Final

Team 10: Yuqi Wang, Rongqing Jia, Xiaoyu Li

5/2/2021

Introduction

The data source we use for this final project comes from the UCI machine learning repository. It contains information regarding the red and white variants of the Portuguese "Vinho Verde" wine. The dataset has 1599 observations and 12 variables, which are the fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. The fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol are independent variables and are continuous. Quality is the response variable and is measured based on a score of 0 to 10. Later, we re-categorized quality to a binary variable called qual. Qual is considered good if the quality score is greater than 5, otherwise is considered poor.

By doing this project, we hope to classify the quality of each observation into either good or poor based on their performance on the physicochemical tests.

Exploratory analysis

In total 855 wines were classified as "Good" quality and 744 as "Poor" quality. The average values for the 11 features for wines of good and poor quality was shown in Table 1. Fixed acidity, volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulphates and alcohol were significantly associated with the wine quality (P-values for t-tests < 0.05), which suggests important predictors.

We also built the density plots to explore the distribution of the 11 continuous variables over "Poor" and "Good" quality of wine (Figure 1). The plots showed that wine with good and poor quality did not differ for PH and residual sugar, while different types of wine differs in other variables, which was consistent with the t-test results.

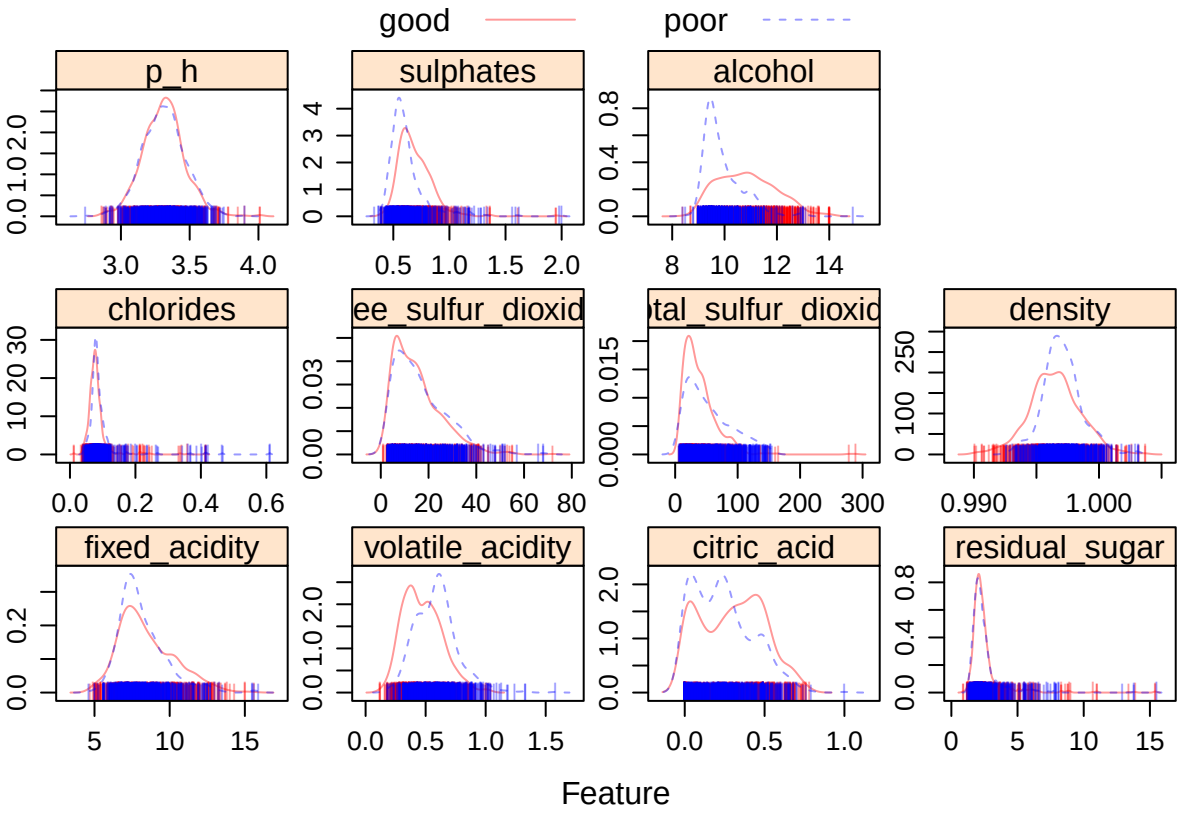

Figure 1. Descriptive plots between wine quality and predictive features.

