 **Baseball Case Study**

**SUBMITTED TO SUBMITTED BY**

**DataTrained Academy Satyanath Das**

**Batch No-1829**

Table Content

1.Introduction

2. Problem Statement

3. Exploratory Data Analysis

4. Data Visualization

5. Building Machine Learning Models.

6. Concluding Remarks.

Introduction:

In this project, I have built and evaluated multiple regression models using Python. I have used scikit-learn to calculate the regression, while using pandas for data manipulation and seaborn for data Visualization

Problem Statement:

This dataset utilizes data from 2014 Major League Baseball seasons in order to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. There are 16 different features that will be used as the inputs to the machine learning and the output will be a value that represents the number of wins.

Input features: Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, and Errors

Output: Number of predicted wins (W)

Attribute Information:

W -- WIN (Target Variable)

R -- Runs

AB -- At Bats

H -- Hits

2B -- Doubles

3B – Triples

HR -- Homeruns

BB -- Walks

SO -- Strikeouts

SB -- Stolen Bases

RA -- Runs Allowed

ER -- Earned Runs

ERA -- Earned Run Average

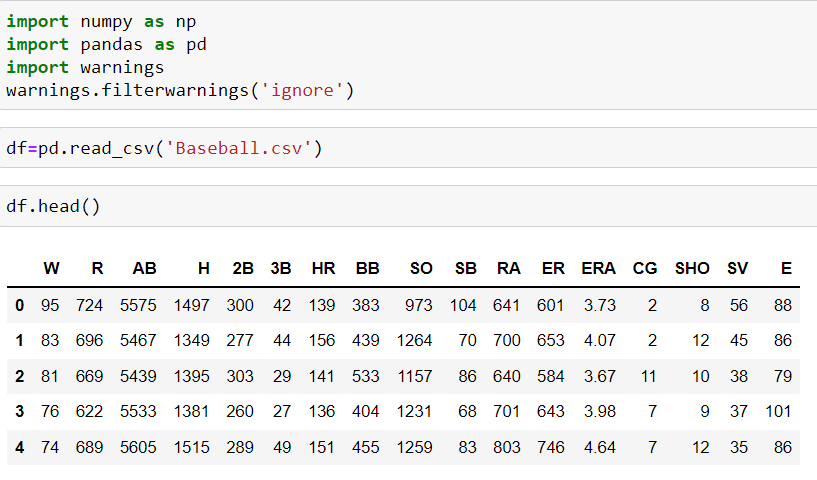
SHO -- Shutouts

SV -- Saves

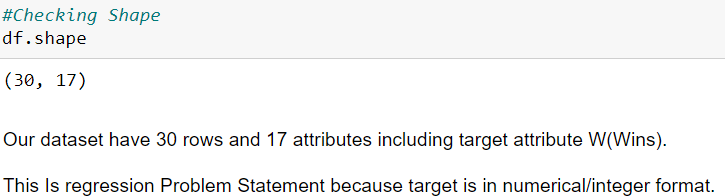
E -- Errors

Importing Dataset and Required Libraries

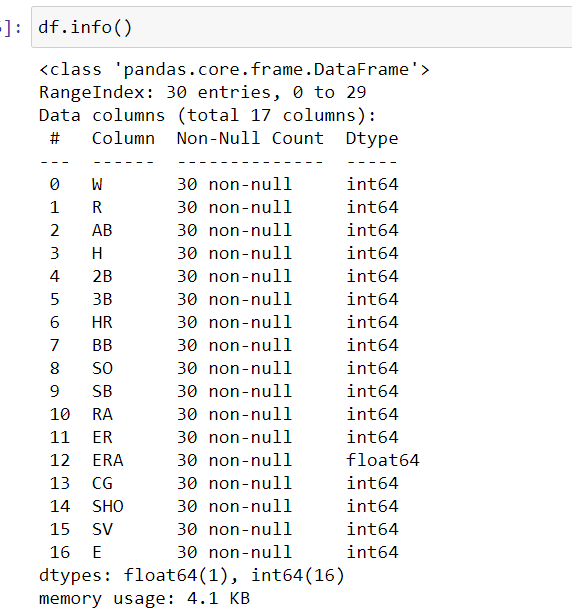
First we will import necessary libraries like ‘Numpy’ and ‘Pandas’ and Dataset



