Machine Learning application — Baseball Case Study

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

**The Dataset:**

The dataset provided to us contains 30 rows, and 17 different independent features. We aim to predict number of wins. Since the data predicts continuous values this clearly is a regression problem, and we will train the regression models to predict the desired outputs.

Mentioned below are the details of the features provided to us, which we will be feeding to our regression model to train it.

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

**Output:** Number of predicted wins (W)

To understand the columns meaning, follow the link given below to understand the baseball statistics: https://en.wikipedia.org/wiki/Baseball\_statistics

**Contents of the article:**

The following information/steps will be covered further in the article –

1. Exploratory data analysis

2. Data modeling

3. Outlier detection and skewness treatment

4. Scaling the data — Standard scaler

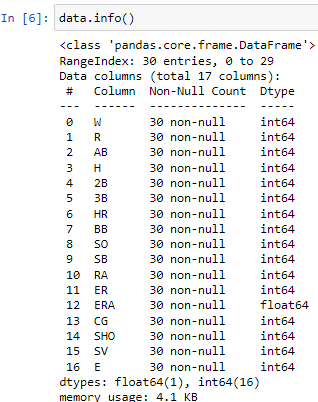
5. Fitting the machine learning models

6. Cross-validation of the selected model

7. Saving the final model and prediction using saved model

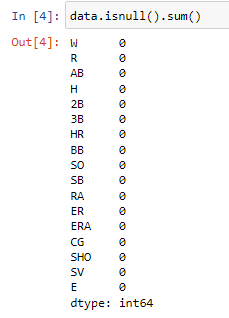
# Exploratory Data Analysis:

# The first step that we do is to check the information about our data. We see the results shown in the image below:

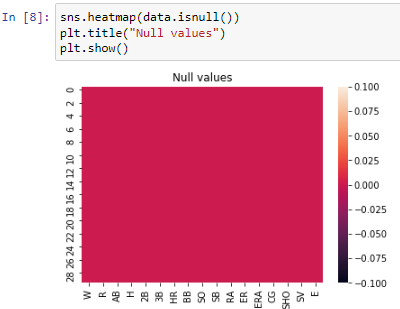


We see that we have all numeric columns. We have 16 integer columns and 1 float type columns. We observe that the count of entries is 30 for all columns, hence no NaN values are present in our dataset.

We confirm this assumption using data.isnull().sum() command –

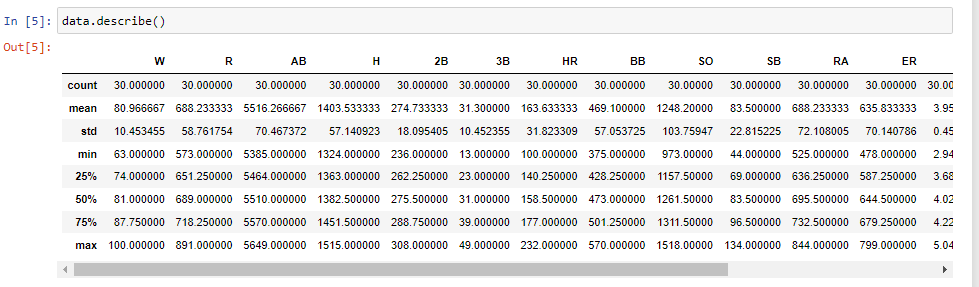


For better understanding, We can also show this with help of heat map



It show that there is no null value present in our dataset

We also check the numerical statistics of our data using data.describe () command –



Following observations are made in this step –

- The ‘W’ column has a range of 63 to 100.

- The ‘R’ column has a minimum value of 573 and maximum value of 891

- The ‘AB’ number has a range of 5385 to 5649

- The ‘H’ starts from 1324 and ends at 1515

- The ‘2B’ starts at 236 and ends at 308

- ‘3B’ range between 13–49.

- The ‘HR’ number has a range of 100 to 232

- The ‘BB’ starts from 375 and ends at 570

- The ‘SO’ starts at 973 and ends at 1518

- ‘SB’ range between 44–134.

