Red Wine Quality Prediction

Jiafei Li, jl5548; Gaotong Liu, gl2677; Yuchen Qi, yq2279

2020/5/16

Introduction

This final project aims to find an optimal model in order to better predict red wine quality based on physicochemical tests. The original dataset contains 1599 observations of 11 predictors from physicochemical test (such as PH value, density, etc.), and 1 response variable (wine quality), describing features of the Portuguese red wine "Vinho Verde".

Data Preparation

Missing data and Variable type

We first examined variable types and missing values of the dataset. As a result, all variables are numeric and there is no variable containing missing data. We converted the response variable "quality" to factor variable of 6 levels, and collapsed them into 2 levels ("poor", "good") to balance the total count of different levels. Then we cleaned the dataset, and partitioned the dataset to get training and testing data.

Exploratory analysis

Correlations

11 predictors are all numerical variables. There is no strong correlation (>0.7) between the predictors.

Density plot

As shown in the plots, the predictors ph, sulphates, free sulfur dioxide, and residual sugar have similar distributions for the good and poor quality, and other predictors which do not may be important predictors.

Models

All the 11 numeric predictors were included in the models. 8 classification models were used to predict the quality of the red wine (poor and good).

Linear Discrimininant analysis (LDA)

LDA projects the feature space onto a smaller subspace while maintaining the class discriminatory information. It has the linear boundary and assumes the same covariance matrix in each class. It is quite robust to the distribution of the classification data when the sample size is small.

Quadratic Discrimininant analysis (QDA)

QDA has the quadratic boundary and assumes different covariance matrix in each class.

k-Nearest-Neighbor classifiers (KNN)

It predicts class label given x_0 by finding k nearest points in distance to x_0 and then classify x_0 using majority vote among the kneighbors. The tuning parameter is k with optimal value 41 in the knn model.

Classification tree

Tree-based method uses recursive binary splitting to segment the predictor space into simple regions according to the largest reduction of total varaince across the k classes(Gini index), then predict the class labels by the majority vote in the simple regions. Although single tree can have small bias, the variance is quite large.

CART approach is used to prune the classification tree. A large tree is grown at first and then prune it back by penalty for tree complexity(α controls). The tuning parameter is cp(α) with optimal value 0.003583 in the single classificaton tree model(CART).

Random Forest

Random Forest is one of the ensemble methods which uses collections of single trees to get better predictive performance(lower variance). Random Forest generate B different bootstrapped training data sets and the split in each tree is considered a random selection of m out of p(full set) predictors, then it predicts the class labels by majority vote among B trees.

The tuning parameters are mtry (m) with optimal value 1, min.node.size(minimal node size) with optimal value 1 in random forest model.

Boositng(AdaBoost)

Boosting grows tree uses information from previously grown trees. AdaBoost repeatedly fit classicication trees to weighted versions of training data and update the weights to better classify.

The tuning parameter are n.trees(the number of trees) with optimal value 5000, interaction.depth(the complexity of boosted ensemble) with optimal value 23, shrinkage(the rate of boosting learn) with optimal value 0.005, n.minobsinnode(minimal node size) with optimal value 1 in boosting model.

Support Vector Machine Linear kernel(SVML)

SVM finds a hyperplane to separate the class in feature space and C, as a regularization parameter, controls the margin size and shows the tolerance for observations on the wrong side. Linear kernel has linear boundary. The tuning parameter is cost(C) with optimal value 0.01019 in support vector machine with linear kernel.

Support Vector Machine Radial kernel(SVMR)

The best tuning parameter is sigma = 0.050, C = 54.598, the train error is 1 - 0.8565 = 0.1435, and the test error is 1- 0.7598 = 0.2402. Both the train and test errors are smaller than the linear kernel SVM.

Different from the linear kernel, radial kernel can construct nonlinear classification boundaries. The tuning parameters are cost(C) with optimal value 20.0855, sigma(γ, local behavior) 0.04978 in support vector machine with radinal kernel.

