**Blog**

on

# Baseball Case Study



# Problem Definition

# Overview: -

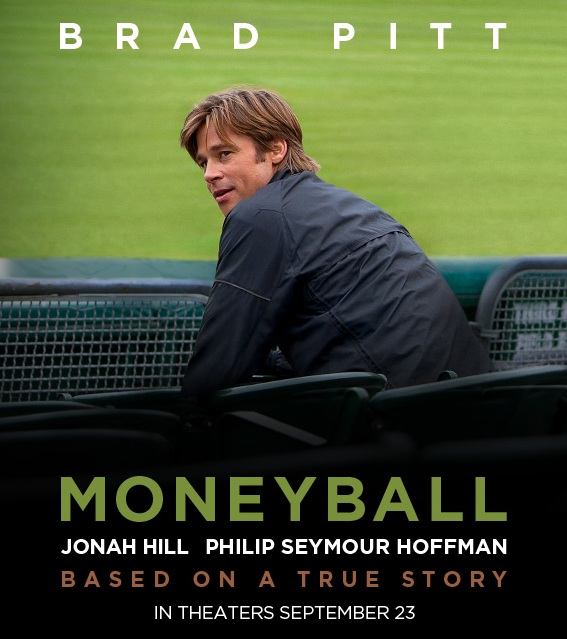
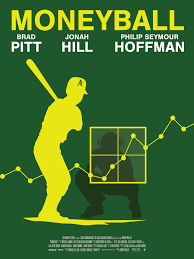
Major League Baseball, a professional baseball league in the US and Canada, is one of the most popular sports leagues in North America. Partially because of its popularity and the wide availability of data from games, baseball has become the subject of significant statistical and mathematical analysis. Machine learning techniques are more mathematically complex and able to fit variations in baseball game outcomes better.

There is indeed a relationship between the statistics to a game and its outcome & baseball statistics can be collected and analysed in such a way that provides accurate answers to specific questions. On the basis of predictions baseball teams experience various patterns and react to factors in a repetitive manner, this ultimately affects their in-game performance and increases probability of wins.

This project is incomplete without discussing about baseball sports analytics movie Moneyball which was released in 2011.

**Money Ball Movie: -**

Moneyball is a 2011 American [biographical](https://en.wikipedia.org/wiki/Biographical_film) [sports](https://en.wikipedia.org/wiki/Sports_film) [drama film](https://en.wikipedia.org/wiki/Drama_(genre)) directed by [Bennett Miller](https://en.wikipedia.org/wiki/Bennett_Miller) and written by [Steven Zaillian](https://en.wikipedia.org/wiki/Steven_Zaillian) and [Aaron Sorkin](https://en.wikipedia.org/wiki/Aaron_Sorkin). The film is based on [Michael Lewis](https://en.wikipedia.org/wiki/Michael_Lewis)'s 2003 [nonfiction book of the same name](https://en.wikipedia.org/wiki/Moneyball), an account of the [Oakland Athletics](https://en.wikipedia.org/wiki/Oakland_Athletics) [baseball](https://en.wikipedia.org/wiki/Baseball) team's [2002 season](https://en.wikipedia.org/wiki/2002_Oakland_Athletics_season) and their general manager [Billy Beane](https://en.wikipedia.org/wiki/Billy_Beane)'s attempts to assemble a competitive team. In the film, Beane ([Brad Pitt](https://en.wikipedia.org/wiki/Brad_Pitt)) and assistant general manager [Peter Brand](https://en.wikipedia.org/wiki/Paul_DePodesta#Moneyball) ([Jonah Hill](https://en.wikipedia.org/wiki/Jonah_Hill)), faced with the franchise's limited budget for players, build a team of undervalued talent by taking a sophisticated [sabermetric](https://en.wikipedia.org/wiki/Sabermetric) approach to [scouting](https://en.wikipedia.org/wiki/Scout_(sport)) and analysing players.

**Need: -** Predicting the winner of a particular Major League Baseball (MLB) game is an interesting and challenging task. Previously there was no definitive formula for determining what factors will conduct a team to victory, but through the analysis of many years of historical records many trends could emerge. Recent studies concentrated on using and generating new

statistics in order to rank teams and players according to their perceived strengths and consequently applying these rankings to forecast specific games. This model approach uses past data of Major League Baseball (MLB) 2014, when making a prediction.

Major League Baseball (MLB) is a multi-billion dollar statistically filled business, and many people are strongly interested in developing systems with the aim of providing the best prediction of the winner in many specific baseball games. Large quantities of historic baseball data are currently available (often publicly available) from different sources in the form of numerically or symbolically represented statistics (e.g., general season information, play-by-play, game logs, players line-up, etc.).

# Aim of this Project: -

# In this project, we will test out several machine learning models to predict the number of wins for a given baseball team in the 2015 season based on several different indicators of success. 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. This Project is a Regression based problem as output variable Number of predicted wins (W) is continuous type of data in nature. So, we will use multiple regression algorithms to build our model.