- The ‘RA’ number has a range of 525 to 844

- The ‘ER’ starts from 778 and ends at 799

- The ‘ERA’ starts at 2.94 and ends at 5.04

- ‘CG’ range between 0–11.

- The ‘SHO’ starts from 4 and ends at 21

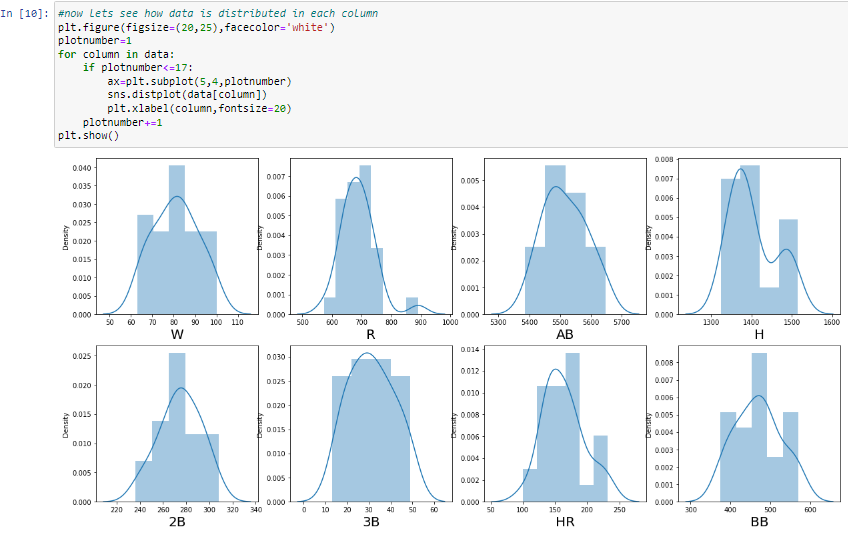
- The ‘SV’ starts at 28 and ends at 62

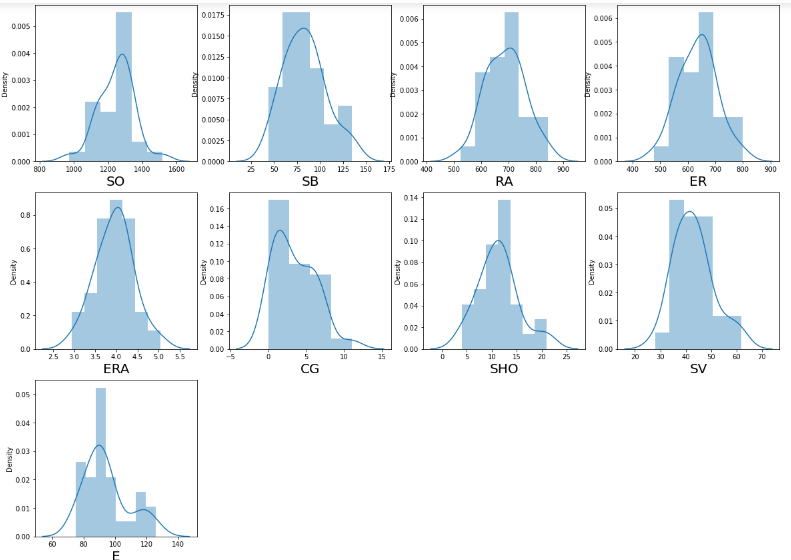
- ‘E’ range between 75–126.

