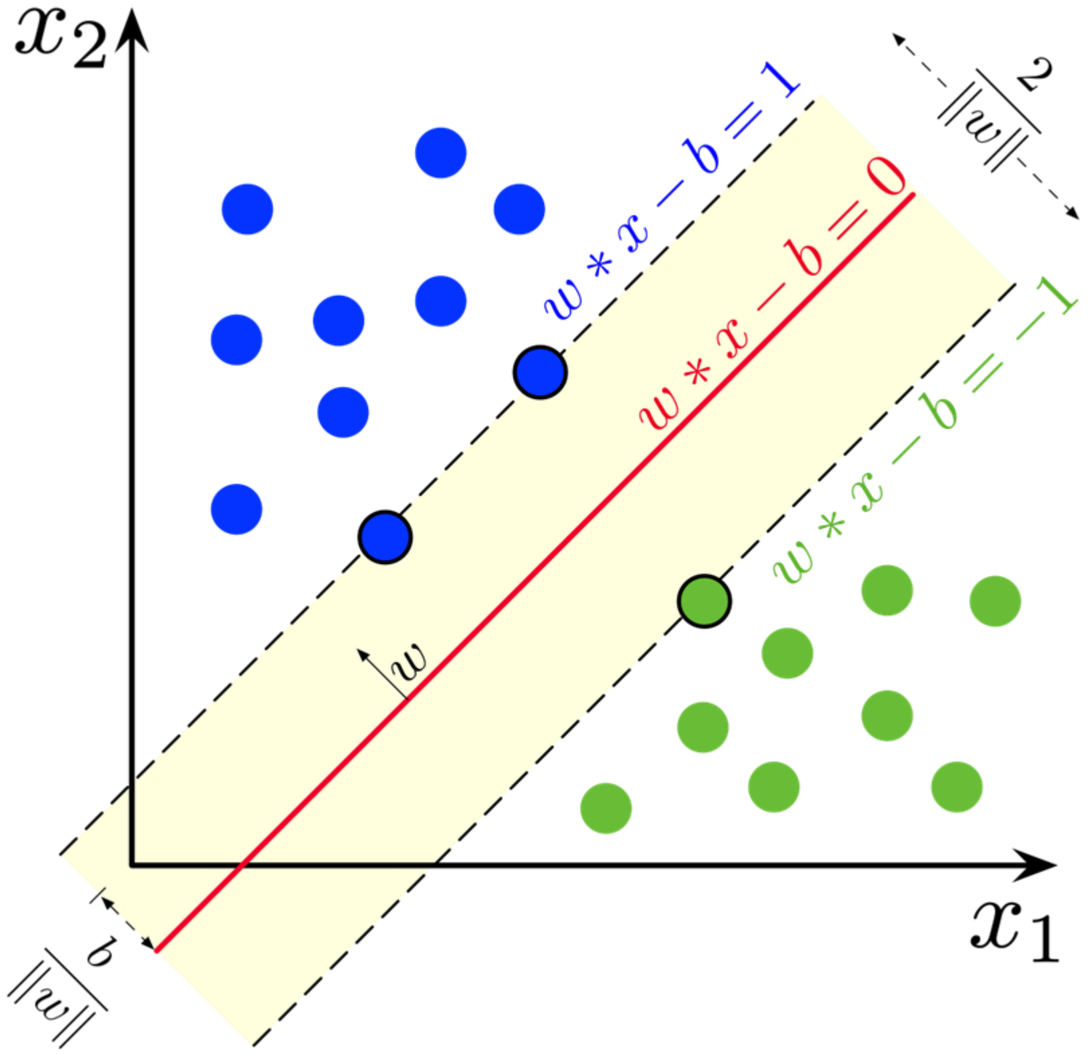
$$R(\hat{\boldsymbol{\beta}})^\top (H(\hat{\boldsymbol{\beta}}) I_n^{-1}(\hat{\boldsymbol{\beta}}) H(\hat{\boldsymbol{\beta}})^\top)^{-1} R(\hat{\boldsymbol{\beta}}) \to_d \chi_r^2$$

