Capstone Project 2

Project Title

“Red Wine Quality”

Project Proposed by

Anshul Bhardwaj

Project Mentor

Gritank Dhamija

**Problem Statement**

The quality of any wine depends on various factors such as amount of alcohol, sugar, acidity, pH, density, sulphates, citric acid, chlorides etc.

By using the wine quality data provided by Kaggle, I created a model which helps in categorizing Wine into High Quality and Low Quality on the basis of various input variables such as –

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

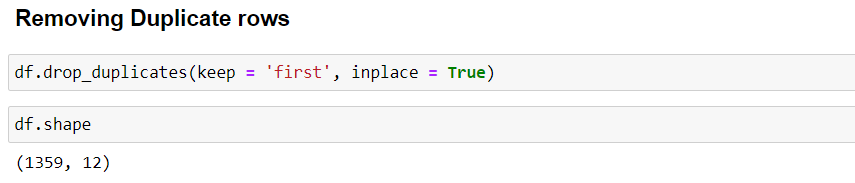
10 - sulphates

11 – alcohol

I used various machine learning algorithms such as Logistics Regression, Random Forest Classifier and Xgboost classifier to build the model.

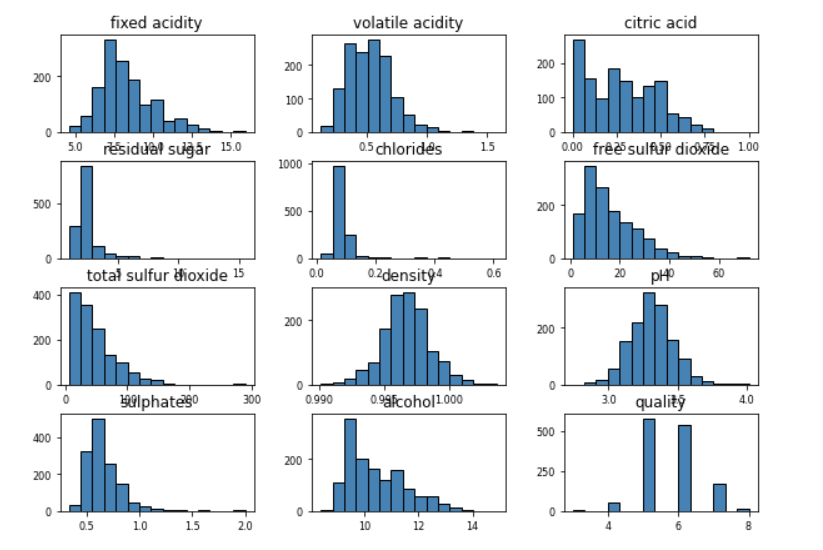
**Data Wrangling**

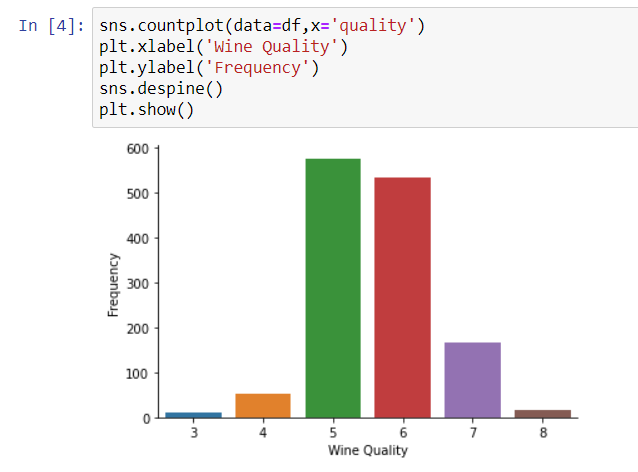
The raw dataset from Kaggle contained 1599 rows and 12 columns. It was already a clean dataset containing no empty or NaN values. It had some duplicate rows which were removed.



After the removal of duplicate rows our Data Frame was left with 1359 rows and 12 columns.

