LINKÖPINGS UNIVERSITET

Predict the football player performance in the next game

732A76 Research Project

Hussnain Khalid

30 December 2023

Contents

Abstract

Football is no doubt one of the most popular sports in the world. Predicting the result of a game has become a fundamental problem in the field of machine learning. Many researchers have tried to predict the result of a game, or to predict the performance of players, or to predict the players who should be selected according to their performance, etc. using machine learning. In this paper I discuss predicting player performance techniques along with comparison among these techniques.

1 Introduction

Predicting player performance for the next game is the game changer, and this technique is used now in almost all the games to help teams build a better team. In this research the goal is to make model which can accurately predict the player performance for the next game. For this purpose, we are using game-by-game data set of performance of every player that has played a game in a simulated football league through Football Manager 2022. Three different performance metrices are used for predicting the player performance i.e., Average Rating, Goals and xG. Reason for choosing these three metrices is discussed in preprocessing part of Method section.

2 Theory

In this section, theoretical background of the techniques is discussed. Detailed concept of these implemented techniques is discussed in Method section. As it is a supervised regression problem regressor models are used and trained to get the most optimal results.

2.1 Linear regression

Linear regression is a basic and most common regression model used to describe the relationship between dependent and a set of independent variables. This model is used to predict the dependent variable using independent variables. And both variables are considered linearly related with each other. The equation of simple linear regression is,

$$y = b_0 + b_1 x$$

where bo is intercept, b1 is slop, x is independent variable and y is dependent variable.

2.2 Neural Network

Neural network is an artificial intelligence method, used for supervised and reinforcement learning. It is mostly used to solve the non-linear problems. A neural network contains various layers. Every layer contains different nodes. These nodes multiply the inputs to weights, add the values and pass it to activation function and then output is passed to next layer nodes. The same is done there too unless values reach to final output layer. The main task of neural network is to minimize the loss which is done as backward propagation.

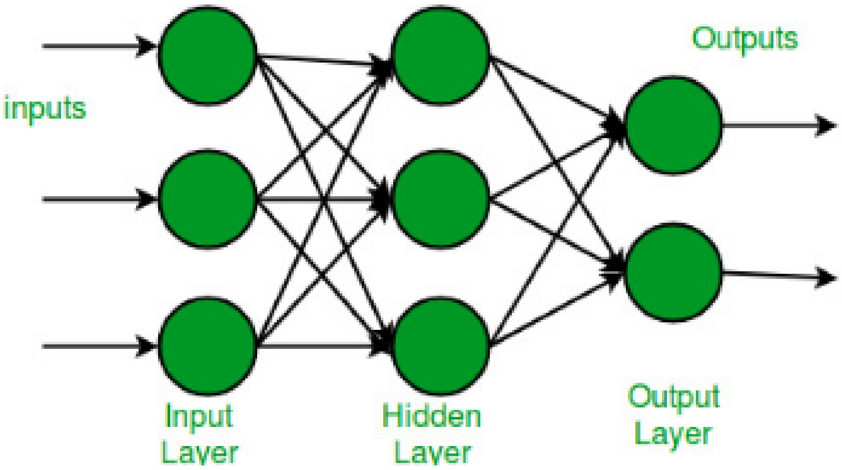


Figure 1: Neural Network topology

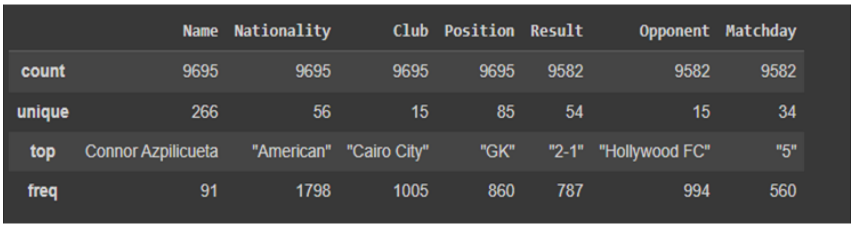
2.3 XGBoost

XGBoost stands for Extreme Gradient Boosting, is a distributed gradient-boosted decision tree. It is a supervised learning technique that uses an ensemble approach based on gradient boosting algorithm. It is used for regression, classification, and ranking purposes.