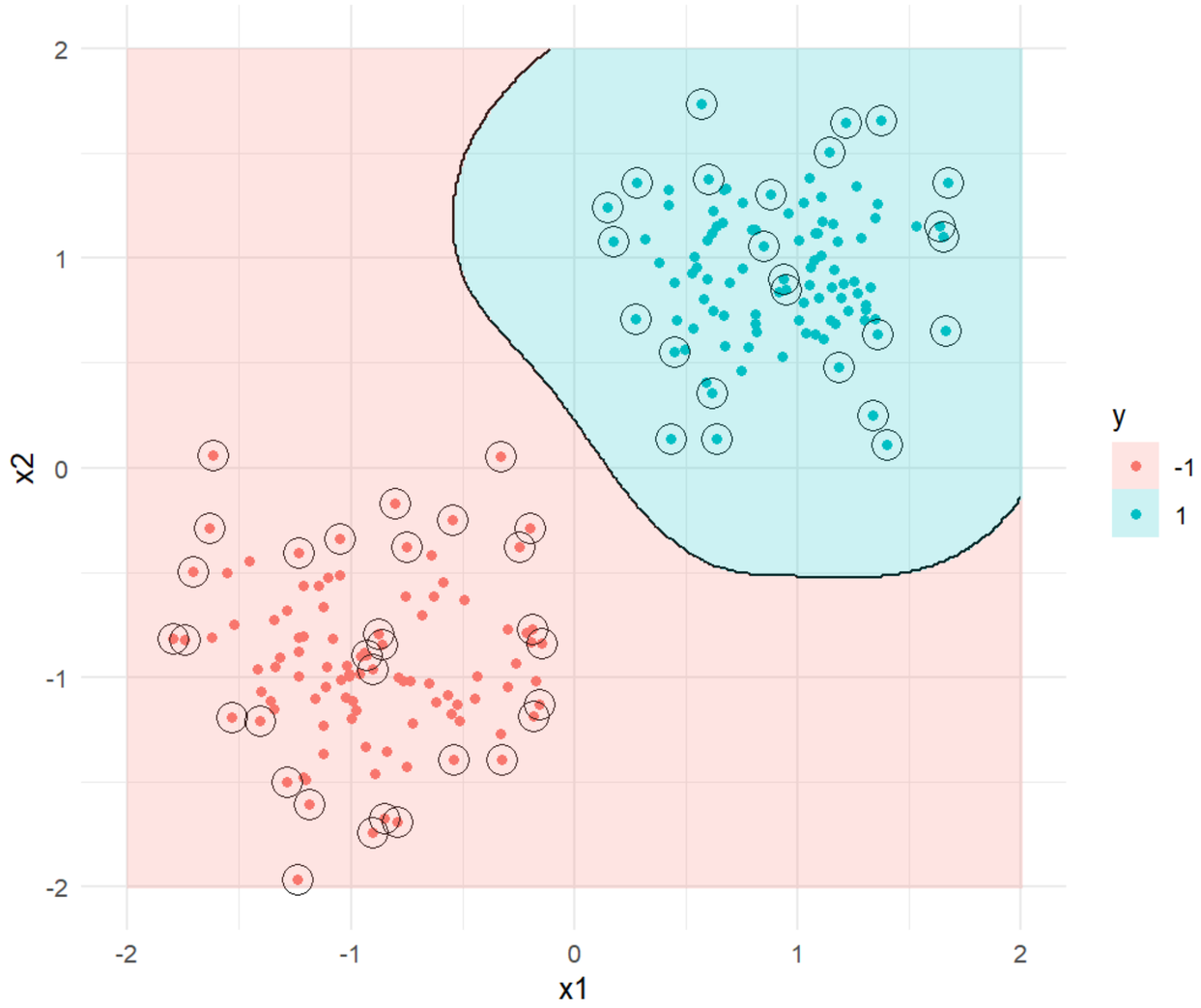
$$\frac{(\hat{\beta}_i)^2}{[(\mathbf{X}^\top \mathbf{V} \mathbf{X}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}})^{-1}]_{ii}} \to_d \chi_1^2$$



