



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Regression Model to Estimate Goals Scored By Football Players

Neel Niraj Patel¹, Mohanraam S²

Student, Department of Computer Engineering, SRM University, Chennai, India¹

Student, Department of Computer Engineering, SRM University, Chennai, India²

ABSTRACT: Predictive systems have been employed to predict events and results in virtually all walks of life. Football goal scorer prediction in particular has gained popularity in recent years. Statistical approaches have shown complex and low prediction results. Our main objective of this paper is to develop a goal predictor using data mining techniques without any human intelligence and intuition. We constructed a more comprehensive system with an improved prediction accuracy by using the features that directly can predict the numbers of goals scored by a particular player by using the various attributes needed to determine that. Our prediction system for predicting the goal scorer was implemented using linear regression with Weka as a data mining tool. The technique yielded 85% - 93% prediction accuracy for linear regression technique. With this output, it is observed that the prediction accuracy is higher than those of existing systems.

KEYWORDS: Linear Regression, Weka, [1] Correlation Evaluation, weighted average algorithm.

I. INTRODUCTION

Football is a popular game worldwide and a rich source of data. It is a dynamic and unpredictable game that has constant and predictable patterns. Identifying potential goal scorers has been a demanding task for data analysts. The top tier Football clubs nowadays employ experts in order to identify talents and acquire that player before anyone else does. In this paper, a novel approach is taken to help predict the most likely chances of scoring goals in the upcoming games. This can be done in an efficient way using Data Mining. [2]

Mining is the process of finding new, potentially useful and non-trivial knowledge from data. Using this technique new relationship, dependencies, and the effect of an attribute for a goal scoring chance can be determined. Data mining is a process of analyzing data from different perspectives and summarizing it into useful information. Technically, Data mining is the process of finding correlations or patterns among dozens of field in large relational databases.

For this purpose, data of a set of few sample popular footballers with known results are collected and fed to the Weka Tool for mining. The dataset contains various attributes which affect the overall chances of scoring a goal like "shots per game", "shots outside the box", "shots from six yard box", "shots from penalty area", "accuracy of shots". The result of the mining algorithm is used to formulate a custom algorithm and determine the probable outcome given just the above mentioned attributes.

The goal predictor can help to provide the analytics platforms that take the vast range of unstructured and misleading data available around sports, identify the underlying trends, to deliver highly accurate statistics and predictive analytics. Fantasy football managers can use this to track the expected goals to be scored by a player and a team against oppositions in real life.

II. DATA ACQUISITION

The data for the data set is acquired from a trusted source www.whoscored.com. [3] The data set has 50 records, having the details of 350 footballers currently playing in the English Premier League. The data for the shots taken outside the box, shots taken from six yard box and shots taken inside penalty box(excluding shots taken inside six yard

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

box) are chosen particularly to help identify the chances needed for a particular player and a team to score the goals for the club. The accuracy is taken by dividing the shots on target (shots angling into the goal post) by the total shots taken which depicts the accuracy of the player's shots. The Chances created by the team attribute is collected in total, ie, the total chances created at a whole by the respective clubs which will help in the process of predicting the goal scorers. [5]

III. PROPOSED ALGORITHM

Regression is a data mining technique used to fit an equation to a dataset. [4] The simplest form of regression, Linear Regression, uses the formula of a straight line ($y=mx+b$) and determines the appropriate values of m and b to predict the values of y based upon a given value of x . Basically a Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted is called the Dependent variables. The factors that are used to predict the values of a dependent variable are called Independent variables.

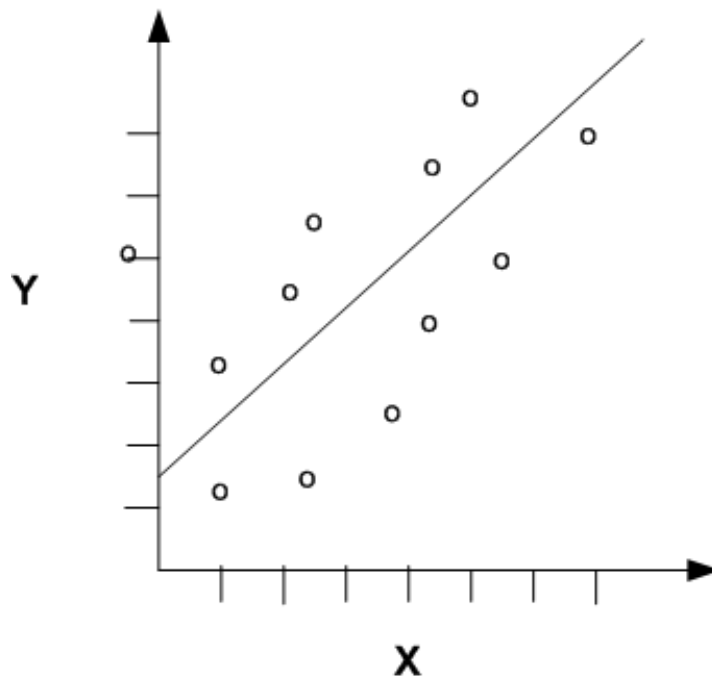


Fig. 1. Logistic relationship between X and Y

In a linear regression scenario with a single predictor ($y = \theta_2x + \theta_1$), the regression parameters (also called coefficients) are:

