Modeling to Predict Wine Quality

Final Project

YoungRi Lee & Jihyun Lee

Last compiled on 2020-11-21

1 Introduction

1.1 Data

This wine quality dataset is from [1]. The dataset includes *vinho verde*, a unique product from the Minho (northwest) region of Portugal. The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of vinho verde. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis.

The outcome variable is `wine quality`, which was measured by a minimum score of three sensory assessors using blind tastes in a scale that ranges from 0 (very bad) to 10 (excellent). There are 11 attributes of the wine based on physicochemical tests: fixed acidity (g(tartaric acid)/dm^3), volatile acidity (g(acetic acid)/dm^3), citric acid (g/dm^3), residual sugar (g/dm^3), chlorides (g(sodium chloride)/dm^3), free sulfur dioxide (mg/dm^3), total sulfur dioxide (mg/dm^3), density (g/dm^3), pH, sulphates (g(potassium sulphate)/dm^3), and alcohol (vol.%). Originally, two datasets were created separately, one for red wine ($n = 1599$) and another for white wine ($n = 4988$). In this report, we use a merged dataset and create a dummy variable to indicate the wine type, `red`. Thus, in total, the dataset includes 11 numerical attributes (covariates), one dummy variable, and one numerical outcome. There is no missing value in this dataset.

Table 1 shows the descriptive statistics of 11 attributes by wine type. [ADD SOME DESCRIPTION]

Figure 1 shows the distribution of wine quality by wine type. Generally, it shows a normal shape distribution and centered around the middle point of the scale. Red wine has fewer observations than

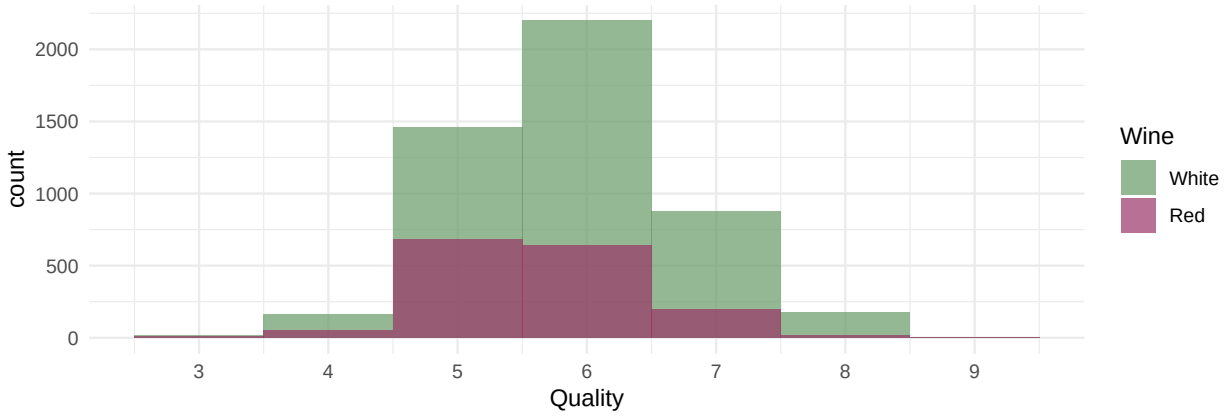
Table 1: Descriptive statistics of 11 attributes

	White			Red		
	Min	Max	Mean	Min	Max	Mean
Fixed acidity	3.8	14.2	6.855	4.6	15.9	8.320
Volatile acidity	0.08	1.10	0.278	0.12	1.58	0.528
Citric acid	0	1.66	0.334	0	1.00	0.271
Residual sugar	0.6	65.8	6.391	0.9	15.5	2.539
Chlorides	0.009	0.346	0.046	0.012	0.611	0.087
Free sulfur dioxide	2	289	35.308	1	72	15.875
Total sulfur dioxide	9	440	138.361	6	289	46.468
Density	0.987	1.039	0.994	0.990	1.004	0.997
pH	2.72	3.82	3.188	2.74	4.01	3.311
Sulphates	0.22	1.08	0.490	0.33	2.00	0.658
Alcohol	8.0	14.2	10.514	8.4	14.9	10.423

Note:
The scale of each attribute is given in the text.

white wine.

Figure 1. Histrogram of wine quality by wine type



1.2 Research Questions

Using this dataset, we would like to (1) build the best model to predict wine quality, (2) examine the most influential set of attributes to predict wine quality.

To achieve this goal, we implemented two classes of techniques: shrinkage approaches (i.e., Ridge and Lasso regressions) and tree-based approaches (i.e., regression tree). This report includes the comparisons of two approaches, model selection procedures (e.g., cross-validation), and proposal of the best model.

2 Methods

We implemented shrinkage approach (i.e., Ridge and Lasso regressions) and tree-based approach (i.e., Regression tree) to build a prediction model.