**Data modeling:**

**Graphical Representation:**

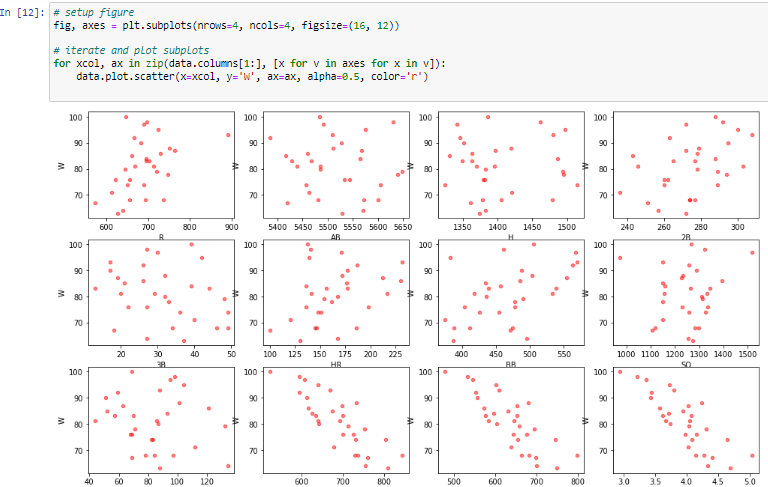
Lets we will do the graphical representation of our dataset, so that we come know that how data is distributed in each column so let’s we will plot distplot

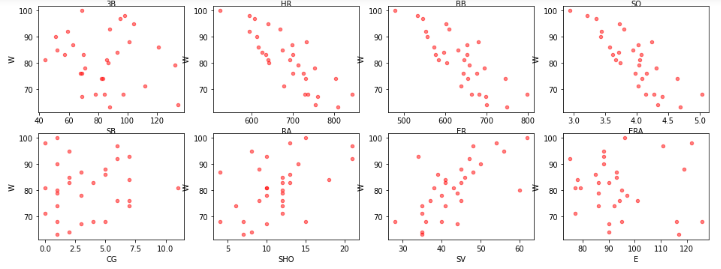




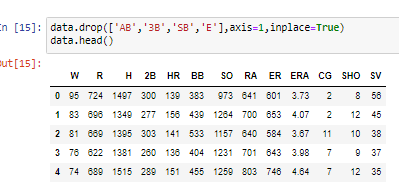
It shows the uniform distribution of data in each column

Now lets we plot the scatter plot so that we come to know that how data is distributed with respect to target column ‘w’ and also will come to know how data is correlated with target column

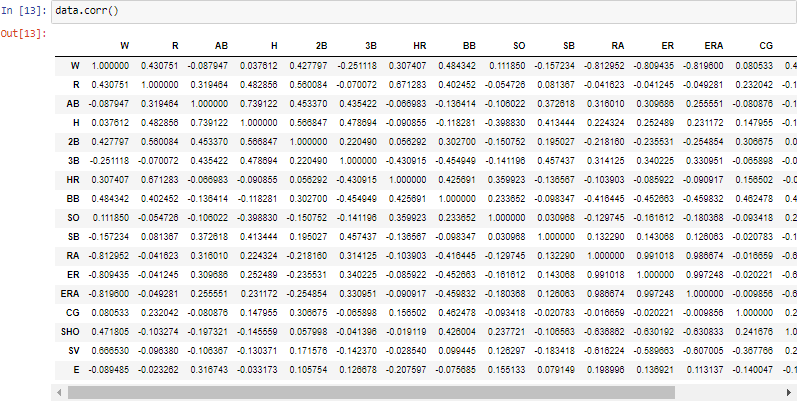




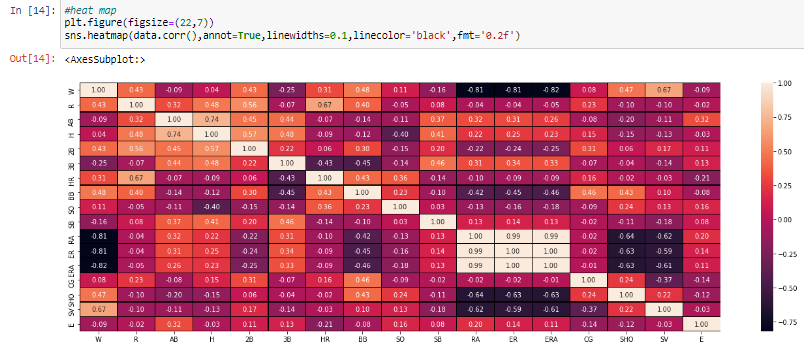
from scatter plot it is seen that column AB , column 3B,column SB , column E shows scattered orientation which show very less dependency over the target variable so let’s will drop that column in further steps, other column shows linear relationship with the target variables



