WINE QUALITY CASE STUDY

INDEX

IMAGE INDEX

1. INTRODUCTION

Two datasets are available of which one dataset is on red wine and have 1599 different varieties and the other is on white wine and have 4898 varieties. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.

1.1 OBJECTIVE

Prediction of Quality ranking from the chemical properties of the wine.
A predictive model developed on this data is expected to provide guidance to vineyards regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters.

DATA SETS GIVEN

Two data sets are given at the start of the study, containing chemical compositions in RED and WHITE Wines respectively. Both of these data sets are combined together to make a new data set called DATA

1) winequality-red.csv
2) winequality-white.csv
   both of these data sets are combined to get
3) data.csv

1.4) ABOUT THE DATA SET

Given data set contains quality of 6497 different wines along with their chemical composition. Following are the attributes given in data which are all numeric: -

1.4.1) VARIABLES

1) fixed acidity
2)volatile acidity
3)citric acid
4) residual sugar
5) Chlorides
6) Free Sulphur Dioxide
7) total Sulphur dioxide
8) Density
9) pH
10) Sulphates
11) Alcohol

TARGET VARIABLE

1) QUALITY

1. APPROACHES TAKEN FOR ANALYSING DATA

For analysis of given data R-Programming is used which is a programming language used mainly for the purpose of analysis.

The following analytical approaches are taken:

- **Regression Tree**
    - Random Forest
    - C50

2. EXPLORATORY DATA ANALYSIS

Basic summary statistics for both red and white data sets shows that ranges for variables in red dataset varies from variable ranges in white wine data set.so before joining two data sets, first pre-processing has to be done. Following graph shows outliers in red and white data sets respectively.

3.1 OUTLIERS

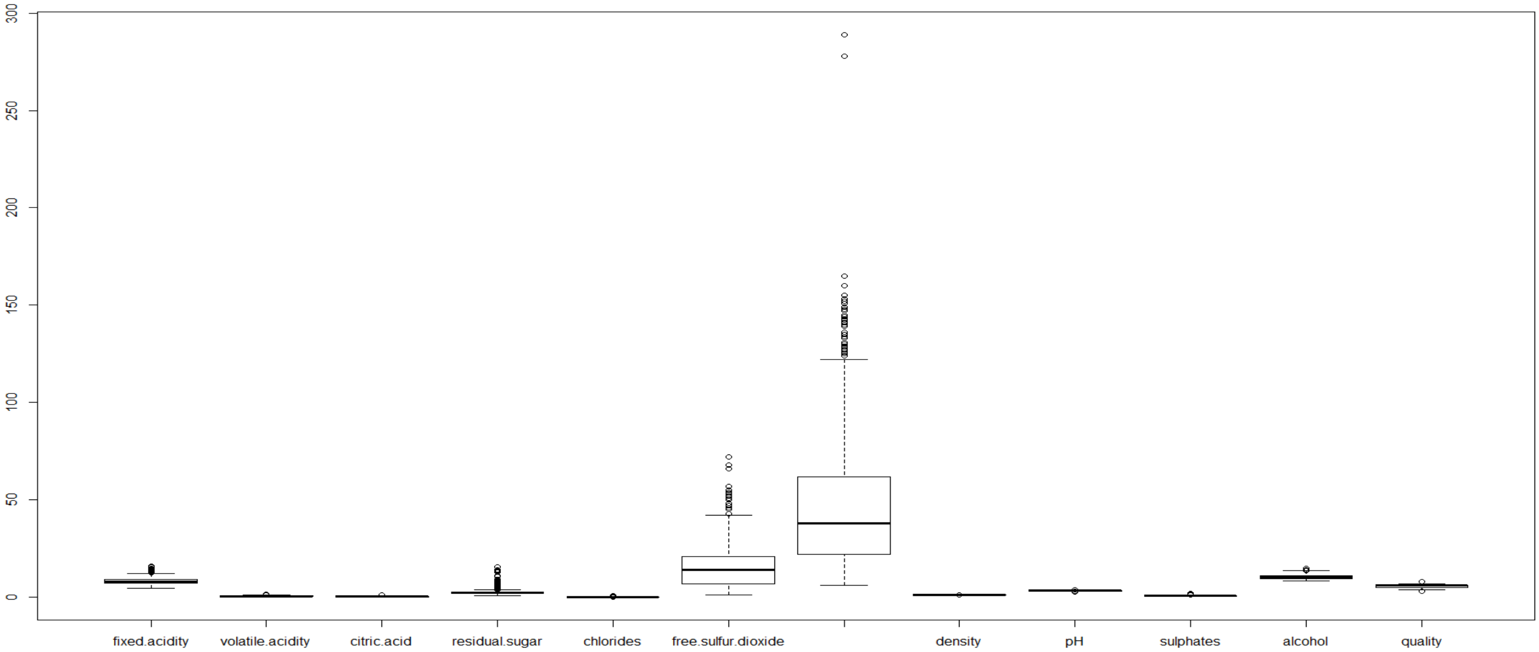
Outliers are variables which lies far away from mean. They can be calculated as,