2.1 Shrinkage approaches

The shrinkage method uses all possible covariates while shrinking the covariates' coefficients towards zero that does not associate with the outcome as strong as other predictors. The advantage of this method is to reduce the variance by reducing the relatively unimportant attributes and selecting variables. Thus, important covariates in the model can be emphasized. However, it may lead to the bias of the model by reducing the variance (*bias-variance trade-off*). Generally preferred shrinkage methods are *ridge regression* and *lasso*.

2.1.1 Ridge regression

The purpose of ridge regression is to reduce variance of the predictions by minimizing the residual sum of squares (RSS) of the model as well as the shrinkage penalty:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$ and the second term indicates the shrinkage penalty with a *tuning parameter* λ. As the coefficients of the attributes become close to zero, the model fit of ridge regression will be desirable. The value of λ controls the impact of shrinkage penalty. As the value of λ increases, the impact of the shrinkage penalty grows and the ridge regression coefficients will be close to zero. Generally, the value of λ is selected using cross-validation to find the optimal value. Ridge regression has a drawback in using all variables in the data and cannot select or subset the important variable, because the coefficients *shrinks toward* zero, not become exact zero. This can lead a challenge in interpretation of the model with large set of variables.

2.1.2 Lasso

Lasso is similar to Ridge regression as it shrinks the coefficients toward zero but has a more stringent shrinkage penalty that can force the coefficients to be equal to zero:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j^2|.$$

By pushing the coefficient to be zero, lasso performs variable selection identifying the variables that substantially impact the outcome. Thus, the model using lasso might be more parsimonious than that using ridge regression, making the model results more interpretable. Like ridge regression, the value of λ in lasso is critical and will be determined by the cross-validation.

2.2 Tree-based approaches

The tree-based approaches can be used for regression and classification. We will focus on regression trees because the outcome variable (*wine quality*) is a numeric in our data.

2.2.1 Regression trees

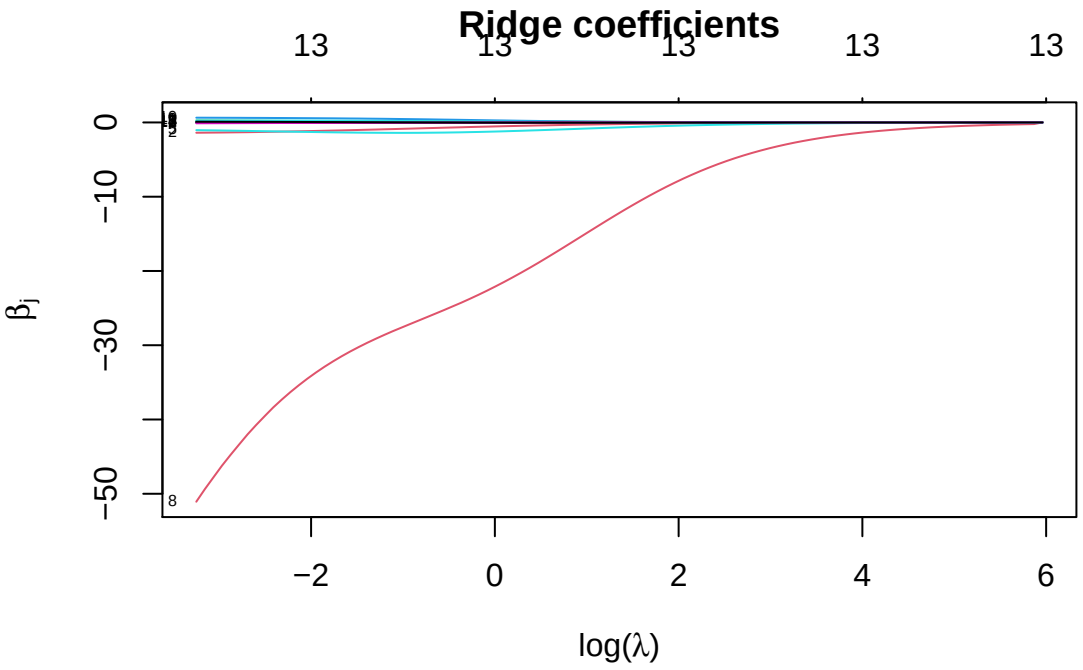
Tree-based approaches segment the predictor space into several simple regions. In a tree, each rule that split the segment will be summarized, called decision-tree methods. After splitting the predictor space into a number of regions (a leaf node; $R_1, R_2, ...R_J$) using the selected predictors X_j, the average of the outcome within each region ($\hat{y}_{R_j}$) will be calculated. The model fit of the regression tree aims to minimize the RSS:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