Checking Shape of Dataset



Checking Data Types and Null Values

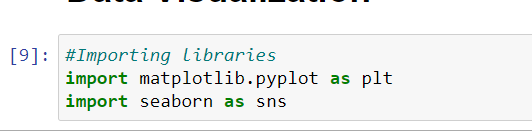


So we see that All the columns have Numerical Value and There are No Null Values.

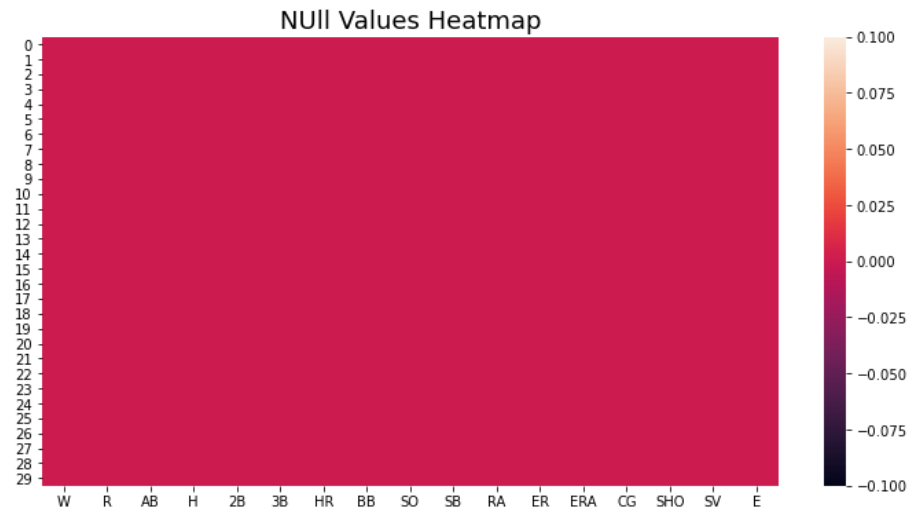
And we also see that the Target Attribute is Numerical Type.So we can say that This is a Regression Type of Problem.

**DATA VISUALIZATION:**

Importing visualization libraries like ‘Matplotlib’ and ‘Seaborn’



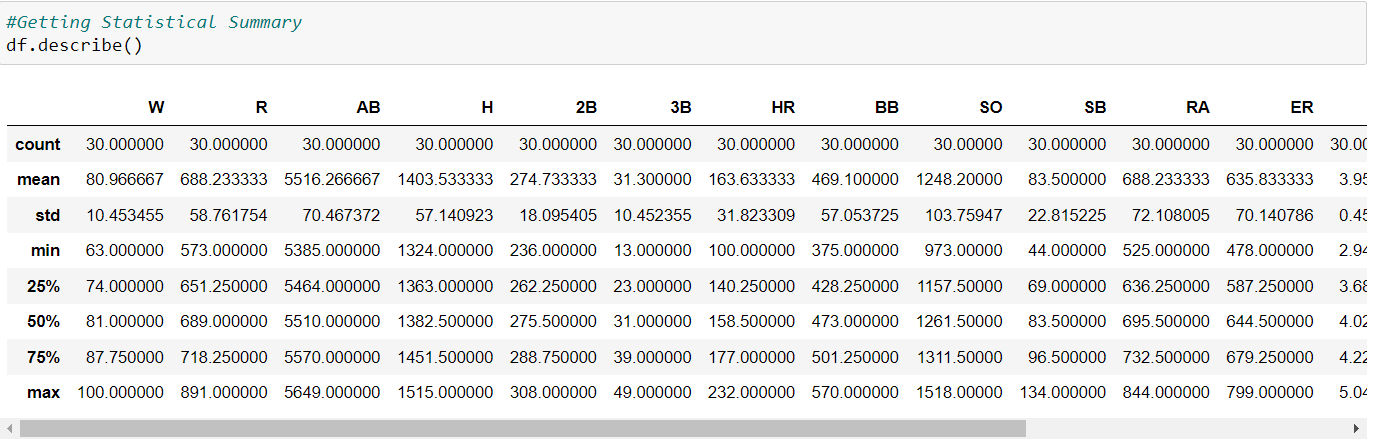
Checking Null Values Using Heatmap:



As we know that ‘There are no Null Values Present’

Statistical Summary:

Statistical Summary is very import information in order to know about distribution/skewness/outliers present in our dataset. Statistical summary generates information only for integer/float datatype.



We can infer from this that Most of attributes data are normally distributed.

In some attributes such as 'H','HR', Mean value is greater than Median value which means data is right skewed in Nature.

We can observe from the above table that for most of attributes Mean and Median (50%) is approximately same.

It seems from above table that possibility of outliers are very less in our dataset.

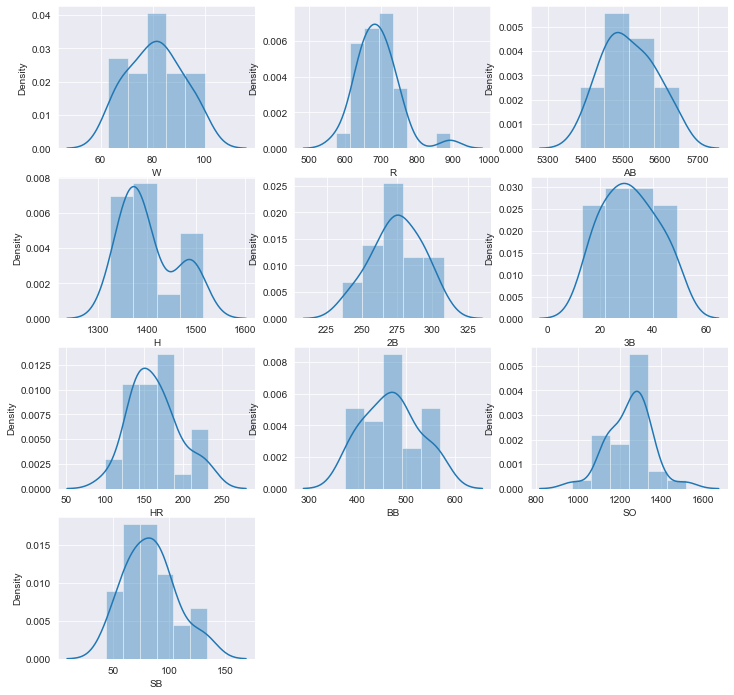
**Data Visualization and Analysis:**

Here we will do univariate and bi-variate analysis with the help of data visualization methods.

Univariate Analysis- Univariate is a term commonly used in statistics to describe a type of data which consists of observations on only a single characteristic or attribute.

