Prediction of a Football Team Rating Using Machine learning Regression Algorithms

Ashok Ramavath [1,a)] Vandana Bhattacharjee [2,b)] Sanjay Kumar [3,c)]

[1,2,3)]*Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India(IND)*

[a)] btech10476.21@bitmesra.ac.in [b)] vbhattacharya@bitmesra.ac.in [c)] sanjaykumarcse@bitmesra.ac.in

Abstract:- Football is a famous sport played around the world that involves various actions from defending to attacking the ball to score goals. Various rules like half-side, No hands, Throw-ins Direct and Indirect kicks, etc. are followed. In a game, players might get yellow-cards and red-cards which serves as warning and elimination from that particular game. There are-many factors like shots, possession, pass accuracy, yellow and red cards, etc. that decide which team wins in a match. In the data that we worked on,the Rating is calculated from the parameters which are mentioned. Machine learning regression algorithms have been applied to predict the Rating feature. The goal of this paper is to know which regression algorithm performs better to predict the Rating variable. Evaluation metrics like Root mean square error (RMSE), Mean absolute error (MAE), and R square (R2) are used to identify the best algorithm among the applied algorithms.

The rest of the paper is divided into the following sections: 1) Introduction 2) Brief overview of the regression algorithms 3) Experimental Setup 4) Results and analysis 5) Conclusion 6) References.

1. INTRODUCTION

Football is a common sport played around the world. It is considered as very popular game in the family of sports[1]. Predicting the results of a football match is an interesting challenge [2]. The major football leagues which are played around the world are Laliga, Bundesliga, English Premier league and others etc. [3]. An International governing body FIFA is an organization that controls International football and organizes the World Cup. Machine learning approaches have been used to solve several real life problems like diabetes prediction etc..(4-5). Various researchershave worked on the prediction of football matches result using deep learning Neural Network techniques like Artificial Neural Networks (ANN) and Constitutional Neural Network (CNN). Machine learning algorithms like Random Forest, Support Vector Regression (SVM) etc.,[6] are used extensively to predict the rating of a team which wins in a match [7]. Machine learning algorithms can handle large amount of data and are useful in building prediction models. Since the number of domestic and International football matches played in a day are increasing, there is a need of study on the football sport so that a better prediction can be made before the start of the match and gives a deep insights to the team [8]. Analyzing the important parameters like shots per goal, players pass accuracy, team possession and others, which helps the team to identify the weak areas of a team [9]. One such way to predict the team performance is to calculate the rating of a team is based on above mentioned regression algorithms.

2. BRIEF OVERVIEW OF THE REGRESSION ALGORITHMS

2.1 LINEAR REGRESSION

Linear regression is a statistical method used to predict the output variable dependent on one input variable. The dependent and Independent variables are plotted on the X and Y axis of the graph. A relationship is obtained between two variables using the below-mentioned formulae.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

ε = Residual error Y_i = Dependent variable β_0 = Constant / Intercept

$\beta 1$ = Slope/Coefficient X_i = Independent variable

2.2 MULTIPLE LINEAR REGRESSION

Multiple linear regression is similar to linear regression. It is a statistical method used to predict the output variable. Instead of using one Independent variable, we can use one or more Independent variables for a better relationship while plotting a slope line.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots . \beta_p X_{ip} + \varepsilon$$

ε = Residual error Y_i = Dependent variable β_0 = Constant / Intercept

β_p = Slope/Coefficient for each Independent variable X_i = Independent variables

2.3 SUPPORT VECTOR REGRESSION

Support Vector Regression is a supervised learning method used to find the best fit line in N-dimensional space(N - number of features) .In Support Vector Regression , the best fit line is called hyperplane. The data points are classified on either side of the hyperplane. Maximum data points are present in the hyperplane. The closest ones to the hyperplane are called Support Vectors. Hyper-parameters like hyperplane, kernel, and Boundary lines are used in Support Vector Regression. A kernel is an input where data is modified into a required form using mathematical functions to find a hyperplane.