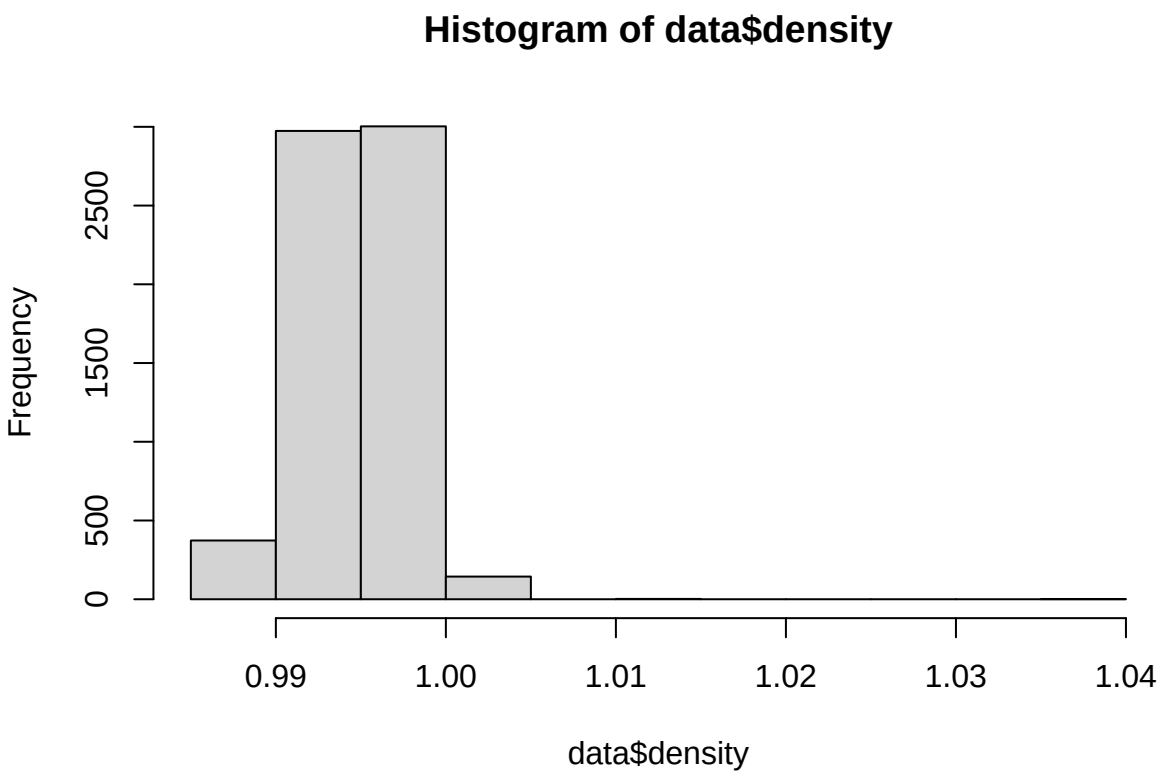
Thus, the regression trees split the predictor space and create the two new branches only if the new split decreases RSS. Since the regression trees use this top-down, greedy approach to select the model, the final model might not be the true optimal model. However, the regression trees have the advantage of taking into account the nonlinearity or interaction of covariates. Also, the regression tree is useful in interpreting the results.