Performance

From summary table of training cross validation performence, random forest has largest mean ROC. In addition, from the plot of AUC using test data, random forest has the best test performance.

Variable importance

The top three variables which play important roles of predicting red wine quality are `alcohol`, `sulphates` and `volatile_acidity`.

PDP

The most important variable `alcohol` is chosen to investigate the typical influence on red wine quality across all observations. From the partial dependence plot, the higher the alcohol, the lower the quality after averaging all the effects of other predictors.

ICE

From the individual conditional expectations plot, the higher the alcohol, the lower the quality after ignoring the effects of other predictors. ICE and PDP plots are quite similar, so the alcohol is independent of other predictors.

Explain prediction

After fitting a simple model around a single observation that mimic how th global model behaves at that locality. The prediction of two new observation can be explained by three features.

The first new observation is labeled as good quality with probability 0.21. This observation has alcohol smaller than 9.5 and this feature associates with poor quality. This observation has sulphates smaller than 0.55 and this feature associates with poor quality.This observation has chlorides smaller than 0.07 and this feature associates with good quality.

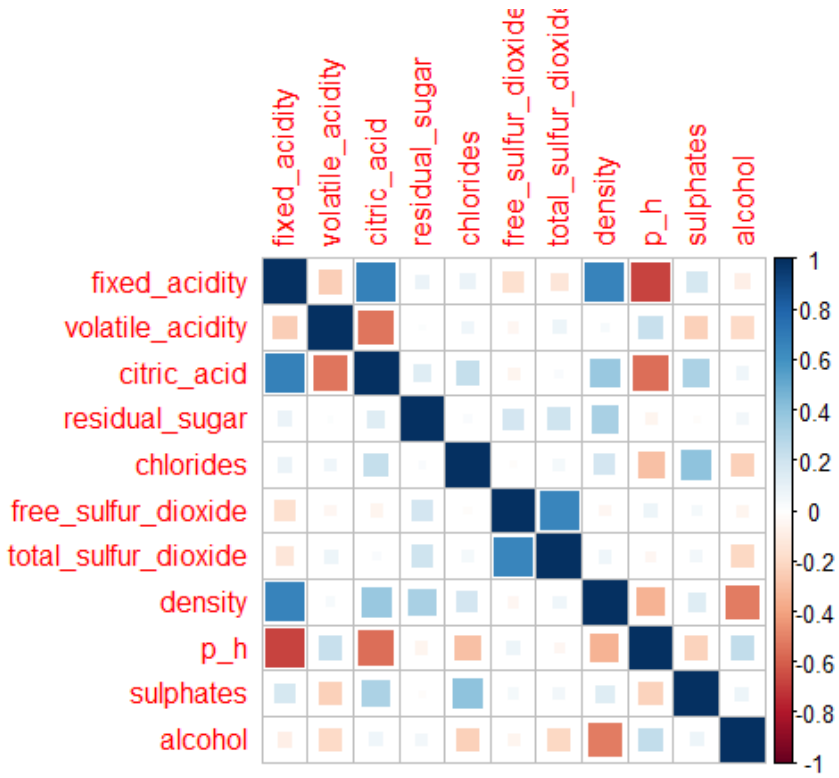
The second new observation is labeled as good quality with probability 0.38. This observation has sulphates smaller than 0.55 and this feature associates with poor quality. This observation has density smaller than 0.996 and this feature associates with good quality.This observation has volatile_acidity larger than 0.64 and this feature associates with poor quality.