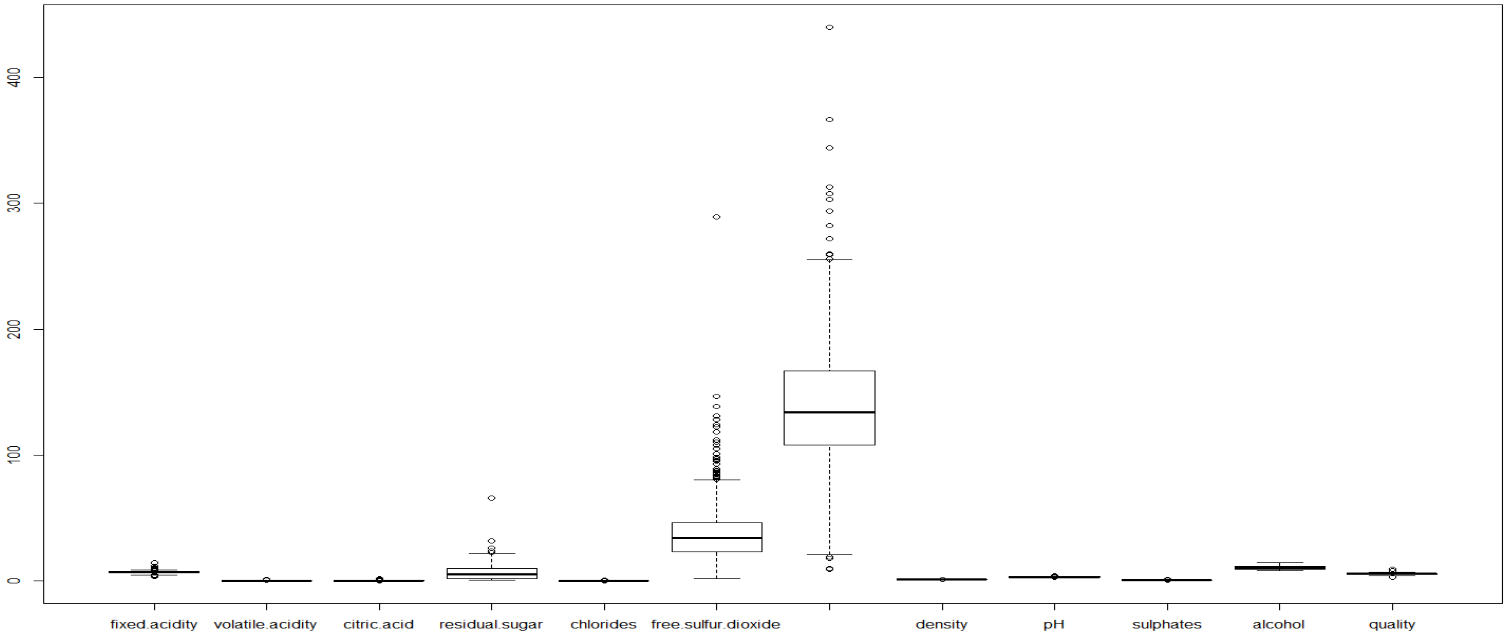
Outlier range <- Q1 – 1.5*(Q3 – Q1) To Q3 + 1.5 * (Q3-Q1)

Any points lying outside above range is called an outlier.

