Wine Quality Data Analysis

RMSC 4002 Group Project

LI	Wai Yin	1155063766
YUEN	Kam Ho	1155077809
YEUNG	Wang Fung	1155077537

1. Abstract

In this project, We would like to figure out the rules of classifying two kinds of wine: 1) **whether the wine is red or white**, 2) the **wine quality** based on the physicochemical test in regard to the dataset from http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

First, normalization with z-score and outliers detection have been done. After these process, there are three sets of data 1) **raw**, 2) **z-scored**, and 3) **outlier dropped**, for the under models to fit.

Second, logistic regression have been fitted with the three sets of data. In this part, binary classification have been done with the variable 'red' and the Multinomial Logit (MNL) have been done with the variable 'quality'.

Third, Classification Tree have been done to classify 'red' and 'quality' with the three sets of data.

Finally, we have compared with the above models to find out an efficient classification way to identify the wines' characteristic.

2. Introduction

In this paper, we use the two datasets which are related to red and white wine of Portuguese "Vinho Verde" wine. 1600 samples of red wine and 4899 samples of white wine are included. Then, they are merged into a single dataset with a binary variable 'red' which is equal to 1 if the data is from the red wine set, else equal to 0.

We identified the quality of wine based on 13 variables. Here are the 13 variables with definition.

1. **Fixed acidity**: most acids involved with wine or fixed or nonvolatile
2. **Volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. **Citric acid**: add 'freshness' and flavor to wines
4. **Residual sugar**: the amount of sugar remaining after fermentation stops
5. **Chlorides**: the amount of salt in the wine
6. **Free sulfur dioxide**: the ratio between free form of SO_2 between molecular SO_2 (as a dissolved gas) and bisulfite ion in equilibrium; it prevents microbial growth and the oxidation of wine
7. **Total sulfur dioxide**: amount of free and bound forms of SO_2; in low concentrations, SO2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. **Density**: water depending on the percent alcohol and sugar content
9. **pH**: describes how acidic of a wine is on a scale from 0 to 14 ; most wines are between 3-4 on the pH scale

10. **Sulphates**: a wine additive which can contribute to sulfur dioxide gas (SO_2) levels
11. **Alcohol**: the percent alcohol content of the wine
12. **Quality**: output variable (based on sensory data, score between 0 and 10)
13. **Red:** binary variable (0 if red, 1 if white)

3. Data prepartion

1. Z-transformation

We have done the z-score transformation first. We made a creation of shifted and scaled versions of the datasets, where the intention is allowing the comparison of corresponding z-score for different datasets in a way that eliminates the effects of certain gross influences.

2. outlier direction

We have used Mahalanobis distance to detect the outlier. We have thrown 4368 - 4327 = 41 cases from the raw data to form a cleaned dataset by 99% chi square distribution with degree of freedom equal to 13 (with 13 variables in the data).

4. Logistic regression

1. Binary classification with ‘red’ or ‘white’

We have built a binary classification to classify whether the wine is red or not with all other variables. Although we have used step function to find better regression models, the results are not good. So, we kept the raw model and regress with the three sets of data.

a. raw dataset

pr (training)	0	1	pr (testing)	0	1
FALSE	3246	18	FALSE	1457	8
TRUE	10	1094	TRUE	5	479

The above tables are classification tables with raw data. The error rate is 0.006157 and F1 score is 0.987365 with training set, and the error rate is 0.00667 and F1 score is 0.986612 with testing set. It seems the model provide a good classification result.

b. normalized with z score

pr (training)	0	1	pr (testing)	0	1
FALSE	3246	18	FALSE	1457	8
TRUE	10	1094	TRUE	5	479

The above tables are classification tables with z score data. The result is same as that of raw data. The error rate is 0.006157 and F1 score is 0.987365 with training set, and the error rate is 0.00667 and F1 score is 0.986612 with testing set. So, in this situation, normalization cannot provide a better insight for us to classify whether the wine is red or not.

c. dropout outlier.

pr (training)	0	1	pr	0	1
FALSE	3338	1	FALSE	1453	7
TRUE	1	987	TRUE	9	479

The above tables are classification tables with cleaned dataset. The error rate is 0.000462 and F1 score is 0.998988 with training set, and the error rate is 0.009749 and F1 score is 0.980671 with testing set. Although this fitted better with the training data, the testing result is worse than that of raw data. So, the outlier detection may not be useful in this situation.