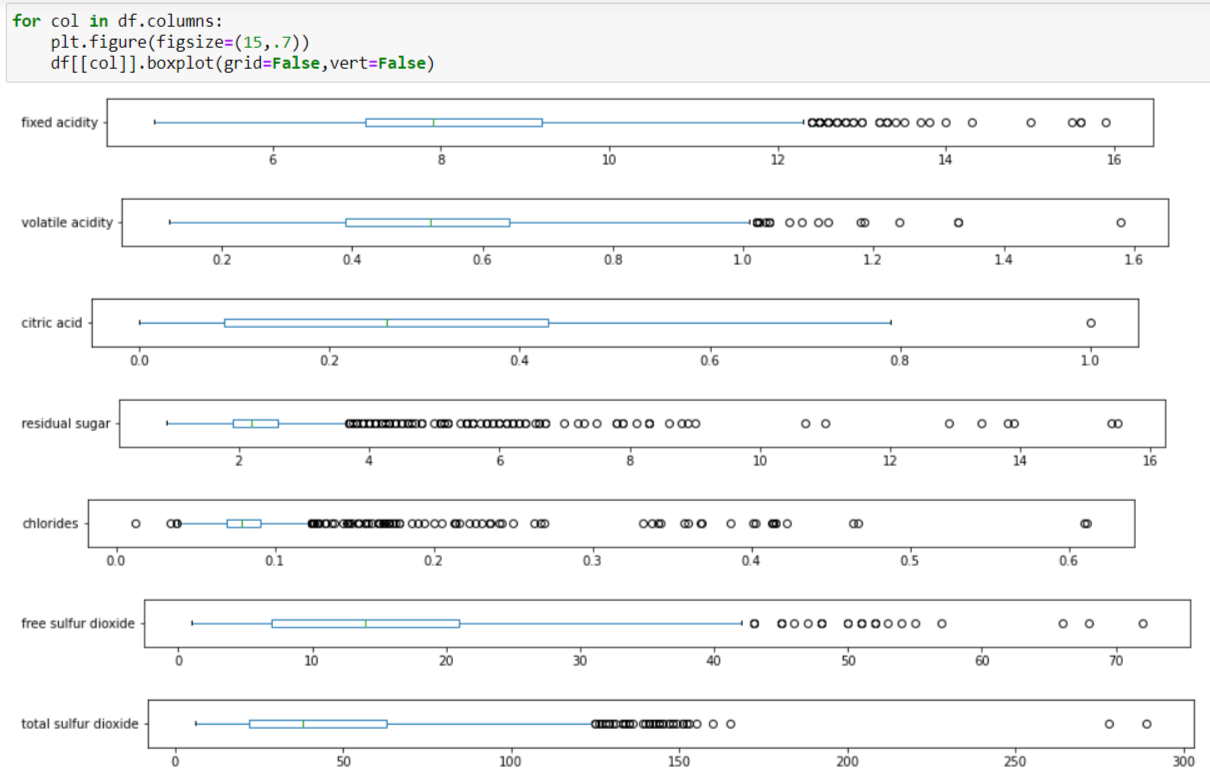
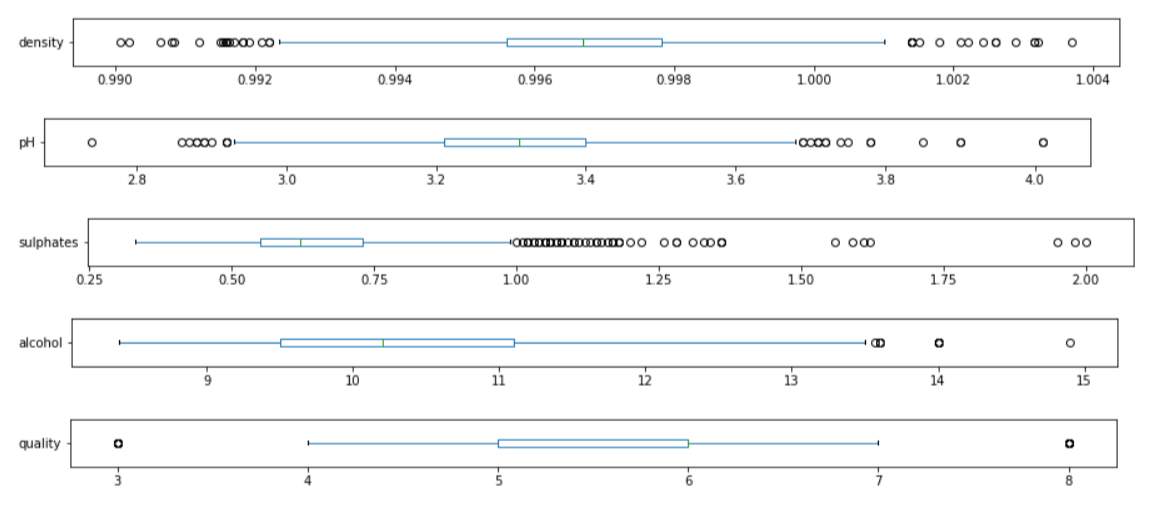
**Exploratory Data Analysis**

