WINE QUALITY ANALYSIS AND PREDICTION

- PRAKASH RANA

Introduction:

We need to predict the quality of red and white wine with respect to the given variables. At the first glance the variable provided looks fairly chemical composition in wine.

We are given two dataset containing red and white wine data. So let’s merge them together with row bind.

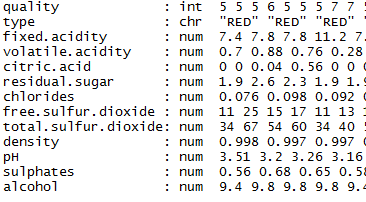
Before merging the data one ‘type’ column is created to differentiate red from white wine.

The column are also rearranged with type data at first. We have 12 variables and 1 target variable.

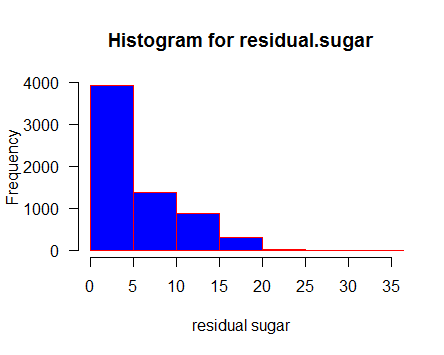
Now we will use this data for further analysis.

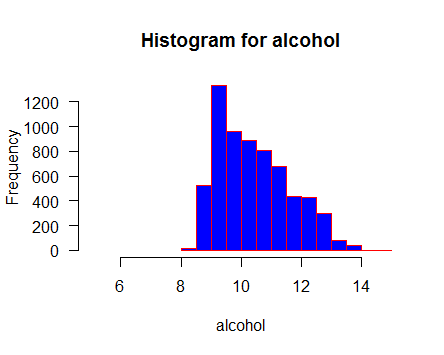
* Data Exploration:

If we see the structure of our data we can see that except the target variable ‘quality’ and variable ‘type’ all other are numeric (continuous)

.

Now the quality ranges from 3 to 9 with 7 levels so changing quality to factor. And also the ‘type’ to factor with red and white as the levels



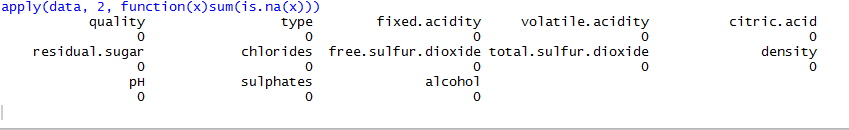


We can look at the distribution of data according to the variables. The data for residual. Sugar looks highly skewed.

Missing value analysis:

Now let’s check for missing values in each variable.

Here we will create a function which checks the number of missing values (NA) in each variable



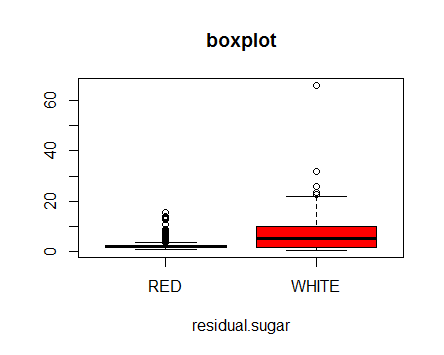
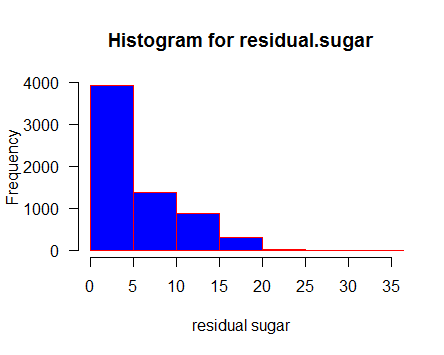
Here, we can see there is no missing values in any of the variable so we are good to go.

But may be the data contains empty values that are not detected by our function which checks only NA so we convert all empty data to NA and check again for missing values. Even after doing that we got no missing values so the data is complete.

Outlier Detection:

Outlier detection is done to remove the faulty data.

Let’s take the residual.Sugar and see the data distribution:



Now, if we observe the histogram and the boxplot of the ‘residual.sugar ‘. Even though the data for residual.sugar is positive skewed, we can see it contains an outlier. This is removed in next step.