During analysis it was found that there were lot of outliers masked by other outliers, hence process of removing outliers has to be done multiple times.

After removing outliers, size of data sets was reduced, for red data set from 1599 to 923 and for white data set from 4898 to 3659.


0
50
100
150
200
250
300
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
density
pH
sulphates
alcohol
quality


0
100
200
300
400
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
density
pH
sulphates
alcohol
quality

3.2 FEATURE ENGINEERING

1)During background study on wines it was found that variable

free sulfur dioxide/ total sulfur dioxide

plays an important role in deciding test of wine hence it was added.

2)since not only ranges of variables differ across data sets, behavior of variables against quality variable also changes to account for it, a new variable was introduced called "is.red" which was "1" if wine is red and "0" if wine is white

3)After normalizing the data, both data sets were combined together to form a data set called "wine", and then quality variable is binned with values "0" for quality less than or equal to 5, otherwise "1".

3.3 NORMALIZING DATA

It is important to convert all variables in data set in equal or near to equal range to improve model performance.

There are different ways to get variables in near equal ranges

1)standardization/z-score Method <- it converts data in terms of standard deviation

2)normalizing data/Min-max Method <- it compresses data into a range of 0 to 1

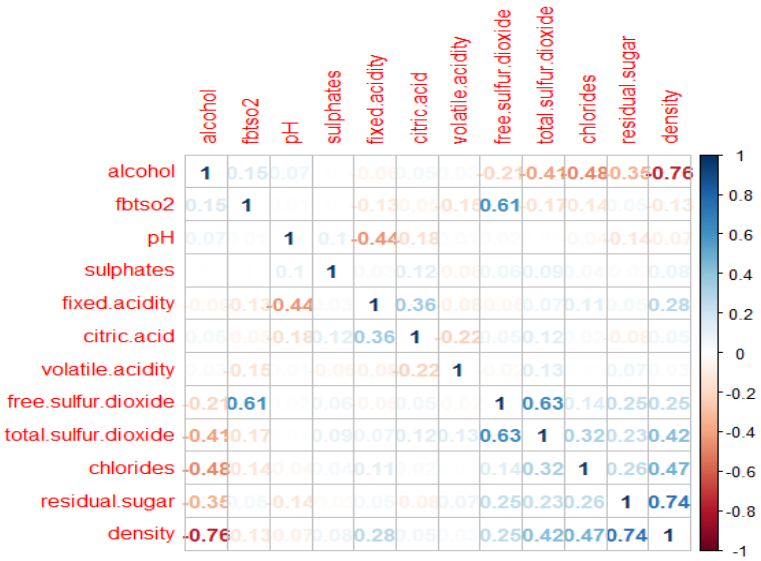
Out of above two min-max method is chosen as it gave maximum efficiency.

ClusterSim package was used for this.

3.4 CORRELATED VARIABLES

Variables which are highly correlated inflates the outcome, hence they have to be removed, this is done using "corrplot" package.

Following diagram shows correlation matrix of all independent variables on each other



Free sulfur oxide and density variables are removed since they were highly correlated with alcohol, total sulfur dioxide, residual sugar.

3. TRAIN AND TEST DATA SET

There are different ways to sample data set to create train and test data sets
1) Simple Random sampling
2) Stratified Sampling
3) Cluster Sampling
4) Systematic Sampling

We have combined two data sets in which variables have somewhat different relationship with target variable, also number of observations in both data sets are different. So to get training data set with equal amount of both red wine samples and white wine samples, Stratified Sampling technique is used with "Sampling" library.

