P8106 Midterm Project Report -- Prediction of Wine Quality Level

Xiaoluo Jiao (xj2278)

Introduction

In recent years, wine is increasingly enjoyed by a large range of consumers. To meet the growing demand, the wine industry is investing in new technologies for making better quality wines. The quality assessment of wine is a key element for the wine-making process since it can be used to improve wine-making techniques and help to stratify different levels of wine and set prices. Thus, identifying the most influential physicochemical factors for wine quality is important.

Dataset

In this project, we are interested in which physicochemical properties are critical in allowing the wine to have higher quality and building a model to predict the quality of the wine based on those physiochemical features. Our dataset is related to the red variant of the Portuguese "Vinho Verde" wine from the north of Portugal. It is built with 1599 red wine examples, and 11 physicochemical statistics related to each red wine example are included. The response variable is quality. The physicochemical features are:

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- p_h
- sulphates
- alcohol
- quality: based on sensory data, a score between 0 and 10

Data Preparation

Before building up a predictive model, we need to perform data cleaning to the raw dataset first. Firstly, for the outcome variable quality, I set a cutoff making a 7 or higher quality score gets classified as "good" and the remainder as "not good" because I am interested in a classification model for selecting "high-quality wine" in this project. Then I convert `quality` into a factor variable with binary responses. There is no missing data.

After the data cleaning, I split the dataset into two parts: 70% of it goes into the training data, and 30% goes into the test data. The training data contains 1120 observations and the test data contains 479 observations.

Exploratory analysis

The dataset has 1599 observations and 12 variables. The outcome variable is quality, and the predictors are fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, p_h, sulphates, alcohol. Among these variables, only the response variable, quality, is categorical, and the others are continuous.

The feature plot contains the overlayed density plots for each predictor is shown in Figure.1. We can see that a better quality score might be associated with a higher amount of sulphates, higher percent alcohol content, a lower amount of chlorides, lower volatile acidity, and a higher amount of citric acid.

Models

I keep all 11 variables in the original dataset as predictors. Since we want to predict a binary response variable, either "good" or "not good", it is appropriate to select classification models on our dataset and use ROC and AUC as our model evaluation metrics after training. In this project, I choose logistics regression, penalized logistics regression, generalized additive model (GAM), multivariate adaptive regression splines (MARS), and linear discriminant analysis (LDA) models to train the data for classification with 5-fold cross-validation. Fitting each model requires different assumptions. The logistic regression model assumes that the observations are independent; A linear relationship between predictors and the logit of the response variable is also required; the logistic regression also assumes that no multicollinearity exists among predictors, and the fourth assumption is that no extreme outliers are present. GAM and MARS models do not make assumptions. LDA model assumes equal variance.

Logistic regression and LDA does not have tuning parameters. Since I have 11 predictors, I choose to tune the order from 1 to 3 and prune from 8 to 15 and decide the final values through cross-validation for MARS. In the case of penalized logistics regression and GAM, the best tuning parameters are also selected via 5-fold cross-validations.

The boxplots of ROC of the five models I trained based on the training data are shown in Figure.2. By resampling our training data, among the five models, the GAM model is found to have the highest ROC value, which indicates that it has the best training performance at distinguishing between good quality wine and others.

Using the test data, new ROC curves and the corresponding AUC are shown in Figure.3. The GAM model is also found to have a smoother ROC curve which is closer to the upper left corner of the graph, and it has the highest AUC value (0.866), which indicates that it has the best testing performance. Since the GAM model has the highest ROC and AUC values for both training and test data, I would select the GAM model to be the optimal model to predict the quality level of the wine.

Based on the GAM model, at a significant level of 0.05, free_sulfur_dioxide, alcohol, residual_sugar, sulphates, volatile_acidity, total_sulfur_dioxide, and density appear to be statistically significant in predicting response. Partial dependence plots of each continuous predictor in the model are shown in Figure.4, reflecting the marginal effects of each predictor.

These partial dependence plots show similar trends that free_sulfur_dioxide, alcohol, residual_sugar, sulphates, volatile_acidity, total_sulfur_dioxide, and density might contribute more significantly to the response variable. The adjusted R^2 of the GAM model is 0.419. The AUC of the model is 86.6%, suggesting that there is an 86.6% chance that the model will distinguish between good quality and not-good quality correctly.

One disadvantage of GAM is that the model is less interpretable. In a GAM model, each predictor has converted into a smooth covariate which has a nonlinear function, which complicates the interpretation of each parameter.

