March 8, 2016
Ilker Karakasoglu

Predicting Wine Quality

Problem description:

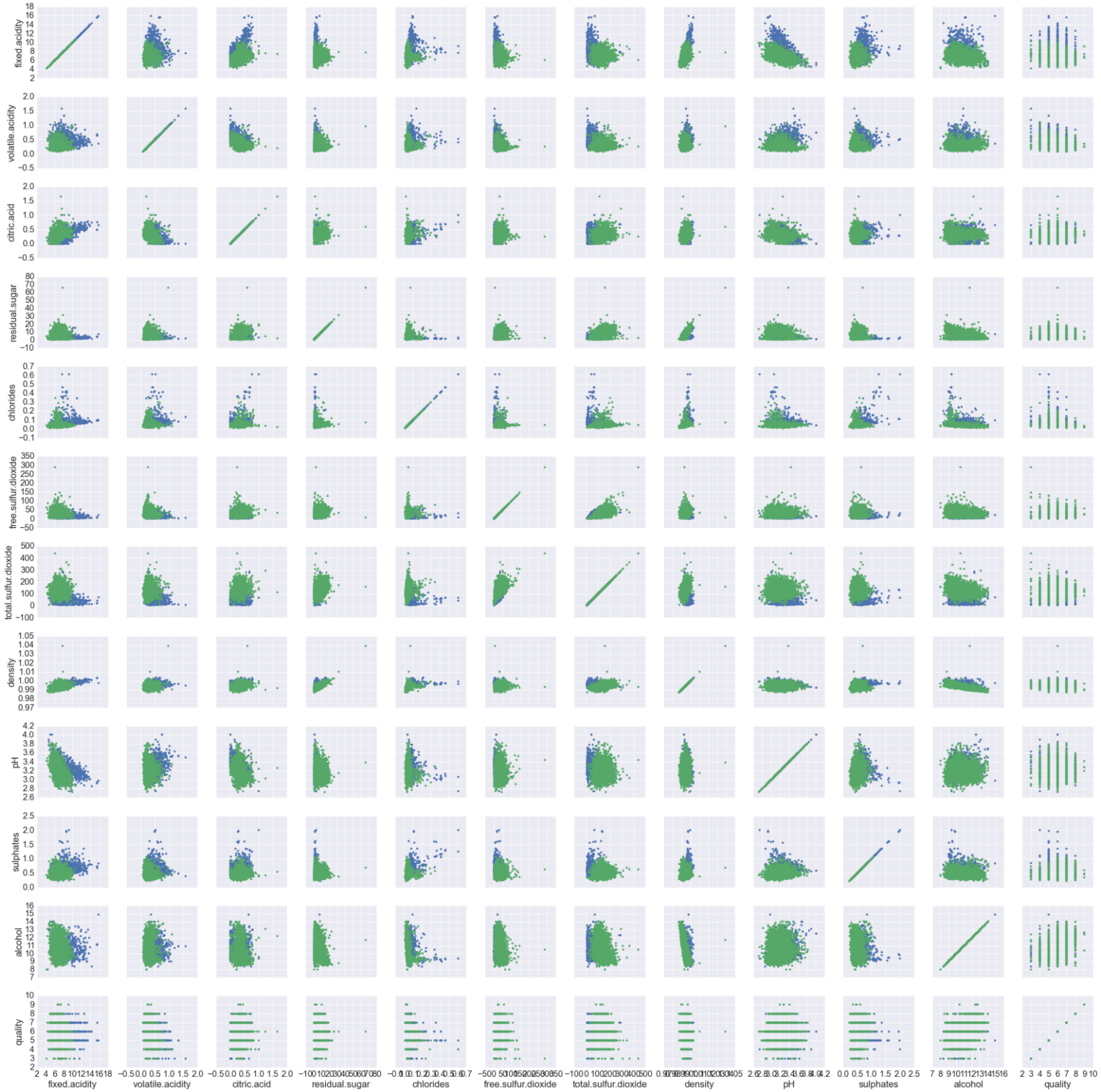
You have been retained as a statistical consultant for a wine co-operative, and have been asked to analyze these data. Each row represents data on a particular Portuguese wine, and the columns are attributes. The last column is the response quality, which is a quantitative (integer) score between 0 (very bad) and 10 (excellent) made by wine experts (in our data there was no wine lower than a 3, and none higher than 9). Your clients are interested in predicting the quality score based on the attributes. They would also like to get some sense of which attributes are more important for this task, and their role in the prediction procedure.

