**Wine Quality Prediction Report**

**Prajwal R**

**30-12-2018**

**Contents**

**1. Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Exploratory Data Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3

**2. Methodology**

2.1 Pre Processing. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4

2.1.1 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . .4

2.1.2 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5

2.1.3 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.1.4 Feature Scaling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2.1 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2.2 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2.3 Gradient boosting . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2.4 XGBoost . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2.5 Linear Regression. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

2.2.6 KNN. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

2.2.7 Support Vector Machine. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

**Appendix**

Extra Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

**References**

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

Objective of this problem is to predict the quality of red wine based on the amount/volume of different components that are present in a wine. Quality score ranges from 0 to 10. Wines with high scores have better qualities when compared to others.

* 1. **Data**

There are 12 variables in this data set out of which 11 are independent variables and 1 dependent variable. This problem can be solved as both classification and regression. But here we are considering the target variable as continuous variable and the problem will be solved by regression models.

**Variables Information:**

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

**1.3 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyze data sets to summarize their main characteristics. In the given data set there are 12 variables and 1599 observations. Data types of all variables are float64 and int64.

RangeIndex: 1599 entries, 0 to 1598

Data columns (total 12 columns):

fixed acidity 1599 non-null float64

volatile acidity 1599 non-null float64

citric acid 1599 non-null float64

residual sugar 1599 non-null float64

chlorides 1599 non-null float64

free sulfur dioxide 1599 non-null float64

total sulfur dioxide 1599 non-null float64

density 1599 non-null float64

pH 1599 non-null float64

sulphates 1599 non-null float64

alcohol 1599 non-null float64

quality 1599 non-null int64

dtypes: float64(11), int64(1)

From EDA we came to know that all the variables are continuous variables.

**Target variable:** quality

**Chapter 2**

**Methodology**

Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project. Few pre-processing techniques are appliedon the data set to bring it to proper shape.

**2.1 Pre-Processing**

Below data pre-processing techniques are applied on the data set before it is fed to the models.

**2.1.1 Missing Value Analysis**

Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of missing values, either we ignore the entire column or we ignore those observations.