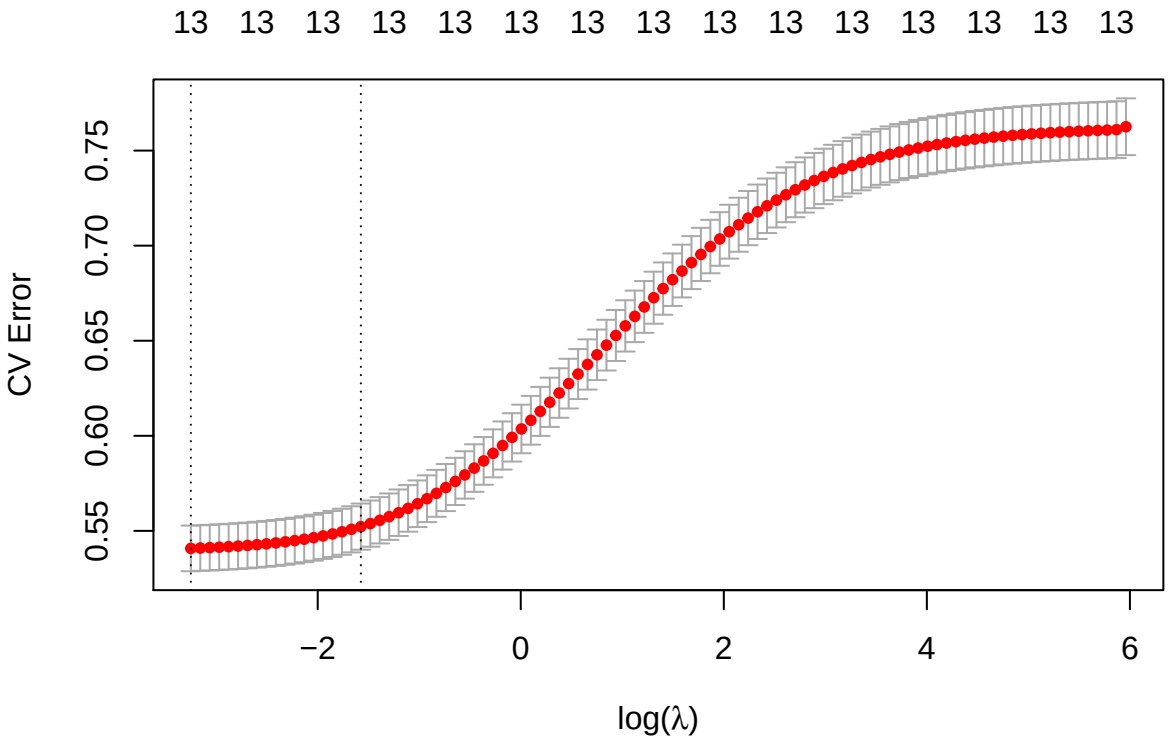
3 Results

3.1 Ridge regression

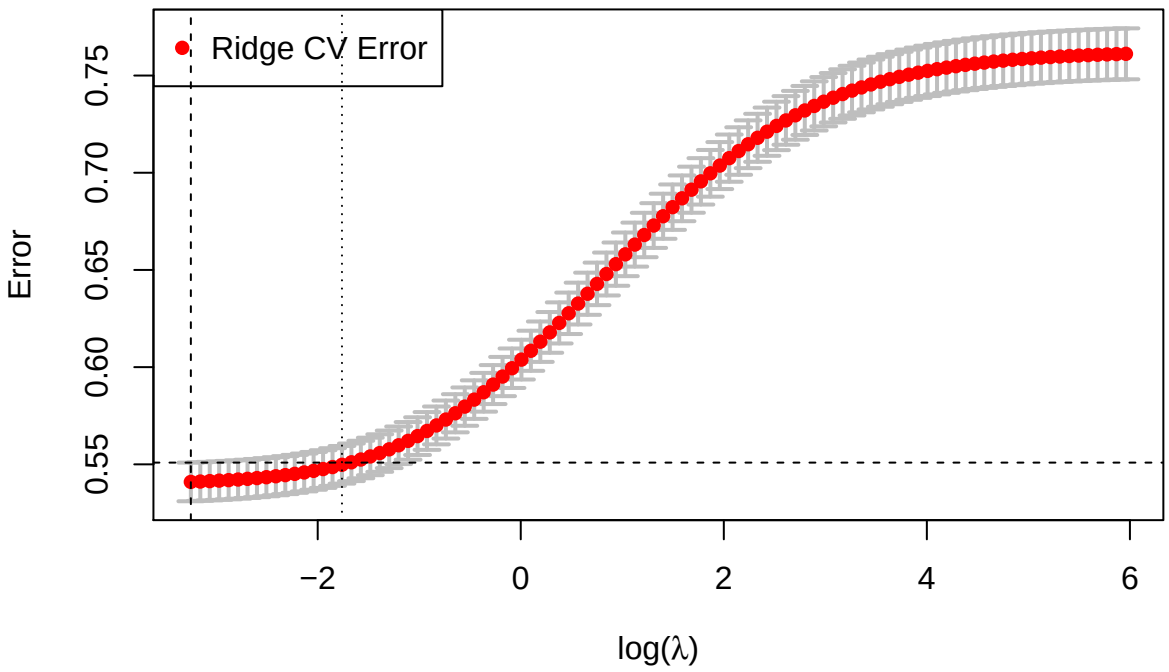


X_8 is `density`. (Humm...)