The file wine.test.ho.csv consists of 1300 wines where the quality score is omitted. Use your model to predict the quality score for each of these wines.

Solution:

We first visualize the data to get a better understanding of it. Below is a pairplot which illustrates all variables and output quality plotted among each other. We color code red wines with *green* and white wines with *blue*. We observe two things.

First, white and red wines do not exactly have same attributes. In many variables, there are distinctions between them. They even look separable. Therefore, it may make more sense to predict the wine quality separately for reds and whites. Second, there are strong correlations among certain variables. Some of these correlated variables can be left out.

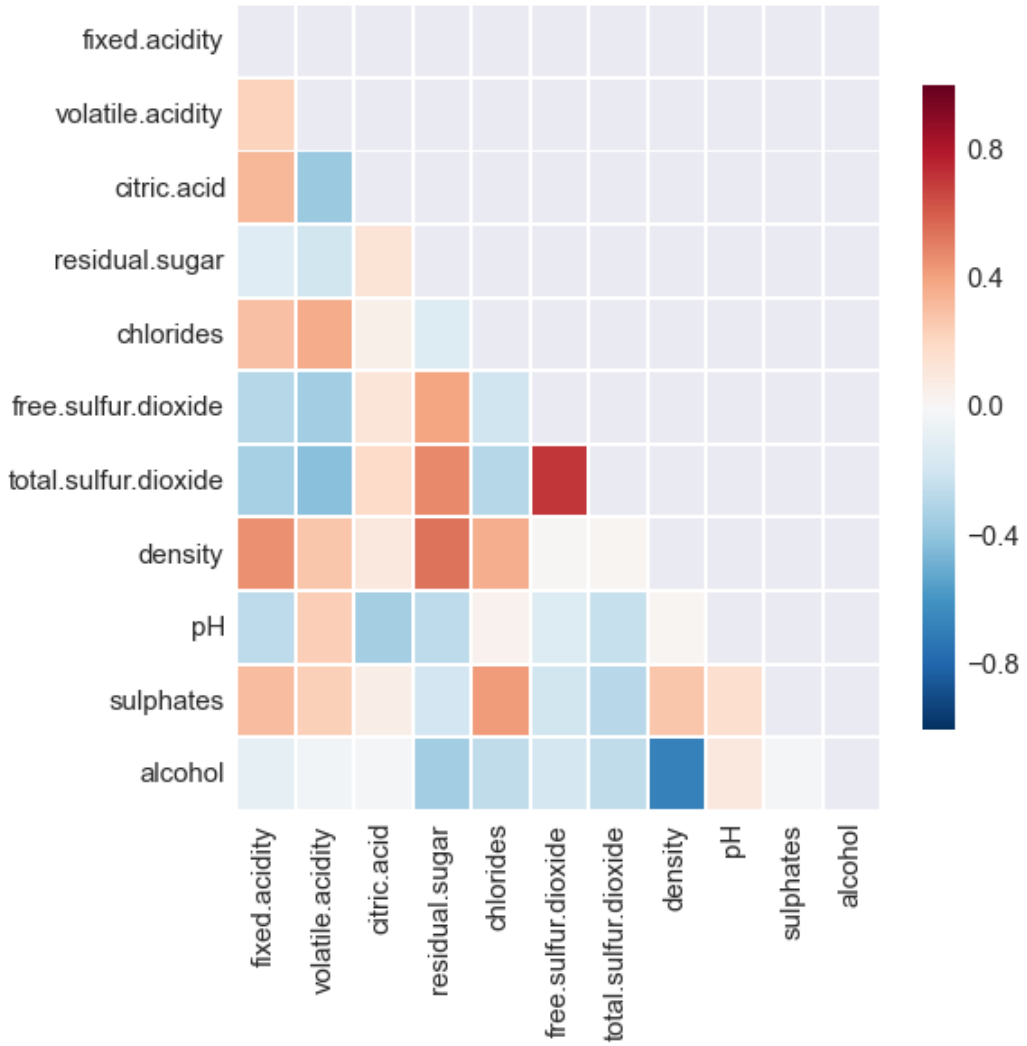


In order to better see the correlations, a heatmap of correlations is illustrated below. Warm colors indicate a positive correlation, while cold colors indicate a negative one. In an ideal case of all variables being normal to each other, we would get a white map.