The following figure shows the distribution of various features of Dataset. 

The following figure shows the distribution of Wine Quality. 

This shows that in the Wine Quality is ranging from 3 to 8 where 3 signifying the worst and 8 signifying the best quality wine.

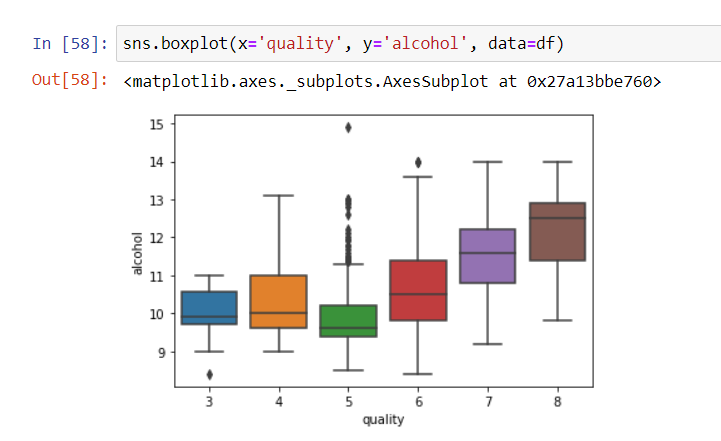
The following figure shows the boxplots for all the features in Wine Quality Data Frame.

The presence of potential outliers can be clearly seen.

#### The following figures show the correlation between features. Observations:

* There is high correlation between fixed acidity and citric acid, fixed acidity and density, alcohol and quality, free sulfur dioxide and total sulfur dioxide.
* High Negative correlation is observed between fixed acidity and pH, volatile acidity and quality, volatile acidity and citric acid, citric acid and pH, density and alcohol.
* Very weak correlation between residual sugar and quality

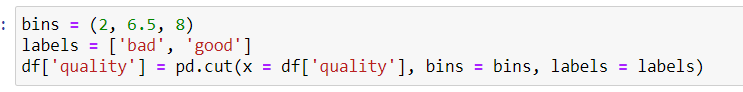
The following figure shows that Higher amount of alcohol leads to better quality. 

The following figure shows that High volatile acidity is bad for Wine Quality.

**Preprocessing**

Next, I create a new Quality feature from old Quality feature by splitting the old quality in two as follows -