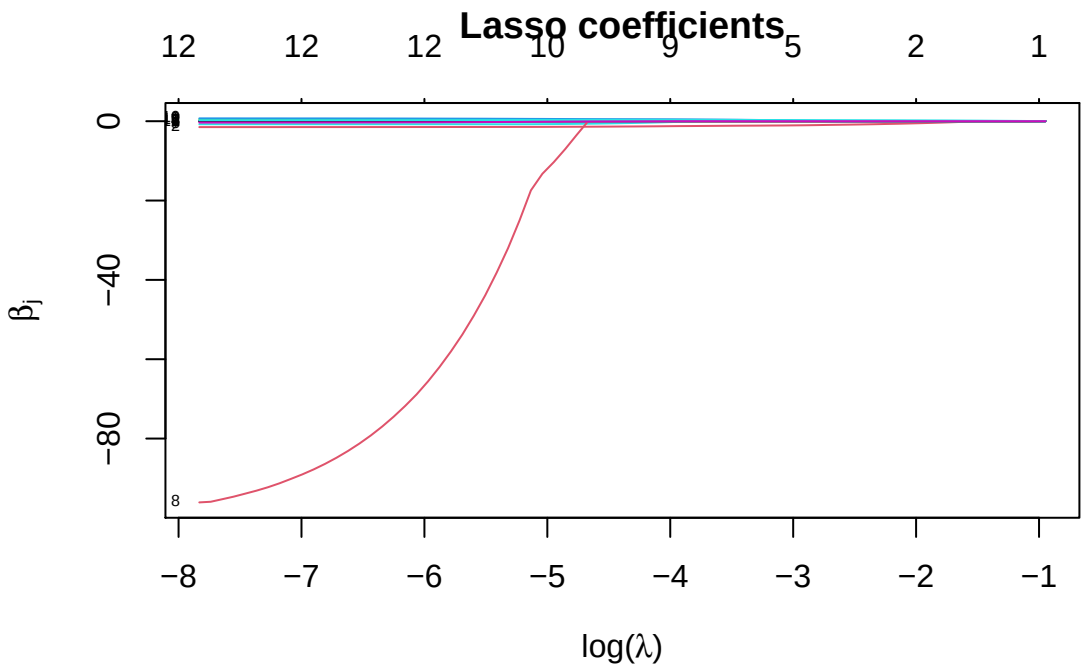

Histogram of data$density
Frequency
0
500
1500
2500
0.99
1.00
1.01
1.02
1.03
1.04
data$density