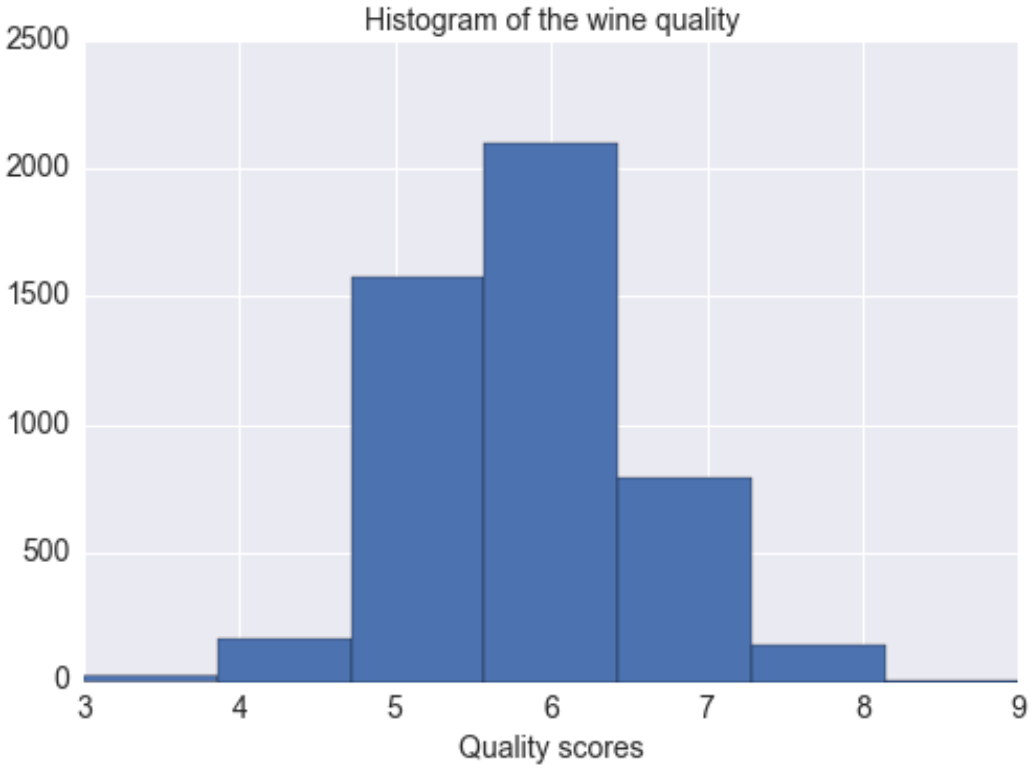
In this one, we observe various amounts of correlations among variables. For instance, there is a very strong positive correlation between total sulfur dioxide and free sulfur dioxide. Similarly, density is positively

correlated with fixed acidity and residual sugar. On the other hand, it has a strong negative correlation with alcohol.

All these correlations intuitively make sense. Strong correlations mean that these correlated variables should be handled carefully in a learning model. Possibly, some of them can be dropped out depending on which ones carry the highest importance.



Next, we visualize the distribution of the outcome variable- wine quality. We see a strong concentration on average wine scores- 5 and 6. This unbalanced distribution is a challenge for a learning model, since most predictions would center around 5 and 6. So other scores may get harder to predict.



Having seen the basic properties of the data, we continue with developing a learning model for quality prediction. We prefer a regression instead of a classification because the quality is inherently ordered. After regressing, we round up the scores to the closest digits.

We split the training set into to get a smaller training set and a validation set. We use the validation set to judge our performance on the real test set.