Conclusion

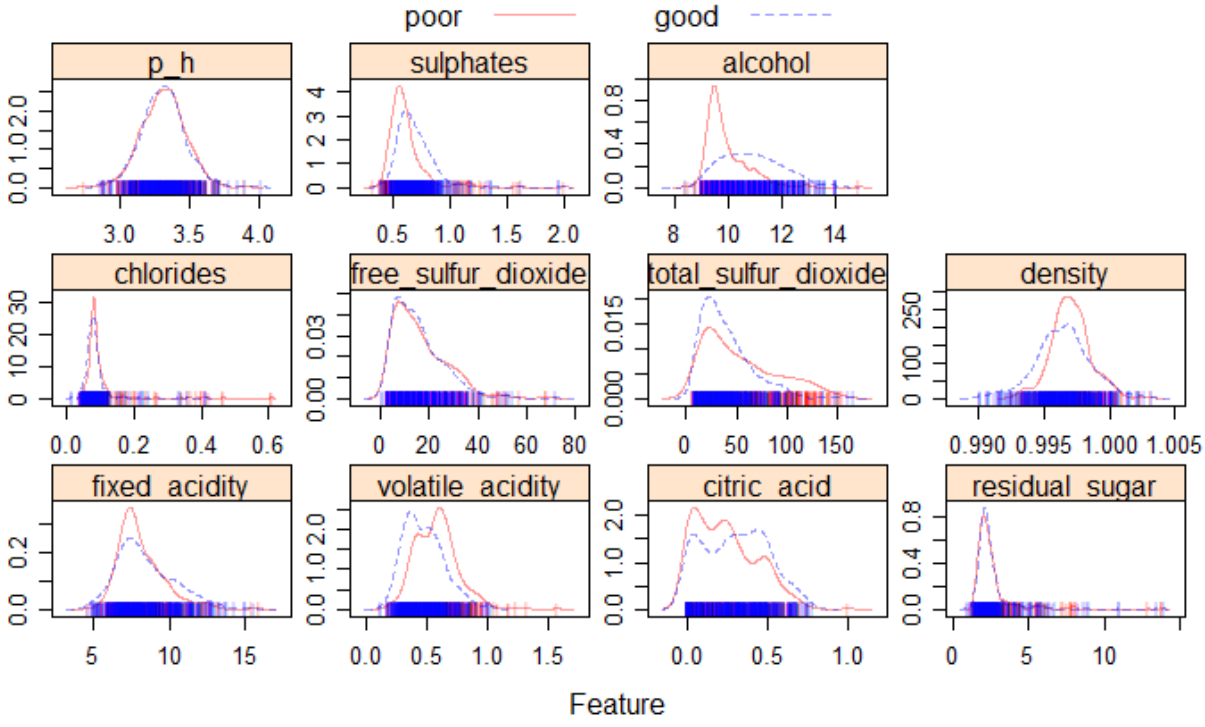
The performance of our models is evaluated based on ROC. The random forest model gives the best prediction performance with the ROC mean 0.8797 based on training data. Thus, for the red wine quality, we may choose random forest to predict whether it is poor or good. The test classification error rate is 0.1932 based on the confusion matrix below.

We expect that predictors having different distributions in the good and poor quality groups are important predictors, and among those predictors `alcohol`, `total_sulfur_dioxide`, `volatile_acidity` do play important roles of predicting red wine quality. As for the model, it is expected that the boosting model will perform best, and the lack of sufficient parameter tuning may account for the choice of random forest instead. We tried different tuning grids, and to find the best tuning parameters we need to expand the tuning grid, but it exceeds our computer capacity. Alternatively, random forest may be the optimal solution for this dataset.