Bi-variate Analysis- Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

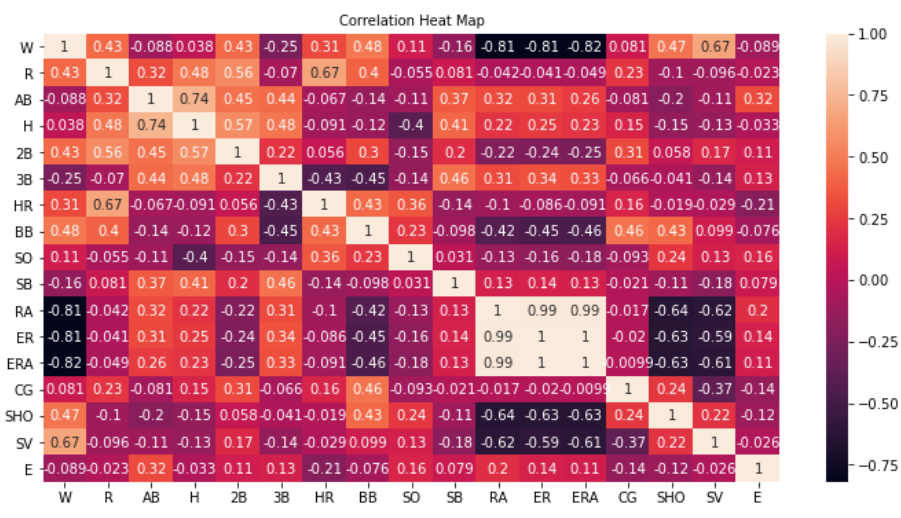
**Checking Data Distribution:**

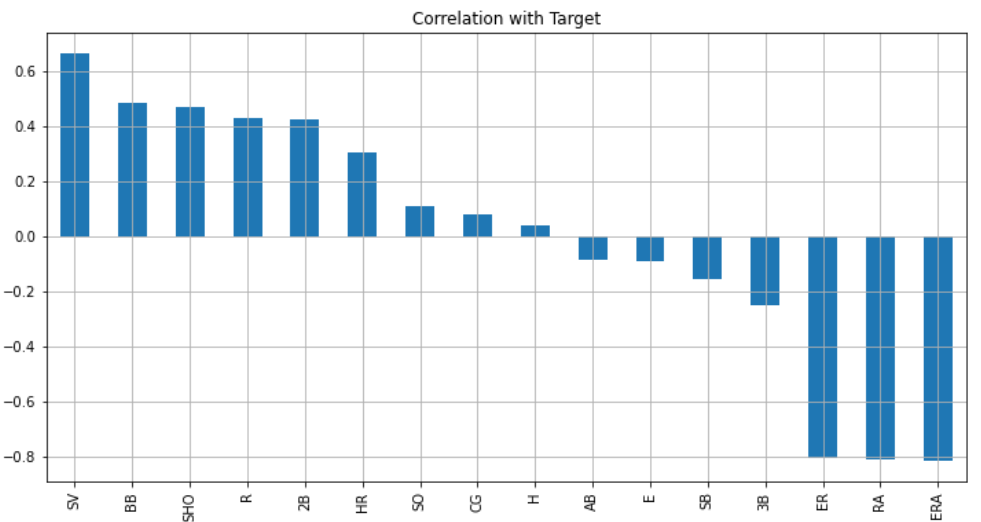


From the above graph we can say that all the columns are normally distributed. One of the column have lowest value and there is less Skewness in the Dataset.

**Checking Correlation of Dataset:**

Correlation plot: Correlation plots can be used to quickly find insights. It is used to investigate the dependence between multiple variables at the same time and to highlight the most correlated variables in a data table. In this visual, correlation coefficients are colored according to the value.

****

****

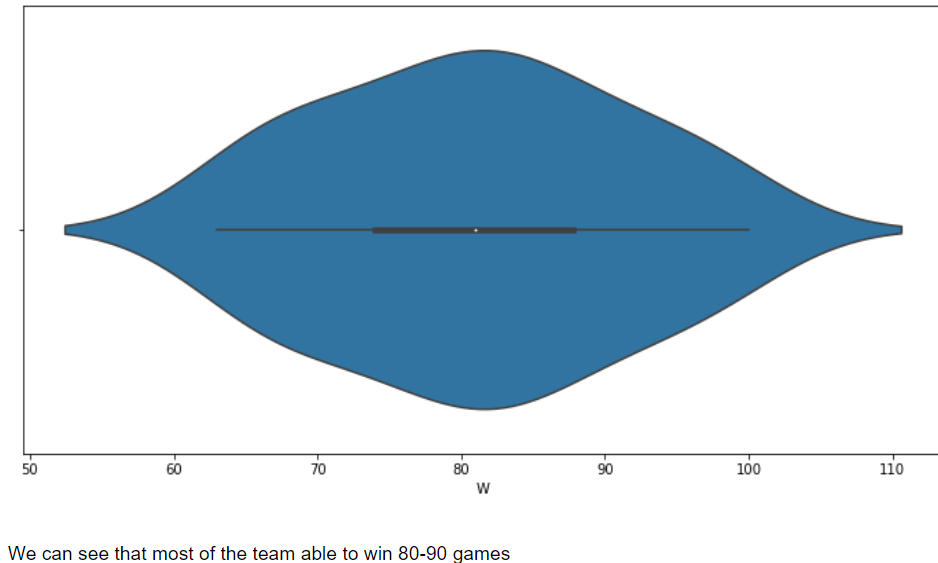
From the above correlation graph we can say that:

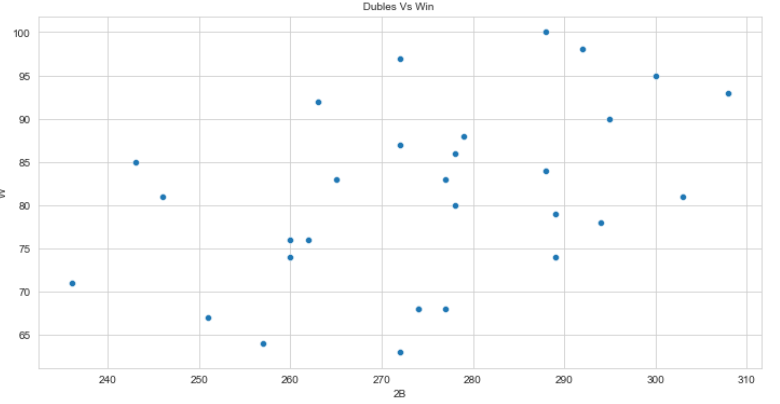
SV, BB, SHO, R, 2B, HR are highly Positively Correlated with Target Variable (Win).

SO, CG, H,AB, E,SB, 3B are Slightly correlated with Target Variable.

ER, RA, ERA are Highly Negatively Correlated with Target Variable.

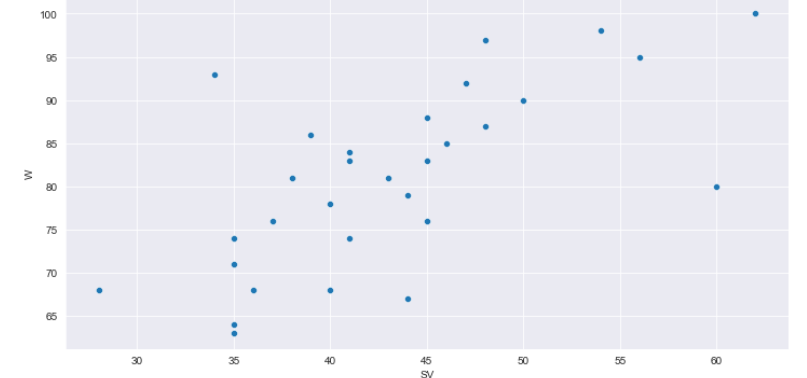
Checking Inside of Data using Visualization:



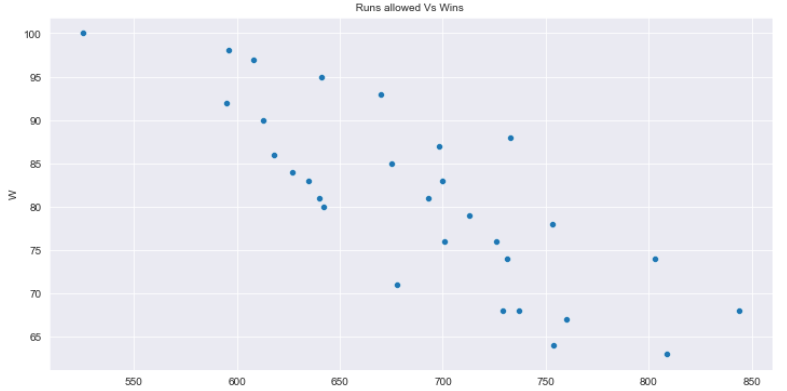
****

There is some positive linear relation we can see here in this plot.

If any team take more double in the match/ Tournament , the number of winning games increases.