2.4 DECISION TREE REGRESSION

Decision tree regression is used to predict the data where Input data is trained and transformed into a tree like structure to obtain the output by comparing the nodes of a tree. Python libraries are used to form a best tree structure such that the output data can be obtained by comparing the input variables from the head node of the tree. The end nodes of the tree are result nodes of Input values. Decision tree is mostly used for classification because continuous data values can be classified easily [10].

Decision tree regression involves the following steps:
1. Importing the required libraries and dataset.
2. Separating the features and the target variable.
3. The data set will be divided into a train set and a test set.
4. Fitting the training dataset to the model.

5. Figuring out the loss after training.
6. Creating a decision tree image.

2.5 RANDOM FOREST REGRESSION

A supervised learning system called Random Forest is built on several Decision Trees and the ensemble learning approach. Because Random Forest uses a bagging method, all computations are performed concurrently and there is no interaction between the Decision Trees as they are constructed. Both Classification and Regression tasks can be solved with RF. The term "Random Forest" refers to the Bagging concept of randomizing data and creating several Decision Trees (Forest). Overall, it is an effective machine learning technique that reduces the drawbacks of a Decision Tree model. Therefore, you wish to include K Decision Trees in our ensemble together with your original dataset D. You also have a number N; you will build a tree until each node has less than or equal to N samples (for the Regression, task N is usually equal to 5). Additionally, each node of the decision tree will include a random feature chosen from a pool of F characteristics. From these F features, the feature that will be used to split the node is chosen (for the Regression job, F is often equal to square root (number of features of the original dataset D) The rest is quite straightforward. K subsets of the data are created by Random Forest from the original dataset[11].

3. EXPERIMENTAL SETUP

EVALUATION METRICS

3.1 MEAN ABSOLUTE ERROR (MAE) - The degree of discrepancy between an observation's predicted value and its actual value

$$\mathbf{MAE} = \Sigma_1^n |Y_i - X_i| / N$$

N = total number of samples Y_i = Predicted values X_i =Tested values

3.2 ROOT MEAN SQUARE ERROR (RMSE) - The square root of the residuals' variance is the RMSE. It shows how well the observed data points match the values predicted by the model, which is known as the model's absolute fit to the data.

$$\mathbf{RMSE} = \sqrt{\Sigma_1^n |Y_i - X_i| / N}$$

N = total number of samples Y_i = Predicted values X_i =Tested values

3.3 COEFFICIENT OF DETERMINATION - R-squared (R2) is a statistical metric that depictsthe percentage of a dependent variable's variation that is explained by one or more independent variables in a regression analysis.

$$R^2 = 1 - TSS/RSS$$

R^2 = COEFFICIENT OF DETERMINATION TSS = SUM OF SQUARE OF RESIDUALS

RSS = TOTAL SUM OF SQUARES

DATASET DESCRIPITION:

TABLE 1. DATASET DESCRIPITION

NUMBER OF SAMPLES	NUMBER OF FEATURES USED	OUTPUT VARIABLE
121	07	RATING OF TEAM

FEATURES -

- **3.3.1 GOALS**
- **3.3.2 SHOTS PER GOAL**
- **3.3.3 YELLOW CARDS**
- **3.3.4 RED CARDS**
- **3.3.5 POSSESSIONO PERCENTAGE**
- **3.3.6 AERIALS WON**

4. RESULTS AND ANALYSIS

Football teams dataset has 121 rows and 7 columns. The evaluation metrics are applied on the tested and predicted data. First a correlation matrix is computed, as shown in Figure 1.