**Columns Count**

fixed acidity 0

volatile acidity 0

citric acid 0

residual sugar 0

chlorides 0

free sulfur dioxide 0

total sulfur dioxide 0

density 0

pH 0

sulphates 0

alcohol 0

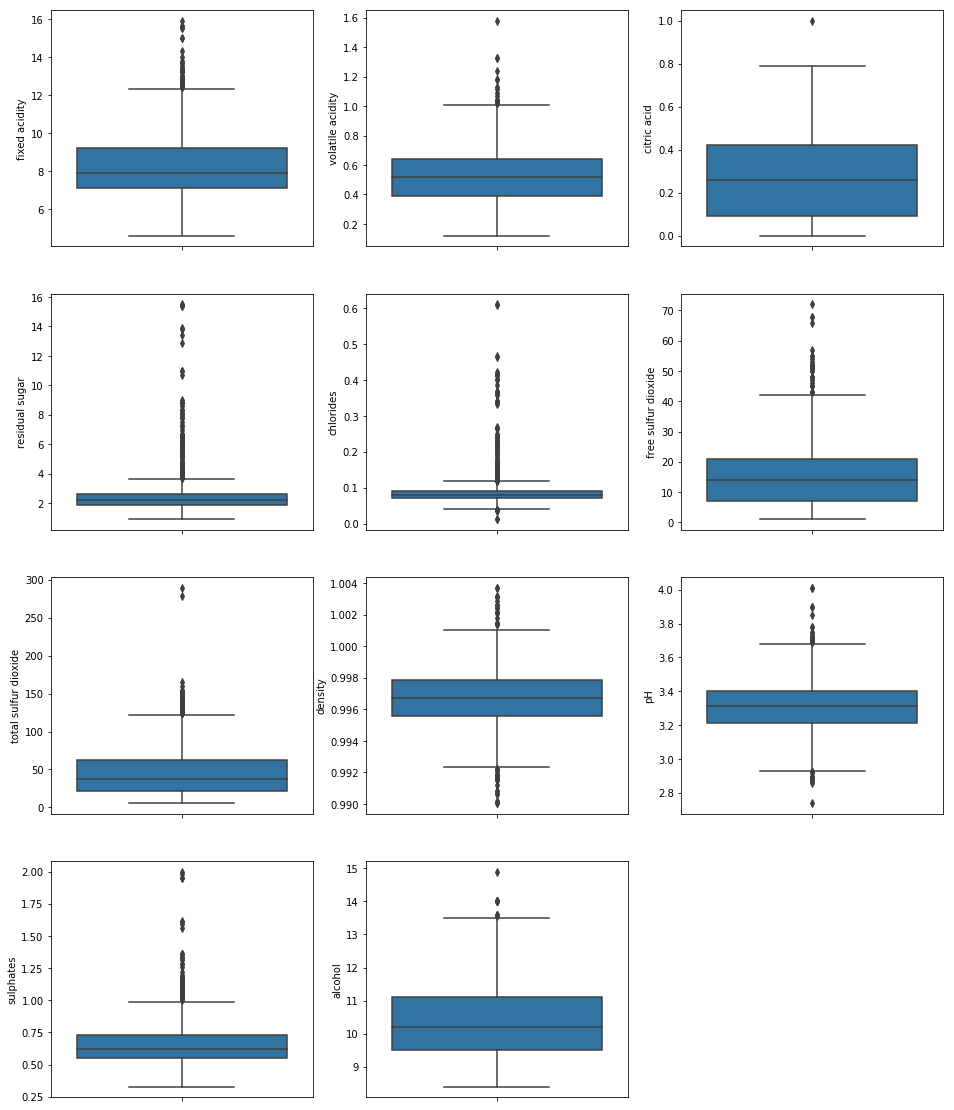
quality 0

Since there are no missing values in the data set, we can proceed without any missing value analysis.

**2.1.2 Outlier Analysis**

In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

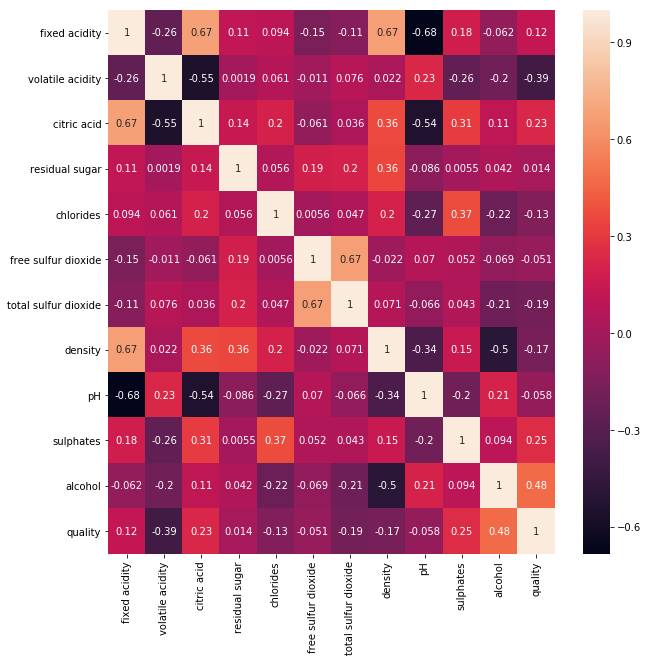
Boxplots are plotted for each of the continuous variables and based on the insights and proper understanding about the data, no outliers are removed from the variables.



From the boxplot,we see that all the variables have outliers. But all the values which are depicted as outliers cannot be considered as outliers since all the observations fall in the possible range.

**2.1.3 Feature Selection**

In this project we have selected **Correlation** Analysis to check the relationship between numerical variables and to remove multi collinearity from the data set.