# Steps used in this project: -

1- Define the Problem

2- Data Gathering

3- Data Cleaning

4- Data Exploration and Visualization

5- Train the algorithm

6- Evaluate our model using evaluation metrics & etc.

# Dataset Information: -

Dataset provided in \*.csv format from 2014 Major League Baseball season.

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.

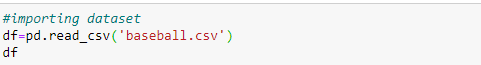
Input features: Runs, At Bats, Hits, Doubles, Triples, Homeruns, Walks, Strikeouts, Stolen Bases, Runs Allowed, Earned Runs, Earned Run Average (ERA), Shutouts, Saves, Complete Games and Errors

-- Output: Number of predicted wins (W)

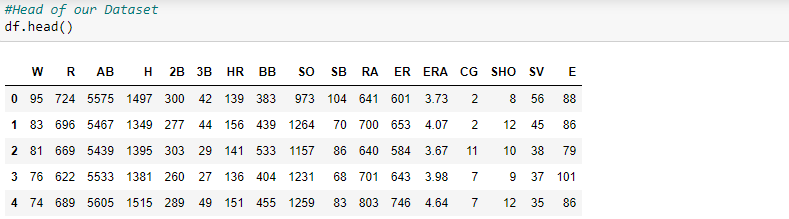
**Data Analysis**

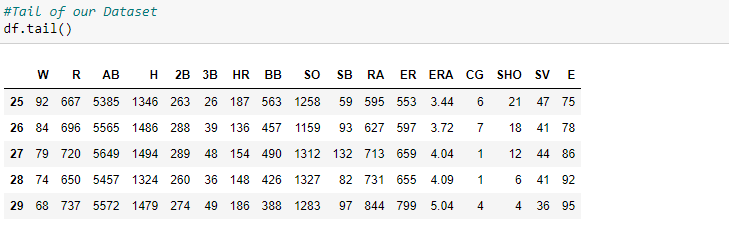
First step in any project is importing basic libraries to do data analysis and import dataset from data source on which we will work, so we are importing it.



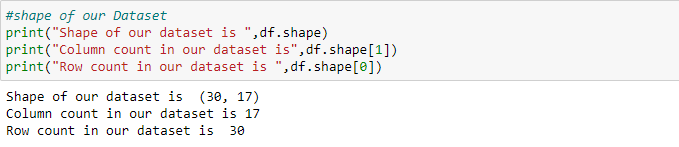


Now we will check Head and Tail of our Dataset: -

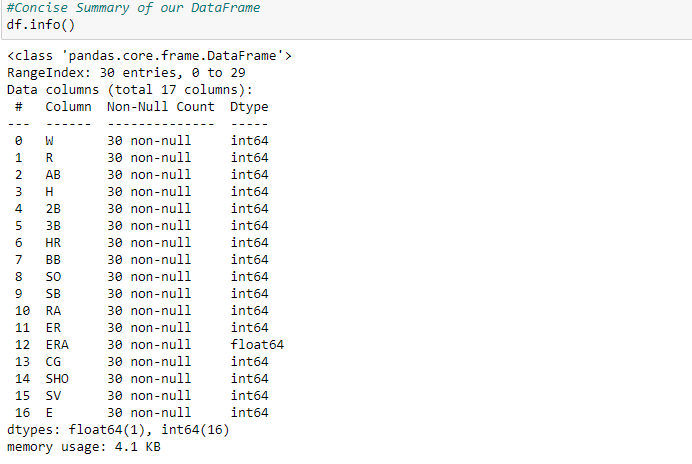




Now we will check shape of our dataset means rows and columns count in our dataset.



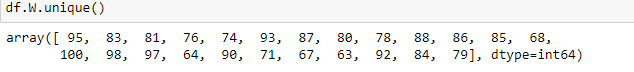
Now we will check Concise Summary of our Dataset. It will show us column information, datatype of all features, also length of each column and most important presence of null values.



Our Dataset has no missing values and no column have object type data.

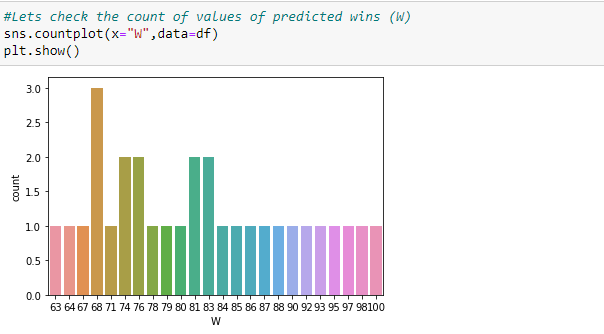
# Exploring Target Variable: -

Target variable Number of predicted wins (W) is continuous type of data in nature.





Now we will do analysis of our Target variable using Count Plot.