The algorithm we choose is Support Vector Machines for Regression (SVR). SVR has four parameters to choose. First one is the kernel. We use a radial kernel (rbf). This kernel trick is a powerful method used to transform input data into a higher dimensional space while not increasing the computational cost. We prefer the radial kernel over the other popular choice linear kernel after seeing the better performance of the radial kernel in this data set. The other three parameters are gamma, epsilon and C. Each affects the bias-variance tradeoff. Higher values of gamma makes the radial kernel more localized. The kernel doesn't expand much onto all data points. It rather only sample around the given observation. The higher the gamma, the less bias but the more variance we would get. Epsilon determines the "epsilon-insensitive" where there are at most epsilon deviations from the actually obtained target values for all the training data. The higher the epsilon, the less variance we can get. C is the cost parameter, which is positive and controls the tradeoff between the model complexity and the amount up to which deviations greater than epsilon are tolerated. Similar to the epsilon, the higher C leads to less variance.

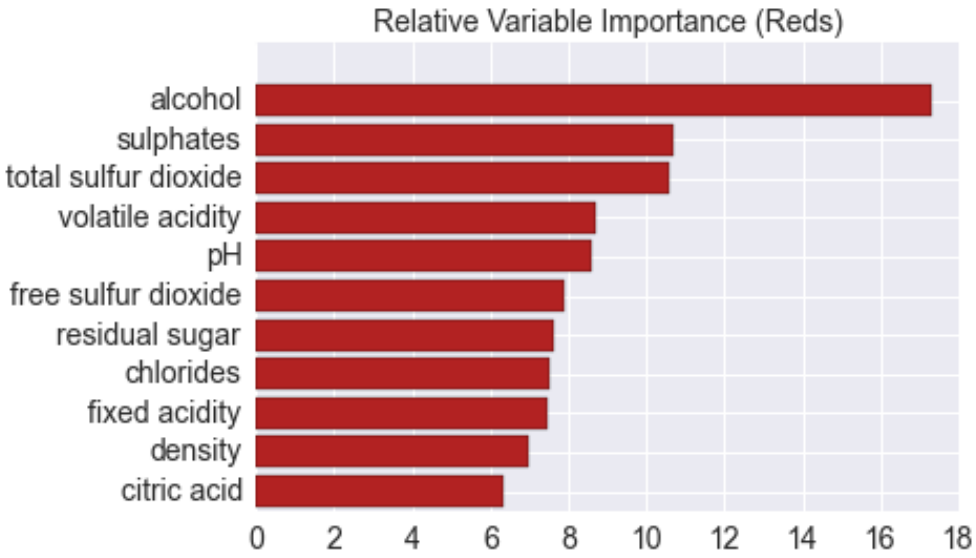
The radial kernel is a nonlinear and flexible one. Therefore, it may give rise to overfitting. To avoid this problem, other parameters- gamma, epsilon and C- must be chosen carefully. We do this using cross validation. Since there are 3 parameters, we use a grid of them to choose the best tuple. We employ the scikit-learn's GridSearchCV function, which does an exhaustive search to find the tuple giving the highest cross validation score. 5-fold cross validation is used. We test for the values of C: [0.1, 1, 3, 10], gamma: [0.001, 0.01, 0.1, 1], epsilon: [0.01, 0.1, 1]. The best tuple giving the highest cross validation score is {epsilon: 0.1, C: 10, gamma: 0.01.}.

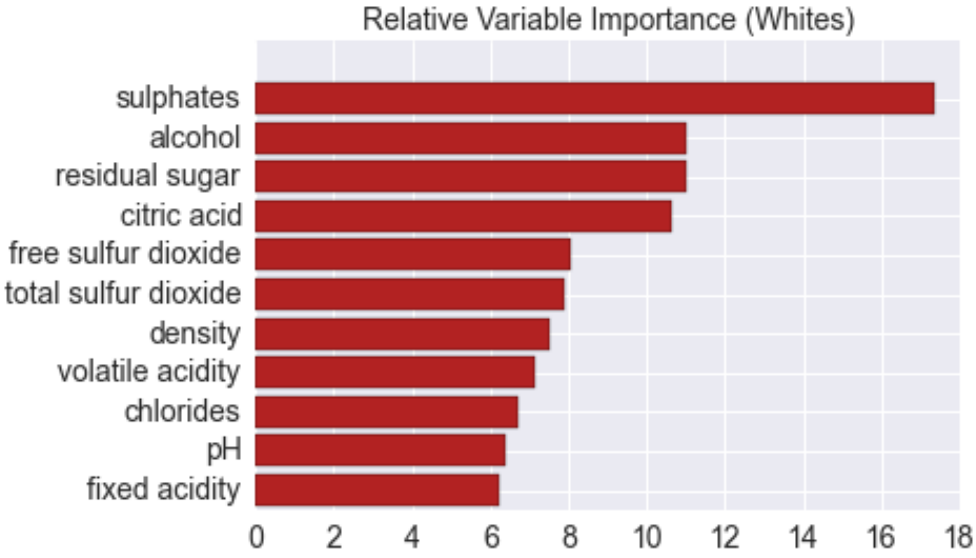
We run 3 different regressions: for reds, for whites and for two combined. CV scores indicate that separate regressions for reds and whites give better results. Still the combined set performance is close. In the combined regression we use the color of the wine as dummy variable.

