Midterm Project Report

Ziqi Zhou

4/3/2020

Introduction

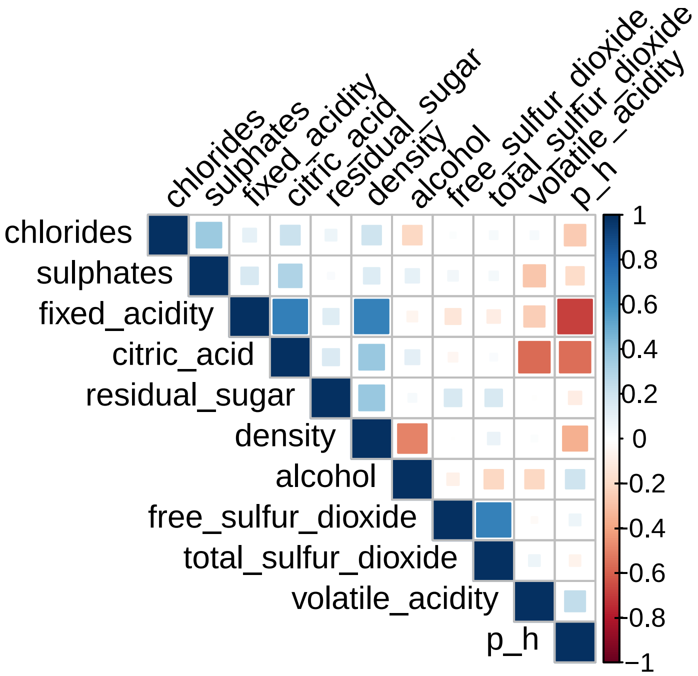
The red wine is one of the most popular alcoholic beverages in the world. Most of our team members enjoy drinking red wine and have a great interest in factors that can affect the quality of the red wine. We want to figure out what determined the quality of the wine. Based on this motivation, we choose this dataset "Red Wine Quality".

This dataset has 12 columns and 1599 rows. And after omitting the NA data, we still got 1599 rows which means there is no missing value in this dataset. There is 1 outcome which is the quality score of the wine(from 0 to 10) and 11 predictors including fixed acidity volatile, acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

We would like to know what factors influence the quality of the wine and try to build a model to predict the quality of wine given the specific factors.

I uesd "janitor::clean_names()" to clean the dataset and separate it into training data and test data "createDataPartition(wine$quality, p = .75, list = F)".

Exploratory Data Analysis

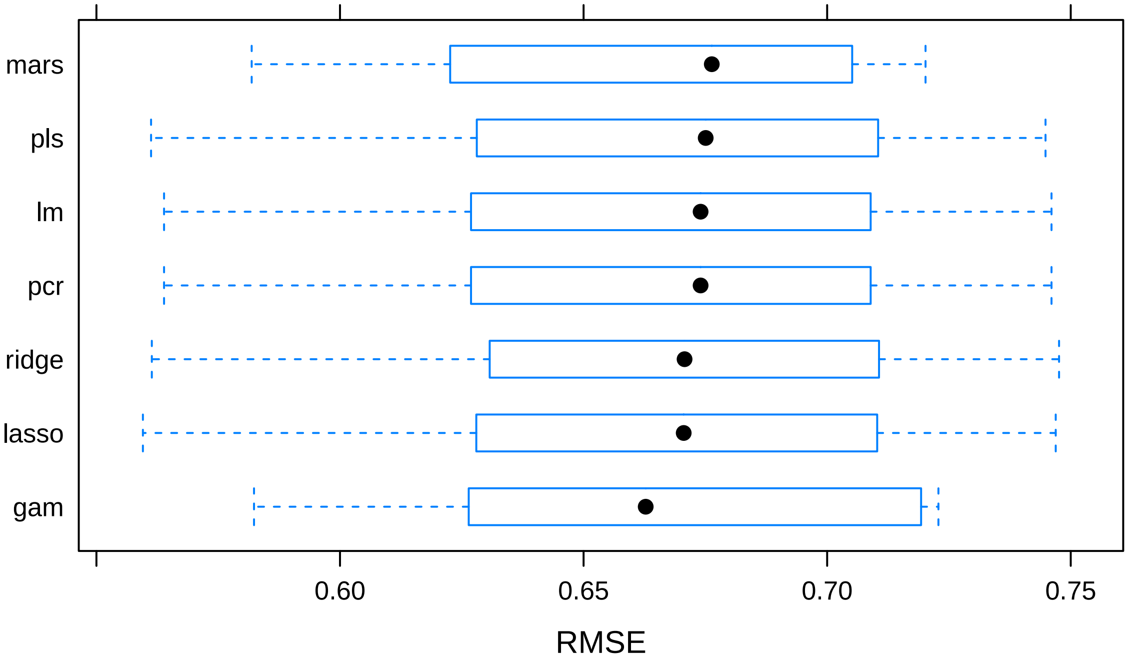


Description

We could find in this plot that the fixed acidity and citric acidity are highly positive correlated. Fixed acidity is highly positive correlated with density. Volatile_acidity is negative correlated with fixed acidity and citric acidity. What's more the pH is negative correlated with fixed acidity and citric acidity. It is easy to interpret since the pH is describes how acidic or basic a wine is. The factors might influence each other somehow.

However, high correlations might cause problem. So I consider use lasso or ridge method to penalize them.

Models



I used the Linear Regression, Ridge, Lasso, PCR, PLS, GAM and MARS to fit the data. I used "caret" package to make cross-validation, resamp() to compare the model and decided to use GAM method to build the model based on the biggest R-square and relatively smallest RMSE.

Method	R-square	RMSE
Linear Regression	0.3420745	0.6668210
Ridge	0.3421780	0.6665864
Lasso	0.3441676	0.6654834
MARS	0.3521095	0.6623480
GAM	**0.3531965**	**0.6633735**
PCR	0.3420745	0.6668210
PLS	0.3424195	0.6666716

GAM being taken to include any quadratilcally oenalized GLM and a variety of other models estimated by a quadratically penalised likelihood type approach. And I use the GCV.cp to estimate the smoothing parameter. In this GAM model we conclude all the predictors.

The information about the final Model is as below.The mse of this model is 0.0345439.

```
gam.fit$finalModel
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(free_sulfur_dioxide) + s(alcohol) + s(citric_acid) +
##     s(residual_sugar) + s(p_h) + s(fixed_acidity) + s(sulphates) +
##     s(volatile_acidity) + s(chlorides) + s(total_sulfur_dioxide) +
##     s(density)
##
## Estimated degrees of freedom:
## 2.54 3.47 3.04 6.62 1.00 5.78 3.74
## 2.25 6.65 7.37 1.61  total = 45.07
##
## GCV score: 0.40868
```

Limitation: the outcome of the data is a classified variable with order. However, in my report I just treated the outcome as continuous variable. It might be more reasonable to use LDA or other method of classification to analysis. Since I used the caret package to generate the GAM cv, I may lose some flexibility in mgcv.

Conclusions

All the 11 predictors are related with the outcome (quality). Some predictors might have correlations. GAM was relatively suitable to be used to build the predict model about the quality of wine.

Nowadays, the quality of wine is decided by the human tasters. So it is so subjective. With the predict model, the producer could control the process of production so that they could produce good wine effectively.