Support Vector Machines

Linear and Radial Kernels

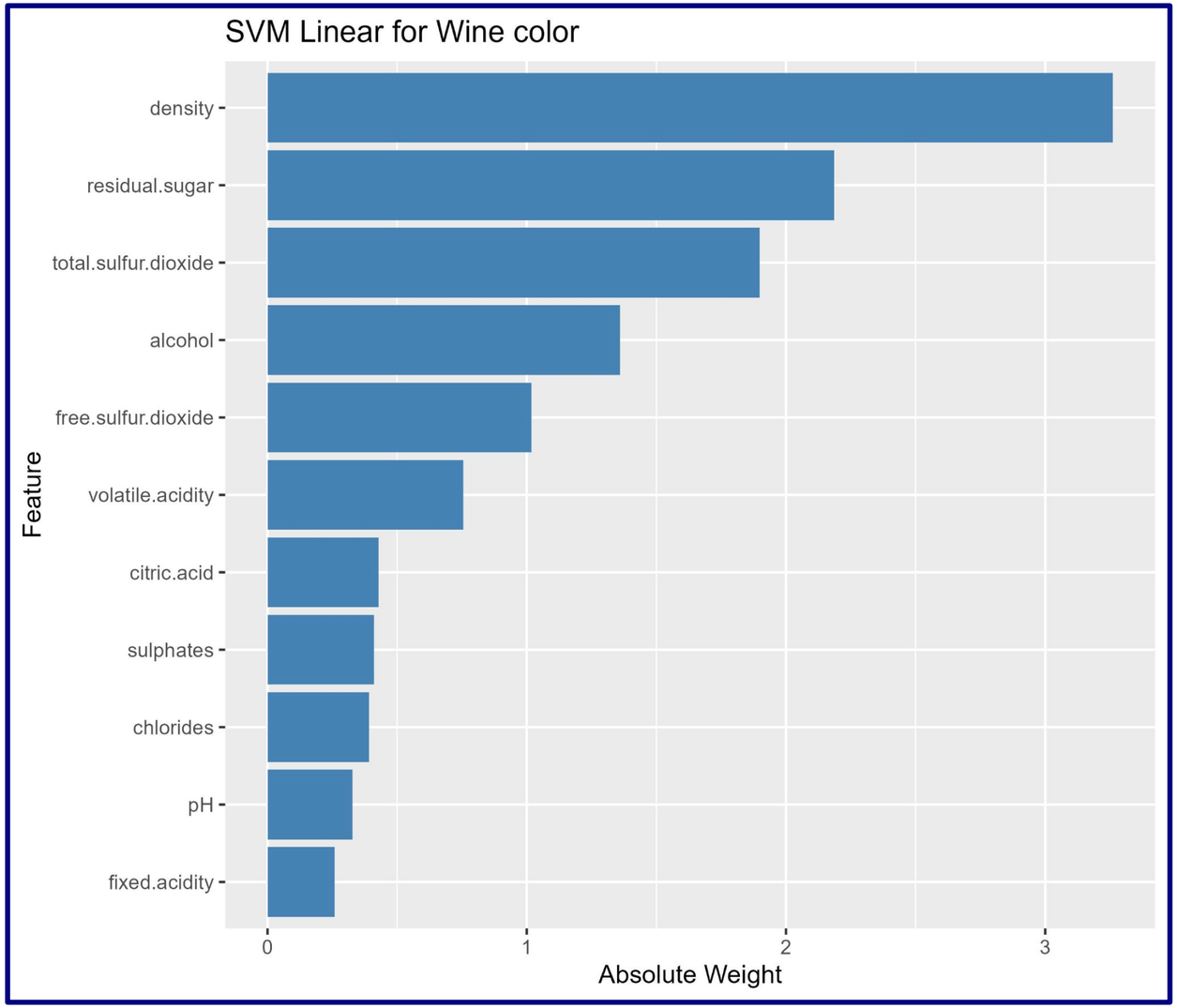




Outcome: Wine Color, Linear Kernel

Table: Confusion Matrix

Predicted	True Observed	
	Red	White
Red	318	0
White	2	980



Outcome: Wine Quality, Linear Kernel