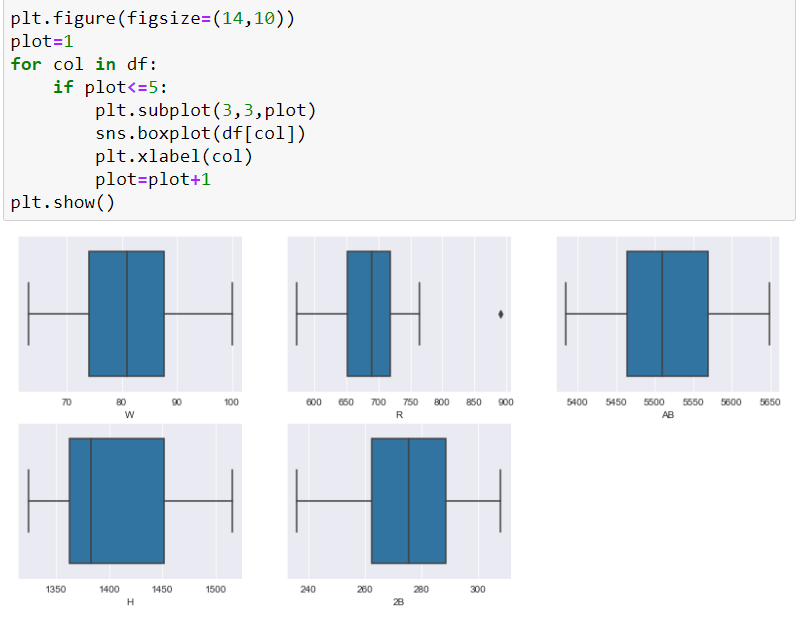
From above scater plot it shows as Number of saves increase the Game winning Chances(numbers) also get increase.



As Runs allowed increases we can see that the number of wins are getting less.

Checking Outliers:

Outliers are those values which differs from the other observations in the Dataset. Outliers can relatively effect the model performance. So we have to check for Outlier and remove them. We use **Boxplot** for outlier detection .



As we can see that there are little Outlier available and outlier is close to Boxplot Whiskers, so we can ignore that Outlier.

If There are outliers available we can use Below Technique to remove the outlier,

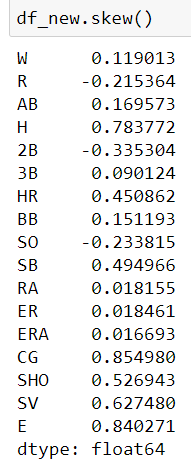
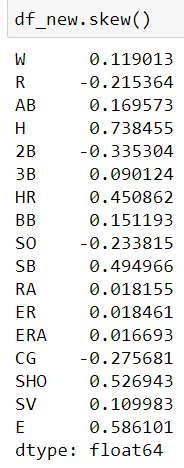
After detecting outliers, it is very important to treat those outliers by taking proper action. Here I will share two method.

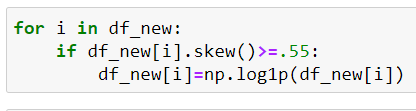
IQR Method—When our data distribution is skewed (right skewed/left skewed) we will use IQR method to remove the outliers.

Z-score Method—We apply Z-score method when our data distribution is normally distributed. When z score is more than 3, we consider those points as an outlier.

**Checking Skewness:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

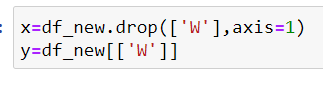




Skewness Removing Method

Before Skewness After Skewness

Splitting data into Input and Output Variable:



**Feature Scaling The Data:**

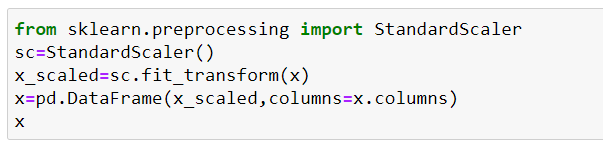
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. ... If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Standard Scaling:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Why scaling?