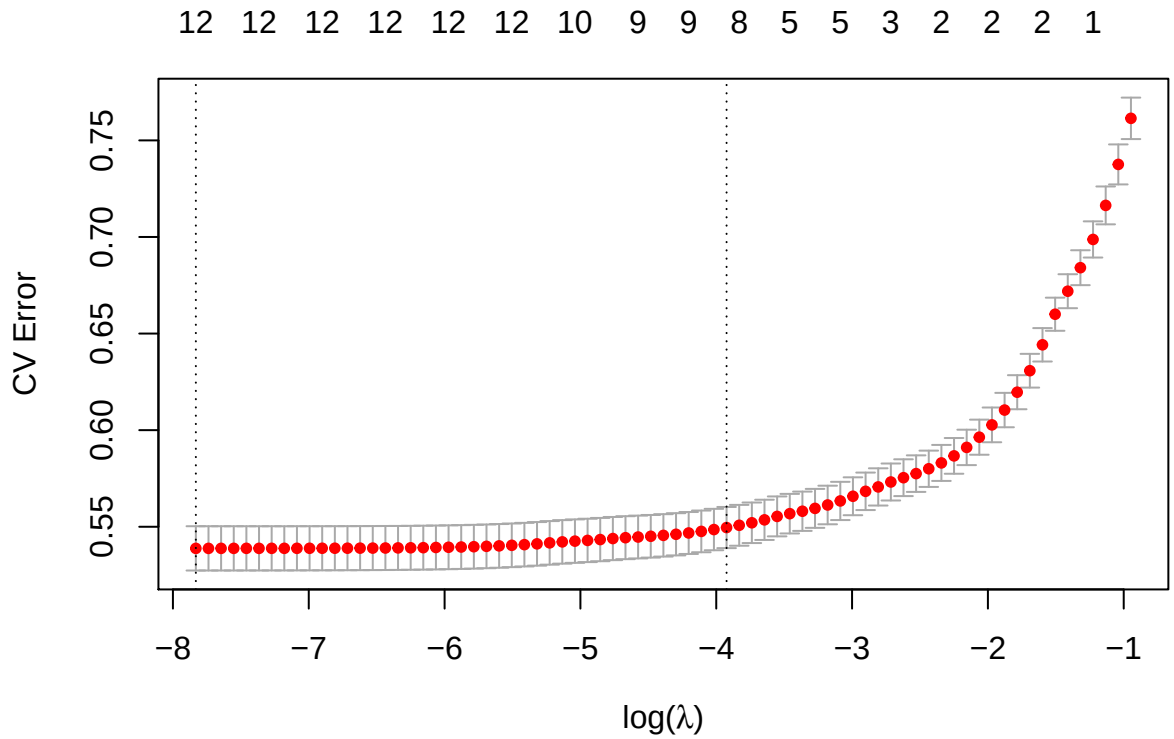


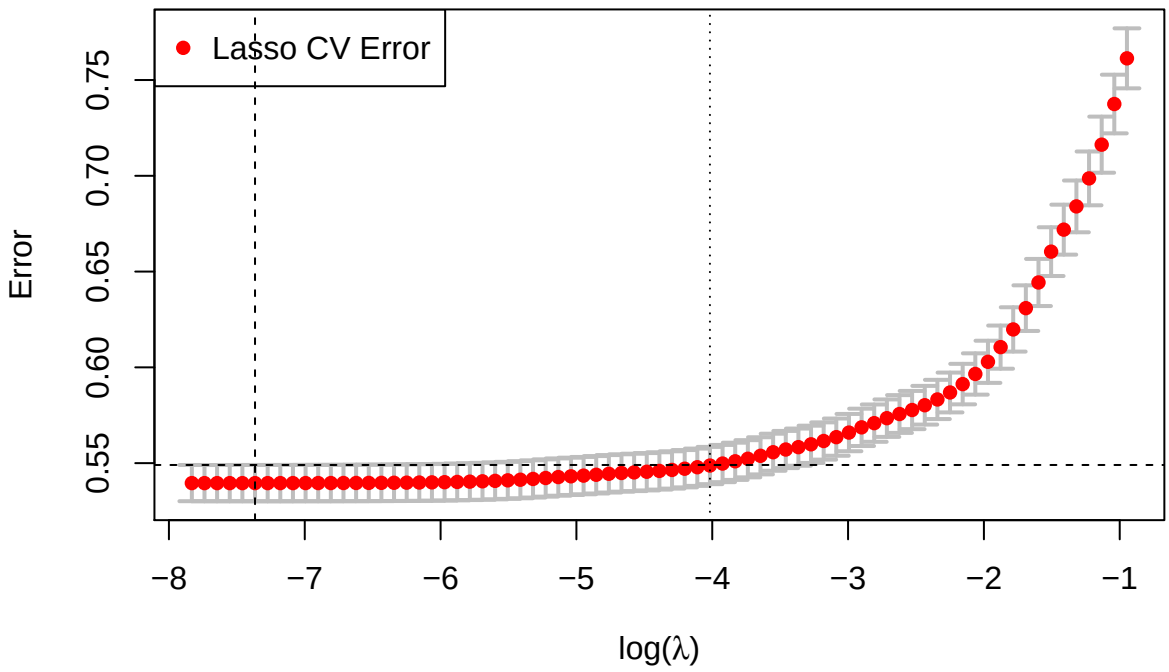
The smallest cross-validation error minimizes with $\lambda = 0.0388$.