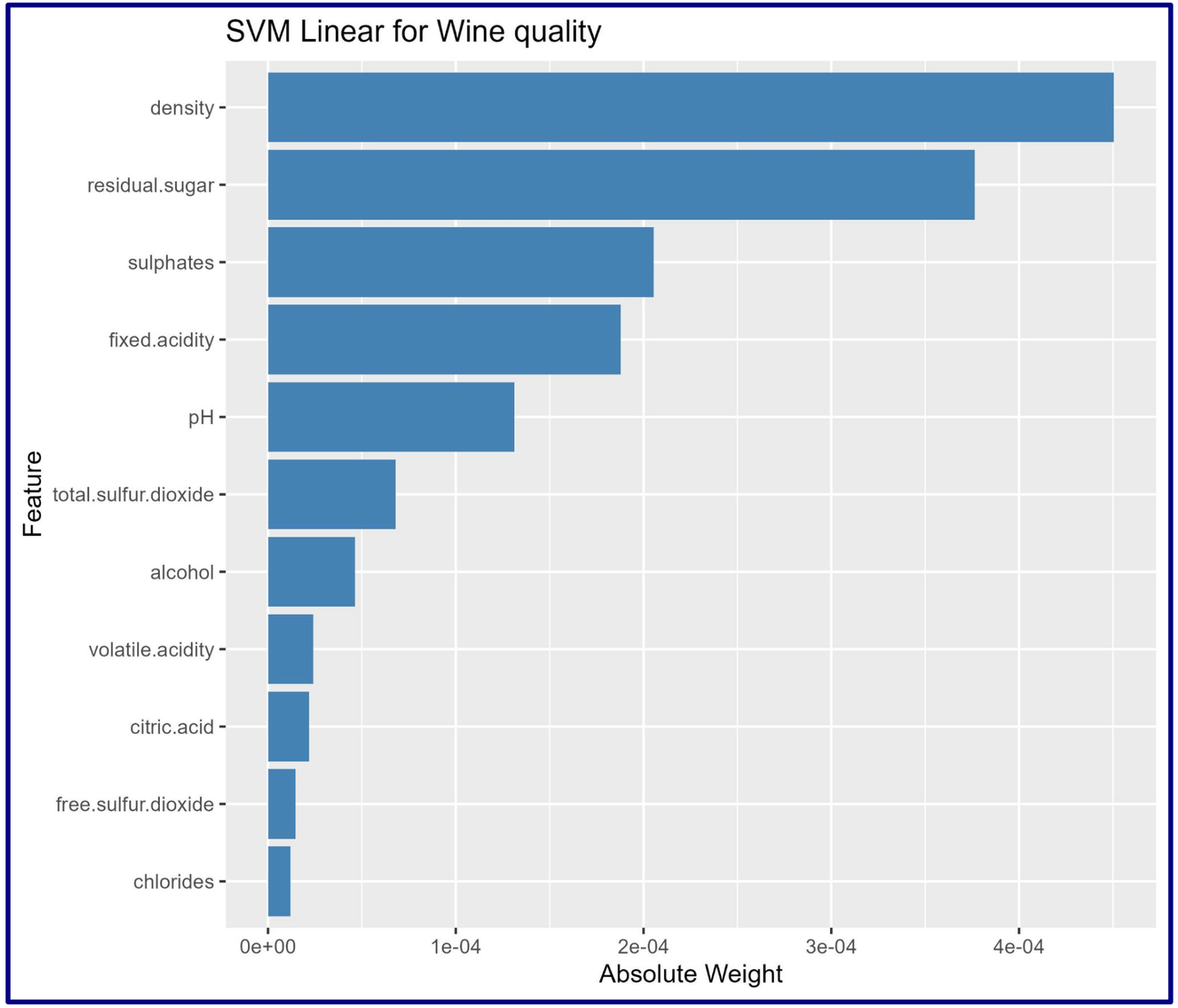


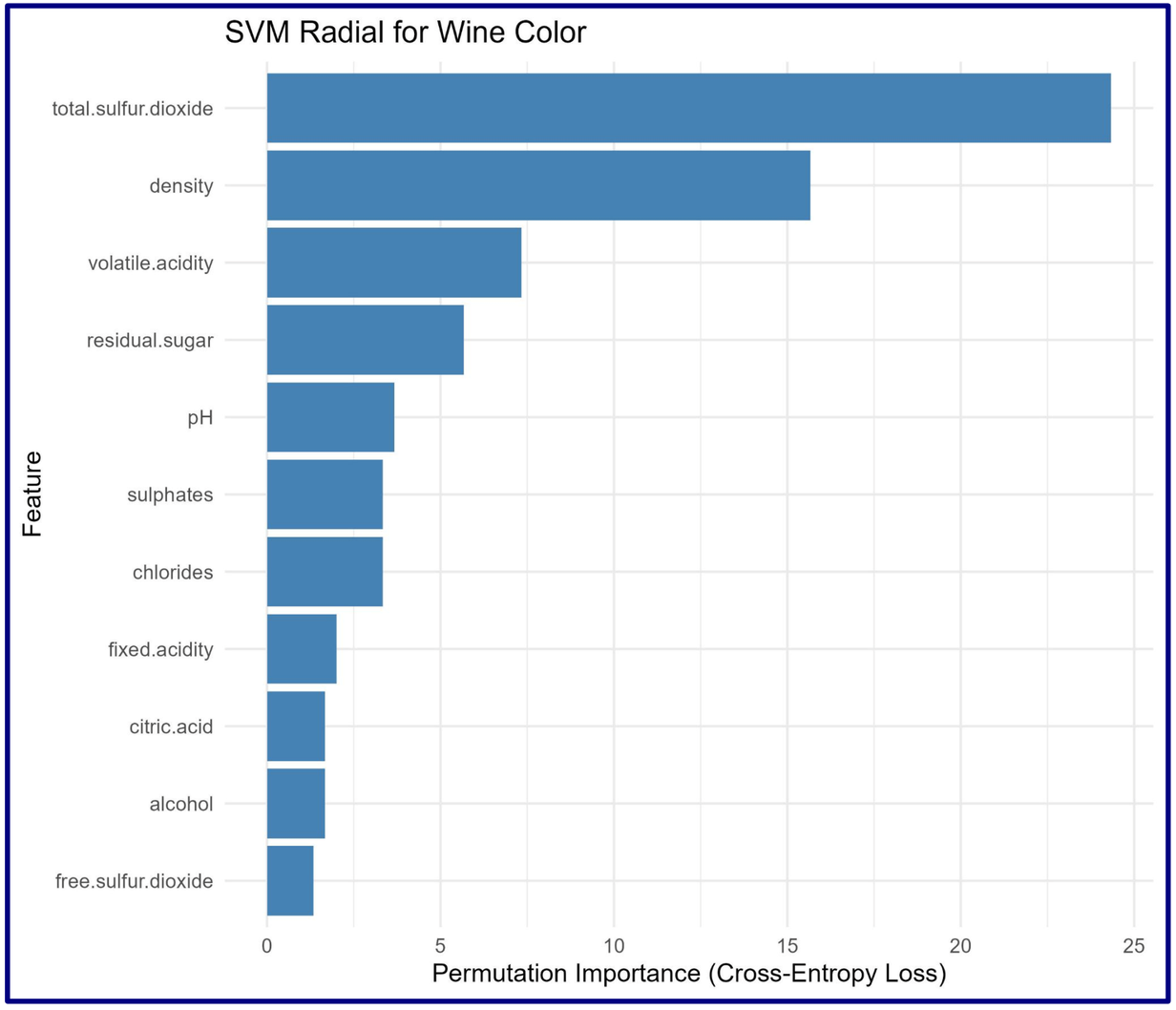
Table: Confusion Matrix

Predicted	True Observed	
	< 7	≥ 7
< 7	1041	259
≥ 7	0	0

Outcome: Wine Color, Radial Kernel

Table: Confusion Matrix

Predicted	True Observed	
	Red	White
Red	318	0
White	2	980