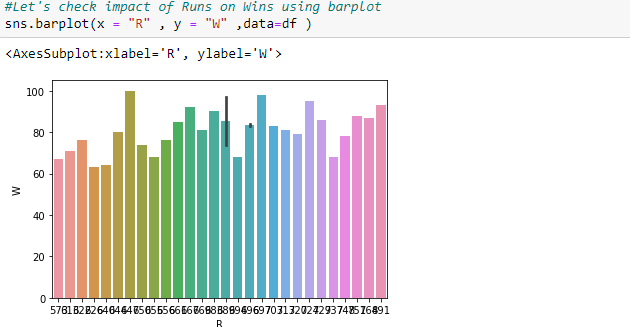
Observations: -

1- Number of predicted wins (W) has most values concentrated in the categories of 68,74,76,81,83.

2- In remaining categories, few observations are present and mostly are same.

**Impact of input features on Number of predicted wins (W): -**

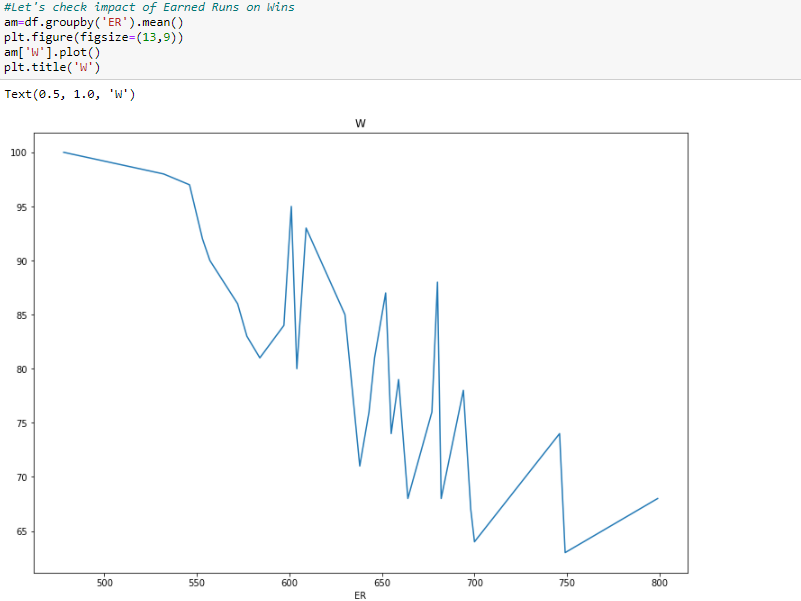
1. Runs- We will check impact of Runs on Number of predicted wins.

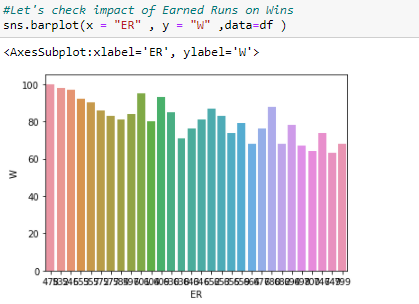


Observation:

At Runs 647-win prediction is 100.

2. Earned Runs: - We will check impact of Earned Runs on Number of predicted wins.

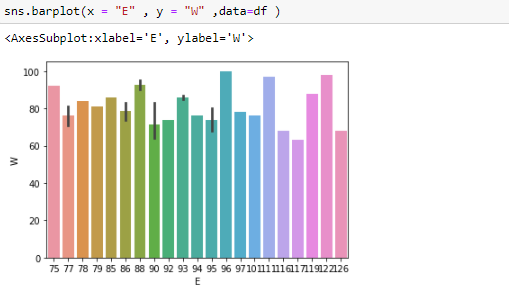


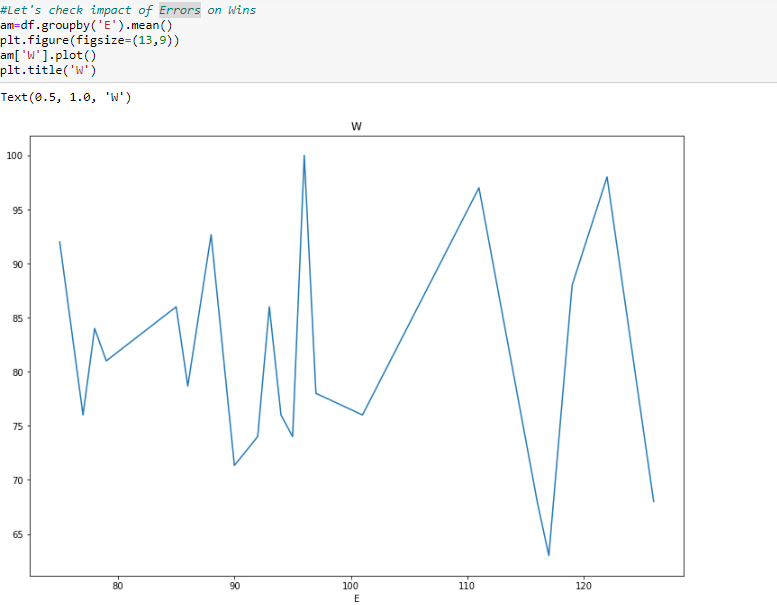


Observation:

On basis of above plots and calculations, win prediction is 100 at 478 ER.