Table 1. Basic characteristics of wines over good and poor quality.

	Stratified by qual			
	good	poor	p	test
n	855	744		
fixed_acidity (mean (SD))	8.47 (1.86)	8.14 (1.57)	<0.001	
volatile_acidity (mean (SD))	0.47 (0.16)	0.59 (0.18)	<0.001	
citric_acid (mean (SD))	0.30 (0.20)	0.24 (0.18)	<0.001	
residual_sugar (mean (SD))	2.54 (1.42)	2.54 (1.39)	0.931	
chlorides (mean (SD))	0.08 (0.04)	0.09 (0.06)	<0.001	
free_sulfur_dioxide (mean (SD))	15.27 (10.04)	16.57 (10.89)	0.014	
total_sulfur_dioxide (mean (SD))	39.35 (27.25)	54.65 (36.72)	<0.001	
density (mean (SD))	1.00 (0.00)	1.00 (0.00)	<0.001	
p_h (mean (SD))	3.31 (0.15)	3.31 (0.15)	0.896	
sulphates (mean (SD))	0.69 (0.16)	0.62 (0.18)	<0.001	
alcohol (mean (SD))	10.86 (1.11)	9.93 (0.76)	<0.001	

Model Building

We randomly selected 70% of the observations as the training data and the rest as the test data. All of the 11 predictors were included into analysis. We performed linear methods, non-linear methods, the tree method and SVM to predict the classification of wine quality. For linear methods, we trained (penalized) logistic regression model and linear discriminant analysis (LDA). The assumptions for logistic regression includes observations being independent of each other and the linearity of independent variables and log odds. LDA and QDA assumes normally distributed features, that is, predictor variables are normally distributed for both "good" and "poor" quality of wine. For nonlinear models, we performed generalized additive model

(GAM), multivariate adaptive regression splines (MARS), KNN model and quadratic discriminant analysis (QDA). For tree models, we conducted classification tree and random forest model. SVM with linear and radial kernels were also performed. We calculated the ROC and accuracy for model selection, and also investigated the variable importance. 10-fold cross-validation (CV) were used for all model buildings.

Linear models The multiple logistic regression showed that among the 11 predictors, volatile acidity, citric acid, free sulfur dioxide, total sulfur dioxide, sulphates and alcohol were significantly associated with wine quality (P-values < 0.05), explaining 25.1% of the total variance in wine quality. When applying this model to the test data, the accuracy is 0.75 (95%CI: 0.71-0.79) and the ROC is 0.818, which suggests relatively good fit for the data. When performing the penalized logistic regression, we found that when maximizing the ROC, the best tuning parameter was alpha=1 and lambda=0.00086, the accuracy was 0.75 (95%CI: 0.71-0.79) and the ROC was also 0.818. Since lambda was close to zero and the ROC was the same as the full logistic regression model, the penalization was relatively small, which suggested that the full logistic regression model was simple enough for classification.

However, since logistic regression requires there to be little or no multicollinearity among the independent variables, the model may be disturbed by collinearity between the 11 predictors, if there was any. As for LDA, when applying the model to the test data, the ROC was 0.819 and the accuracy was 0.762 (95%CI: 0.72-0.80). The most important variables in predicting wine quality were alcohol, volatile acidity and sulphates. Compared to the logistic regression models, LDA is more helpful when the sample size is small or when the classes are well separated, under the condition that normal assumptions are met.