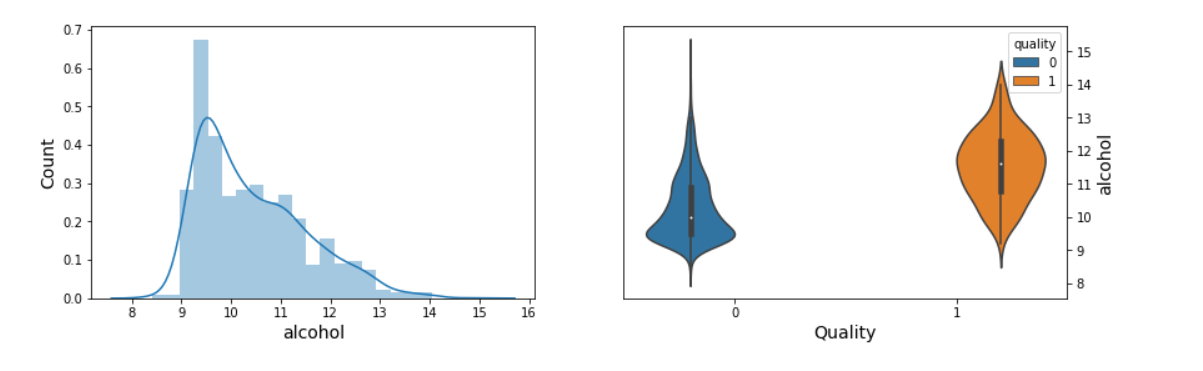
1,2,3,4,5 -> Bad Quality

6,7 -> Good Quality. 

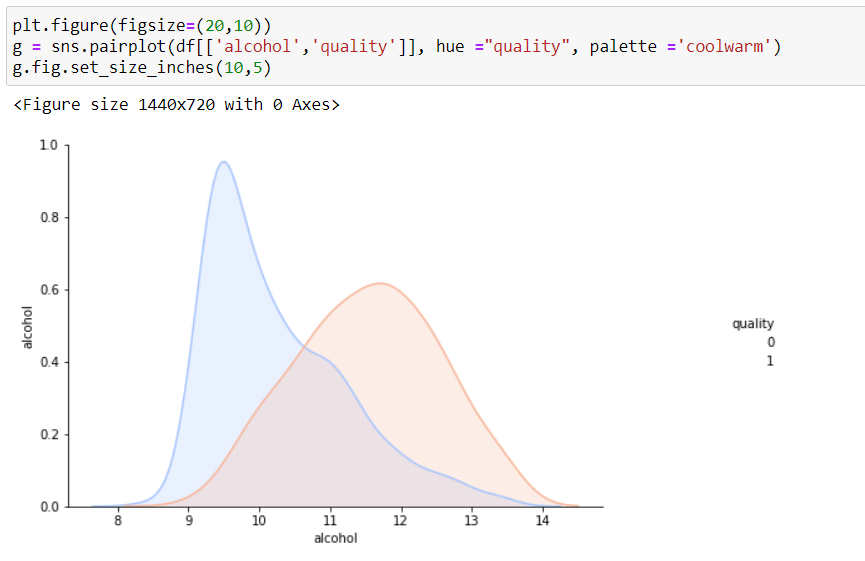


Out of the total 1359 rows, total of 1175 had bad quality label and rest of 184 had good quality label.

The following figure shows the violin plots for alcohol columns for both label good and bad quality.



It can be observed that violin plot for good quality is very much different from violine plot for good quality. Hence, I decided to explore this particular feature further.

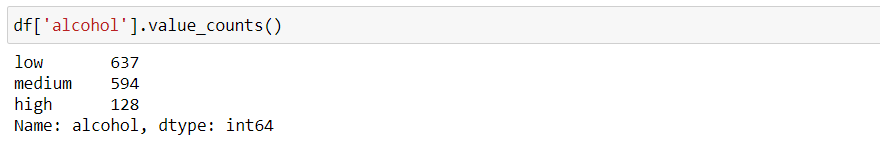
The figure below depicts the pair plot of alcohol.

It can be observed from the figure that wine is very much likely to be good if alcohol amount in it is greater than 12 units. It can also be observed that wine quality is likely to be bad if alcohol amount in it is less than 10 units.

But it will be very ambiguous to say about wine quality if amount of alcohol lies between 10 and 12 units.

