Midterm

ZiqianHe

3/26/2022

Introduction

This datasets is related to red variants of the Portuguese "Vinho Verde" wine from the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price.
The `wine data` is a dataset with 1599 observations, with 11 variables and 1 response.

response:

- `quality`: score between 0 and 10, > 6.5 is good

variables

- `fixed acidity`: most acids involved with wine or fixed or nonvolatile

- `volatile acidity`: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

- `citric acid`: found in small quantities, citric acid can add 'freshness' and flavor to wines

- `residual sugar`: the amount of sugar remaining after fermentation stops,

- `chlorides`: the amount of salt in the wine

- `free sulfur dioxide`: the free form of SO2 exists in equilibrium between molecular SO2 and bisulfite ions

- `total sulfur dioxide`: amount of free and bound forms of S02

- `density`: the density of water is close to that of water depending on the percent alcohol and sugar content

- `pH`: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)

- `sulphates`: a wine additive which can contribute to sulfur dioxide gas (S02) levels
- `alcohol`: alcohol concentration

To understand the relationship between the quality and other variables. I split the quality into two groups with the requirement of the data and then change it to factor. The dataset was randomly split into traning and testing datasets(80% vs 20%) and will fit different models.

Table 1: Data summary

Name	wine_data
Number of rows	1599
Number of columns	12
Column type frequency:	
factor	1
numeric	11
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
quality	0	1	FALSE	2	bad: 1382, goo: 217

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
fixed_acidity	0	1	8.32	1.74	4.60	7.10	7.90	9.20	15.90
volatile_acidity	0	1	0.53	0.18	0.12	0.39	0.52	0.64	1.58
citric_acid	0	1	0.27	0.19	0.00	0.09	0.26	0.42	1.00
residual_sugar	0	1	2.54	1.41	0.90	1.90	2.20	2.60	15.50
chlorides	0	1	0.09	0.05	0.01	0.07	0.08	0.09	0.61
free_sulfur_dioxide	0	1	15.87	10.46	1.00	7.00	14.00	21.00	72.00
total_sulfur_dioxide	0	1	46.47	32.90	6.00	22.00	38.00	62.00	289.00
density	0	1	1.00	0.00	0.99	1.00	1.00	1.00	1.00
ph	0	1	3.31	0.15	2.74	3.21	3.31	3.40	4.01
sulphates	0	1	0.66	0.17	0.33	0.55	0.62	0.73	2.00
alcohol	0	1	10.42	1.07	8.40	9.50	10.20	11.10	14.90

Exploratory analysis

From the `Figure 1.` we can find that `alcohol`, `citric acid` and `volatile acidity` may be statistical significant for the model. It seems that the quality increase with the increase of them.

Models

We choose `GLM`, `GLMNET`, `GAM` `MARS`, `LDA`, `RIDGE` and `ELASTIC` to train the data with 5-fold cross validation. The linear regression model was first fitted, the use GENERALIZED ADDITIVE MODEL (GAM) and MULTIVARIATE ADPTIVE REGRESSION SPLINES MODEL(MARS) to capture the non-linear relationship between the response and the variables. Figure 2-4. are some of the plot of the models.

Comparison

Through the resampling (`Figure 2.`), the `GAM` has the highest ROC though our model have similar ROC performance. The model is used for quality forecasting, so I pick the top three models to draw a plot of sensitivity and found that the `MARS` and GAM have the similar sensitivity (`Figure 3.`). Considering both, `GAM` is chosen as the model.

From the test data performance we find that `alcohol`, `residual_sugar`, `fixed_acidity`, `sulpates`, `volatile_acidity`, `total_sulfur_dioxide` and `density` are statistically significant.