Outcome: Wine Quality, Radial Kernel

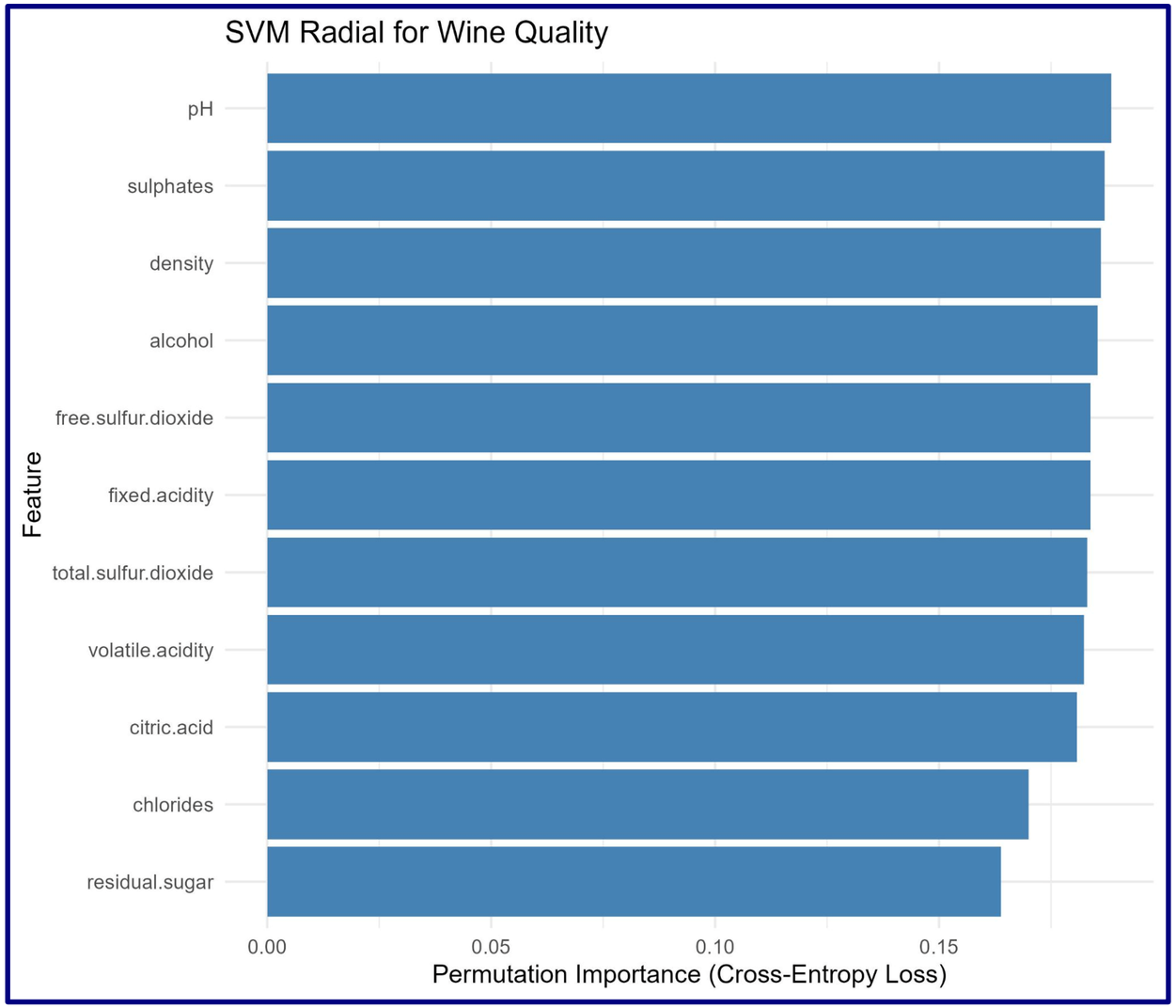


Table: Confusion Matrix – Radial

Predicted	True Observed	
	< 7	≥ 7
< 7	1008	173
≥ 7	33	86

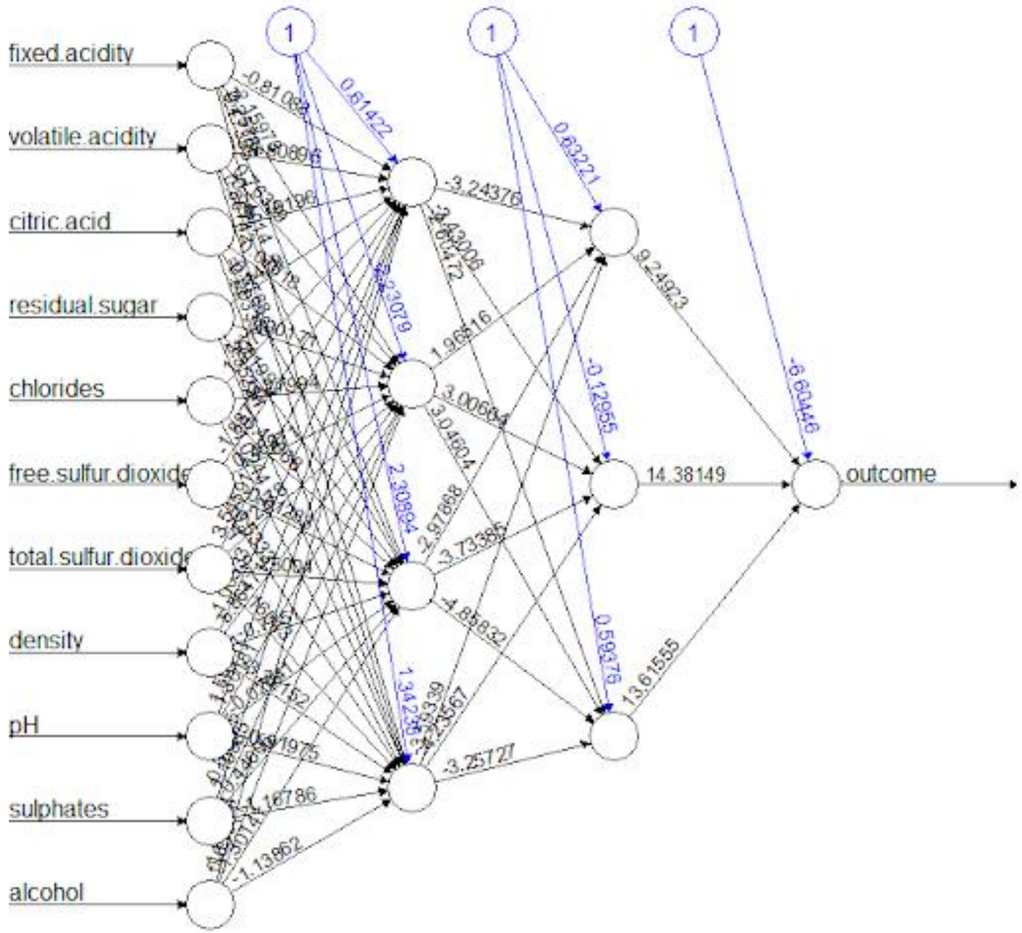
Multi-layer Perceptron (MLP)

Wine Color

1. Nine network architectures with two hidden layers:
    - Layer 1 ∈ {5, 4, 3}
    - Layer 2 ∈ {3, 2, 1}

   ⇒ Then, employ logistic activation function.
2. Evaluate each with 5-fold CV optimizing ROC
    - Model with the hidden unit (4, 3) achieved the highest cross validation performance