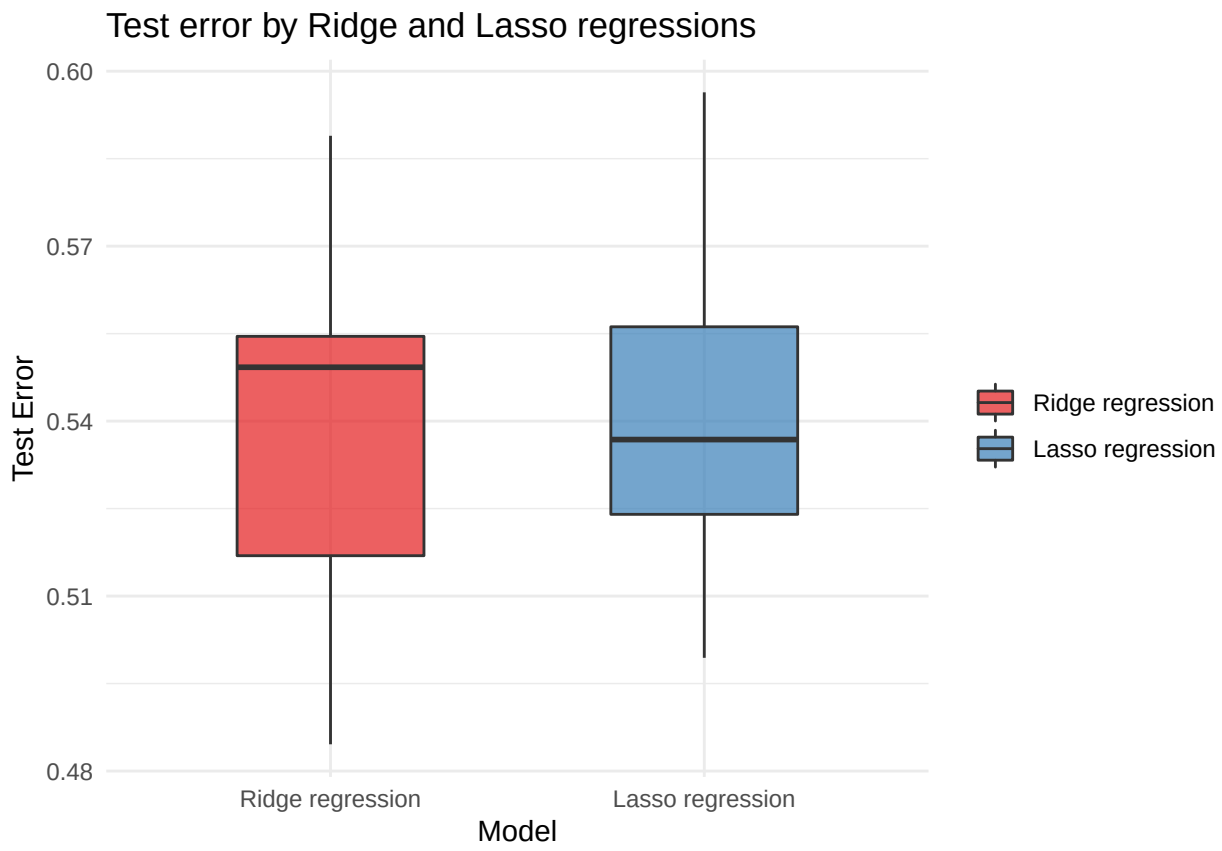

Ridge CV Error
Error
0.55
0.60
0.65
0.70
0.75
−2
0
2
4
6
log(λ)