Weights are assigned to all independent variables and then fed to trees to predict the result. Weights of the variables are increased for the wrong prediction and then fed to other trees. And at the end these trees are ensemble to generate strong and precise model.

3 Data

The data used in this research paper is taken from a game simulation called Football Manager 2022. The raw data is first structured with the help of provided legends information. There are 9695 tuples and 93 columns. And then some exploration is performed to study the raw data like checking unique counts and the frequency of most occurred tuple in whole data shown below.



	Name	Nationality	Club	Position	Result	Opponent	Matchday
count	9695	9695	9695	9695	9582	9582	9582
unique	266	56	15	85	54	15	34
top	Connor Azpilicueta	"American"	"Cairo City"	"GK"	"2-1"	"Hollywood FC"	"5"
freq	91	1798	1005	860	787	994	560

Figure 2: Sample data

The data is then further cleaned and processed according to the required procedure which is discussed in the Method section.

4 Method

As mentioned before three different performance metrices are used for predicting the player performance i.e., Average Rating, Goals and xG. Because of that three different data sets are prepared and then used for the prediction of player performance. All further machine learning techniques are implemented on all these data sets to find the most optimal prediction for the player performance.

4.1 Data preparation

Two different data sets are prepared for these three prediction matrices. Mention below.

1. Data set for xG and Goals matrices
2. Data set for Average rating

4.1.1 Pre-processing

1. Data set for xG and Goals matrices

   As expected goal (xG) is a measure of how many goals the player was expected to score given their chances and goals is the number of goals a player scored. So according to our data a player with the position 'GK' never scored a goal. The goals and xG measures are always zero for this specific player. That means a goalkeeper never scored a goal and the expectation of score a goal for the goalkeeper is also always zero. So that's why data set for these measures is created separately without goalkeeper data and features.

2. Data set for Average rating

   Data set for this specific measure includes all the player attributes and tuples from the raw data.

4.1.2 Feature engineering

While doing some feature engineering I found some features useless for xG and Goals metrics measure as they are goalkeeper's specific attributes, and they are always zero for everyone else. Mentioned below are the following features which are excluded for these measures, but they are considered for Average Rating metric.

1. Aer - Aerial Reach
2. Cmd - Command of Area
3. Com - Communication
4. Ecc - Eccentricity
5. Han - Handling
6. Kic - Kicking
7. 1v1 - One on One
8. Pun - Tendency to Punch
9. Ref - Reflexes
10. TRO - Tendency to Rush

11. Thr – Throwing

After removing unwanted features, we have 37 features for xG and Goals metrics. For Average Rating metric we have 48 features. Correlation heat map is used to find potential relationships between variables and to understand the strength of these relationships.

4.2 Model training

As it is a supervised learning regression problem, so I choose different regression models, explained below.

4.2.1 Linear regression (baseline)

Linear regression is taken as a baseline model in this research as it is most common and simple model to deal with regression problems.

4.2.2 Neural Network (simple)

To build neural network following configurations are taken: 4 hidden layers, 100 nodes, dropout rate of 0.01 and 'relu' activation function in most of them except the last layer where 1 node and 'linear ' activation function is used. The metrics which is used here is mean absolute error and optimizer is 'rmsprop' with batch size = 10 and epochs = 100.

4.2.3 Neural Network (optimized)

Grid search from sklearn is used to tune the hyper parameters and a neural network with the best parameters [1] is trained

1. For xG data set the following settings are used batch size = 4, epochs = 1000 and 'adam' optimizer.
2. For Goals metrics data set the following settings are used batch size = 4, epochs = 1000 and 'adam' optimizer.
3. And for Average Rating metric data set the following settings are used batch size = 4, epochs = 500 and 'sgd' optimizer.

4.2.4 XGBoost

For extreme gradient boosting model, I used following configurations[2]

1. learning rate = 0.1
2. max depth = 15 (maximum trees for base learners)
3. alpha = 15 (penalty factor responsible for L1 regularization)
4. n estimators = 40 (number of gradient boosted trees)

Cross validation is also performed here with 10 splits.