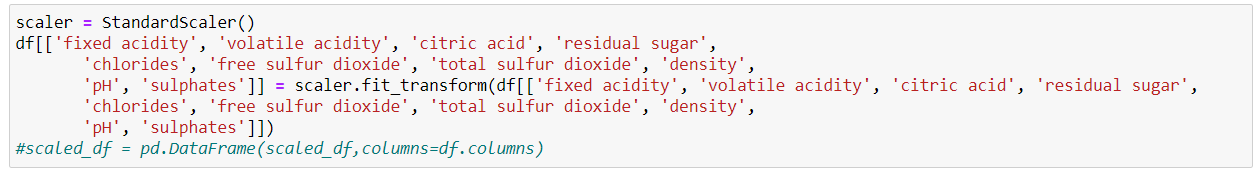
Hence, I decided to categorize alcohol feature into 3 categories low, medium and High.





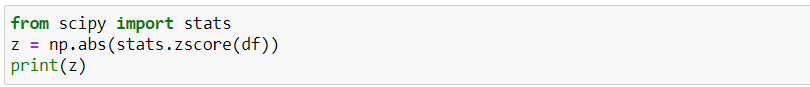
**Standardization**

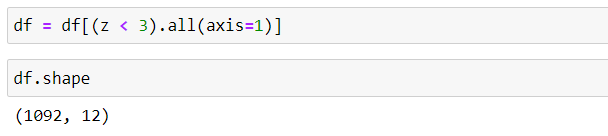
Features were standardized using StandardScaler()



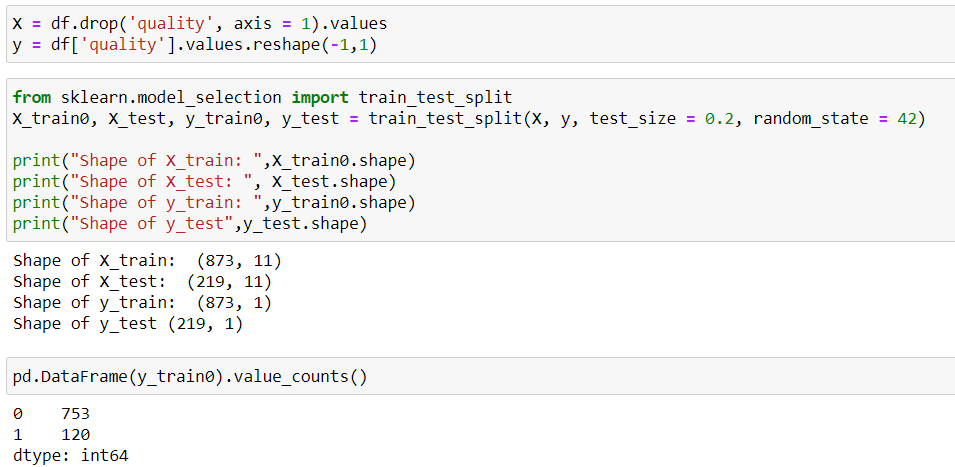
**Outlier Treatment**

Any row containing a value with Z score high than 3 was removed as part of outlier treatment.

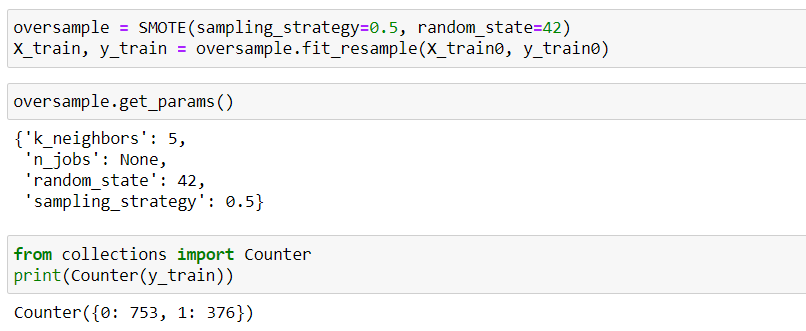


**Train-Test Split**

****

**SMOTE for Balancing Data**

After train test split, the training set has 753 rows for bad quality wine and 120 rows for good quality wine. This is an unbalanced Data set. Hence, I used SMOTE for increasing the number of rows for good quality wine such that the we have half the rows for good quality wine than those that are available for Bad quality wine.



Finally, we have 753 rows for bad quality wine and 376 rows for good quality wine.