From the importance plot (`Figure 4.`), the `residual_sugar` has low importance to AUC loss and other variables mentioned above have high importance to AUC loss.

```
## Setting levels: control = bad, case = good
```

```
## Setting direction: controls < cases
```

```
## Welcome to DALEX (version: 2.4.0).
## Find examples and detailed introduction at: http://ema.drwhy.ai/
```

```
##
## Attaching package: 'DALEX'
```

```
## The following object is masked from 'package:dplyr':
##
##     explain
```

Conclusion

GAM model has higher sensitivity and predictability. `alcohol`, `residual_sugar`, `fixed_acidity`, `sulpates`, `volatile_acidity`, `total_sulfur_dioxide` and `density` are statistically significant. Which align with the original thought. If possible, we can select significant variables and remodel the models, in which case the accuracy may increase. Also we can try to find other models which can fit the dataset better such as the Naive Baye and stuff.

Appendix

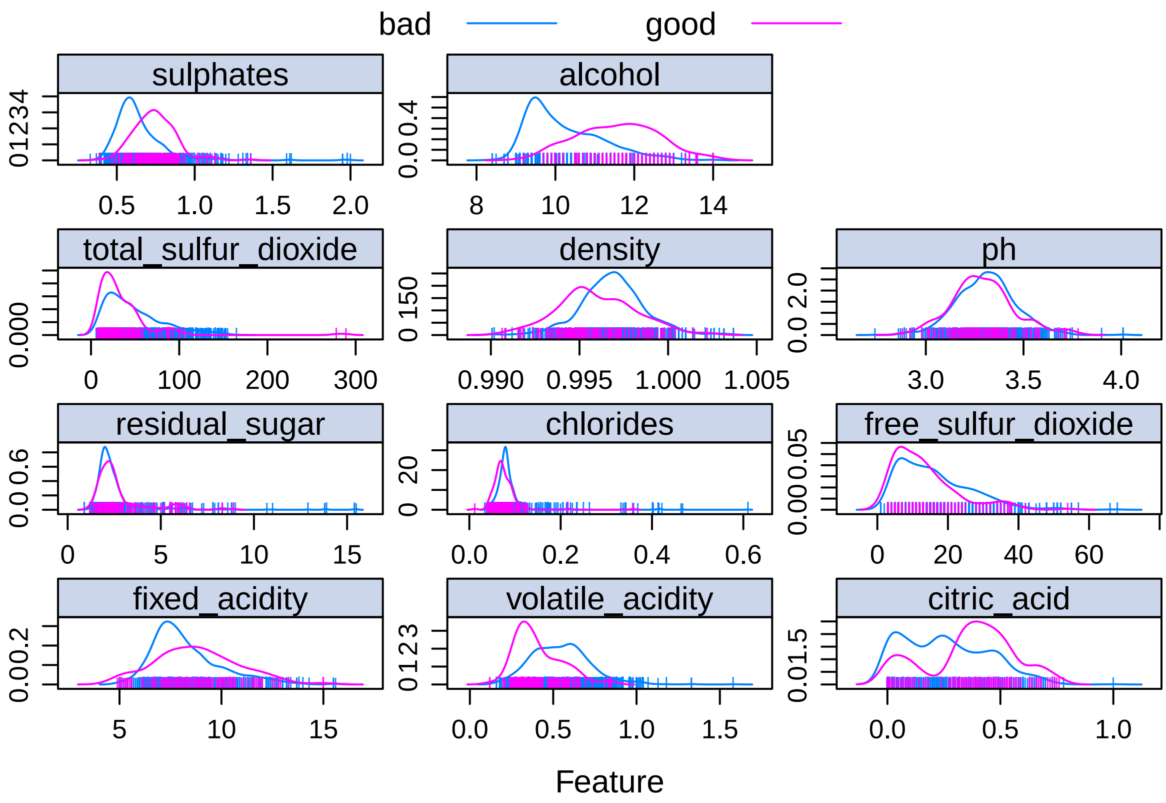

Figure 1. Plot of the feature Plot.