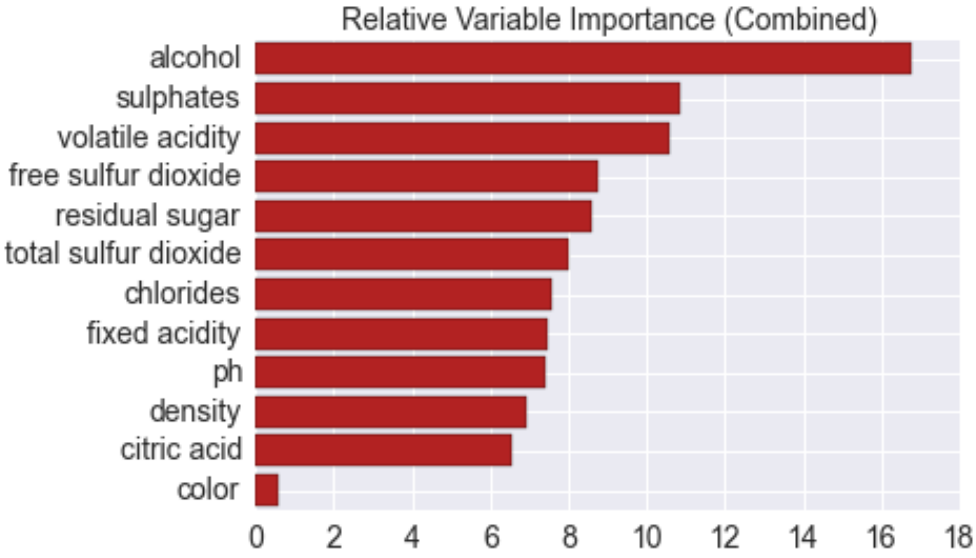
We obtain a validation set score of RMS = 0.71.

Next we assess the variable importance. To do that, in built scikit-learn property of feature_importances_ is used. This is a function which automatically ranks the variables in terms of their significance. Relative importance a variable is assessed by the high variance it produces in data. We plot 3 importance charts for the 3 regressions. In the combined one we see that the color is not significant at all. This explains the combined dataset's close performance to reds and whites alone. Alcohol and sulphates are the most important variables. Density and pH are not very important and previously they were found to be correlated with other variables. Therefore, they can be good candidates to be dropped from the feature space.

A helpful analysis would the confusion matrix rather than the RMS to assess the model (although we haven't implemented yet). As previously shown on the histogram, outcome variable is very skewed around 5 and 6. Therefore, it is likely that scores other than 5 and 6 will have lower recall and precision rates. F1 score and ROC curves could be helpful to summarize the precision and recalls considerations.


Relative Variable Importance (Reds)
alcohol
sulphates
total sulfur dioxide
volatile acidity
pH
free sulfur dioxide
residual sugar
chlorides
fixed acidity
density
citric acid
0
2
4
6
8
10
12
14
16
18


Relative Variable Importance (Whites)
sulphates
alcohol
residual sugar
citric acid
free sulfur dioxide
total sulfur dioxide
density
volatile acidity
chlorides
pH
fixed acidity
0
2
4
6
8
10
12
14
16
18


Relative Variable Importance (Combined)
alcohol
sulphates
volatile acidity
free sulfur dioxide
residual sugar
total sulfur dioxide
chlorides
fixed acidity
ph
density
citric acid
color
0
2
4
6
8
10
12
14
16
18