Wine quality Predictions

Dieudonne Ouedraogo

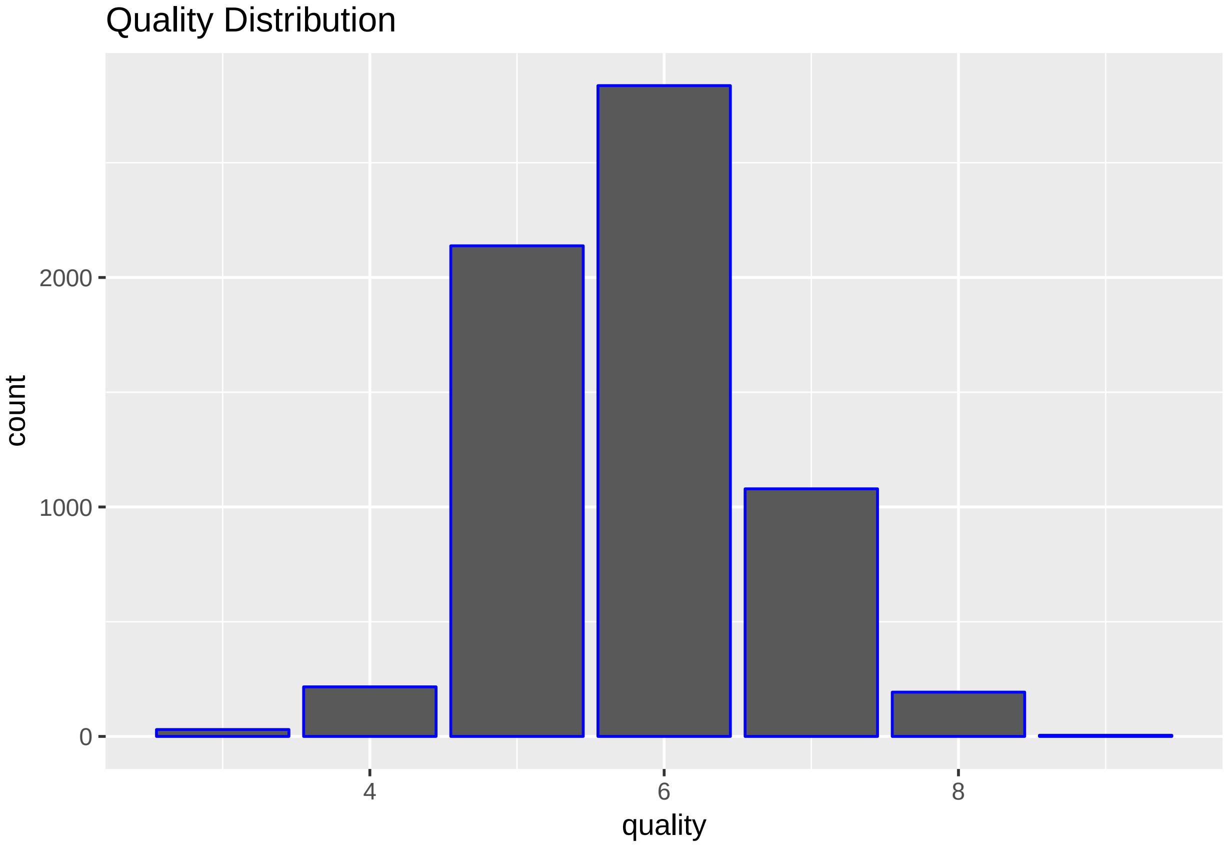
3/22/2019

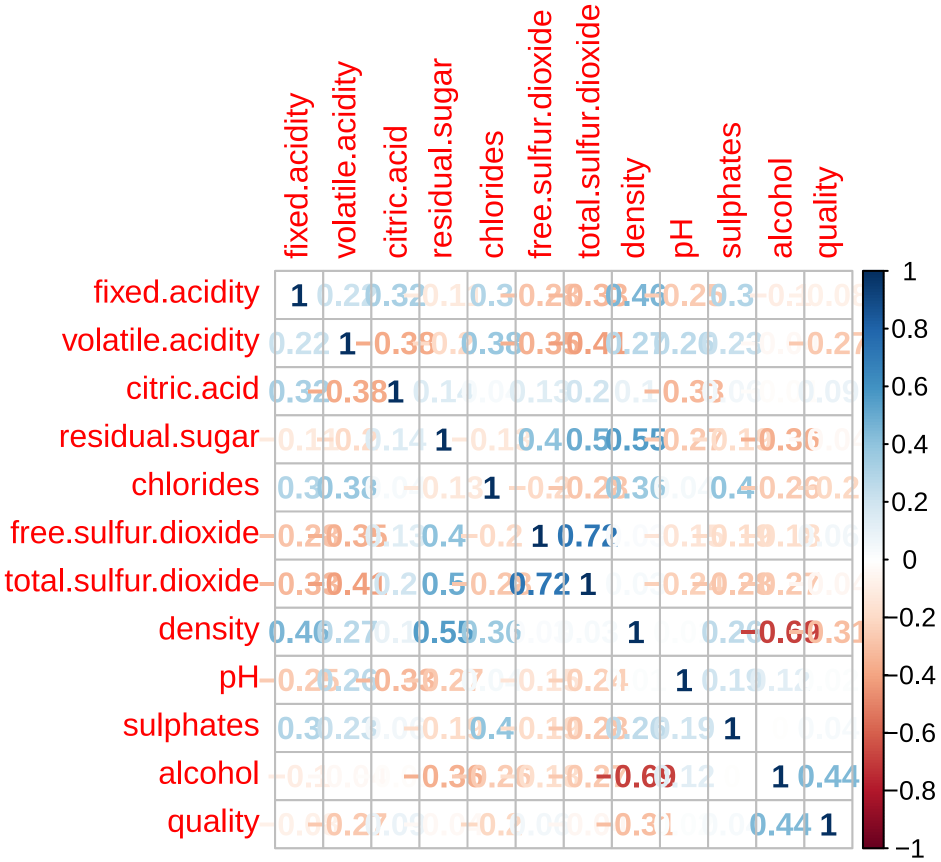
1. Introduction

This project aims to train plsRglm models to predict wine quality score. There are two data sets; red and white wine types. There eleven features in each dataset which are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol; and one output variable which is the quality of a wine.

2. Data Exploration

All features are of real, no strings, factors, and categories. So, no need to create dummy variables or to convert variables to numerical. There are no missing values, so no need to impute data. Also, no feature has zero or near-zero variance. There are outliers in the data where we do a boxplot on features. The target 'quality' is numeric; the data is appropriate for regression analysis. The red wine data and the white data are combined to produce a unique dataset.





The correlations between variables are above.

3. Data processing and feature engineering

Outliers in datasets may have impact on model accuracy to some extent. However, outlier deletion is not always a good approach, if the data is highly dimensional with few observations it is not good to remove outliers. But in our case, the data has many observation and a relatively small amount of features so removal could either help in train time or accuracy. Outlier detection was done using a univariate approach.

Features have various metric units (mg/cmˆ3, g/dmˆ3 etc.). Also, they have different concentrations and therefore have varying ranges. So, normalizing the dataset will make plots consistent for visualization and avoid feature dominance.

4. Models

The two data sets are concatenated, and models are trained for both red and white wine types. The data is split into training and testing with the ratio 80% training and 20% testing.

4.1. Model without manual processing