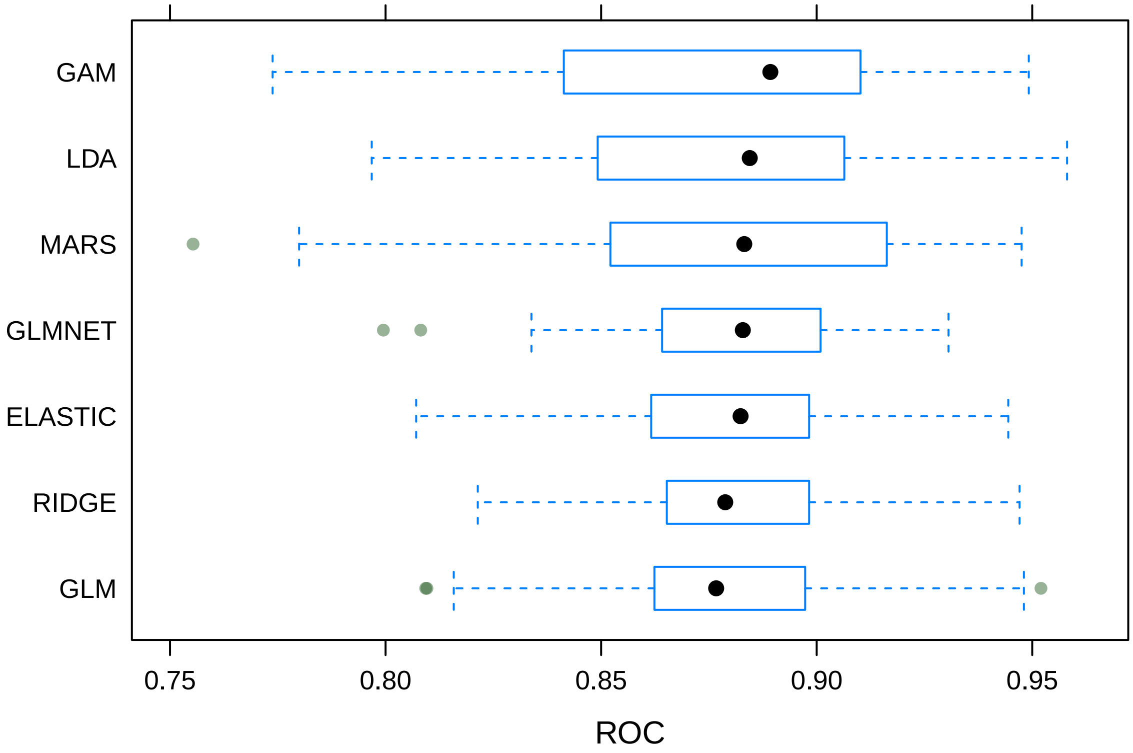

Figure 2. Plot of the ROC.

