Predicting Wine Quality

Sandra & Chris & Evan

12/8/2022

Contents

Executive Summary:

We aimed to predict the quality of 4547 red and white vinho verde wine samples, from the north of Portugal. Using a data set of 1950 wines, we analysed the data to determine the best subset of predictors and fit them to several different models. Using RMSE as a measure of model performance, we chose a Lasso model to make the final predictions.

Background

Background and Goals:

We received data on 1464 red and white vinho verde wine samples, from the north of Portugal. We have been given 13 different variables to help predict quality of wine (0-10 scale). Note that although the scale of wine is 0-10, all observations have a scale from 3-9.

Our goal is to build a model that will accurately predict wine quality from an additional 4547 red and white vinho verde wine samples, from the north of Portugal.

Analysis

We began by splitting the data given to us into a training and test set, this resampling of data was done randomly. The training set has 75% of our data and the test set has the remaining 25%. We will be trying to create a model based on the 75% of the data to predict the final 25%. Once we have a plausible model that accurately predicts the remaining 25% of our data, we will fit the model on the entire data set and predict wine quality of the remaining 4547 wine samples that we have not touched.

Visualization

To explore our data, we started with visualization. This gives us some understanding of the relationships existing within all the variables. We will observe characteristics that will provide valuable information, as well as checking if there exists some non-linear relationships. Given that they exists, the appropriate transformations will be applied to the respective variable.

Correlations

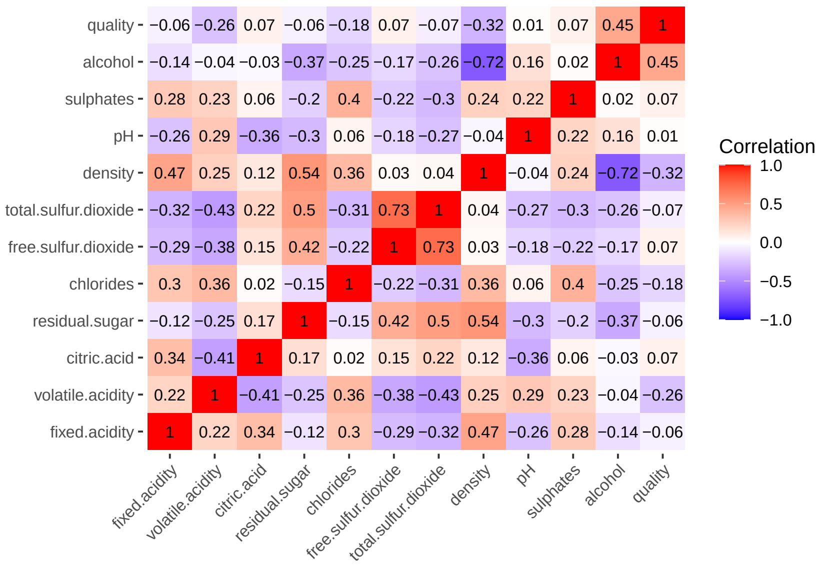

Figure 1: Correlation heatmap.

Figure 1 displays the correlations between each pairwise combination of variables. The variable accounting for wine type is categorical- thus not included in calculating the correlations. We observe variables having multiple notable relationships with other variables; `residual sugar`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, and `alcohol`. For our response variable, `quality` is most related to `alcohol`. These relationships are kept in mind as we proceed forward in our data analysis.

Input Variable versus Output Variable

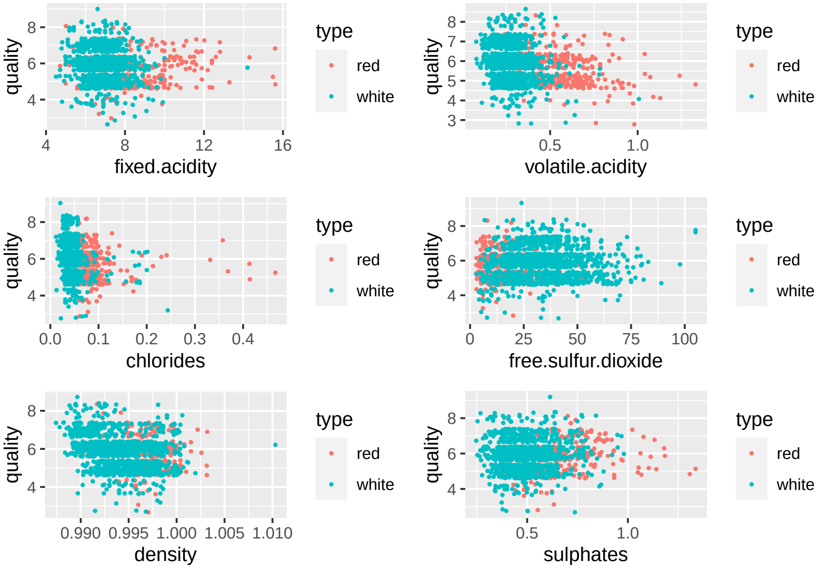

Figure 2: Variables with non-linear relationships.

Scatter plots for all the input variables against `quality` were visualized. Figure 2 includes plots for variables we suspect might have a possible non-linear relationship. Although we suspect a possible fault in our observation, we deemed a quadratic relationship to be fairly or most present within these selected variables.

Base Model

A linear model was trained using all input variables, with no transformation, to seek a CV RMSE value that will be referred to as the error we are trying to improve. The error we will attempt to improve is **0.7196**.

Viewing the plots with variables we suspected had non-linear relationships, we observe some pattern. Although, we still caution that our observations may not be accurate. Now, we will proceed implementing the transformations from our initial observation.