	Reference	
Prediction	red	white
red	316	0
white	4	979

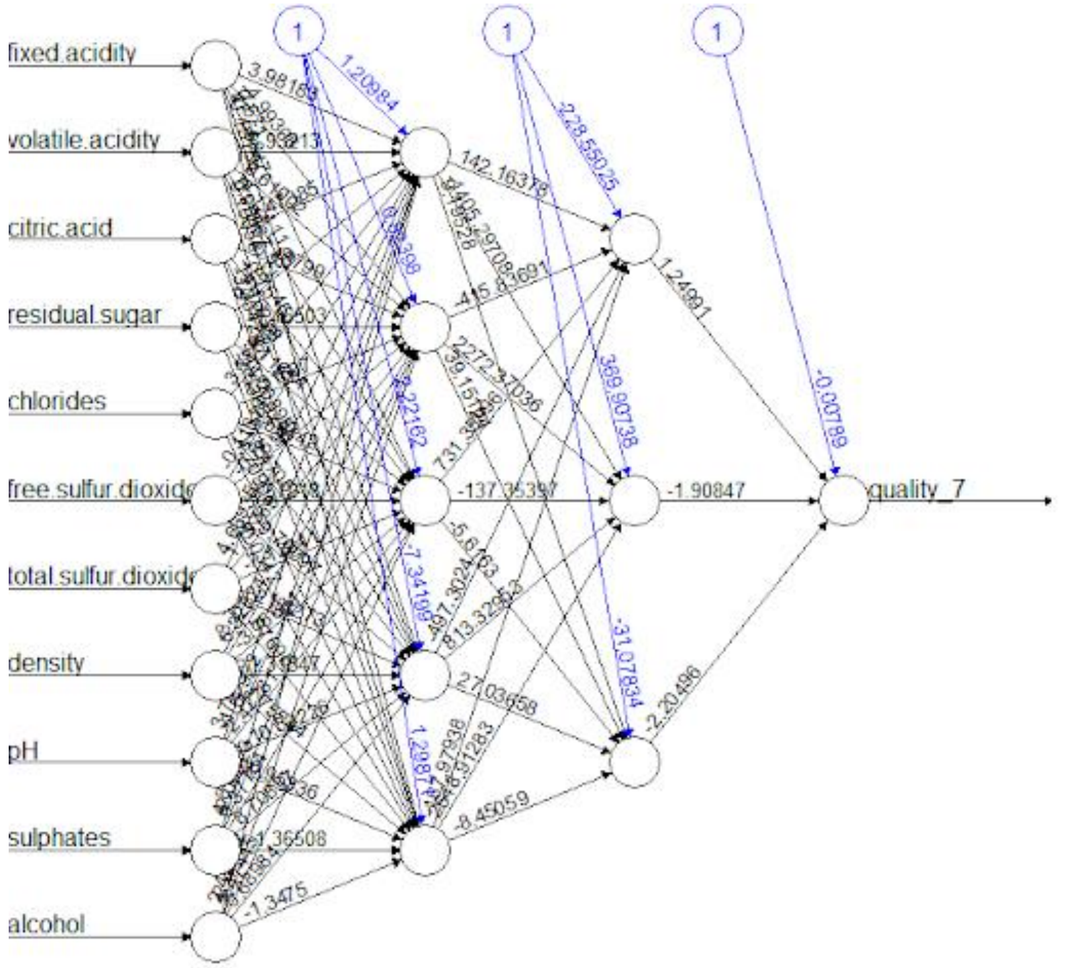


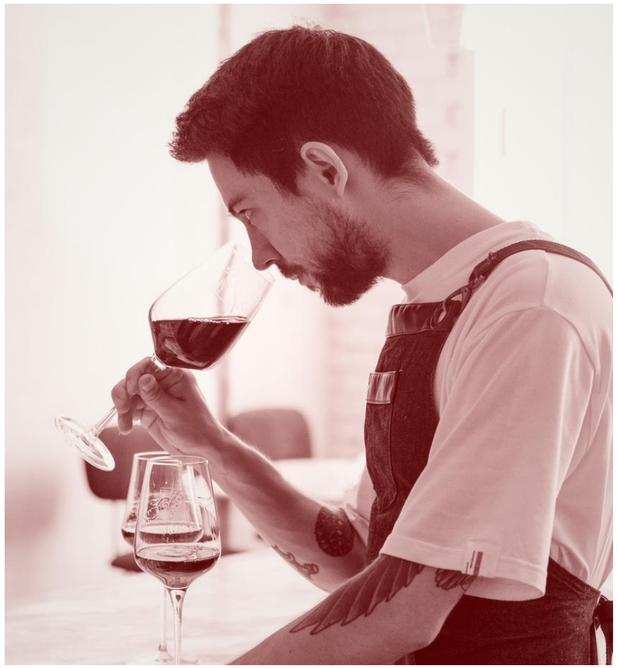
Wine Quality

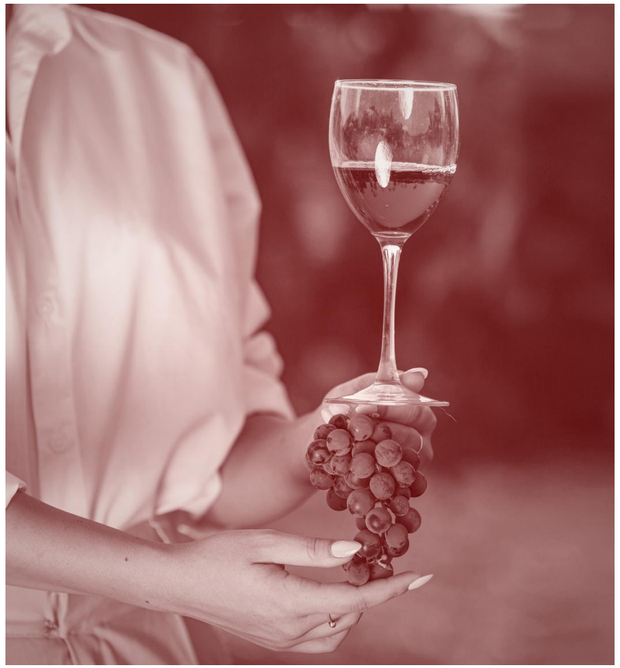
1. Nine network architectures with two hidden layers:
    - Layer 1 ∈ {5, 4, 3}
    - Layer 2 ∈ {3, 2, 1}

   ⇒ Then, employ logistic activation function.
2. Evaluate each with 5-fold CV optimizing ROC
    - Model with the hidden unit (5, 3) achieved the highest cross validation performance

Prediction	Reference < 7	Reference ≥ 7
< 7	982	169
≥ 7	59	60







01

Introduction

- The Problem
- The Solution
- Data Description

02

Methodology

- Logistic Regression
- Bayesian MCMC
- Newton Raphson
- SVM & MLP

03

Conclusion

- Comparison
- Conclusions
- Future Work

Comparison of Methods

Method	Wine Quality			Wine Color		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Logistic Regression	0.8085	0.5431	0.2432	0.9969	0.9969	0.9906
Bayesian Logistic Regression	Inf: 0.7723 No: 0.8023	Inf: 0.8061 No: 0.8096	Inf: 0.9424 No: 0.9846	0.9577	0.9623	0.9816
SVM	L: 0.8008 R: 0.8415	L: NA R: 0.7227	L: 0 R: 0.3320	L: 0.9985 R: 0.9985	L: 1 R: 1	L: 0.9938 R: 0.9938
MLP	0.8246	0.6040	0.3475	0.9969	1	0.9875

Conclusion

Color: All models demonstrated strong predictive ability.

Best: Logistic regression was selected for its simplicity and interpretability.

Quality: Overall, predicting quality proved more challenging.