TABLE 2: COMPARATIVE PERFORMANCE OF REGRESSION MODELS

REGRESSION MODEL	COEFFICIENT OF DETERMINATION	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARE ERROR
LINEAR REGRESSION	**0.7432**	**0.0431**	**0.0451**
MULTIPLE REGRESSION	**0.8903**	**0.0332**	**0.0409**
SVM REGRESSION	**0.8196**	**0.0477**	**0.0039**
RANDOM FOREST REGRESSION	**0.8064**	**0.0557**	**0.0654**
DECISION TREE REGRESSION	**0.4646**	**0.0776**	**0.1010**

4.1.1.1.1. LINEAR REGRESSION -

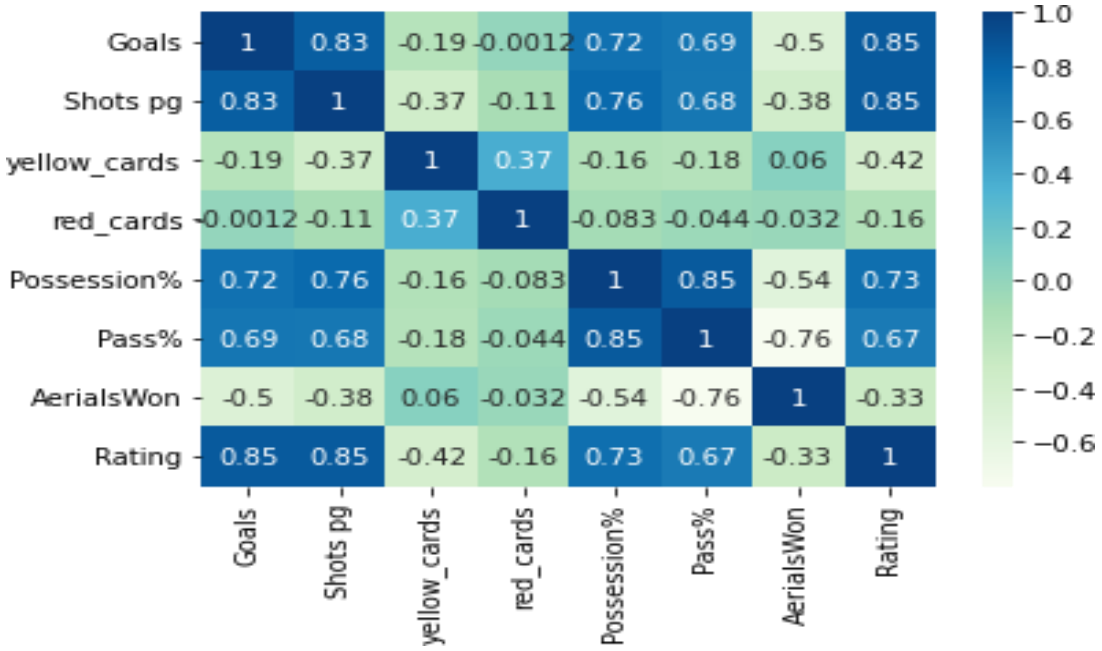
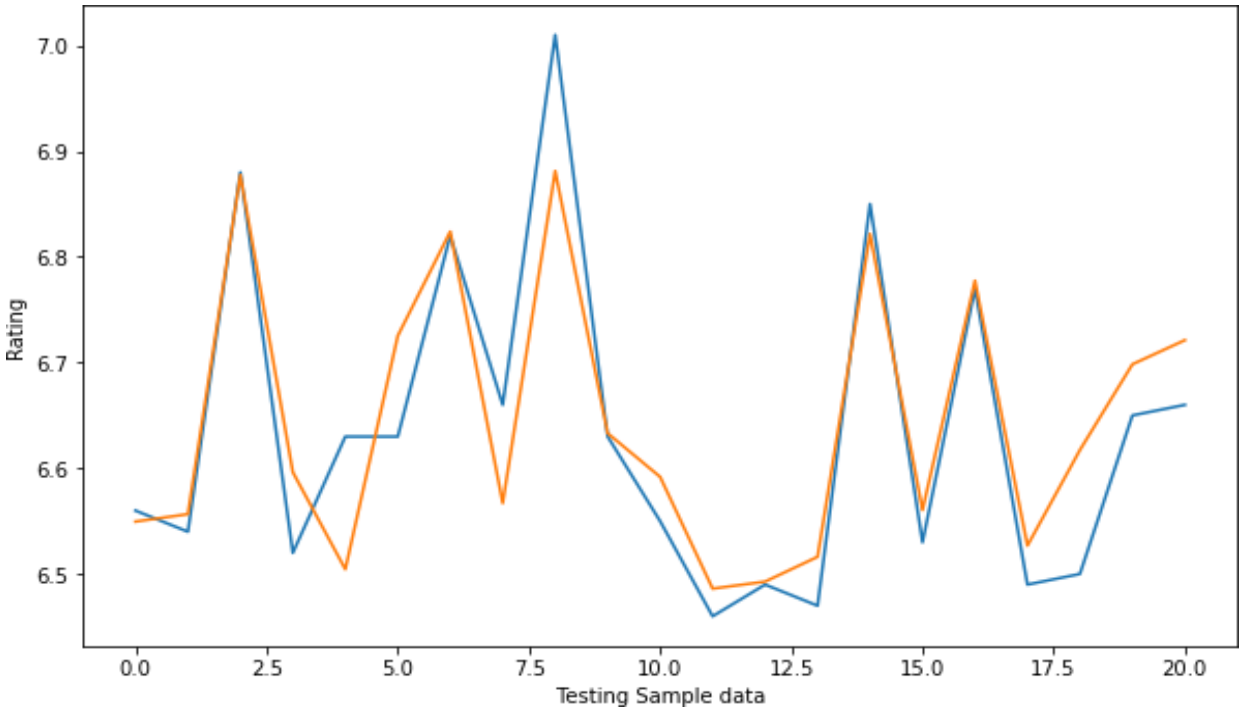