4. CLASSIFICATION MODELS

There are different types classification models available, out of which C5.0 and random forest were chosen, since they create easily understandable models whose rules can be easily interpreted and are also one of the most efficient models.

5.1 CARET PACAKGE

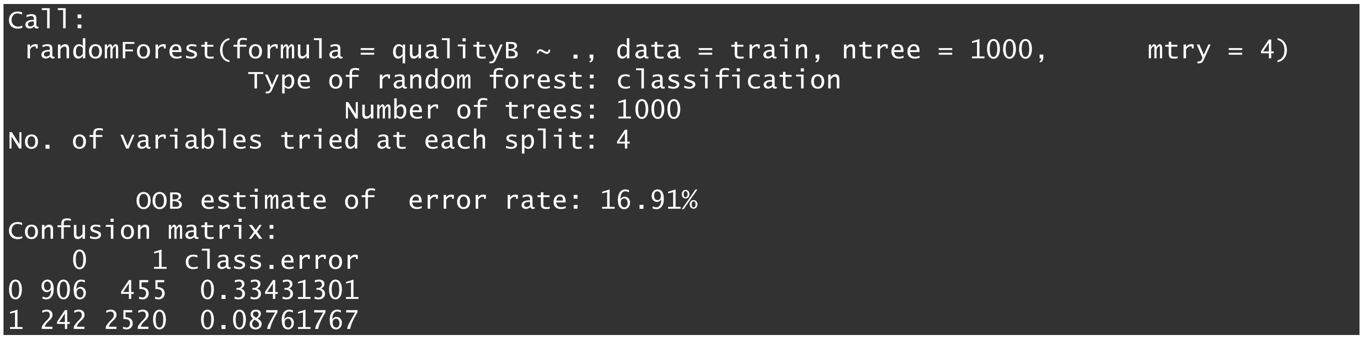
Caret package is used to create primary models. Caret package provides a very useful function called train which calculates optimum values for certain attributes giving us a model which is optimized.

Both random forest and c5.0 methods were used using train function but could not get more that 81% accuracy. Hence they are modified.

5.2 IMPROVING PERFORMANCE OF MODEL

Max. accuracy in c5.0 model after improvement was 82.35%

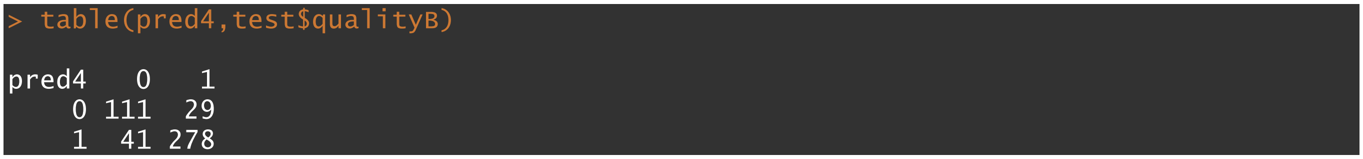
Accuracy of random forest was improved from 82% to 85% using different attribute values.



```
Call:
 randomForest(formula = qualityB ~ ., data = train, ntree = 1000,      mtry = 4)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 16.91%
Confusion matrix:
    0    1 class.error
0 906  455  0.33431301
1 242 2520  0.08761767
```

When tested against test data set we got following confusion matrix



```
> table(pred4,test$qualityB)

pred4   0   1
    0 111  29
    1  41 278
```

To calculate accuracy <- (111+278)/459 <- .847

i.e. 84.7% Accuracy.

6. RULES

Following are the few rules selected based upon their error rate

Error rate <= 15%

Rules :-

1) Alcohol quantity should be more than 5.5 units

2) If volatile acidity is less than .22 then Sulphates should be more than 0.499

3) if Chlorides are less than 0.0403 then Sulphates should be more than 0.499

4) if volatile acidity is less than 0.23 then alcohol should be more than 4.45