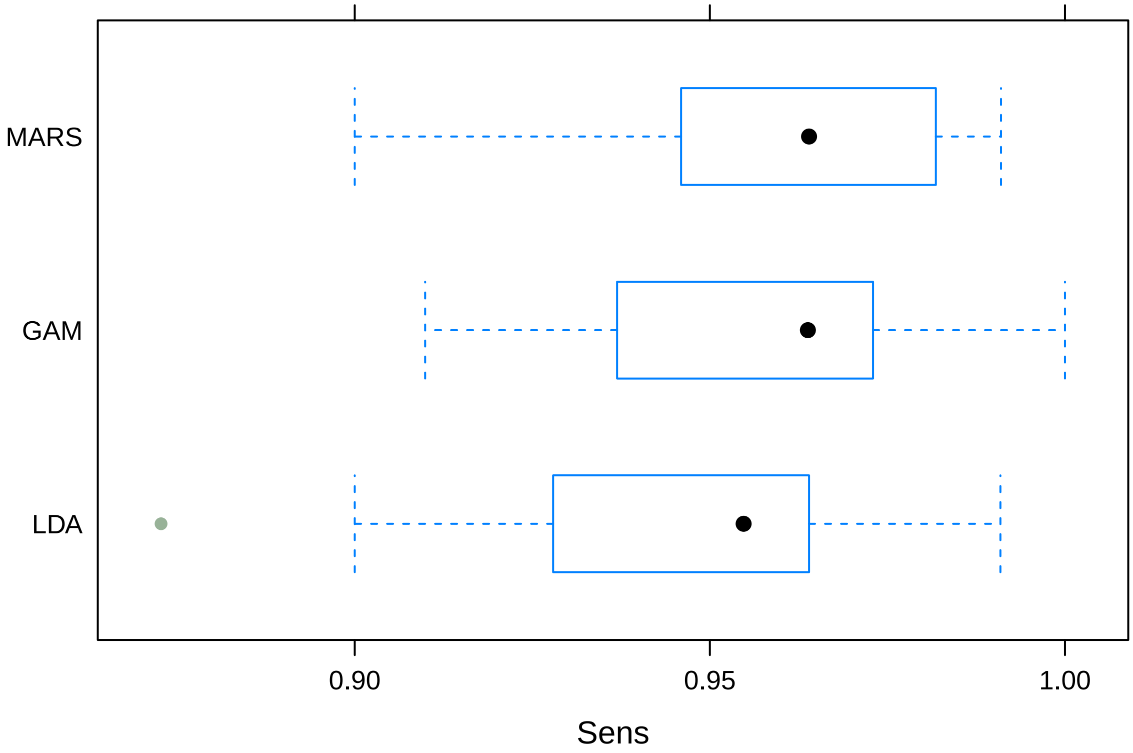

Figure 3. Plot of the Sensitivity.

```
plot(gam_int)
```

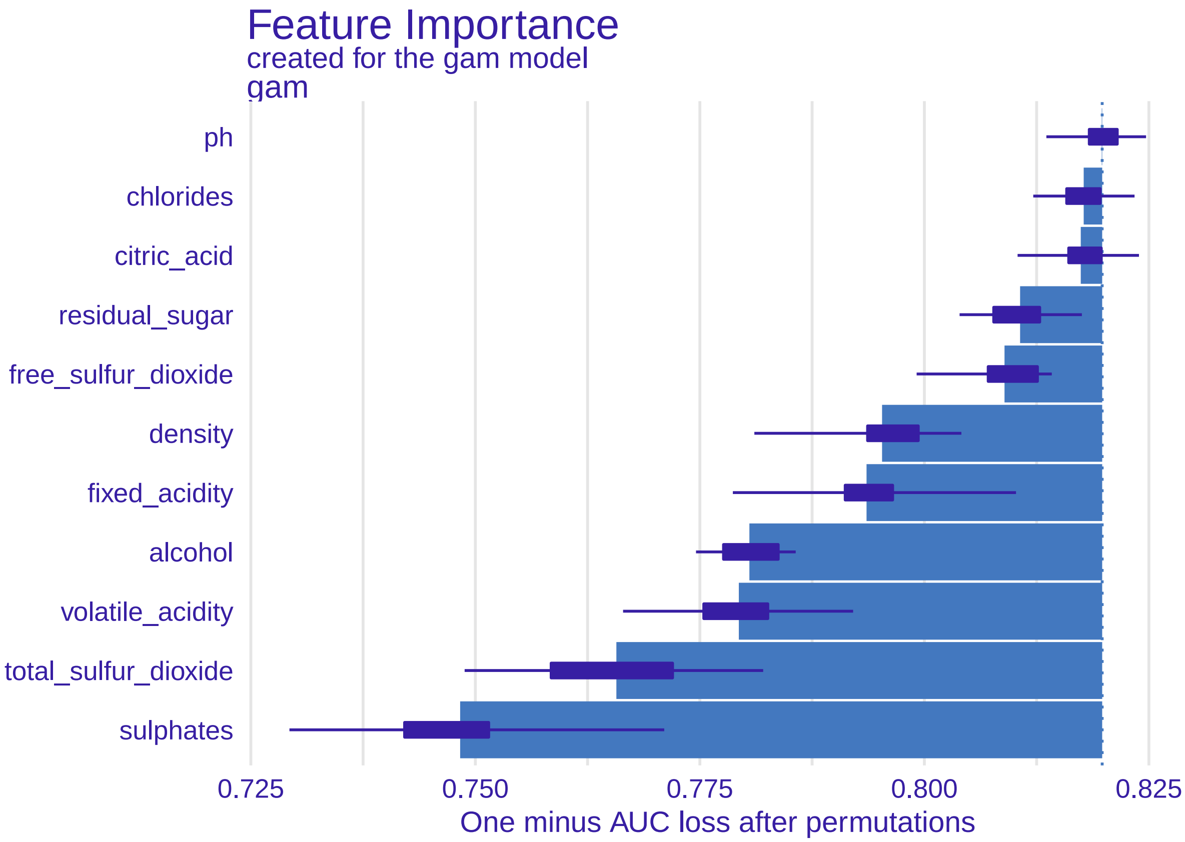

