Data Analytics Assignment 7: 2 pm Session

Shiwei Huang

**1. QUESTION ONE**

**Exploratory Data Analysis (3%) Explore the statistical aspects of both datasets. Analyze the distributions and provide summaries of the relevant statistics. Perform any cleaning, transformations, interpolations, smoothing, outlier detection/ removal, etc. required on the data. Include figures and descriptions of this exploration and a short description of what you concluded (e.g. nature of distribution, indication of suitable model approaches you would try, etc.). Min.1 page text + graphics (required).**

Dataset1: Red wine Quality Data

Dataset2: White wine Quality Data

I choose to perform the wine quality data set for the data analysis project. For the wine quality data set, the characteristics is multivariate; and the attribute characteristics is real. Moreover, it is better to perform regression and classification models on this set of data.

Regarding the characteristics of the data sets, they are related to red and white variants of the Portuguese “Vinho Verde” wine. According to the description, these datasets can be viewed as classification or regression tasks. There are twelve attributes related to the dataset which is:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. Quality

A screenshot of a social media post

Description automatically generated

Similarly with the red wine data set, I also did an summary analysis on the white whine data set. These are the summary based on the dataset attributes:

A screenshot of a social media post

Description automatically generated

I separated the data sets from one column to multiple columns:

The following is the wine red analysis:

A screenshot of a social media post

Description automatically generated

The following the wine white dataset:

A screenshot of a social media post

Description automatically generated

I made a ggplot on red wine with alcohol count, this is my finding:

A screenshot of a cell phone

Description automatically generated

I made a ggplot on citric.acid with Ph level for the red wine:

ggplot(winered, aes\_string(y=winered$pH, x=as.factor(winered$citric.acid)),)+

geom\_boxplot() + xlab('citric.acid') + ylab('pH')

A screenshot of a computer

Description automatically generated

This is the ggplot on pH data with citric.acid:

A screenshot of a cell phone

Description automatically generated

In addition, I wanted to understand what is the range for pH level for red wine, citric acid range for red wine, volatile acidity on red wine and chlorides range for red wine.

I performed analyses on the range function:

This is my finding:

> range(winered$pH)

[1] 2.74 4.01

> range(winered$pH)

[1] 2.74 4.01

> range(winered$citric.acid)

[1] 0 1

> range(winered$volatile.acidity)

[1] 0.12 1.58

> range(winered$chlorides)

[1] 0.012 0.611

This is the range for wine white pH level, citric acid, volatile acidity and chlorides:

> range(winewhite$pH)

[1] 2.72 3.82

> range(winewhite$citric.acid)

[1] 0.00 1.66

> range(winewhite$volatile.acidity)

[1] 0.08 1.10

> range(winewhite$chlorides)

[1] 0.009 0.346

**2. Question Two: Model Development, Validation, Optimization and Tuning (14%) Choose two (4000-level\*) or three (6000-level) or more different models (e.g. a model with a different set/ number of variables/ features in a regression, or classification, etc. does NOT count as a different model). Explain why you chose them. Construct the models, test/ validate them. Explain the validation approach. You can use any method(s) covered in the course. Include your code in your submission. Compare model results if applicable. Report the results of the model (fits, coefficients, graphs, trees, other measures of fit/ importance, etc.), predictors, and summary statistics. Min. 4 pages of text + graphics (required). \* 4000-level will receive extra credit for 6000-level responses.**

Dataset 1: red wine

Model 1: Multivariate linear regression

Multivariable linear regression on redwine:

Reasons I chose this model is I want to understand the dependency of the variables:

Result:

A close up of text on a white background

Description automatically generated

A screenshot of a social media post

Description automatically generated

White wine data:

A screenshot of text

Description automatically generated

A close up of a piece of paper

Description automatically generated

Model 2: decision tree model

Because the decision tree model can be used to solve regression and classification problems.

Redwine:

A screenshot of text

Description automatically generated

Whitewine:

A screenshot of a cell phone

Description automatically generated

Model 3: random forest model

Random forest can handle the missing values and prune the data accuracy.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

**3. Decisions (3%) Describe your conclusions in regard to the model fit, predictions and how well (or not) it could be used for decisions and why. Min. 1 page of text + graphics.**

My conclusions is that the dataset 1 and 3 provides a good summary on each variable. Dataset number 2 does not work well with the wine data.

**A screenshot of a cell phone

Description automatically generated**

I performed summary on winered and wine white:

This the summary on winered:

A close up of text on a white background

Description automatically generated

This is the summary on winewhite:

A screenshot of a cell phone

Description automatically generated

Doing the comparison, we could find the median summary on winered is 7.9 while the median on whitewhite is 6.8; in terms of the volatile acidity, the median for winered is 0.52 while the median for winewhite is 0.26; in terms of citric acid, the median for wine red is 0.26 while the median for wine white is 0.32; in terms of the residual sugar, wine red median is 2.2; wine white median is 5.2; for chlorides, the medium for winered is 0.079, the median for wine white is 0.043; in terms of free sulfur dioxide, the medium for wine red is 14 while the medium for wine white is 34; in terms of total sulfur dioxide, the medium for wine red is 38 while the median for wine white is 134; in terms of density, the median for wine red is 0.996 while wine white density is slightly less which is 0.9937.