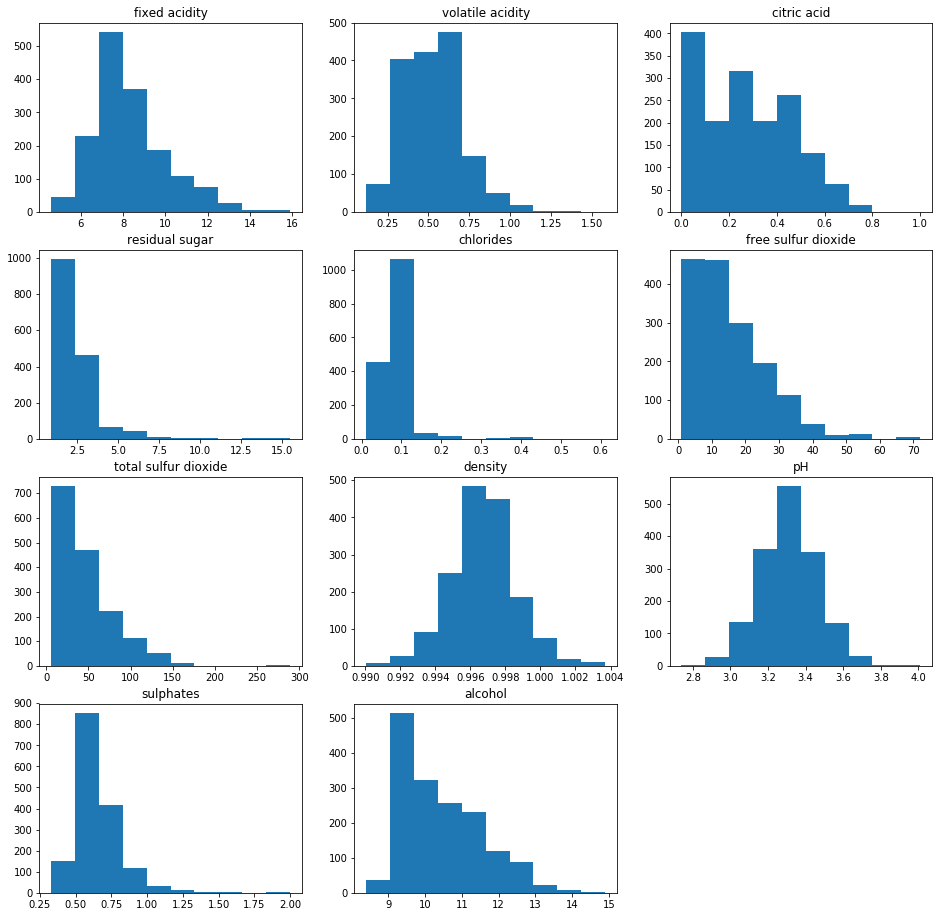


From correlation analysis, it is evident that there is no multi collinearity in the data set. So none of the variables are removed from the data set.

**2.1.4 Feature Scaling**

We have plotted the histogram to check the distribution of data in all the continuous variables.

Since most of the variables are not uniformly distributed, data is fed to the model as is.



After all the data pre processing, the data will be fed to the models without any changes. Will freeze this dataset and apply different regression algorithms.

**2.2 Modeling**

After a thorough preprocessing we will be training the data with regression models to predict the target variable. Models are trained using train dataset and evaluated using test dataset.

After the data is trained using regression models, different error metrics are calculated in both R and python to evaluate the performance of models.

**2.2.1 Decision Tree**

|  |  |  |
| --- | --- | --- |
| Decision Tree | R | PYTHON |
| MAE | 0.53 | 0.56 |
| MAPE | 9.54 | 9.68 |

**2.2.2 Random Forest**

|  |  |  |
| --- | --- | --- |
| Random Forest | R | PYTHON |
| MAE | 0.28 | 0.41 |
| MAPE | 5.22 | 7.6 |

**2.2.3 Gradient boosting**

|  |  |  |
| --- | --- | --- |
| Gradient boosting | R | PYTHON |
| MAE | 0.4 | 0.44 |
| MAPE | 7.27 | 7.97 |

**2.2.4 XGBoost**

|  |  |  |
| --- | --- | --- |
| XGboost | R | PYTHON |
| MAE | 0.36 | 0.45 |
| MAPE | 6.62 | 8.19 |

**2.2.5 Linear regression**

|  |  |  |
| --- | --- | --- |
| Linear regression | R | PYTHON |
| MAE | 0.41 | 0.53 |
| MAPE | 7.38 | 9.71 |

**2.2.6 KNN**

|  |  |  |
| --- | --- | --- |
| KNN | R | PYTHON |
| MAE | 2.01 | 0.54 |
| MAPE | 59.19 | 10 |

**2.2.7 Support Vector Machine**

|  |  |  |
| --- | --- | --- |
| SVM | R | PYTHON |
| MAE | 0.39 | 0.49 |
| MAPE | 7.18 | 9.06 |

**Chapter 3**

**Conclusion**

In this chapter we are going to evaluate our models, select the best model for our dataset.