3. Errors: - We will check impact of Errors on Number of predicted wins.

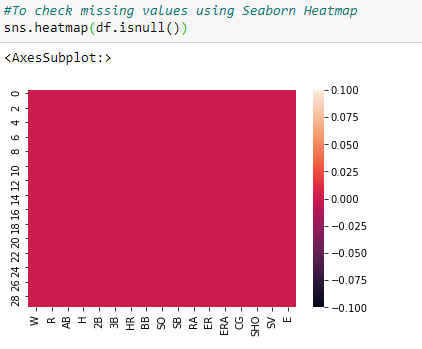




Observation:

On basis of above plots and calculations, win prediction is 100 when Error value is 96.

**Let’s check Null values using Heatmap: -**

****

Observation: -

Dataset has no missing values.

**EDA Concluding Remarks**

In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

# 1- Summary Statistics: - In descriptive statistics, summary statistics is used to summarize set of observations, in order to communicate the largest amount of information as simply as possible. It includes central Tendency, dispersion, skewness, variance, range, deviation etc.

# 

Observation on basis of Summary Statistics: -

The Mean is more than median (50th Percentile) in 7 columns and median is more than mean in 9 columns and for SB column mean & median is same.

# 2-Correlation Matrix: - A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. A correlation matrix consists of rows and columns that show the variables.

# 

# Checking correlation using Heatmap with annotations: -

# 

Observations: -

1- Output feature (W) is highly correlated with SV.

2- Dark shades are highly correlated.

3- Output feature (W) is negatively correlated with features RA, ER, ERA.

4- Earned Runs (ER) & Earned Run Average (ERA) features are very highly correlated.

**3-Plotting Outliers: -** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers should be investigated carefully. So, we will use graphical techniques Box Plot and Scatter Plot for identifying outliers.

**Box Plot-** The box plot is a useful graphical display for describing the behaviour of the data in the middle as well as at the ends of the distributions. The box plot uses the [median](https://www.itl.nist.gov/div898/handbook/eda/section3/eda351.htm) and the lower and upper quartiles (defined as the 25th and 75th [percentiles](https://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm)). A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median

**Scatter Plot-** Scatter plot reveals relationships or association between two variables.  If one point of a scatter plot is farther from the regression line than some other point, then the scatter plot has at least one outlier.  If a number of points are the same farthest distance from the regression line, then all these points are outliers. If all points of the scatter plot are the same distance from the regression line, then there is no outlier.

# A- Univariate Analysis: - Univariate involves the analysis of a single variable. First, we are going to do univariate analysis using Box Plot method.

# 

# 

# Observations on basis of Box Plots: -

# 1-Outliers are present in R, ERA, SHO, SV & E columns. In ERA, SHO, SV columns data is close to whisker as compared to R & E.

# 2- Boxplot of R, ERA, SHO & SV columns shows minimal outliers however column E have maximum outliers present.

# Now we are creating histogram of each input variable to get a broader idea of the distribution.

# 

# 

# Observation: -

# Histograms are a handy way to identify outliers. Presence of unusual values in above histograms & also distribution is not normal in some columns and these things denote the possibility of potential outliers.

# B- Bivariate Analysis: - It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. We are using scatterplot for this purpose.

# 

# .

# Observations: -

# 1- Scatter Plot between W and E & between W and R is shown here.

# 2-From scatter plot, we can look at the interaction between the variables and some variables are negatively related to each other and some plots are showing negative correlations.

# Joint plot to check relation between output variable & Earned Run Average: -

# 

# Using Marker Plot for checking highly correlated values with Output variable W: -

# This function is used to draw points (markers) in a diagram.

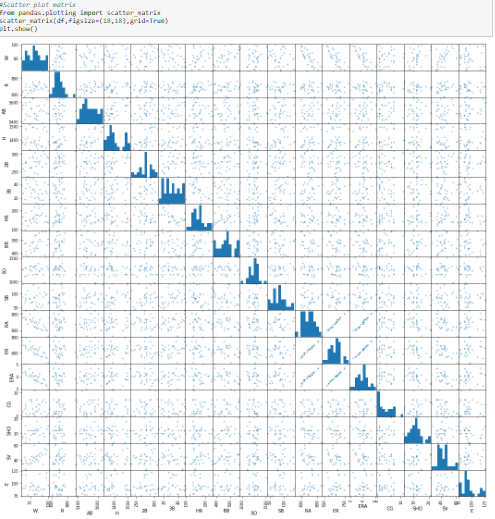
# 

# 

# Using Marker plot, we have marked the data points in our graph using ‘o’(circle) markers to emphasize each point and highly correlated values.