Now will find out the correlation among different columns and also with the target variable it also give positive and negative correlation with target variable



For better understanding this correlation can also shown with the help of heatmap so lets will plot heatmap

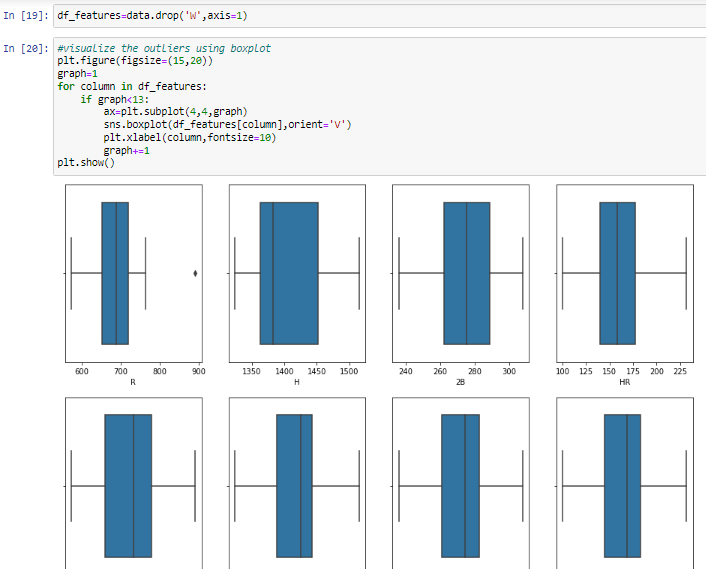


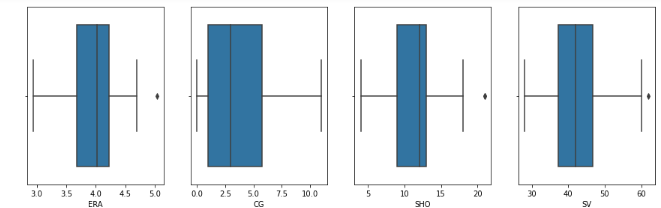
It is found that from heat map the feature R, 2B, HR, BB, SHO, SV shows maximum contribution in ‘w’ column

**Outlier detection and skewness treatment:**

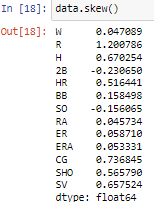
**Outliers detection:**

Lets we will plot the boxplot so that we come to know that whether outliers are presents in dataset or not

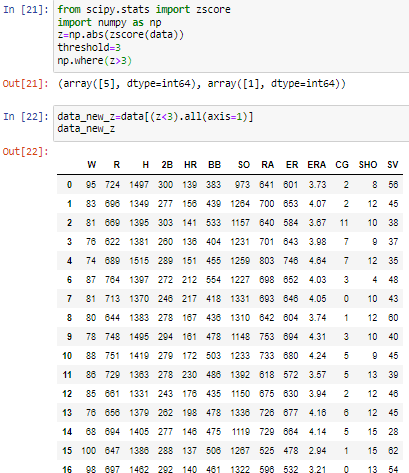




It show that some column contains little amount of outliers in it we will remove it, now we also check skewness