3.2 Lasso regression



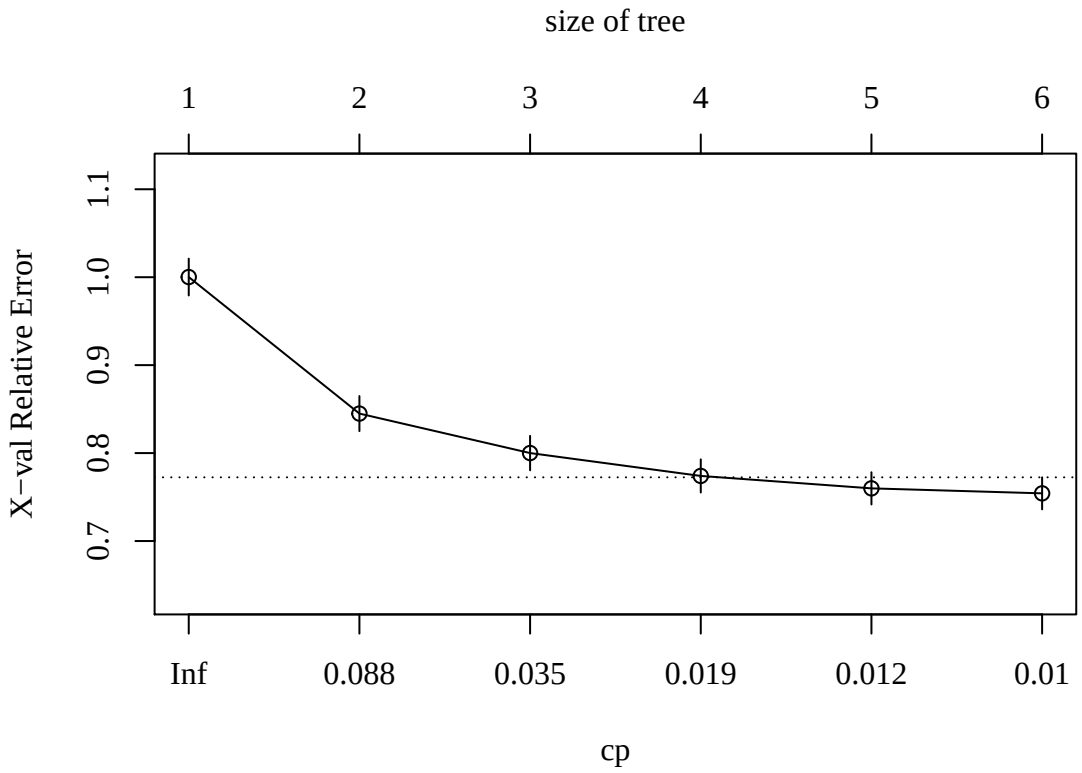


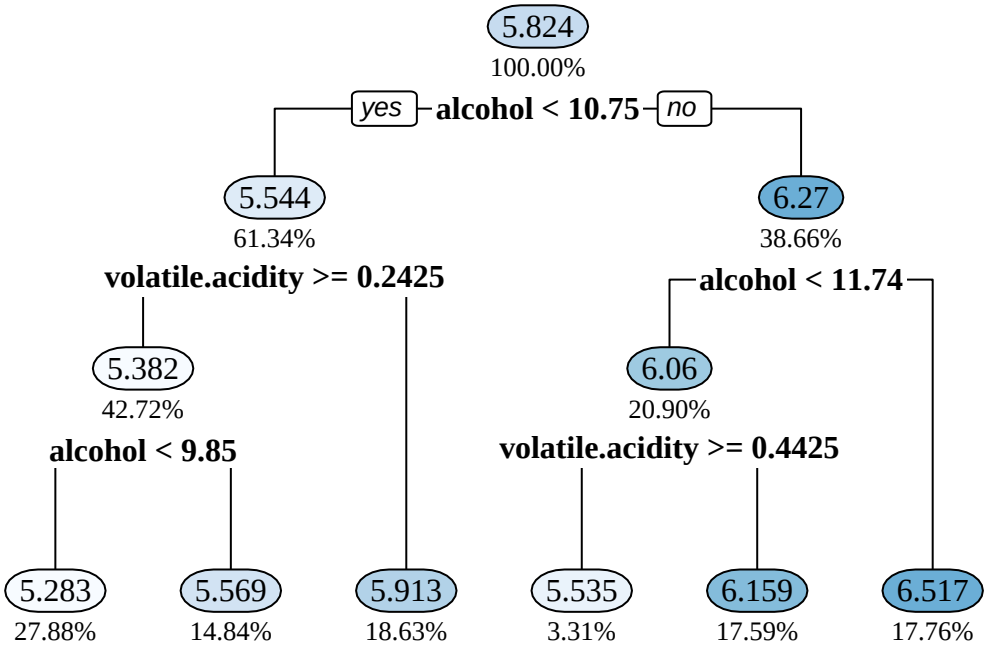

Lasso CV Error
Error
0.55
0.60
0.65
0.70
0.75
−8
−7
−6
−5
−4
−3
−2
−1
log(λ)

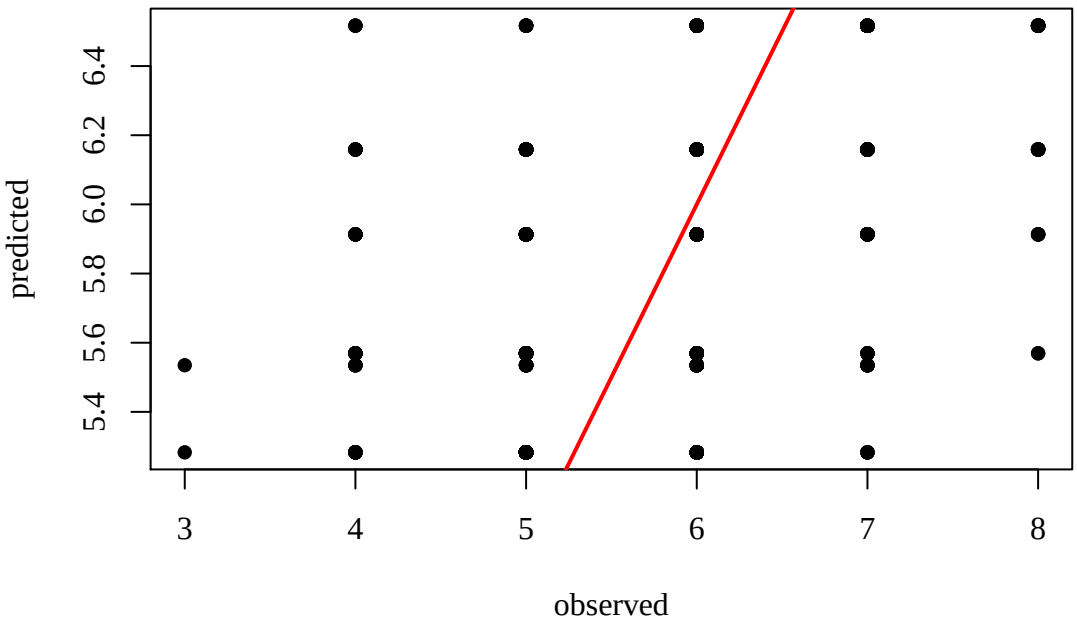


3.3 Tree-based approaches

```
## Size of optimal tree: 6
```


size of tree
1
2
3
4
5
6
X−val Relative Error
0.7
0.8
0.9
1.0
1.1
Inf
0.088
0.035
0.019
0.012
0.01
cp


5.824
100.00%
yes
alcohol < 10.75
no
5.544
61.34%
6.27
38.66%
volatile.acidity >= 0.2425
alcohol < 11.74
5.382
42.72%
6.06
20.90%
alcohol < 9.85
volatile.acidity >= 0.4425
5.283
27.88%
5.569
14.84%
5.913
18.63%
5.535
3.31%
6.159
17.59%
6.517
17.76%



4 Discussion

4.1 Limitations

- Only uses the wine data from *vinho verde* region of Portugal. If a test dataset outside of this region is available, we would be able to test to generalize the results to a broader range of wine.

4.2 Implications

- Wine producers will use this information to produce the better quality of wine considering the selected attributes.
- Wine consumers will be able to choose a good quality of wine without tasting (e.g., via online shopping) if the physicochemical information of wine is available.

References

[1] P. Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009), pp. 547–553.