The slope of the line (θ_2) — the angle between a data point and the regression line

The y -intercept (θ_1) — the point where x crosses the y axis ($x = 0$)

Numeric classes are mostly used today. To calculate weights from training data,

$x = w_0 + w_1a_1 + w_2a_2 + \dots + w_k a_k$ where A is attribute and W is weight given to that attribute.

- Standard matrix problem – Works if there are more instances than attributes
- Nominal attributes – two-valued: just convert to 0 and 1 multi-valued.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Correlation is a bivariate analysis that measures the strengths of association between two variables and the direction of the relationship.[4] In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1 , then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. the direction of the relationship is simply the + (indicating a positive relationship between the variables) or - (indicating a negative relationship between the variables) sign of the correlation. Usually, in statistics, we measure four types of correlations: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.

In the weka tool, we perform correlation evaluation under the select attribute option. CorrelationAttributeEval evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Fig. 2. Proposed Algorithm Formula

IV. DATA ATTRIBUTES

The data set used has 5 attributes, all of which are numeric. [3]

- Total Shots taken
- Shots taken outside the box
- Shots taken from Six Yard Box
- Shots taken inside penalty area
- Accuracy of the Shots

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

V. HYBRID ARCHITECTURE

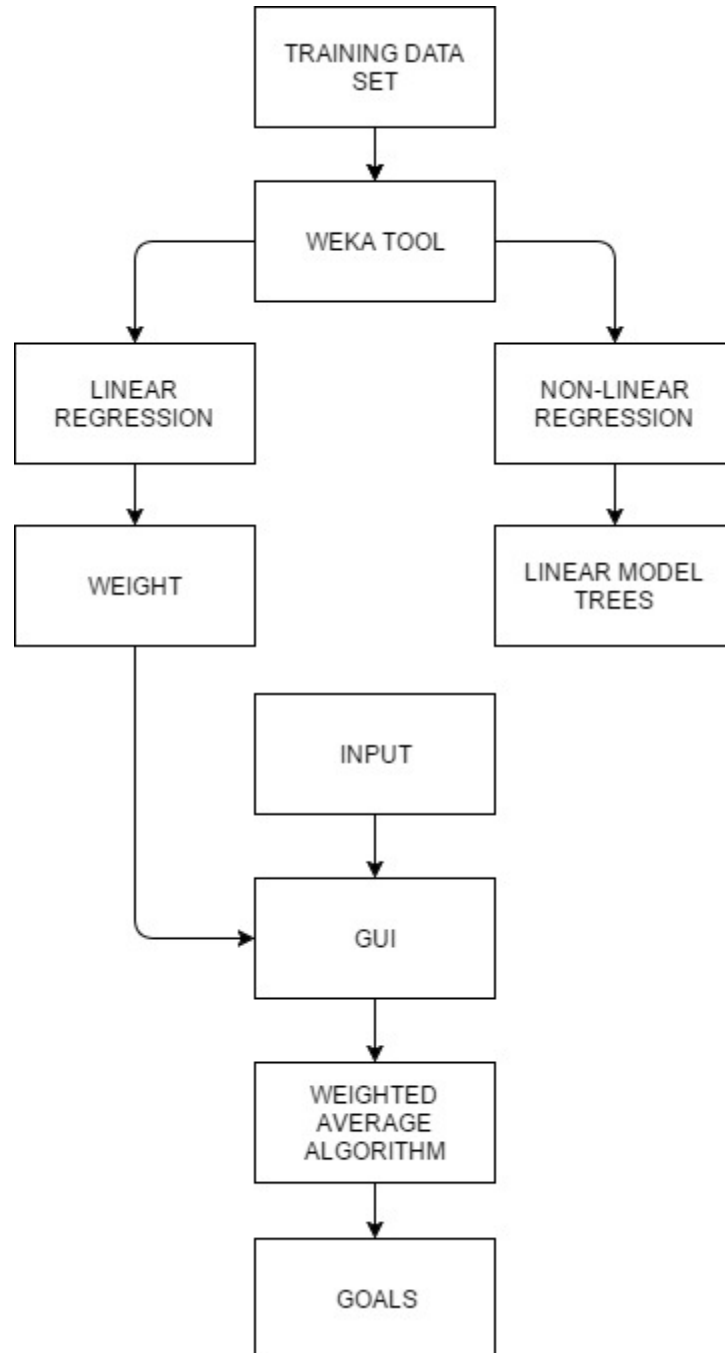


Fig. 3. System Architecture

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 9, September 2017

VI. RESULTS AND INFERENCE

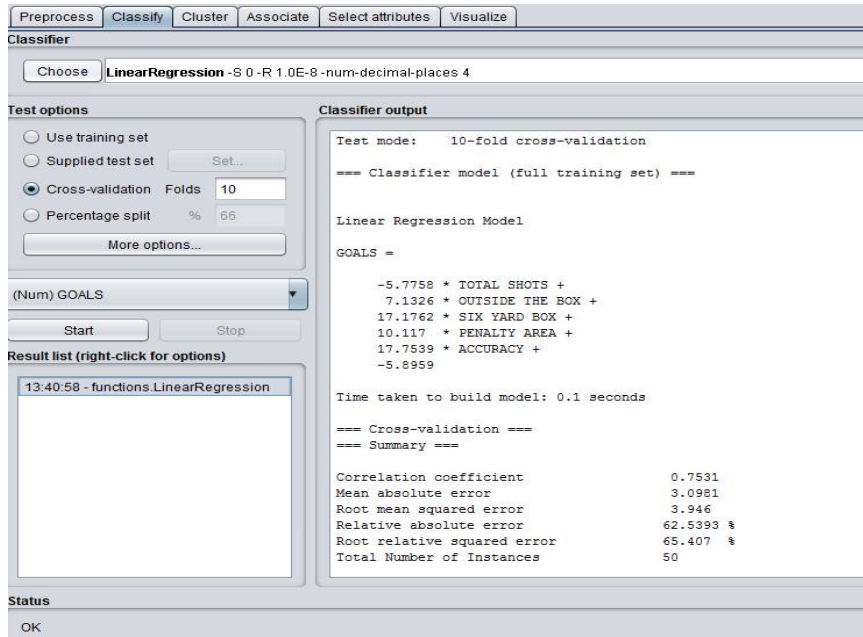


Fig. 4. Linear regression coefficients and correlation

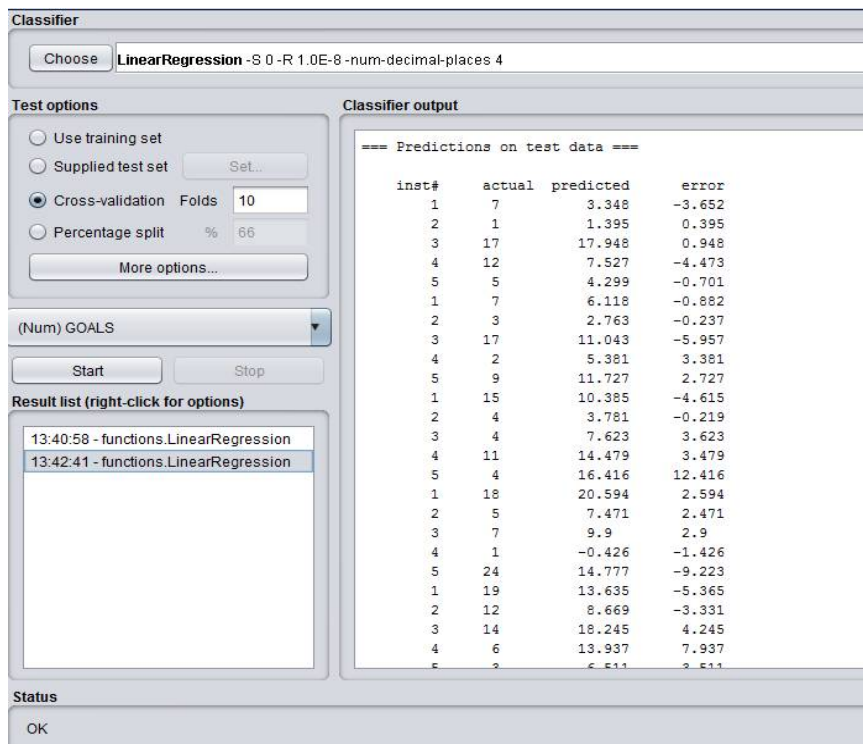


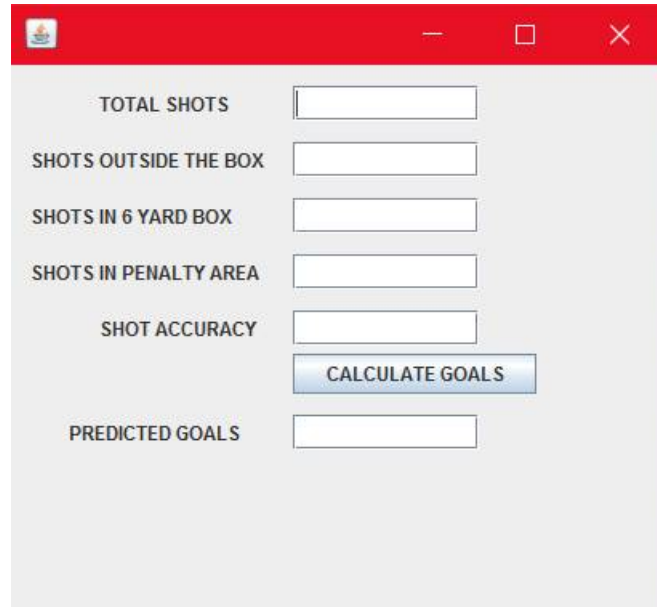
Fig. 5. Actual and Predicted Values

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

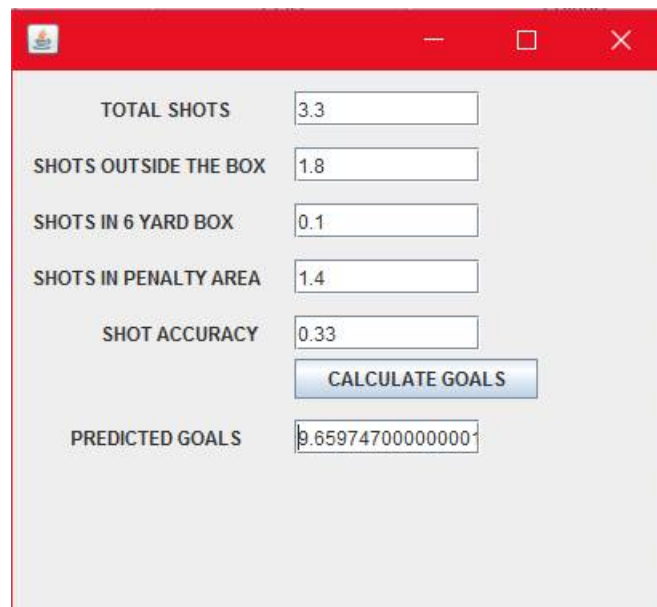
Website: www.ijircce.com

Vol. 5, Issue 9, September 2017



TOTAL SHOTS	<input type="text"/>
SHOTS OUTSIDE THE BOX	<input type="text"/>
SHOTS IN 6 YARD BOX	<input type="text"/>
SHOTS IN PENALTY AREA	<input type="text"/>
SHOT ACCURACY	<input type="text"/>
	<input type="button" value="CALCULATE GOALS"/>
PREDICTED GOALS	<input type="text"/>

Fig. 6. Design



TOTAL SHOTS	3.3
SHOTS OUTSIDE THE BOX	1.8
SHOTS IN 6 YARD BOX	0.1
SHOTS IN PENALTY AREA	1.4
SHOT ACCURACY	0.33
	<input type="button" value="CALCULATE GOALS"/>
PREDICTED GOALS	9.659747000000000

Fig. 2. Output

VII. CONCLUSION

To measure the accuracy of our model, we tested it with our own training data set. A total 299 players out of 350 were correctly classified, hence making our model 85.4% accurate.

The Linear Regression technique predicts a numerical value. Regression performs operations on a dataset where the target values have already been defined. And the result can be extended by adding new information. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

datasets where the target values are not known. The test data's are take to prove the relationship between the predictor and the target variable which is being represented by linear regression equation.

$$y = a +bx$$

In this paper we have successfully devised an algorithm to predict the goals scoring opportunities for a footballer given its various attributes, using data mining techniques. This paper is flexible enough to accommodate an increase in the size of the data set for more accurate results. The algorithm used in this paper can also be applied in various other applications like to estimate the number of runs scored by a cricketer, bowlers potential to take wickets, number of aces recorded by a tennis player, etc.

REFERENCES

- [1] <http://www.cs.waikato.ac.nz/~ml/weka/>
- [2] <http://www.comp.dit.ie/btierney/DataMining/>
- [3] www.whoscored.com
- [4] http://www.camo.com/rt/Resources/simple_linear_regression.html
- [5] <https://en.wikipedia.org/>