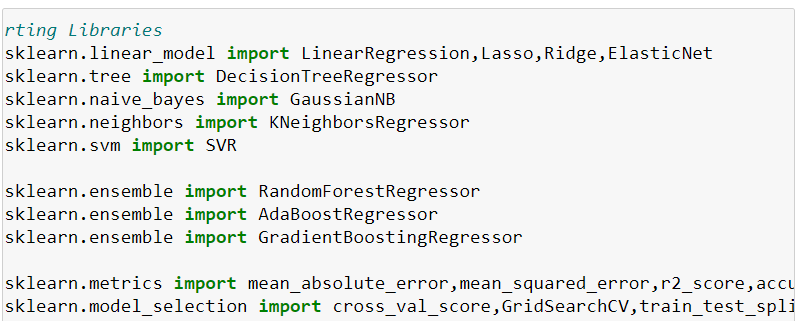
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

****

**Checking Feature Importance:**

****

**Model Building:**

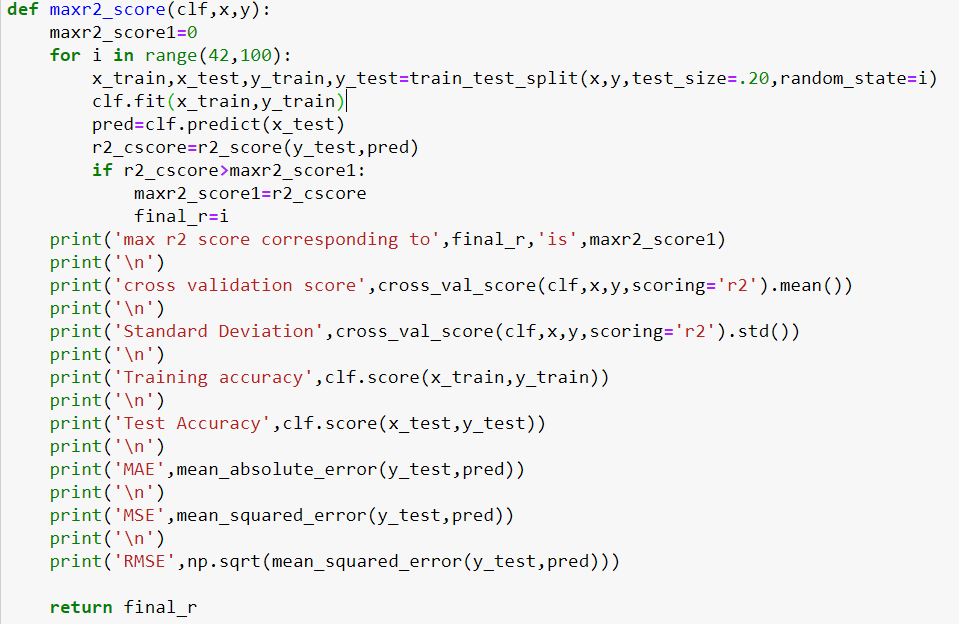


Science it is a Regression Type of problem I choose Linear Regression, Lasso, Ridge, Elasticnet, Decision Tree Classifier, KNEighbors and SVR Algorithms.

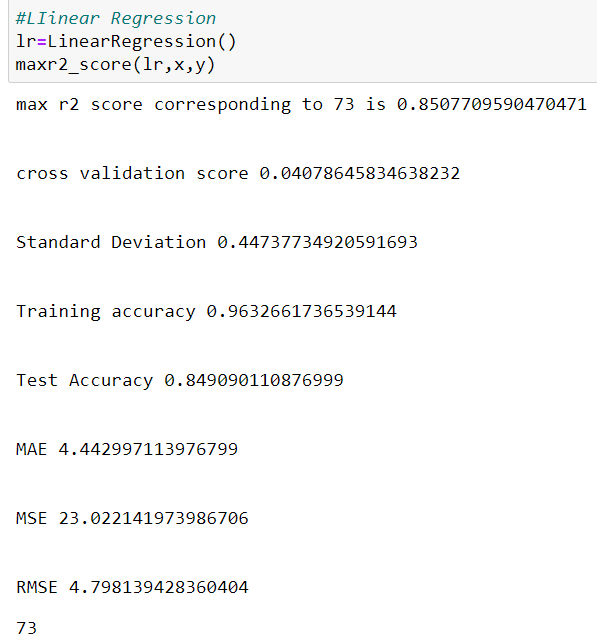
From Sklearn.ensemble I choose Random Forest, AdaBoost and Gradient Boosting.