# C- Multivariate Analysis: - Multivariate analysis is used to study more complex sets of data. It is a statistical method that measures relationships between two or more response variables.

# Scatter plot matrix: -Scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables.



Observation:

Using multivariate analysis, we can look at the interactions between the variables. Scatterplots of all pair of attributes helps us to spot structured relationship between input variables. Diagonal grouping of some pairs of attributes suggests a high correlation and a predictable relationship.

4- **Skewness-** Skewness is a measure of asymmetry or distortion of symmetric distribution. It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as normal distribution. A normal distribution is without any skewness, as it is symmetrical on both sides. Hence, a curve is regarded as skewed if it is shifted towards the right or the left. Skewness is of 2 types: - 1- Positive Skewness 2- Negative Skewness.

# Distplot to check Distribution of Skewness: - Distplot plots a univariate distribution of observations. The distplot () function combines the matplotlib hist function with the seaborn kde plot () and rug plot () functions. For individual columns we are using Distplot.

# 

# 

# Observation: -

# Distribution of Data in columns E, CG, H & R is not normal. Except these columns, some are distributed normally and others have very minimal deviation.

# We can also check skewness using skew () method.

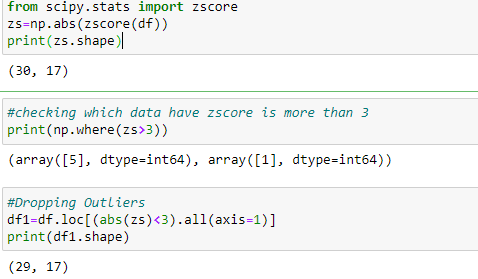
# 

# Observation:

# Columns R, H, CG, SV and E have presence of skewness and R have high skewness.

5- **Removing Outliers: -** Most parametric analysis methods require valid data distribution assumptions, and the existence of outliers very often results in the violation of such assumptions. So, we need to remove outliers as they are present in our dataset.

**Z-score**- Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. We will calculate Z-score and use it in outlier removal.

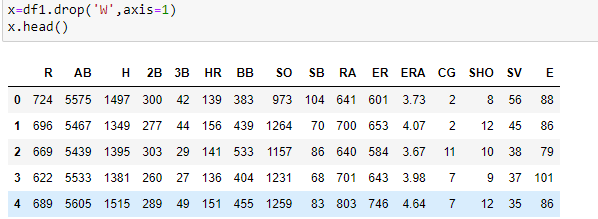


Observation:

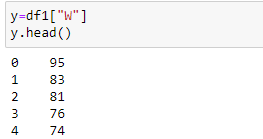
So, as per above shapes: 30-29=1 so 1 outlier is removed and data loss is nearly 3%.

**Pre-processing Pipeline**

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. Pipelines are a simple way to keep our data pre-processing and modelling code organized. Many stages involved in this process are Data Cleaning, Data Integration, Data Transformation & Data Reduction. We will start this process by dividing dataset into input & output.

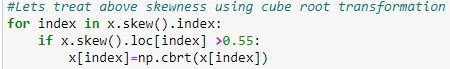
****

Output feature: -

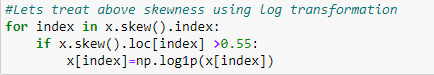
****

**Data Transformation by Removing skewness from input dataset: -** We will use cube root & log transform method to remove skewness in input dataset.

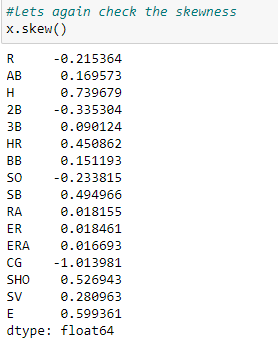
**Cube Root transformation**: The cube root, x to x^ (1/3 is a fairly strong transformation with  
a substantial effect on distribution shape used for reducing right skewness, and has the  
advantage that it can be applied to zero and negative values.



**Log transformation**: - Log transformation is a data transformation method in which it replaces each variable x with a log(x). It is commonly used for reducing right skewness and is often appropriate for measured variables. We will use NumPy log1p () method for this process.



After using transformation methods, we will again check skewness in input dataset.

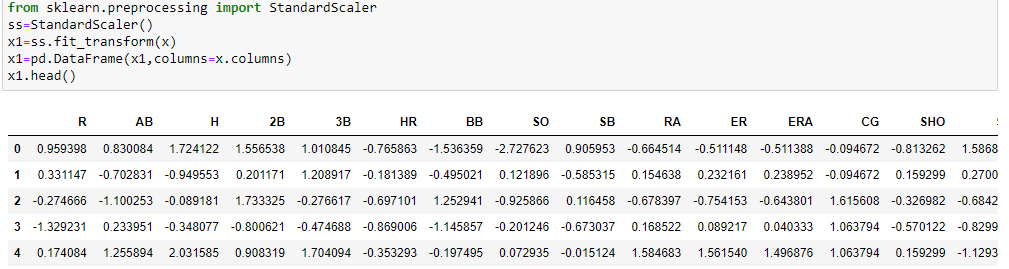


Observation:

We have reduced most of the skewness from input dataset and currently only 2 features have presence of some skewness.