Nonlinear models In the GAM model, only the degree of freedom for volatile acidity was equal to 1, suggesting linear association, while smoothing spline was applied for all other 10 variables. The results showed that alcohol, citric acid, residual sugar, sulphates, fixed acidity, volatile acidity, chlorides and total sulfur dioxide were significant predictors (P-values < 0.05). In total, these variables explained 39.1% of the total variance in wine quality. The confusion matrix using the test data showed that the accuracy for GAM was 0.76 (95%CI: 0.72-0.80) and the ROC was 0.829. The MARS model showed that when maximizing the ROC, we included 5 terms out of 11 predictors, with nprune equal to 5 and degree of 2. In total, these predictors and hinge functions explained 32.2% of the total variance. According to the MARS output, the 3 most important predictors were total sulfur dioxide, alcohol and sulphates. When applying the MARS model to the test data, the accuracy is 0.75 (95%CI: 0.72, 0.80) and the ROC is 0.823. We also performed the KNN model for classification. When k was equal to 22, the ROC was maximized. The accuracy for KNN model was 0.63 (95%CI: 0.59-0.68) and the ROC was 0.672. The QDA model showed that ROC was 0.784 and the accuracy was 0.71 (95%CI: 0.66-0.75). The most important variables in predicting wine quality are alcohol, volatile acidity and sulphates.

The advantage of GAM and MARS is that both two models are nonparametric models and able to deal with highly complex nonlinear relationship. Specifically, MARS model can include potential interaction effects into the model. However, because of the model complexity, time-consuming computation and the high propensity of overfitting are the limitations for the two models. As for the KNN model, when k was large, the prediction may not be accurate.

Tree Methods

Based on the classification tree, the final tree size is 41 when maximizing the AUC. The test error rate is 0.24 and ROC is 0.809. The accuracy of this classification tree is 0.76 (95%CI: 0.72-0.80). We also conducted the random forest method to investigate the variable importance. As a result, alcohol is the most important variable, and followed by sulphates, volatile acidity, total sulfur dioxide, density, chlorides, fixed acidity, citric acid, free sulfur dioxide, and residual sugar. pH is the least important variable. For the random forest model, the test error rate is 0.163, the accuracy is 0.84 (95%CI: 0.80-0.87), and the ROC is 0.900.

One potential limitation for the tree methods is that they are sensitive to the change in data, that is, a small change in data may cause a large change of the classification tree.

SVM

We used SVM with linear kernal and we tuned over cost. We found that the model with maximized ROC had cost = 0.59078. ROC for this model is 0.816, accuracy is 0.75 (test error is 0.25) (95%CI: 0.71-0.79). The most important variable for quality prediction is alcohol; volatile acidity and total sulfur dioxide are also relatively important variables. When performing SVM with radial kernal, we tuned over both cost and sigma, and found that the model with maximized ROC had sigma = 0.0286 and cost = 17.9733. ROC for this model is 0.821, accuracy is 0.75 (test error is 0.25) (95%CI: 0.71-0.79). Same as SVM using linear kernal, the most important variable for quality prediction is alcohol; sulphates and volatile acidity are the second and third most important variables. If the true boundary is non-linear, SVM with radial kernel performs better.

Model Comparison

After model building, we conducted model comparisons based on the training and test performance of all models. The following tables shows the cross-validation classification error rates and ROCs of all the models. In the results, random forest model has the largest AUC value, while KNN has the smallest. Therefore, we selected the random forest model as the best predictive classification model for our data. Based on the random forest model, alcohol, sulphates, volatile acidity, total sulfur dioxide and density are the top 5 important predictors that help us predict the classification of wine quality. Since factors such as alcohol, sulphates and volatile acidity are the ones that may determine the flavor and taste of wines, so such findings meet our expectation.

While looking at the summary of each model, we realize that KNN model has the lowest AUC value and the largest test classification error rate, 0.367. The other nine models have close AUC values that are about 82%.

Table 2. Model comparison of ROCs and R square of all models

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
logistic	0.7512821	0.7829327	0.8112179	0.8141979	0.8472888	0.8769231
penalized_logistic	0.7522436	0.7849359	0.8123397	0.8146170	0.8463454	0.8753205
LDA	0.7477564	0.7834135	0.8134615	0.8141959	0.8463282	0.8772436
GAM	0.7782051	0.7967949	0.8243590	0.8302269	0.8598512	0.8872229
MARS	0.7855769	0.8127653	0.8300481	0.8315484	0.8432692	0.8871795
knn	0.6456731	0.6965946	0.7371863	0.7222954	0.7475742	0.7737179
QDA	0.6971154	0.7500000	0.8036859	0.7912686	0.8302099	0.8512821
ClassTree	0.7546474	0.7943109	0.8202724	0.8058090	0.8245708	0.8274038
RandomForest	0.8397436	0.8540865	0.8700895	0.8754754	0.8869391	0.9349359
SVML	0.7490385	0.7879006	0.8099359	0.8148223	0.8523059	0.8705128
SVMR	0.8060897	0.8112981	0.8203526	0.8398396	0.8709089	0.9041667

