Project Report

Wine Quality Modeling

Hui Li hxl144130

Mengheng Hu mxh144730

Xinrui Shuai xxs140430

Wine Quality Modeling

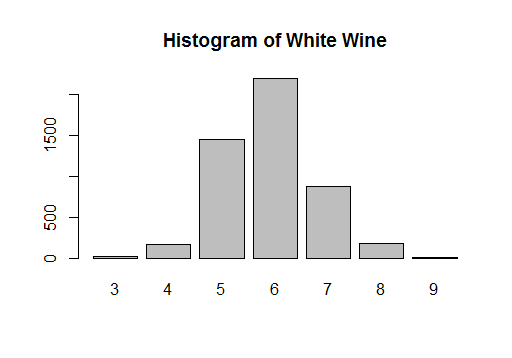
Hui Li(hxl144130), Mengheng Hu(mxh144730), Xinrui Shuai(xxs140430)

**Problem and Data Description:**

The datasets are related to red and white vinho verde wine samples from the north of Portugal.

The goal of this project is to find a good model for wine quality based on the physicochemical measurements.

There are two sub datasets: white wine (4898 instances) and red wine (1599 instances). The response variable is quality which is discrete numerical number range from 3 to 9. There are 13 predictor variables which are all continuously numerical values. The datasets are highly unbalanced which means most of the data points are in the mid quality range and only few poor and excellent instances.

**Approach**:

1. Data preprocessing:

Since the data is highly unbalanced and there are only less than 1% data are in the poor or excellent condition, as shown in the histogram graph. We decide to remove the data points which is has highest and lowest value. Then the data is randomly divided into training set and test set in 8:2.

2. Experiments Setup and Model selection:

We noted that the response variable is numerical but it also only contains a few values. We think this is a good data to test both regression and classification models and compare the accuracy between them. For the regression model, if the prediction value is within the 0.5 range of the real value, we consider it is a correct prediction. In this project, we have applied 5 regression models and 7 classification models.

Regression models:

We applied linear regression using best subset selection method, and also tried both lesso and ridge model. PCR and PLSR are also applied here using 6 components for red wine and 4 components for white wine.

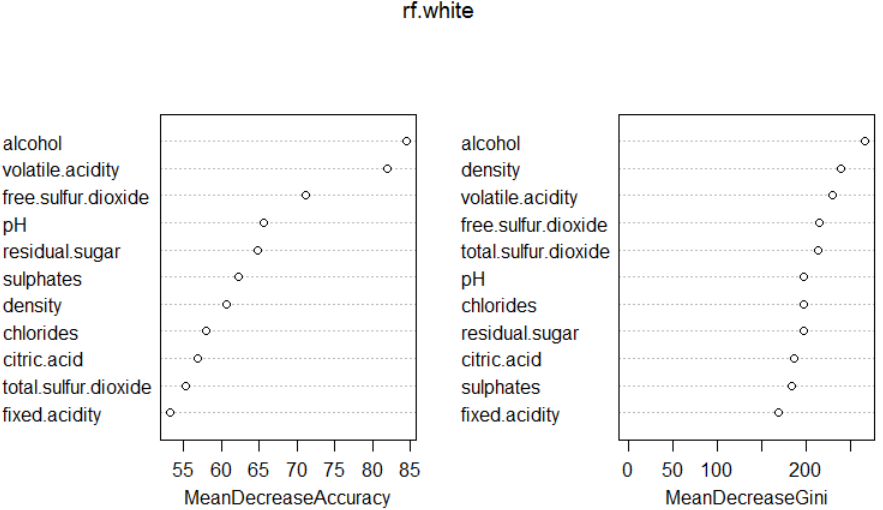
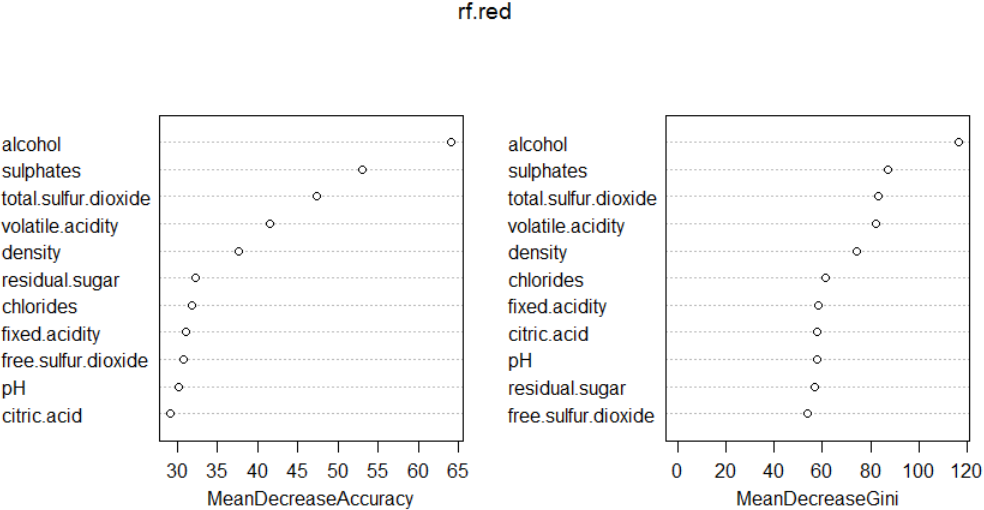
Classification models:

We have applied tree models including normal tree, bagging (500 trees), random forest and boosting (5000 trees) method. KNN is tested using different number of neighbors (1 to 5). Other methods used are: LDA, QDA and SVM.

**Results and Discussion**

1. Correlation of Variables

Correlations between each variable are checked by using correlation matrix and we found two set of variables have internal correlation (correlation value above 0.6): {Density ~ Alcohol, Sugar} and {pH ~ volatile acidity, fixed acidity}.



Importance of each Variable

2. Variables Importance

The variable importance graph indicates the most important variable is alcohol concentration. Red wine did show different variable importance level compared with white wine.

|  |  |  |
| --- | --- | --- |
|  | **Red (%)** | **White (%)** |
| **Linear Regression** | 59.7  (7) | 51.1  (9) |
| **Lesso / Ridge** | 58.7 / 60.2 | 51.3 / 51.5 |
| **PCR /PLSR** | 58.0 /58.4 (6 comp) | 49.1 / 49.1 (4 comp) |
| **Bagging** | 70.5 | 68.5 |
| **Random Forest** | 74.3 | 67.9 |
| **Boosting** | 63.5 | 59.2 |
| **KNN** | 62.2 (1) /58.5(5) | 62.8 (1) / 54.1(5) |
| **LDA** | 63.9 | 52.8 |
| **QDA** | 59.2 | 49.1 |
| **SVM** | 67.1 | 65.1 |

**Table 1**. Predict Accuracy of each Model

3. Model Discussion

The accuracy of each model is shown in table 1. The model shown best accuracy is from Random Forest for red wine (74%) and bagging for white wine (68%). Classification models works much better compared with regression model. This result indicates for response variables with only a few values, it is better to use a classification model compared with regression models.

**Conclusion**

We have built different methods for wine modeling. The result showed that red wine and white wine have different properties in terms of variable importance and model accuracy. There is no model show promised prediction result. The prediction may go higher if we re-categorize the data into fewer classes. Also the wine quality is more related to the grape type and production years. The physicochemical properties may not be enough to do correct prediction.