**Standardization: -** Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance). The purpose of scaling for different feature dimensions is to make the features between different metrics comparable. We are using Standard Scaler for feature scaling purpose so that features will be on same scale.

**Standard Scaler**: - Standard Scaler is a class from Sklearn.preprocessing which is used for standardization. It will Standardize features by removing the mean and scaling to unit variance.

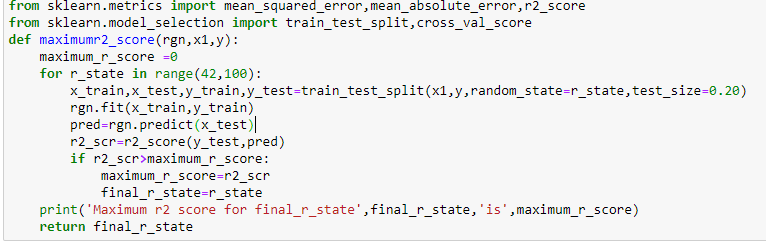


Now our Input and output features are ready to apply machine learning algorithms on them.

**Building Machine Learning Models**

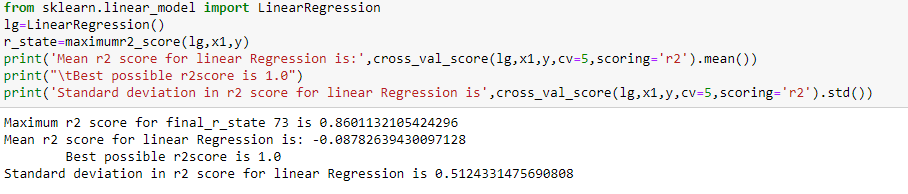
Our Target variable Win is continuous in nature so we will use different regression algorithms to try and find the features that have the best explanation of the target variable. We will use R2 as our metric. Our first step is dividing training and testing data and also use train test split.

Function to calculate maximum R2 score at best random state.



The different regression algorithms used in this project are-

1- **Linear Regression**: - In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. Linear Regression fits a linear model with coefficients w = (w1, …) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.



# 2- k-nearest neighbors: - K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. We will use Grid Search CV method to find n\_neighbors.

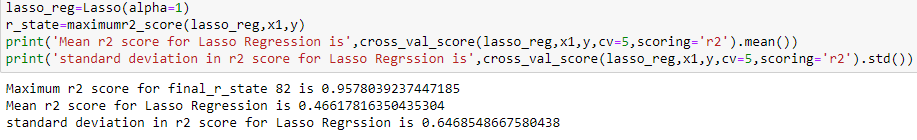
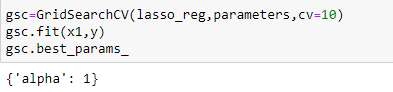
# 

# Grid Search CV: - It helps to loop through predefined hyperparameters and fit our estimator (model) on our training set so in the end we can select best parameters.

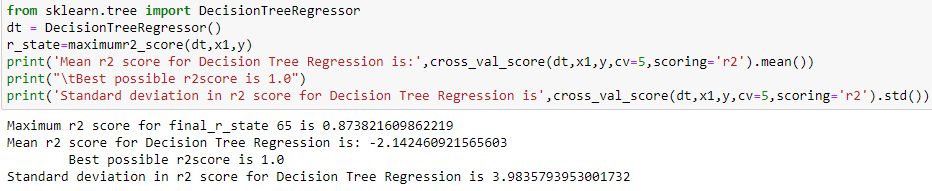
# 

# 

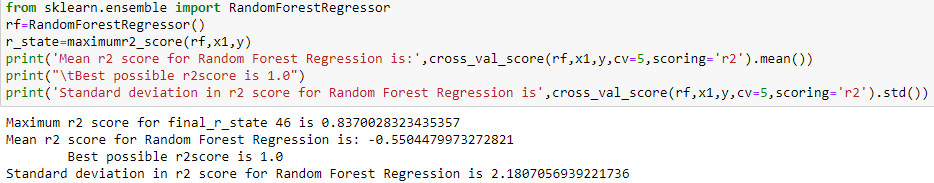
3**- Lasso Regression**: - Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. Again, we will use Grid Search CV method to find best parameters for Lasso regression.

**Grid Search CV for Lasso Regression**: -

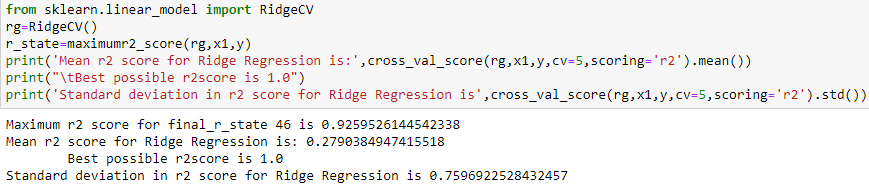
4-**Decision Tree Regression: -** Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output which means that the output is not discrete.