**Model Selection**

I trained three models on this data. These models were as follows -

1. Logistic Regression
2. Random Forest Classifier
3. Xgboost Classifier.

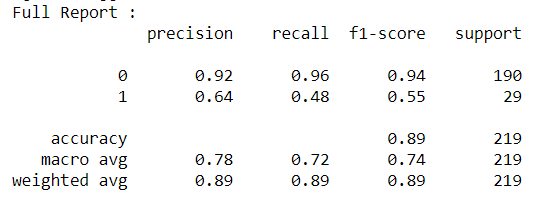
The metric I followed on building my model was roc\_auc score.

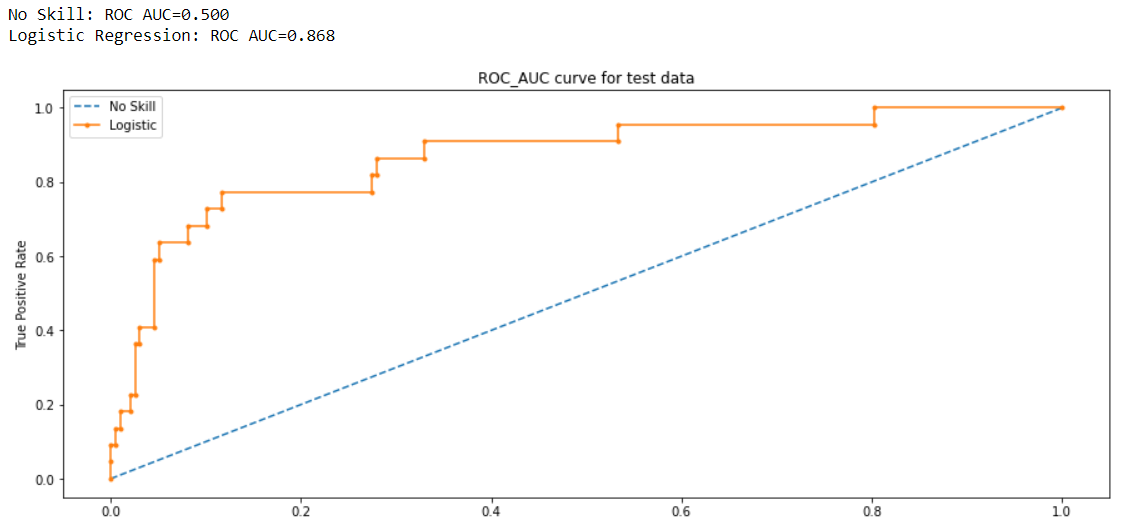
The roc\_auc score for the models were as follows

|  |  |
| --- | --- |
| Model | roc\_auc score |
| Logistic Regression | .868 |
| Random Forest Classifier | .850 |
| Xgboost Classifier | .887 |

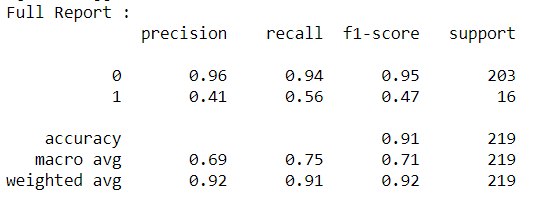
The roc\_auc score for the Xgboost Classifier was the highest.