[1] https://keras.io/guides/training_with_built_in_methods/#many-builtin-optimizers-losses-and-metrics-are-available

[2] https://xgboost.readthedocs.io/en/stable/python/python_api.html

4.3 Model evaluation

For model evaluation at first model is split into training and test data, with 20 percent test data and 80 percent training data to check how the model is performing. Whole data is also scaled using standard scalar from sklearn library. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) is used at the end for evaluation[3] of predicted and actual values. RMSE tells us how far the predicted values are from the original values and on the other hand MAE gives the difference between the original and predicted value by averaged the absolute difference over the data set.

5 Results

Following are the results of all the models implemented on defined three data sets.

Metrics	Linear regression	Neural Network	Neural Network optimized	XGBoost
xG	0.232	0.236	0.238	0.251
Goals	0.414	0.472	0.475	0.460
Avg. Rating	0.563	0.569	0.569	0.551

Table 1: Root Mean Square Error

Metrics	Linear regression	Neural Network	Neural Network optimized	XGBoost
xG	0.142	0.123	0.125	0.143
Goals	0.236	0.158	0.156	0.234
Avg. Rating	0.422	0.421	0.415	0.411

Table 2: Mean Absolute Error

Here, MAE is not used to compare between these three data sets as MAE can't compare results between different data sets or models. On the other hand, we can see average absolute error between actual and predicted values.

6 Discussion

The discussion here is left to find which approach is better to take and what is the better way to predict the performance of the player. As I have chosen to go with three different data sets depending on what we want to predict. If the goal is to find out a player can score a goal or not then Goals metric should be taken, and if the goal is to see the expectation of goal over chances than xG metric is the option but to find overall performance of all player's Average Rating metric is the only option where players with position 'GK' (Goalkeeper) can also be evaluated.

Else than that for model selection its quite hard to decide which model is performing better using RMSE metric as mostly our baseline model is performing much better than our advance level techniques. And using these advance techniques which take a lot of execution power this cost is not worthy compared to result achieved as for the defined settings a single run takes approximately 35 minutes at least without the power of GPU. But we can see some promising result for MAE metric though this metric is not useful for comparing different models still handy to evaluate the error.

[3] https://machinelearningmastery.com/regression-metrics-for-machine-learning/

7 Conclusion

Though neural networks are very popular but still traditional methods are better in a lot of cases, especially when we have limited data. The amount of data can be a reason for neural network performance and more machine power can also help to conclude better results.

Furthermore, data for different seasons is not much too and techniques like data imputation can be applied to cover that up but it can also cause a very unrealistic and biased model. As I feel in sports many factors play their role which can't be calculated like emotional and psychological health of a player can completely change the stats of the player which effects the result of the game.

8 Related works

In this blog[4] Sanjit Patnaik performed some analysis and try to predict match ratings of football players using machine learning.

In this paper [3] researchers used machine learning approaches for player performance and match results prediction for cricket is quite similar task but totally different approach.

Here [2] Mei-Ling Huang * and Yun-Zhi Li uses Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches.

Murtaza Saif research paper [5] on Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics uses same XGBoost approach which is proven quite efficient and resourceful technique.

Another paper [4] by Yiming Ren and Teo Susnjak 'Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index' uses a lot of machine learning techniques to find the best model.

In this paper 'Prediction of Winning Team using Machine Learning' [1] Yash Ajgaonkar and Kunal Bhoyar uses simple machine learning models for the prediction of wining team.

[4] `https://medium.com/p/51bf7a5ab6ad`

References

[1] Yash Ajgaonkar et al. “Prediction of Winning Team using Machine Learning”. In: ().

[2] Mei-Ling Huang and Yun-Zhi Li. “Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches”. In: *Applied Sciences* 11 (May 2021), p. 4499. DOI: 10.3390/app11104499.

[3] Harshal Mittal et al. “A study on Machine Learning Approaches for Player Performance and Match Results Prediction”. In: *ArXiv* abs/2108.10125 (2021).

[4] Yiming Ren and Teo Susnjak. “Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index”. In: *arXiv preprint arXiv:2211.15734* (2022).

[5] Murtaza Saifi. “Implementation of Machine Learning Techniques to Predict Player Performance using Underlying Statistics”. PhD thesis. Dublin, National College of Ireland, 2020.