Best: Bayesian Logistic Regression showed notable improvement compared to other models

Conclusion

Main Research Aims: What characteristics might good quality wine have? What characteristics might red wines have that whites do not? Do these characteristics differ across color?

Table: Important Features for All Models

Models	Outcomes	Top 3 Features
Logistic Regression, SVM with Linear Kernel, Multilayer Perceptron	Color	Density, Residual Sugar, Total Sulfur Dioxide
Bayesian Logistic Regression with Non-informative Prior	Quality (Binary) for Red	pH, Volatile Acidity, Sulphates
	Quality (Binary) for white	pH, Density, Critic Acid
Binomial Regression	Quality (Ordinal)	Alcohol, Volatile Acidity, Residual Sugar

Future Work

Look For Any Possibly Missing Covariates

While the given covariates are expansive, they surely do not cover everything a sommelier might find important in a wine.

Increased Prior Selection for Bayesian Log-Reg

The utilized prior was not the best, we need to consider ways to incorporate information from advanced ML methods without easy parameter estimates.

Differences in Importance Across Colors

So far, we only established a rudimentary grasp of how the same models differ across color.

Drink Lots of Wine

We have not found the time as a group to go out and drink lots of wine and pretend we're wine experts.

Questions?

CREDITS

Slides Carnival

UCI Machine Learning
Data Repository

Dr. Naim Rashid