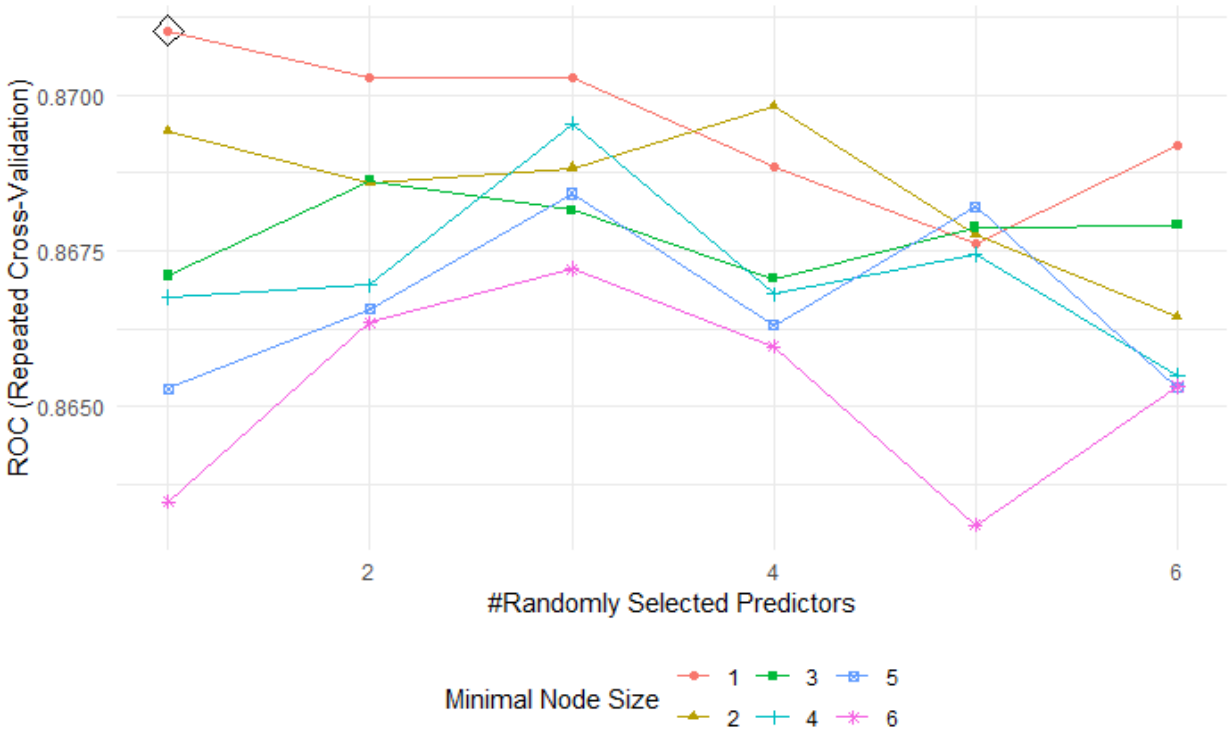
Supplementary figures



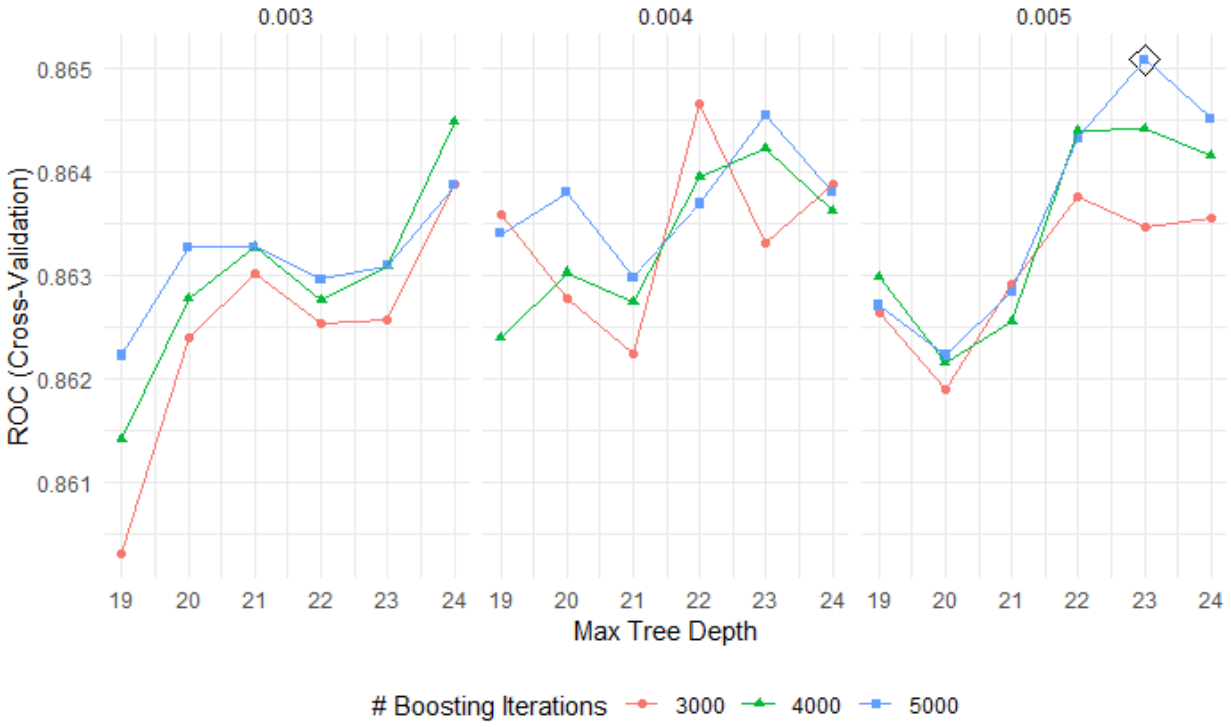