For Matrices I choose mean absolute error, mean squared error , r2 score, accuracy score.

We also divide the dataset into Train and Test and where Test Size=20%



This The Whole code for Testing The Data.

**Model Selection:**

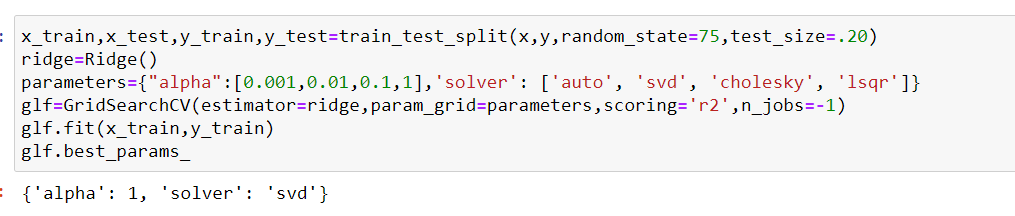
**Above all the Algorithms I saw that Linear Regression Performs well and Give Accuracy score.**

**Training accuracy 0.6182682583105364**

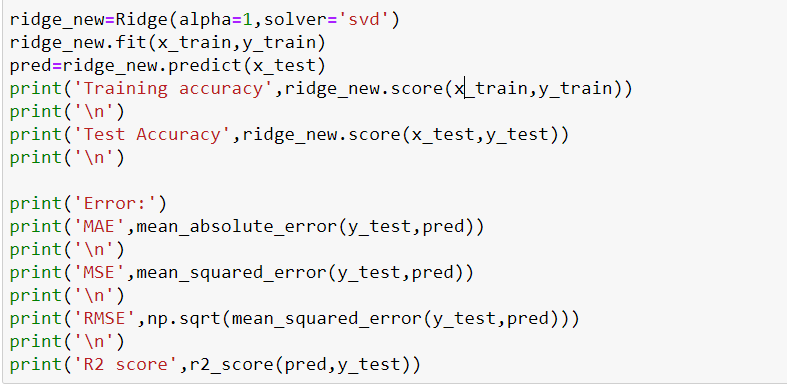
**Test Accuracy 0.6643408594319009**

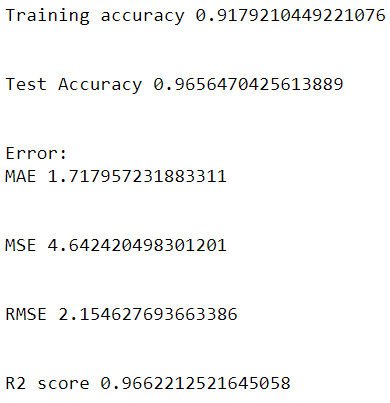
**Hypermeter Tuning:**

Hyperparameters are crucial as they control the overall behavior of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.



Applying Best Parameter:





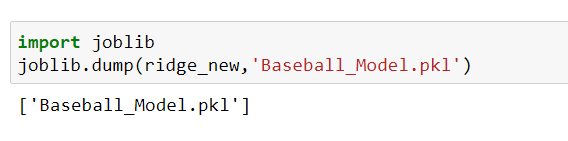
As we can see that after applying Hyperparameter Training we increase the Accuracy of the Model

Current Accuracy:

Training accuracy 0.9179210449221076

Test Accuracy 0.9656470425613889

**SAVING THE MODEL:**

****

Now The Model is Ready To Deploy.

**Conclusion:**

So here we Import the Baseball Dataset into Local System. Then we check the dataset for error and remove them. We understand insight of the dataset. We visualize the various point of the dataset. Then we prepare the dataset and build a predictive model .In this case we Select the Linear Regression Model and save the model in .pkl file

THANK YOU