Figure 4. Plot of the feature importance

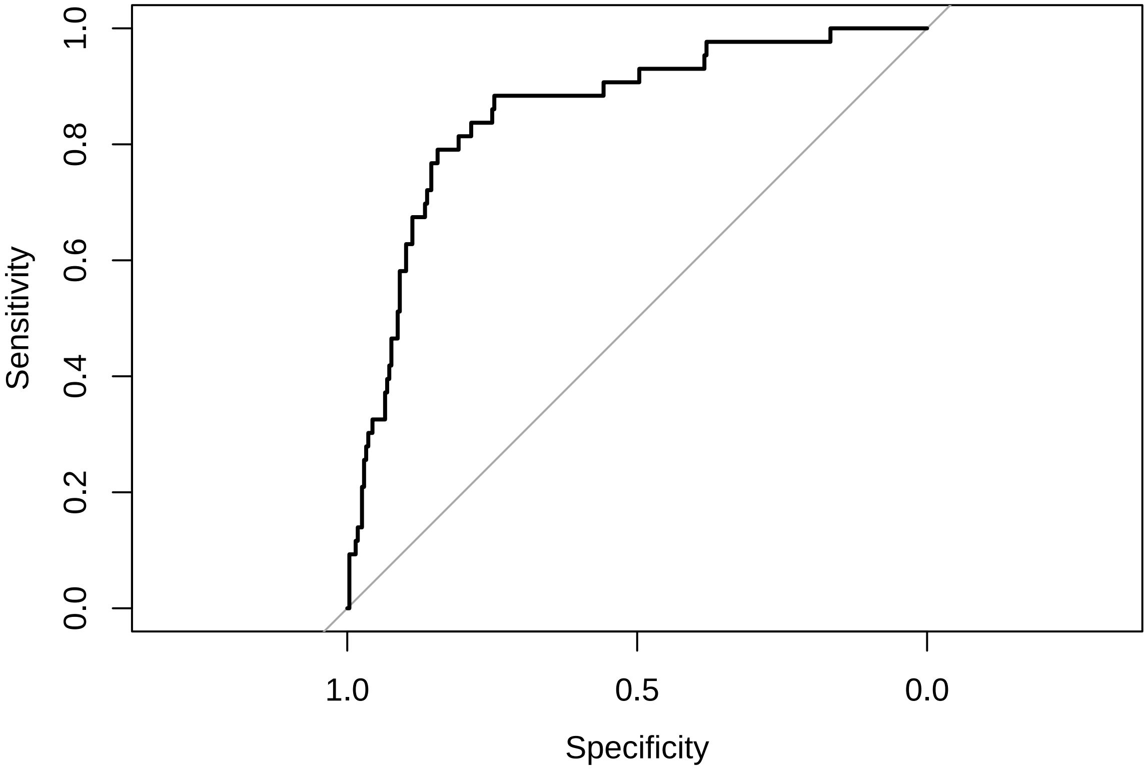

Figure 5. GAM ROC