5- **Random Forest Regression:** - Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.



6- **Ridge Regression: -** Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. This method performs L2 regularization. It reduces the model complexity by coefficient shrinkage.



Observation:

1- We have used 6 algorithms till now, and the best score we have obtained from Lasso regression, Ridge regression & decision Tree Regression methods.

2- Now we will do Hyperparameter tuning using Grid Search CV to find best estimator and parameters from these 3 models.

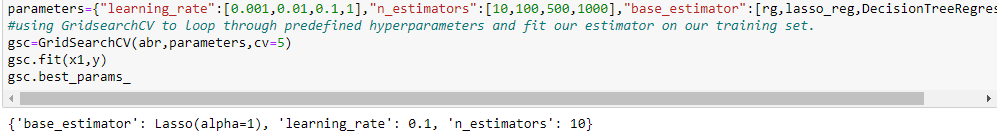
3- We will use best estimator and parameters in AdaBoost regression and check its performance.

7- **AdaBoost Regression:** - An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.



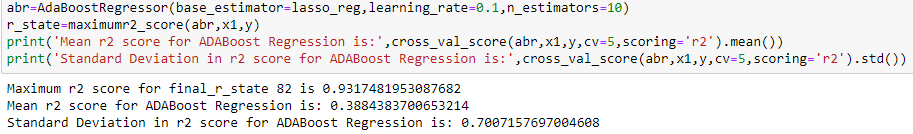
**Hyperparameter Tuning:** - Hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. We are using Grid search approach to do this task.

**Grid Search CV: -** It helps to loop through predefined hyperparameters and fit our estimator on our training set. So, in end, we can select the best parameters from the listed hyperparameters.



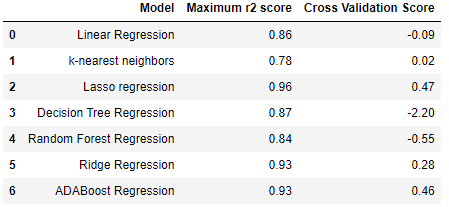
Observation:

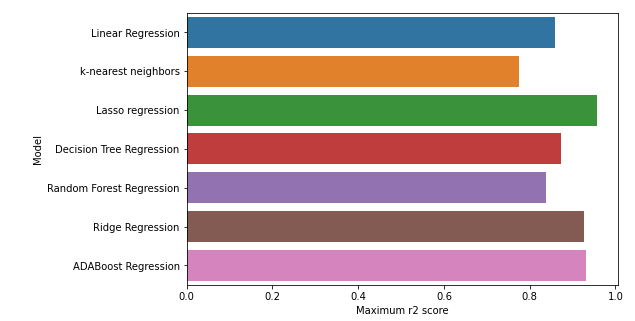
After hyperparameter tuning we have found Lasso regression is the best estimator when alpha is 1 and learning rate is 0.1. So, we will use these parameters in implementing AdaBoost Regression.

After this, we did cross validation for assessing the effectiveness of these models.

**Cross Validation: -** It is amodel validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. When we want to estimate how accurately a predictive model will perform in practice and our goal is prediction then we use cross validation.

After applying different regression algorithms on model and using cross validation we have obtained following coefficient of determination and cross validation scores: -



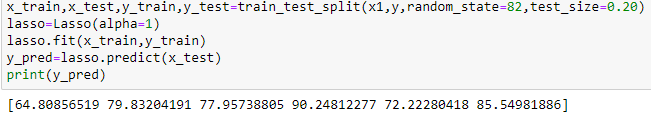
**Bar plot of Maximum r2 score of various models**: -

Observations on basis of Scores & Bar Plot: -

1- After comparing above 7 models, on basis of r2 score, cross validation score and other parameters the Lasso regression gives best results.

2- We already did Hyperparameter tuning to find best parameters using Grid Search CV. Lasso regression model gives best scores when random state is 82 & alpha is 1. So, we will implement Lasso Regression for final model building.

**Building Final model with best parameters: -**

****

**Error metrics for regression model: -**

**1-** **Mean Absolute Error (MAE)-** MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. MAE score is calculated as the average of the absolute error values. MAE can be calculated as follows:



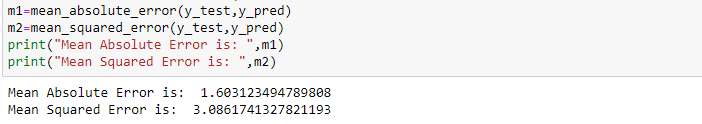
**2- Mean Squared Error (MSE): -** MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset. MSE can be calculated as follows:

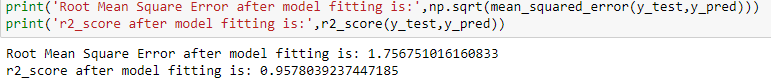


3- **Root Mean Squared Error (RMSE):** - RMSE is an extension of the mean squared error. It’s the square root of the average of squared differences between prediction and actual observation. RMSE can be calculated as follows:

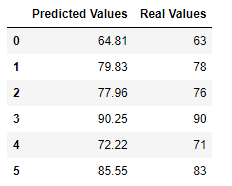


**Let’s check error metrics for our final regression model: -**

****

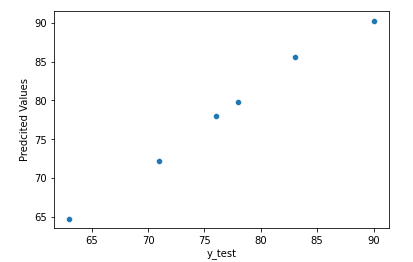
****

**Evaluating Prediction with real values: -**

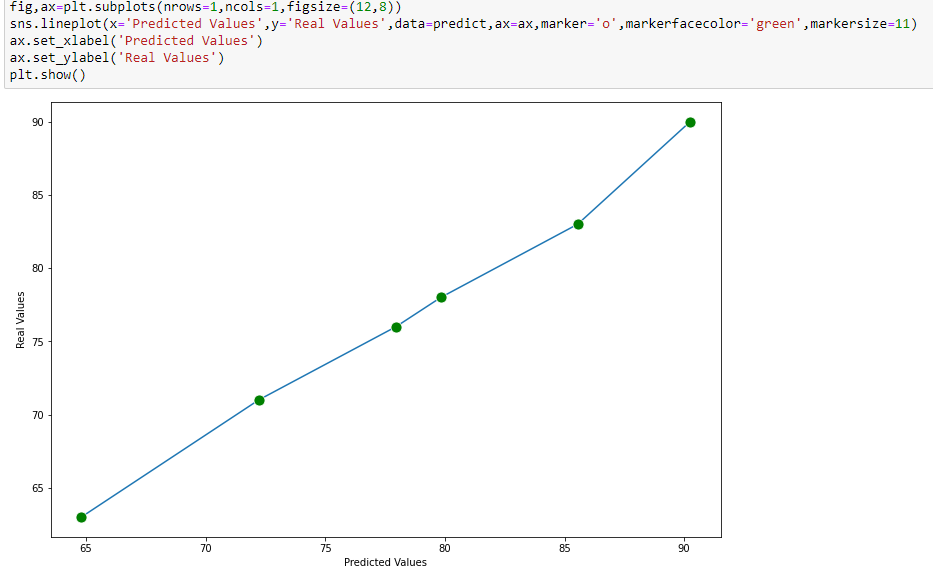
****

**Let's plot Graphs to check relation between predicted and Real Values: -**

**1- Scatter Plot: -**

****

**2- Line Plot: -**

****

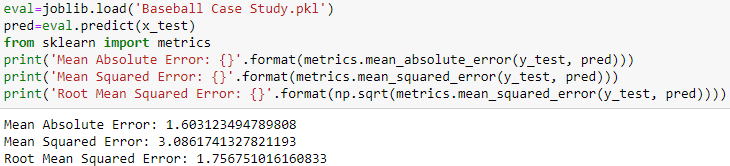
Observation: -

1- Both Plots are linear and this shows best relation between predicted and real values.

2- Predicted and Real values are very much close to each other.

We have saved our final model using JOBLIB and now we are testing the saved model.

1- Loading the saved model and check error metrics: -



2- Checking Predicted values: -



.     **Concluding Remarks**

After building, saving, testing, re-loading our final model, Error metrics & on basis of relation between real and predicted values, we have reached on following conclusions: -

1- We spent a great deal of time organizing your data and carefully choose our features.

2- Our Target variable Win is continuous in nature so we have used different regression algorithms on dataset and calculated different scores for them.

3- We have used Grid Search CV to find best parameters for Lasso Regression, Decision Tree Regression and AdaBoost Regression.

4- We have used best parameters obtained by hyperparameter tuning using Grid Search CV method in building of our final model.

5- We have checked Maximum R2 score, Mean R2 score & Standard deviation for all models.

6- We did cross validation for all models and also calculated cross validation scores.

7- Predicted and Real values are very much close to each other and this shows our model is correct. Also plots between them are linear.

8- We have also checked MAE, MSE, RMSE for our final model and error is minimum while we are using Lasso regression model.

9- While using Lasso regression, R2 score, Cross validation & other scores are maximum.

10- We have also done data analysis to check impact of runs, earned runs & errors on Win and also calculated points where Win is maximum.

11- We have handled outliers in dataset and removed skewness very carefully and data loss was only 3% after outlier removal.

12- During pre-processing we have used Standard Scaler to bring features at common scale.

13- We have tested 7regression algorithms and checked their performance very carefully.

14- Lasso Regression is best choice for our final model.

15- We have tried to achieve our goal at best level as result is good as per our model.

**Thanking You**