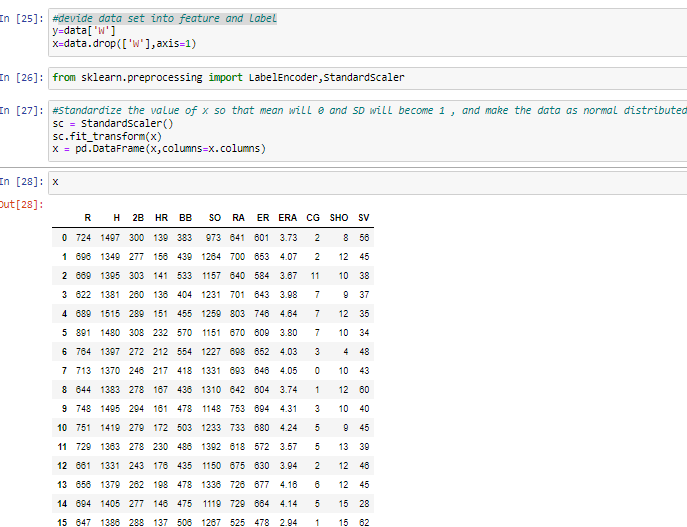


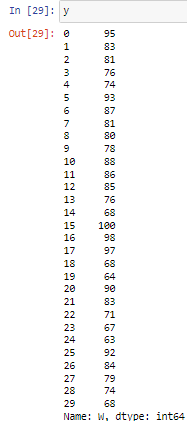
Some amount of skewness present in ‘R’,’H’,’CG’ so lets will perform zscore test by selecting threshold value 3



**Divide data set into feature and label**:

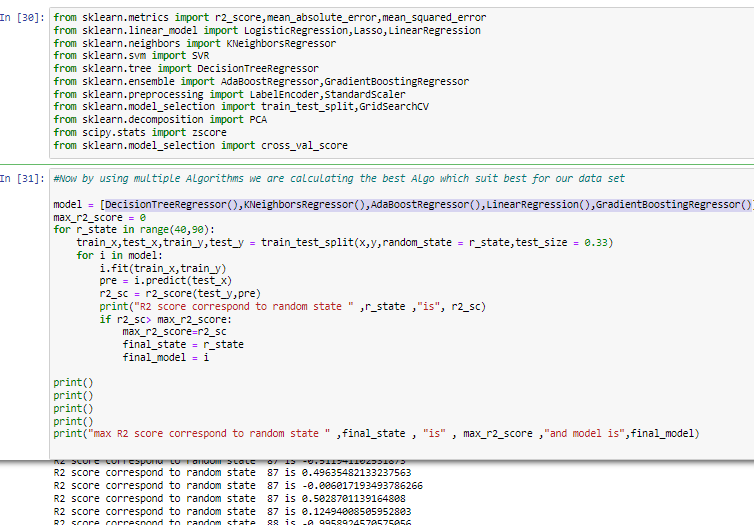
As now have well data so that we will now devide our dataset into feature and target variables for building a model so lets devide data into variable ‘x’ and ‘y’

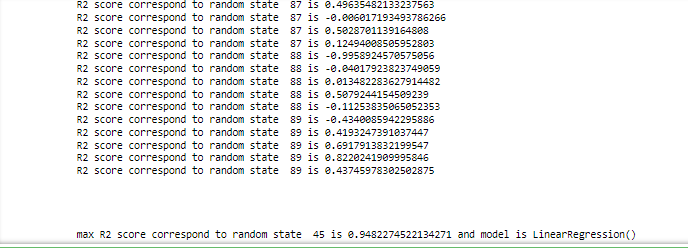




So we have divided whole data set into variable ‘x’ and ‘y’

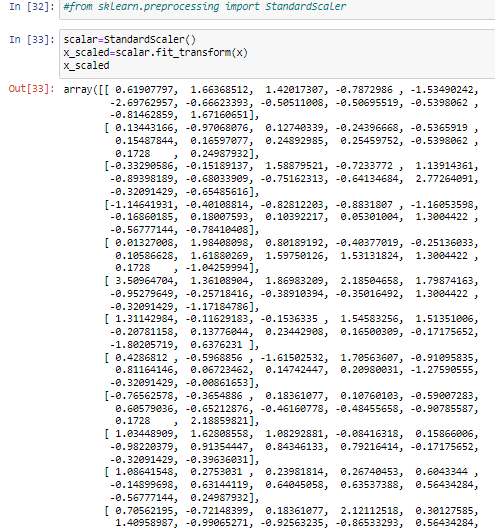
Now by using multiple Algorithms we are calculating the best Algorithm which suit best for our data set for that will import necessary library and will use following different algorithm for our data among that which one will give maximum accuracy will select it, so will use DecisionTreeRegressor, KNeighborsRegressor, AdaBoostRegressor, LinearRegression, GradientBoostingRegressor