Transformations

Following the base model, we included the addition of quadratic variables that were previously mentioned. Using this model with transformed variables, we plotted the residuals to observe any leftover patterns as well as interactions. The plots for each variables were separated by the observation's wine type to show any distinction between them that can provide an indication of interaction.

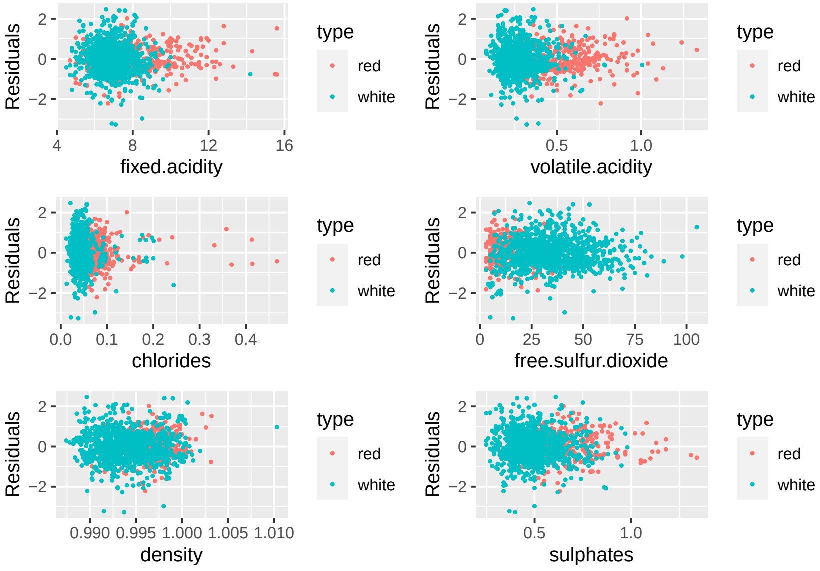

Figure 3: Residual plots from base model.

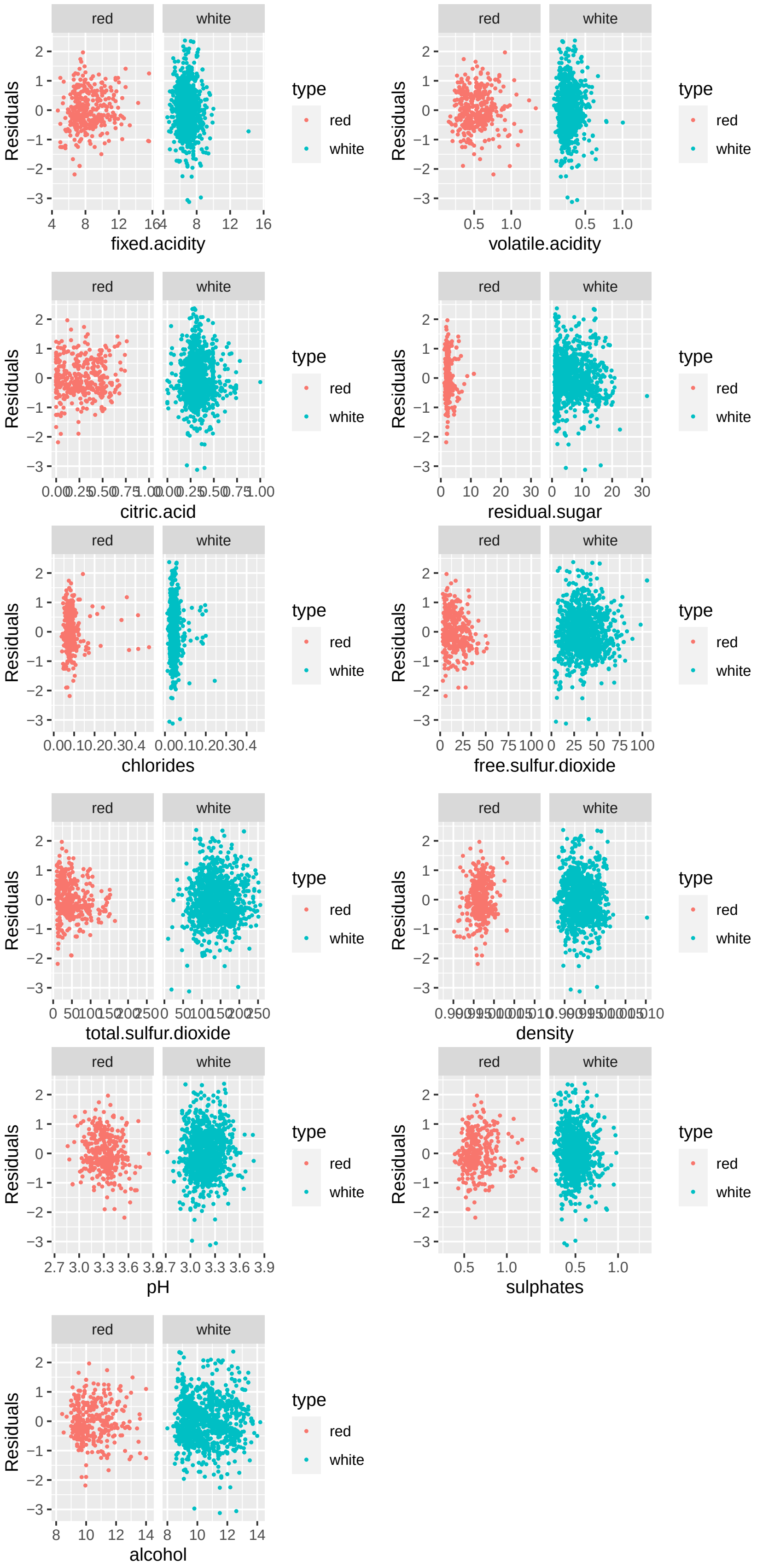

Figure 4: Residual plots from model with transformed variables.