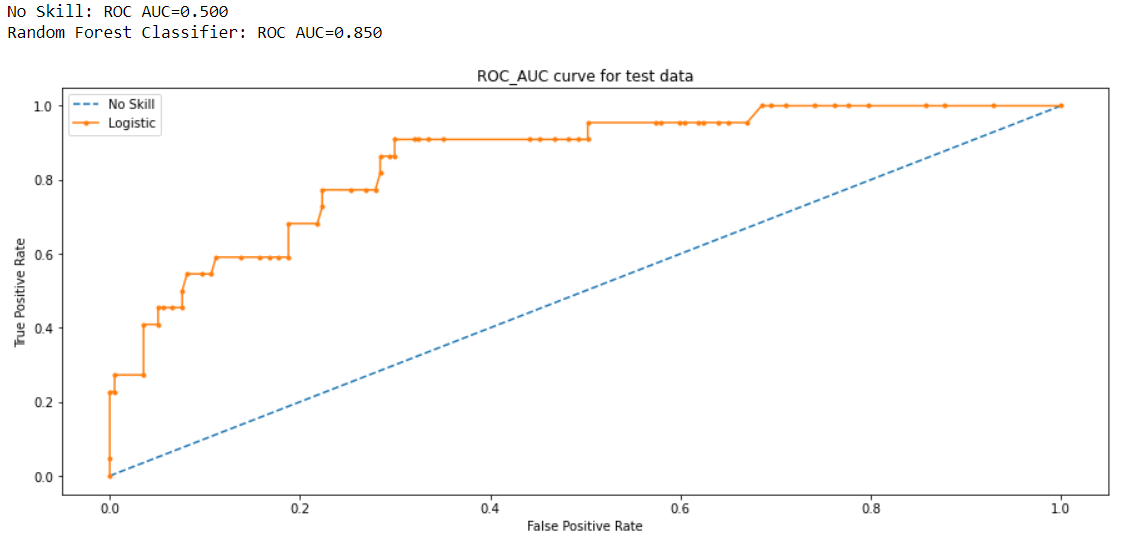
The following is the ROC\_AUC curve for **logistic regression**.



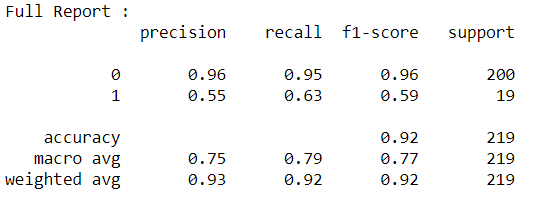


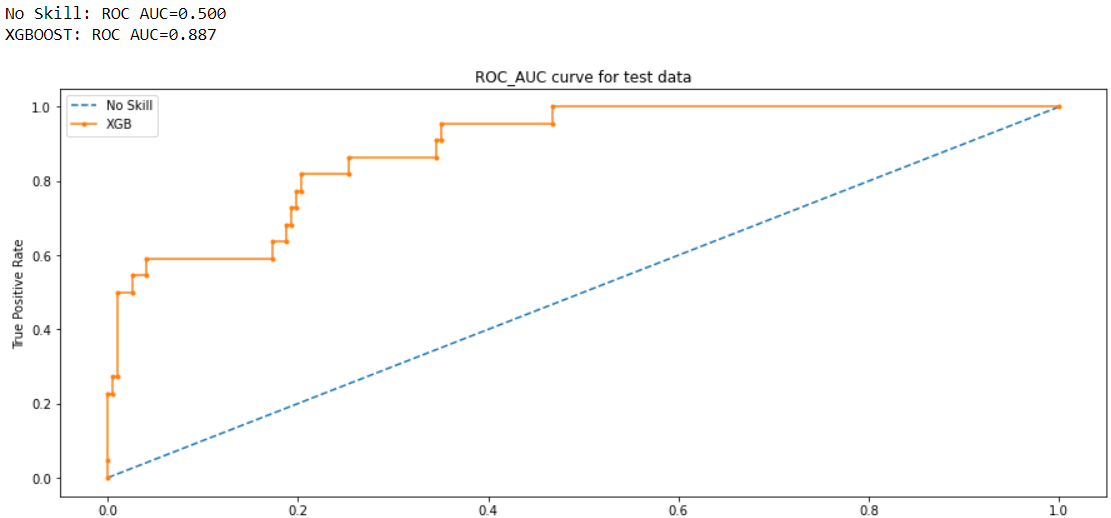
The following is the Model Report and ROC\_AUC curve for **Random Forest Classifier**.





The following is the Model Report and ROC\_AUC curve for **Xgboost Classifier.**





When it came to the models Xgboost performed the best.

**Future Improvements**

1. In future I will try to spending more time on doing better outlier detection and removal.
2. I believe that by doing deep analysis of correlation between various features the model results can be improved.