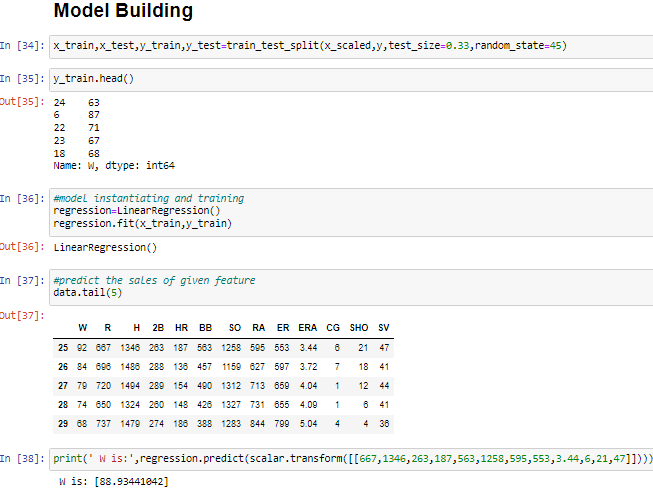
So we got 94.82% accuracy for random state of 45 and model is LinearRegression

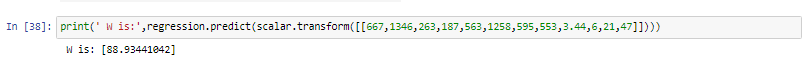
Now we will scale our data with the help of standardscaler



**Model building:**

Now we will built a regression model as it gives better accuracy among other algorithms, we also predict the output with the help of model builded





So we have also predicted the output with the help of model developed

# cross validation of model:

# now we will cross validate our model developed by regularization GridSearchCV

# C:\Users\SAI BABA\Pictures\Screenshots\Screenshot (1372).png

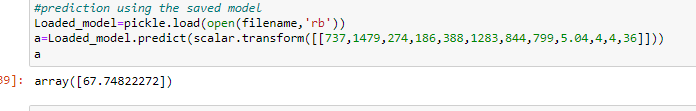
# So we have successfully done cross validation of model now lets we will save the model

# Saving model:

# C:\Users\SAI BABA\Pictures\Screenshots\Screenshot (1374).png

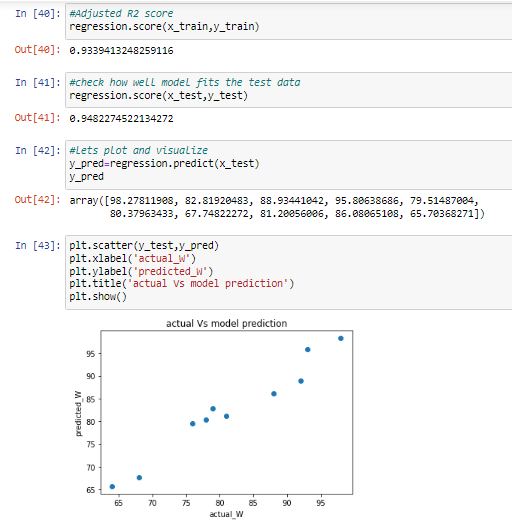
So we have successfully saved our model developed we can use this model for future data prediction

**Conclusion:**

****

So we have tested data by pickle.load and it shows the predicted output so we have successfully developed LinearRegression model that will predict ‘W’

We also check R2score and will also check how well data fit our model developed and also will plot actual vs predicted output graph



So our model developed fit a data well also we have calculated R2score which is 93.39 and we have ploted actual vs model prediction it shows that our prediction is very close to the actual data ,so we can say that we have developed a best model for future data prediction

WRITTEN BY

## Santosh Dharam