FIGURE 1 : THE CORRELATION MATRIX

We can take two Input variable from any of the two Goals and Shots per goal because both of them are equally dependent on the output variable. But Goals is taken as Input variable because it gives better results.

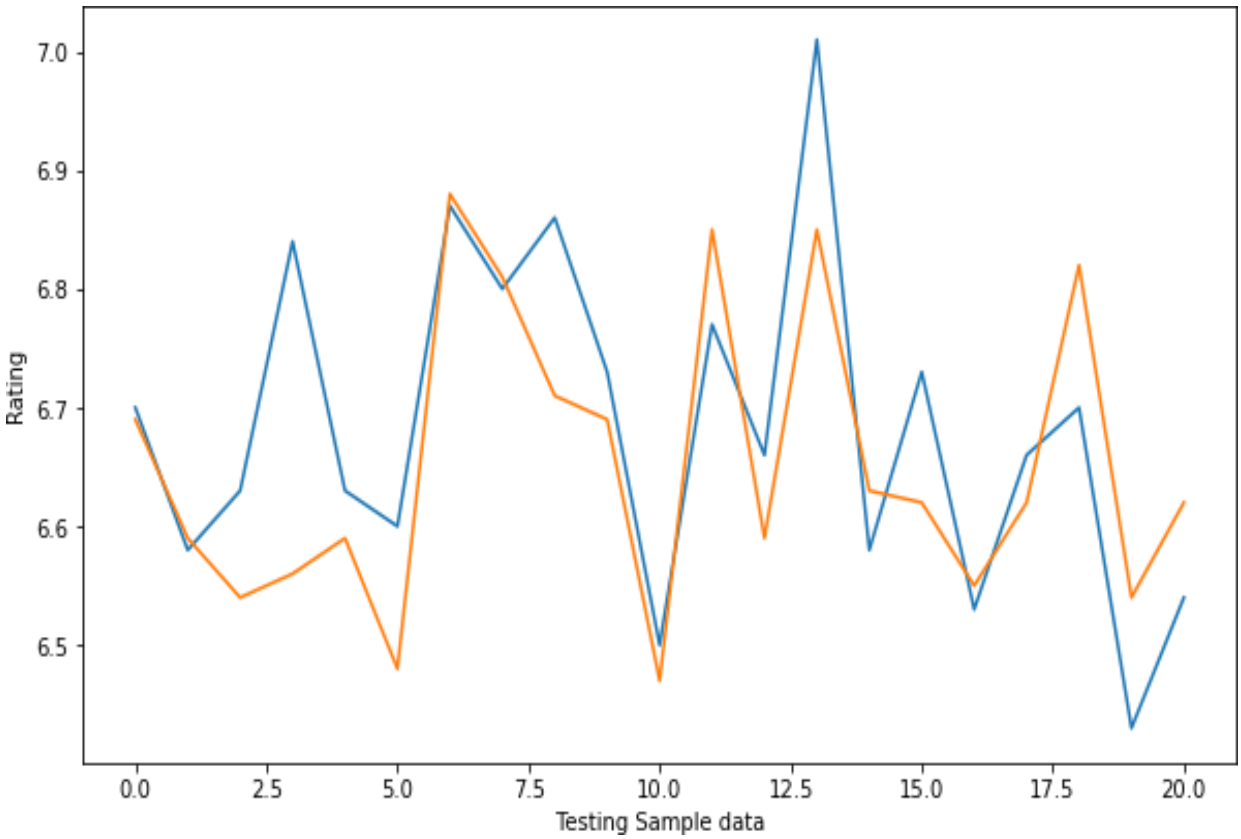
4.1.1.1.2. SUPPORT VECTOR REGRESSION



BLUE LINE - TESTED VALUES YELLOW LINE - PREDICTED VALUES

FIGURE 2 : SUPPORT VECTOR REGRESSION MODEL PERFORMACNE ON TESTED AND PREDICTED VALUES

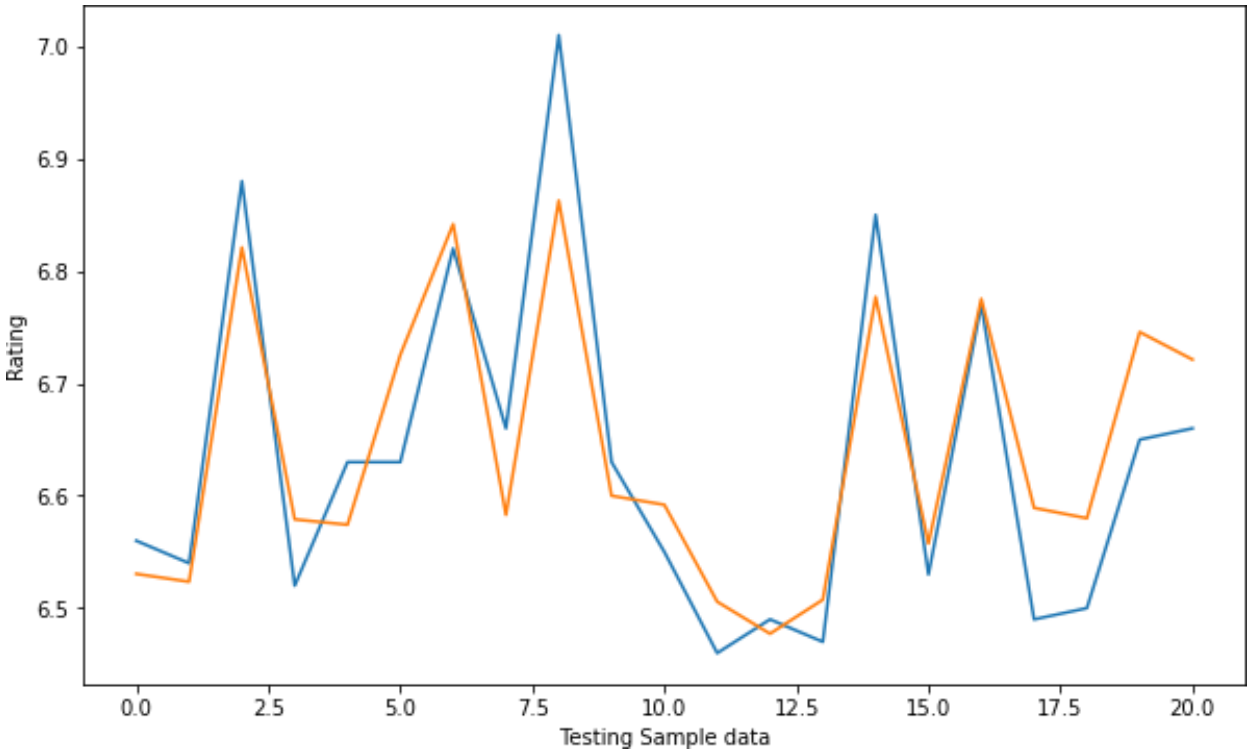
4.1.1.1.3. DECISION TREE REGRESSION



BLUE LINE - TESTED VALUES YELLOW LINE - PREDICTED VALUES

FIGURE 3: DECISION TREE REGRESSION MODEL PERFORMANCE

4.1.1.1.4. RANDOM FOREST REGRESSION



BLUE LINE - TESTED VALUES YELLOW LINE - PREDICTED VALUES

FIGURE 4: RANDOM FOREST REGRESSION MODEL PERFORMANCE

5. CONCLUSION

From the above tables we can see that Multiple Linear Regression has performed the best. It has the high R2 value of (0.8903) and less MAE (0.0332) and RMSE of (0.0039) followed by SVM Regression having R2 score of (0.8196) and MAE and RMSE values of (0.0477) and (0.0039). There is a very slight difference of R2 score between SVM and Random forest Regression. Since the values of MAE and RMSE are least for SVM , so SVM can be ranked as second best fit regression for