2. Multinomial Logit with quality

Also, we have done a Multinomial Logit with quality by regress with all other variables. In this part, we have regressed with raw data and cleaned data set.

a. raw data

training set							
prmnl	3	4	5	6	7	8	9
3	2	0	0	0	0	0	0
4	0	10	5	0	0	0	0
5	12	87	847	418	44	6	0
6	7	54	603	1491	520	111	1
7	1	1	6	115	169	37	2
8	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0

The accuracy of the model is 0.55387 with training set (diagonal over all), which cannot be good fit with the training data. Also, the accuracy of model with testing set is 0.536685, which is not a predictive result.

b. dropout outlier

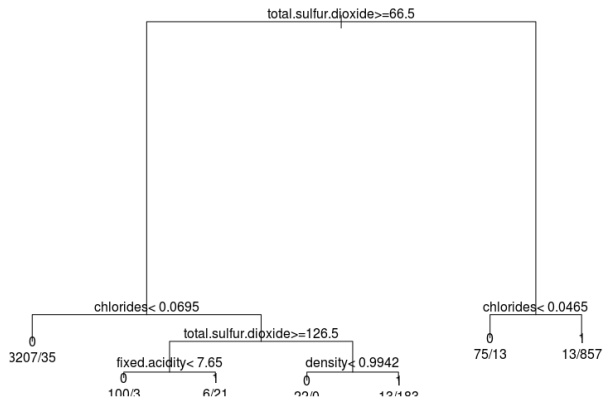
prmnl	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	8	3	0	0	0	0
5	6	74	818	403	43	6	0
6	2	45	564	1418	498	99	1
7	0	0	8	121	170	38	2
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

The accuracy of the model is 0.573274 with training set. Although this is better than that of raw one, it still cannot be good fit with the training data. Also, the accuracy of model with testing set is 0.539764, which is not a predictive result as that of raw data.

This show that logistic regression have good result in binary classification. When the classes of dependent variable increase, the accuracy of logit model may decrease. In this case, we should use a complex model, like neural network, to classify the quality.

5. Classification tree

1. Classification tree with red



The classification tree is applied to analyze the data, mapping the the kind of wine and wine quality to variables including total sulfur dioxide, chlorides, fixed acidity, density, alcohol and volatile acidity. Moreover, using classification tree can provide a clear mode to make prediction.

We do the binary classification with red and white wine first. In the classification tree, there are four variables deciding red or white wine simultaneously.

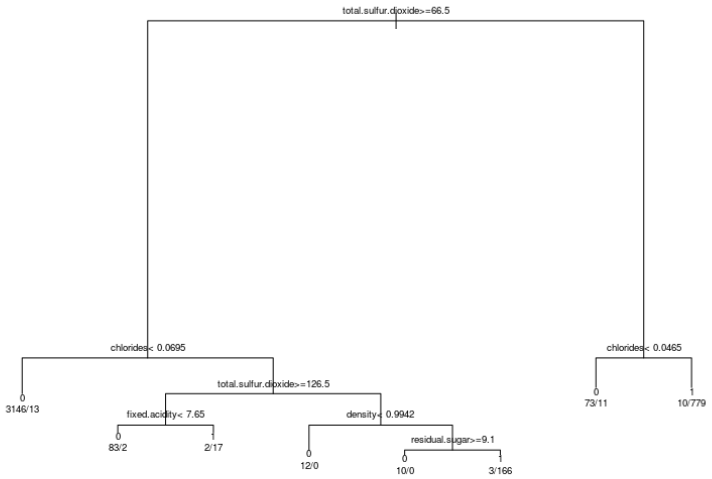
The classification rules are followed:

1. total sulfur dioxide >= 66.5 and chlorides < 0.0695 -> white
2. chlorides > 0.0695, total sulfur dioxide >= 126.5 and fixed acidity < 7.65, -> white
3. chlorides > 0.0695, total sulfur dioxide >= 126.5 and fixed acidity > 7.65 -> red
4. chlorides > 0.0695, total sulfur dioxide >= 126.5 and density < 0.9942 -> white
5. chlorides > 0.0695, total sulfur dioxide >= 126.5 and density < 0.9942 -> red
6. total sulfur dioxide < 66.5 and chlorides < 0.0465 -> white
7. total sulfur dioxide < 66.5 and chlorides < 0.0465 -> red

pr	0	1
FALSE	3404	51
TRUE	32	1061

pr	0	1
FALSE	1444	21
TRUE	18	466

From the classification table of red and white wine, the average error rate is 1.825%, which means the classification tree is accurate. and the F1 score is 0.962358. With the testing set, error rate is 0.02001 and F1 score is 0.959835.



We have built another classification tree with the cleaned dataset.

The classification rules are followed:

1. total sulfur dioxide >= 66.5 and chlorides < 0.0695 -> white
2. chlorides > 0.0695, total sulfur dioxide >= 126.5 and fixed acidity < 7.65, -> white
3. chlorides > 0.0695, total sulfur dioxide >= 126.5 and fixed acidity > 7.65 -> red
4. chlorides > 0.0695, 66.5 <=total sulfur dioxide < 126.5 and density > 0.9942 -> white
5. chlorides > 0.0695, 66.5 <=total sulfur dioxide < 126.5, density > 0.9942 and residual.sugar >= 9.1 -> white
6. chlorides > 0.0695, 66.5 <=total sulfur dioxide < 126.5, density > 0.9942 and residual.sugar < 9.1 -> red
7. total sulfur dioxide < 66.5 and chlorides < 0.0465 -> white
8. total sulfur dioxide < 66.5 and chlorides < 0.0465 -> red

And the rules are quite similar with the classification tree built with raw data. It shows that total sulfur dioxide >= 66.5 is the most significant rules to classify whether the wine is white or red.