Quality is regressed against all variables I ran a partial linear square model using cross-validation and tuning the upper parameters using a grid search. The MAE was ***0.5529662***

Quality is regressed against all variables.

It is worth noting that if we round the results of the predictions, the MAE decreases to ***0.5100154***

4.2. Model with scaled and centered data

When we scale and center the data we get the same MAE as ***0.5529662*** which indicates that using plsRglm method does a step of scaling and centering the data

4.3. Models with removal of outliers and feature engineering

In this step new features are engineered based on the correlation values; the data is then processed, we get a MAE ***0.559248***.

Another model with removal of outliers performs best I obtained a MAE of ***0.5529642***

4.4. Separate models for each dataset

In this case I trained two models on each dataset.

We get a MAE of ***0.4828223*** on the red wine dataset.

The MAE is ***0.587463*** on the white wine.

The quality of the white wine is difficult to predict when compare to the red wine

5. Conclusion

Removing outliers improve the prediction, using cross-validation gives the best model for this dataset, tuning the upper parameters using the grid option offers the best performance.

White wine quality is harder to predict than red wine quality; we can choose to use two different models for each type.

No variables are strongly correlated to the quality; the highest correlation is 0.4 and is between alcohol and the quality. So it is hard to make a strong inference of the quality of the wine based on the features present in the datasets.

Collecting and defining more characteristics for the wine is essential if we want the predictions to be more accurate. Sensor scientists should look into adding and collecting more features to make their conclusions robust.

It is also possible to increase the number of experts and to take their average on the quality, which will give more objective values.