Supplementary Fig 1. Correlation



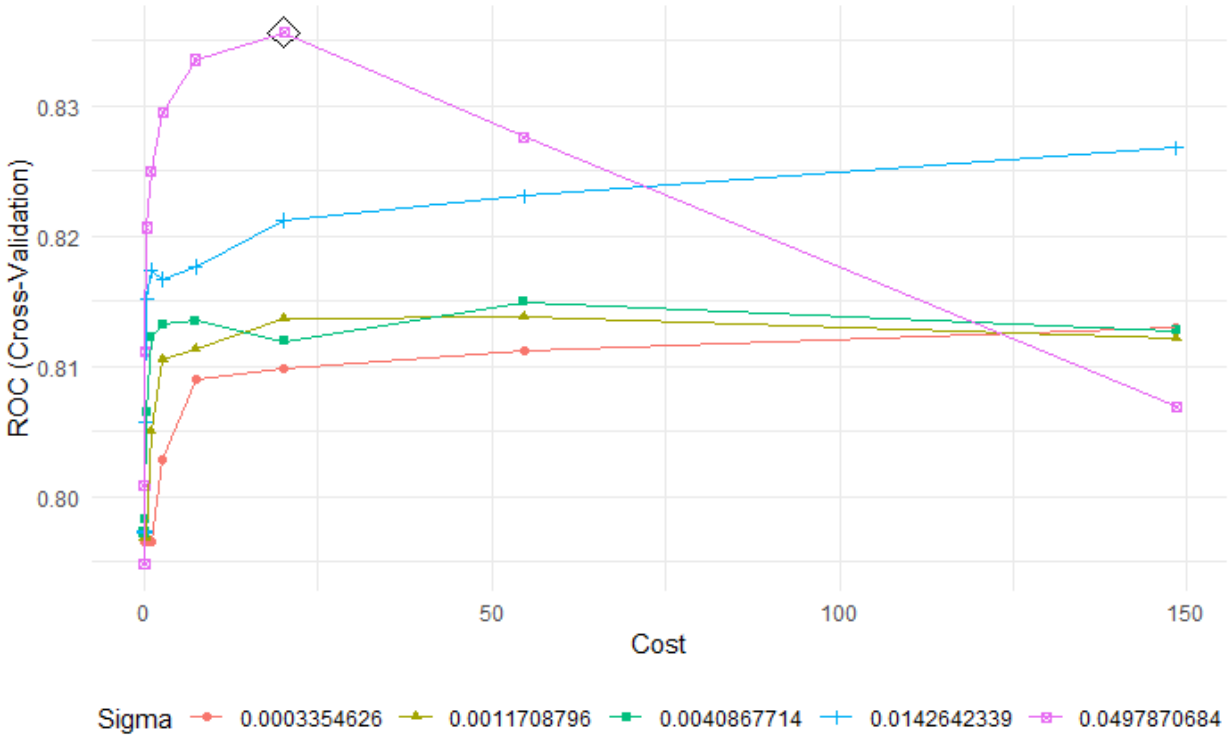
Supplementary Fig 2. Density



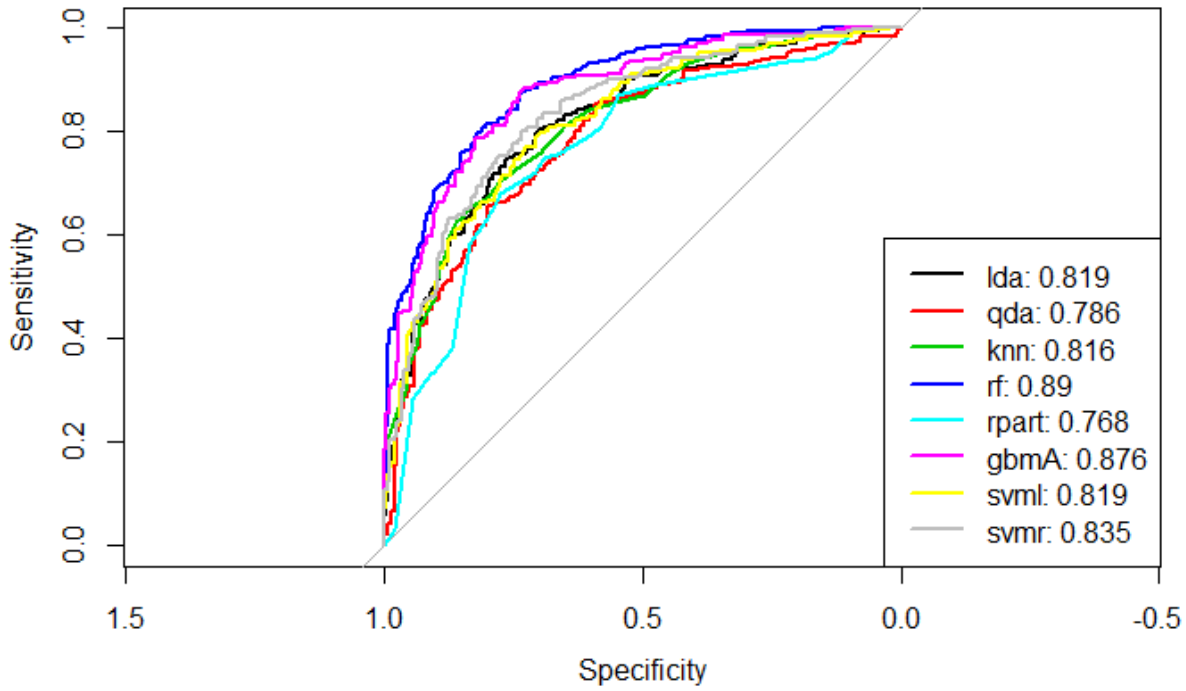