the prediction. The reason for Multiple linear regression outperforming other important algorithms like SVM and Random forest regression is because the values are continuous in nature and there is a minimal difference between any two values of in any Independent variable column. So the multiple linear regression finds the pattern better by calculating the better relation between target variable and all the other Independent variables.

6. REFERENCES

1. Ishan Jawade, Rushikesh Jadhav, Mark Joseph Vaz, and Vaishnavi Yamgekar, "Predicting Football Match Results using Machine Learning", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072 © 2021, IRJET | Impact Factor value: 7.529 | ISO 9001:2008 Certified Journal | Page 177.
2. Josip Hucalijuk and Alen Rakipovic, "Predicting football score techniques using Machine learning techniques" , Proceedings of the 34th International Convention MIPRO, 23-27 May 2011, IEEE Xplore: 28 July 2011, INSPEC Accession Number: 12137640.
3. Karan Bhowmick and Vivek Sarvaiya, "A Comparative Study Of The Different Classification Algorithms On Football Analytics", August 2021International Journal of Advanced Research 9(08):392-407.
4. Sanjay Kumar, Ekta Kumari Gupta and Vandana Bhattacharjee," detailed analysis of Classifiers for prediction of diabetes". International Journal of Engineering Research and Technology (IJERT) ISSN:2278-0181,Vol.11 Issue 09, September-2022.
5. Mitushi Soni and Sunita Verma,"Diabetes Prediction Using Machine Learning Techniques", International Journal of Engineering Research and Technology (IJERT) ISSN:2278-0181,Vol.9 Issue 09, September-2020.
6. Pinar Tufekci ," Prediction of Football Match Results in Turkish Super League Games". Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 pp 515–526.
7. Anand Ganesan and Harini Murugan, "English Football Prediction Using Machine Learning Classifiers",June 2020International Journal of Pure and Applied Mathematics 118(22):533-536.
8. Igiri, Chinwe Peace and Nwachukwu, Enoch Okechukwu, "An Improved Prediction System for Football a Match Result", IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 04, Issue 12 (December 2014), ||V4|| PP 12-20 International organization of Scientific Research 12.
9. Jongho Shin and Robert Gasparyan, "A novel way to Soccer Match Prediction", cs229.stanford.edu/proj2014.
10. Mei-Ling Huang and Yi-Jung Lin," Regression Tree Model for Predicting Game Scores for the Golden State Warriors in the National Basketball Association", May 2020, Symmetry 12(5):835,DOI:10.3390/sym12050835.
11. https://en.wikipedia.org/wiki/Random_forest.