Conclusions

In conclusion, the GAM model is selected to be the best model to predict the quality level of wine because of its highest ROC and AUC value. Similar to what we expected in the exploratory analysis, alcohol, sulphates, volatile_acidity, total_sulfur_dioxide, and density are statistically significant predictors for predicting the quality level; besides, free_sulfur_dioxide and residual_sugar are also important variables. The model provides a good reference for the wine industry on the quality assessment of wine during the wine-making process.

Appendix

Figure.1

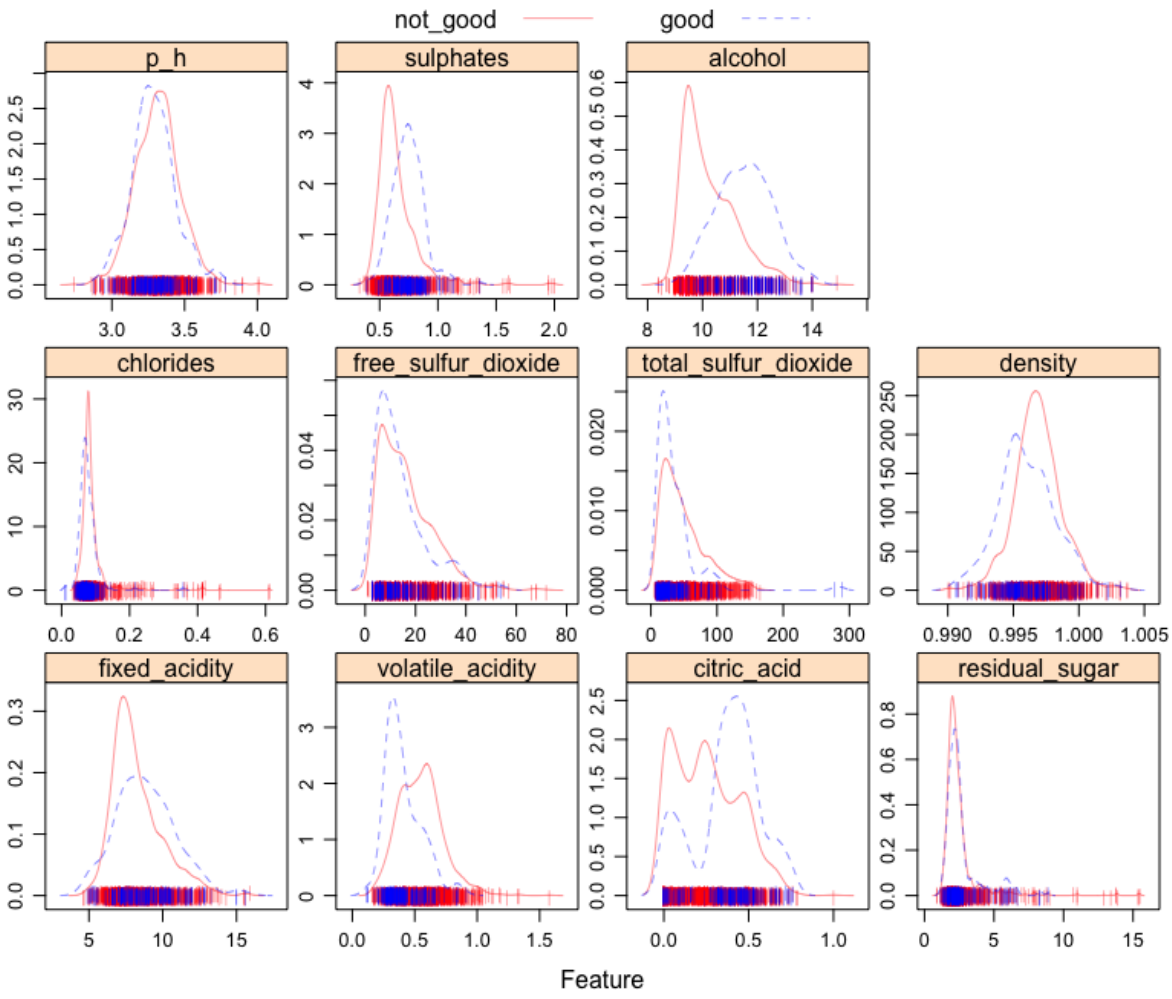


Figure.2

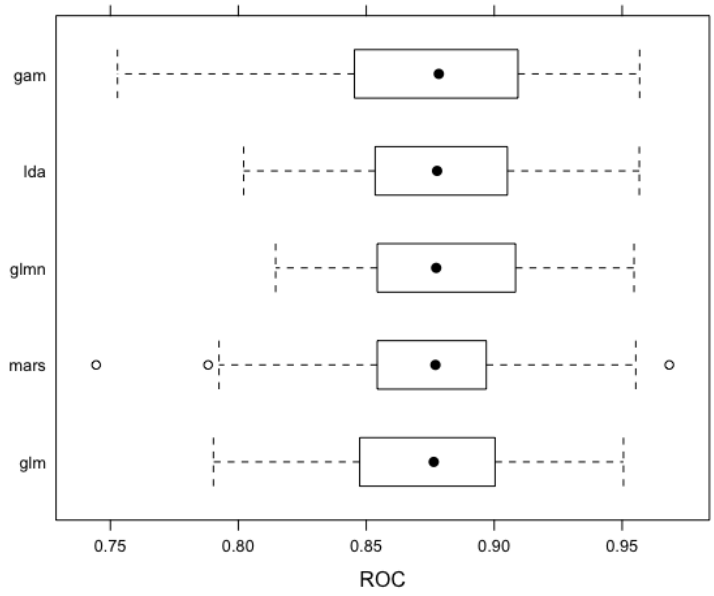


Figure.3

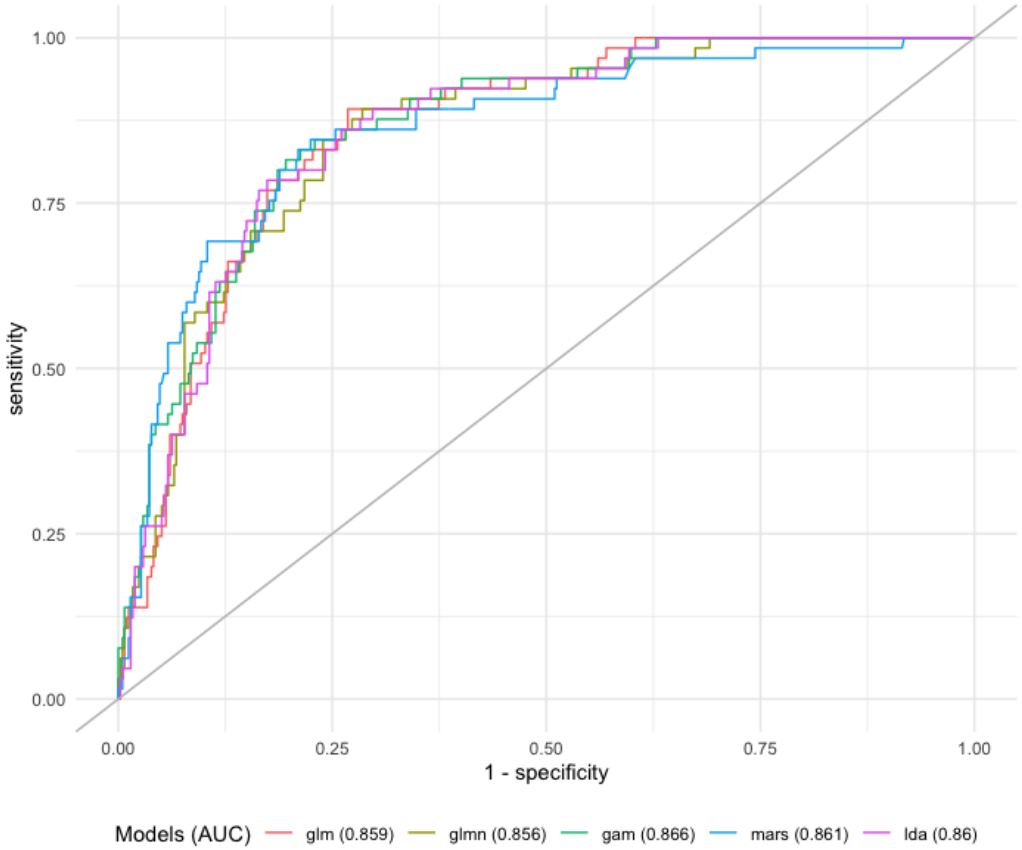
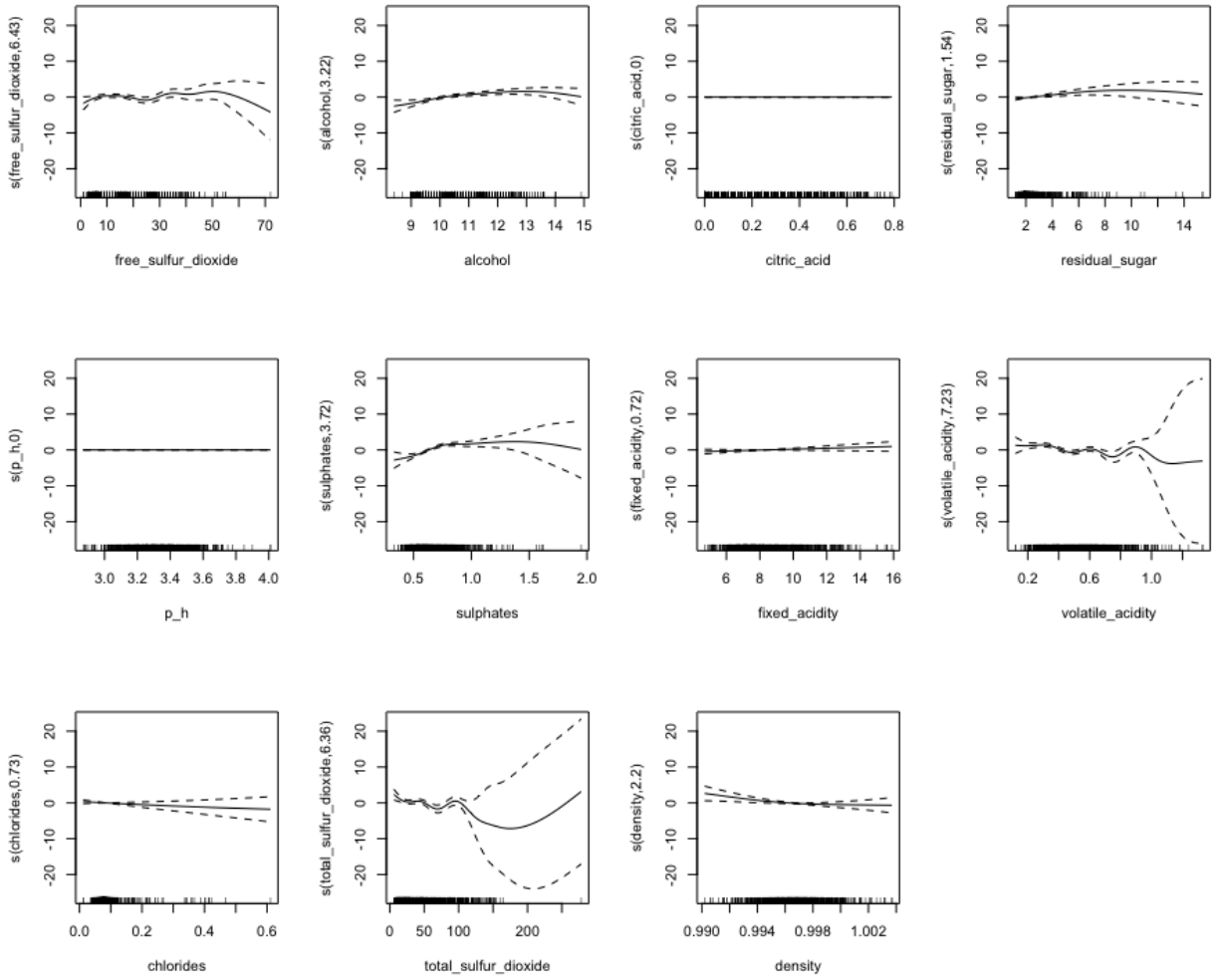


Figure.4


s(free_sulfur_dioxide,6.43)
free_sulfur_dioxide
s(alcohol,3.22)
alcohol
s(citric_acid,0)
citric_acid
s(residual_sugar,1.54)
residual_sugar
s(p_h,0)
p_h
s(sulphates,3.72)
sulphates
s(fixed_acidity,0.72)
fixed_acidity
s(volatile_acidity,7.23)
volatile_acidity
s(chlorides,0.73)
chlorides
s(total_sulfur_dioxide,6.36)
total_sulfur_dioxide
s(density,2.2)
density