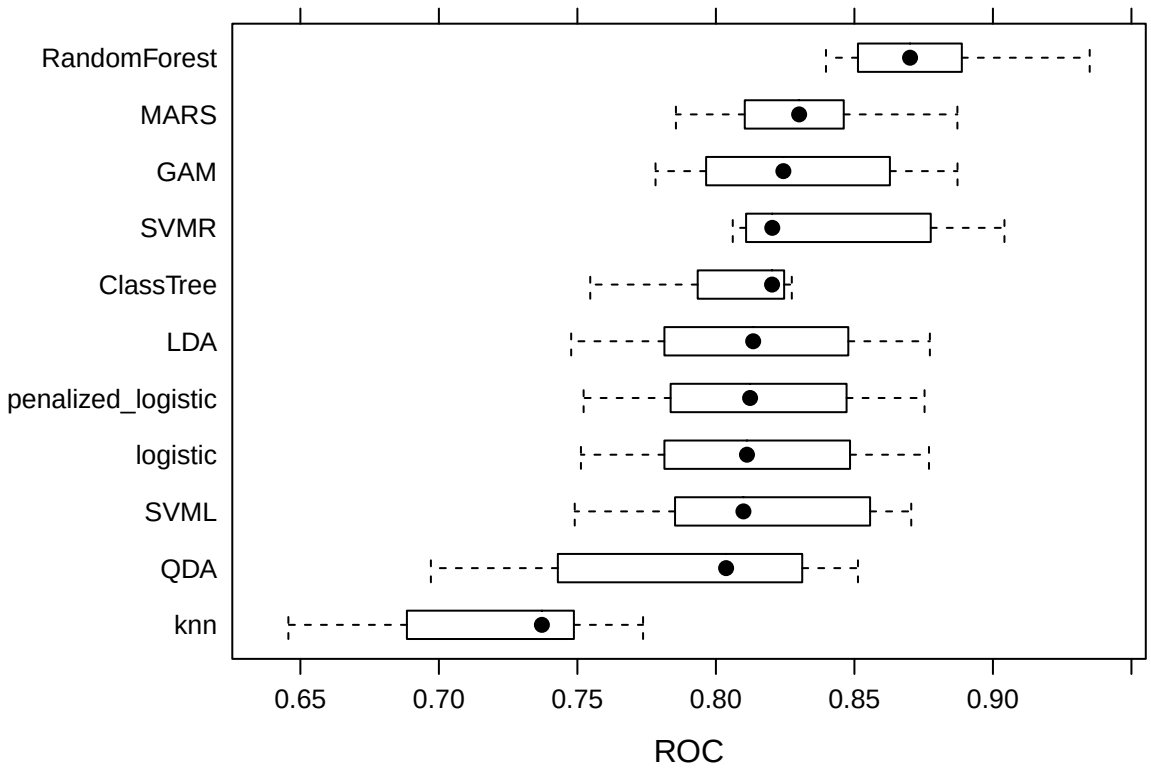

Figure 2. Model comparison: ROC values for all models

Table 3. Test classification error rates, train classification error rate and test ROC values for all models

Model_Name	Train_Error	Test_Error	Test_ROC
logistic regression	0.2553571	0.2505219	0.818
penalized logistic regression	0.2553571	0.2463466	0.818
LDA	0.2571429	0.2379958	0.819
GAM	0.2151786	0.2400835	0.829
MARS	0.2321429	0.2463466	0.823
KNN	0.2928571	0.3674322	0.672
QDA	0.2526786	0.2922756	0.784
Classification Tree	0.1508929	0.2421712	0.809
Random forest	0.0000000	0.1628392	0.900
SVM with linear kernel	0.2589286	0.2505219	0.816
SVM with radial kernel	0.1866071	0.2526096	0.821

Conlcusion

The process of model building shows that in the training dataset, alcohol, sulphates, volatile acidity, total sulfur dioxide and density are the top 5 important predictors for classification of wine quality. We selected the random forest model because of its largest AUC value and lowest classification error rate. This model also performs well in the test dataset. Therefore, this random forest model is an effective method for classification of wine quality.