Supplementary Fig 3. Random forest



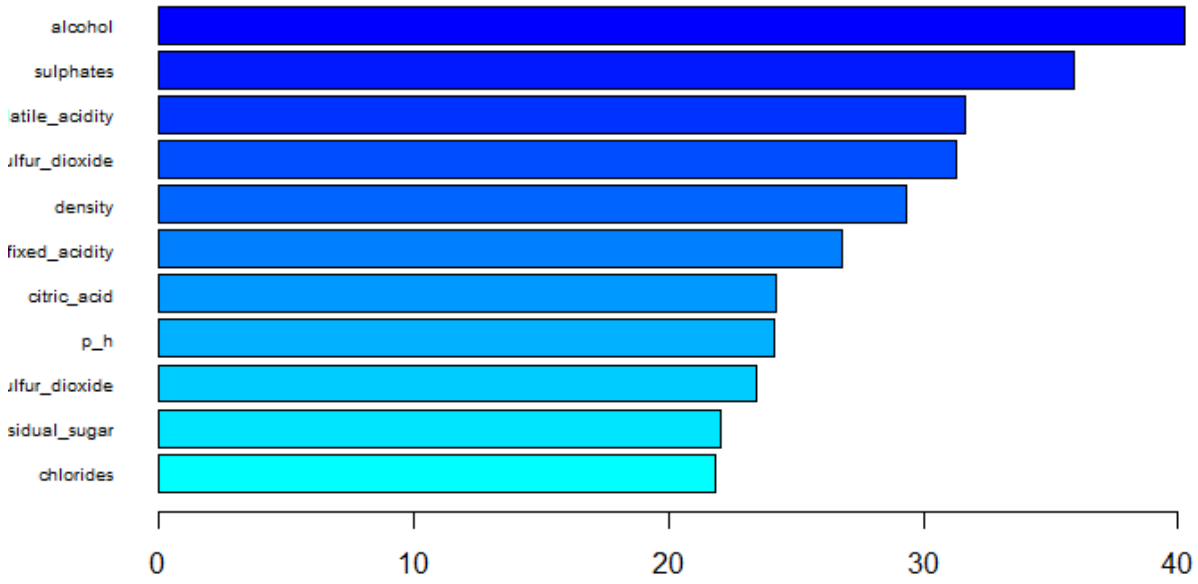
Supplementary Fig 4. Boosting