**3.1 Model Evaluation**

In the previous chapter we have calculated Mean Absolute Error and Mean Absolute Percentage Error of all the models. Model with low MAE and MAPE values, will be treated as the best fit model for this data.

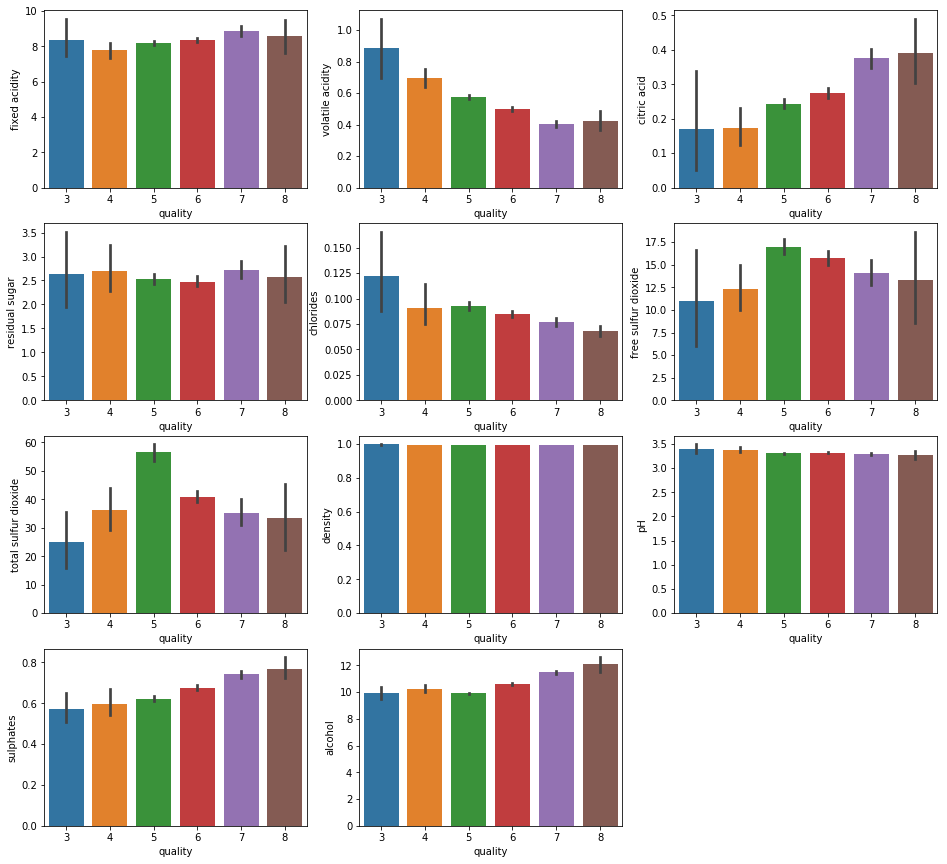
**3.2 Model Selection**

After observing MAE and MAPE values of all the models, we can conclude that Random forest has lower values MAE and MAPE. Therefore, we can freeze Random forest model for this data set.

**Appendix**

**Extra Figures**

Graphs are plotted for all the independent variable against the target variable to find the impact of different predictors on the response.



From the above plot it is evident that some of the independent variables have greater impact on quality of the wine as given below.

1. Higher the citric acid content, higher will be the quality of wine.
2. Wines with lower volatile acidity have higher quality.
3. Higher the chlorides content, lower the wine quality.
4. Wines with high sulphates content have high quality.
5. Higher the alcohol content, higher will be the quality.

**References**

1. Data is extracted from <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-20>

2. For Data Cleaning and Model Development - <https://edwisor.com/career-data-scientist>

3. For XGBoost in R - <https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/>