Now we see for outlier in each variable and delete that observation if needed. We do the outlier detection before the normalization of data because the outlier could affect mean, min and max used in normalization.

For outlier detection we use the outlier function from the outliers’ library.

First a function ‘outt’ is created that detects outliers and deletes that observation.

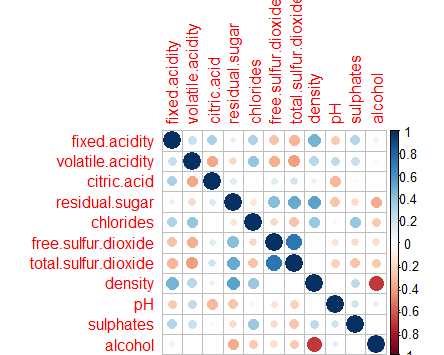
After that a loop is run for each variable except target variable. So the outt function is called in loop for each variable and outlier is detected.

If we observe, the number of observation after outlier detection reduces to 6484 from 6497

Feature Selection and correlation:

Here, we calculate the correlation between the features and observe if any of them are highly correlated so that one can be dropped.

After building the correlation graph for the wine data we get:



Above plot shows the correlation among the different variable. Highly correlated variables share dark blue circle and the colour intensity gradually decreases with the decrease in correlation.

Now if we observe two features: free.sulphur.dioxide and total.sulfur.dioxide have dark blue circle. That means they are positively correlated. Little knowledge and intuition says us that the free.sulphur.dioxide comes inside the total.sulphur.dioxide. So it is obvious that they are correlated.

* Feature Scaling:

Now, feature scaling is done where all feature are brought to same range to make the model unbiased to one feature.

A function normalize is created that is then called in loop for each variable thus making each feature in the range of 0-1.

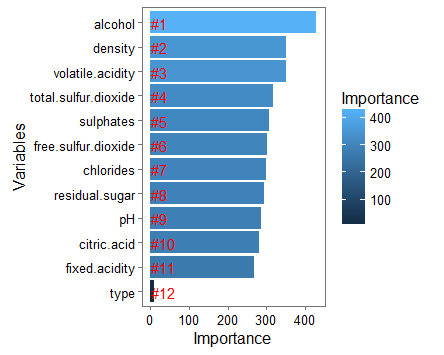
* Dividing the train and test data:

Now let’s divide the data to train and test data. We use 80% of total data for train data and other 20% for test data.

* Model Building:

Now a model is build. We choose Random Forest for this. Random forest creates multiple trees while taking subset of the feature for building each tree then feeding the error of one tree to other.

If we see the importance of the each variable we get:



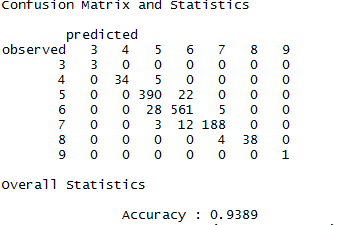
Here we can observe, alcohol in the wine has very large impact on predicting the wine quality. Similarly, the plot shows the importance in decreasing order as density, volatile.acidity and soon with minimum impact of ‘type’ feature we created.

Observation shows the ‘type’ feature could be dropped as it has very negligible importance.

Prediction:

As the model is prepared, let’s test the model with the remaining 20% data we separated from testing.

We check the accuracy using the confusion matrix as shown below:



Here, we can see the predicted values represented by columns and the observed values represented by rows.

Observation shows, if we see the values predicted for quality 5 then 390 test observations were correctly predicted whereas 22 test observation were incorrectly predicted as 6.

Similarly above confusion matrix shows for all level of quality.

Overall accuracy obtained was 93.89% as shown above.

Analysis:

While plotting the importance, we observed that the ‘type’ feature has very negligible effect and can be dropped while modelling. Above accuracy is obtained while neglecting ‘type’. So dropping the ‘type’ will increase the accuracy from 93.35 to 93.89(~94).

* Insights:

The data are not evenly distributed and have outliers and with the outlier removal the prediction increases from 70% to 93.89%.

The major factor determining the quality of wine is Alcohol contain. So greater care should we given while preparation and fermentation.