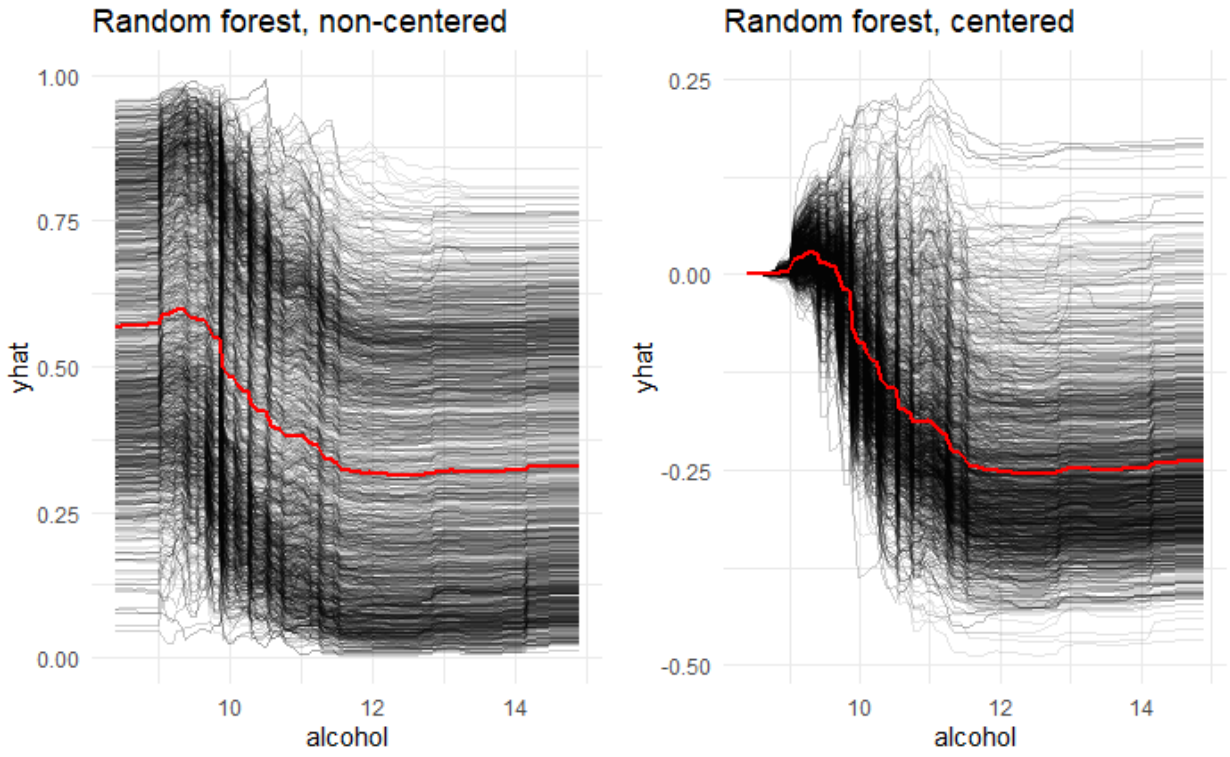
Supplementary Fig 5. SVMR



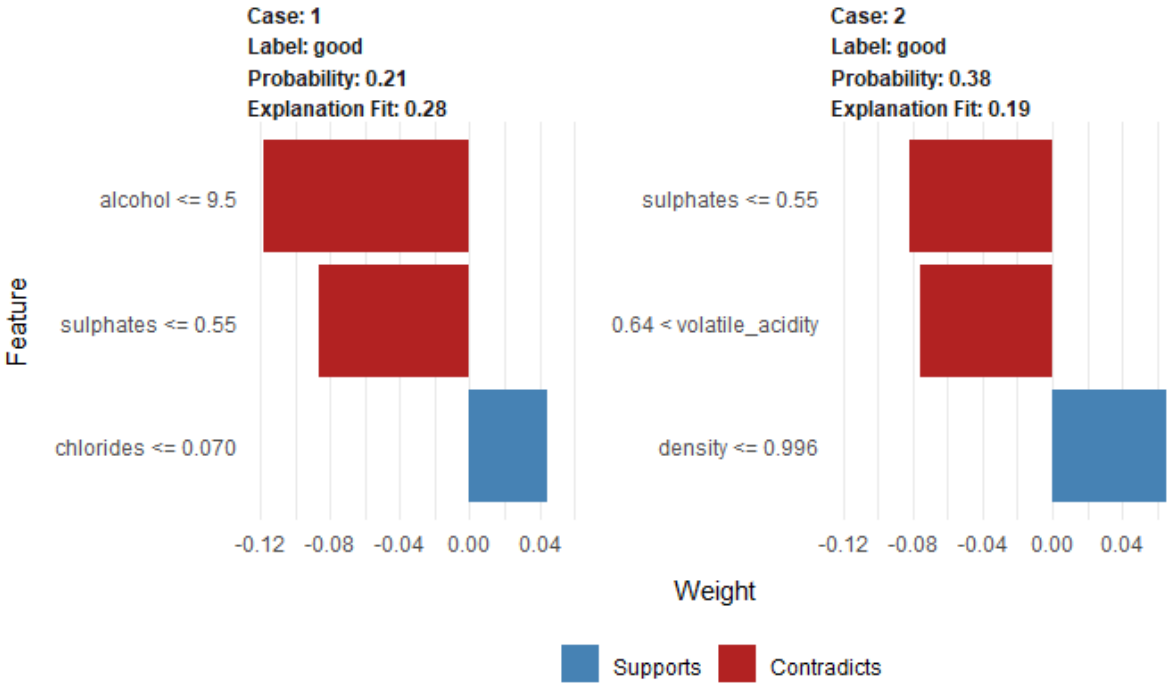
Supplementary Fig 6. Test ROC



Supplementary Fig 7. variable importance



Supplementary Fig 8. PDP



Supplementary Fig 9. ICE