The classifcation table is followed:

pr	0	1	pr	0	1
FALSE	3324	26	FALSE	1449	24
TRUE	15	962	TRUE	13	463

The error rate and F1 score of ctree with dropout training set is 0.009475 and 0.979135, which provide a better result than the ctree with raw data. Also, with testing set, error rate is 0.018984 and F1 score is 0.961578. Both of them is better than c-tree with raw data.

2. Classification Tree with quality

Besides the classification with red and white wine, we also have classified the wine quality. There are two variables deciding the quality level of the wine simultaneously. When the alcohol is larger than 10.12, the level should be 6 and when the it is smaller than 10.12, the

level can be 5 and 6, which implies the importance of alcohol to the wine quality. Moreover, the wine quality level should be 5 if the alcohol is smaller than 10.12 and volatile acidity is larger than and equal to 0.2875. If not, the level should be 6.
From the classification table of wine quality, the average error rate is 47.95%.There is a possible reason to explain the large error rate. Firstly except alcohol, other variables have complex effects on wine quality, so it is hard to fit the model preciously using binary trees.

Remarks: we have also done a classification tree with quality but the result is very bad. Because of the limiting of pages, we skipped to discuss it deeply.

6. Conclusion

To conclude, binary classification which classify red or white done pretty well with both model. The best model in the project is the Logit without cleaned. This reflected that the outlier detection may clean the information also, not only the noise. We may try to use a higher degree of freedom or critical value to prevent the problem. Besides, although both Logit are better than Ctree, Ctree still can prove us an insight which the most important variable is, and this is the Logit cannot provide.

With Testing set	Error Rate	Accuracy
Logit without cleaned	0.00667	0.986612
Logit with cleaned	0.009749	0.980671
Ctree without cleaned	0.02001	0.959835
Ctree with cleaned	0.018984	0.961578

With the multi-class classification, the result is really bad to both model with both cleaned or non-cleaned data. It shows us classifying 6 or more classes is difficult for both logit and ctree, especially to ctree.

Appendices.

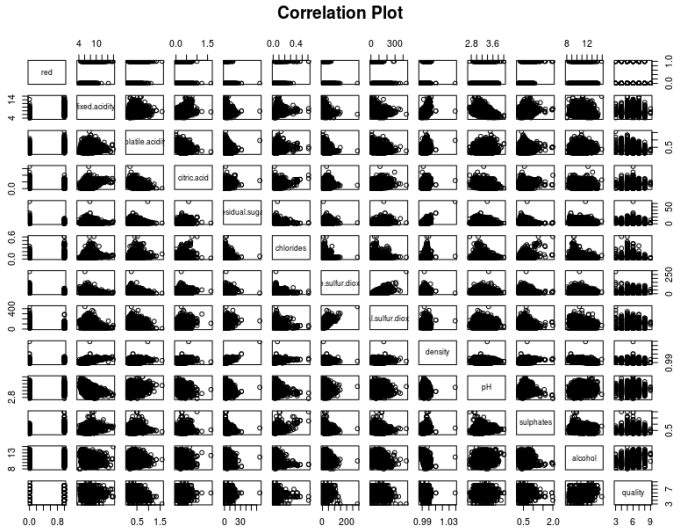


Figure 1. correlation plot with raw data set.

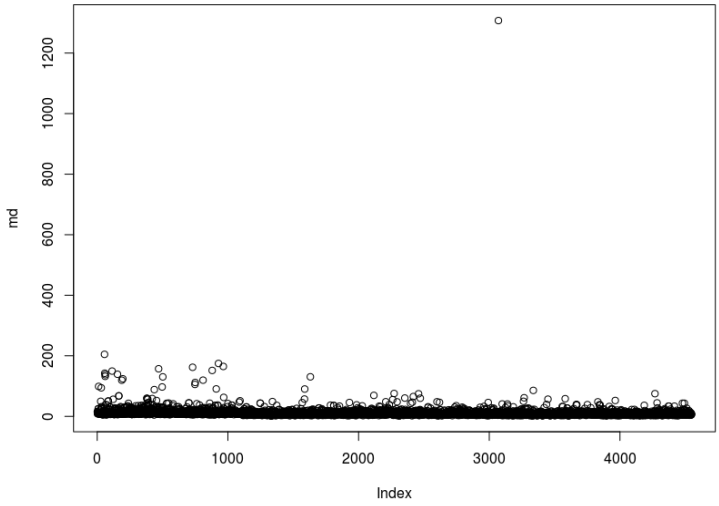


Figure 2. plot of Outlier detection

ctree							
training set							
cl	3	4	5	6	7	8	9
3	8	59	729	327	16	1	0
4	14	91	732	1698	717	153	3
testing set							
cl	3	4	5	6	7	8	9
3	1	27	319	141	11	1	0

4	7	39	358	670	335	38	2
with dropout							
training set							
cl	3	4	5	6	7	8	
3	1	31	600	275	11	0	
4	0	42	616	1413	523	69	
5	0	0	2	72	117	22	
testing set							
cl	3	4	5	6	7	8	9
3	1	27	319	141	11	1	0
4	6	37	351	630	272	29	1
5	1	2	7	40	63	9	1

Table 1. Classification Table of Ctree with quality: