REPORT

Football Player Rating Predictor Model

Date: 10/2/2022

By

Akshit Langeh

Piyush Sandhikar

TO

Prof. Bhavana Tiple

Table Of Contents

Abstract

Evaluating a player's rating is becoming more and more central in football industry. Not only in football but also various fantasy football uses such evaluations to compare performance of professional players. Due to availability of a massive data being generated during a match training dataset to our needs is feasible.

The aim of our model is to help various scouts representing a football team to find the ideal player. With the help of various machine learning algorithm, the model identifies the overall rating of a non – professional player that is yet to be professional. Not only the overall, but the potential that the player carries and also the various positions the player can play in the team.

Introduction

The problem of evaluating the performance of a football player is attracting a lot of companies and scientific communities, thanks to the availability of a massive data being generated during a match. Thus, the football player rating predictor is a consolidated metric that can be used to not only provide the present but also the future potential ratings of a player.

Selecting talents on the entire football players is unfeasible for humans as it is too much time consuming, thus data-driven performance scores could help in selecting a small subset of the best players who meet specific constraints or show some pattern in their performance, thus allowing scouts and clubs to analyse a larger set of players and saving considerable time and economic resources while broadening scouting operations and career opportunities of the talented players.

Also this model is not limited to only football any other sport with similar database can be trained to predict the respective player's ratings and their importance in either the fantasy team or in the scouting process.

Literature Survey

Since many organisations and communities are interested in such metric system that can evaluate player performances, various researches have been done to achieve so. Whether be for a gaming engine or a scouting service or even a fantasy team related applications all require such a metric that can give an array of choices of player for the future parties involved.

PlayerRank is one of such model that was developed to give a detailed analysis of player performances with the help of huge data being generated after each match. Another research was done to find how the age matters in term of performance on the pitch but this research was done specifically for the under 17 women team that was yet to play in the world cup.

Other researches were also made but the one that was of interest to this model was the one that used EA Sports PPI for finding the player performance in the English Premier League. Thus, the EA provided data for the model that when pre-processed, trained and evaluated would give us the results and other various observations for our model.

Methodology

Now, the dataset provided by EA Sports consisted of data about the players at the end of the 2019-2020 season and had about 100+ columns defining the various attributes that has been taken into account to determine the rating of the player and how he will perform on the pitch. The dataset also had many data that was not required for the training purpose and if included could highly impact our algorithm. Thus, a lot of pre-processing techniques were taken into account to get the ideal dataset in order to make it suitable for the machine learning algorithms to be applied.

After the pre-processing step was done the dataset now had no missing values and columns that were not required (about 50+) were dropped. Also the categorical data was encoded with the help of Label Encoder and at last the subset of data was selected for the independent variables that will be necessary for calculating the three dependent variables viz. 'player_positions', 'overall' and 'potential'.

With the proper independent and dependent variables in hand the model was trained with various machine learning algorithms. For 'player_positions' classification algorithms such as K-Neighbours, SVC, Naïve Bayes, Decision Tree, Random Forest were used and for the evaluation purpose confusion matrix was used. Since there were more than 300 classes and the various positions like left mid and left wing back require almost same attributes the accuracy was below 0.5 and the best one was achieved by SVC (rbf kernel) and hence was selected to find the positions a player in a team. Now, for 'overall' and 'potential' regression algorithms were used like Multiple Regression, Polynomial Regression, SVR, Decision Tree, Random Forest. For the evaluation purpose R-Square method was taken into account and by comparing the accuracies of the models Polynomial Regression was selected.

To get the insights from the dataset K-Means clustering algorithm was used. Insights such as how age and pace of a player are related or the footballer with high pace can also have a high shooting rating.

Observations

Algorithm Used	Accuracy Score
K Neighbours	0.470726
SVC	0.495726
Naïve Bayes	0.347222
Decision Tree	0.357265
Random Forest	0.477778

^***Dependent Variable – 'player_positions' (classification)***

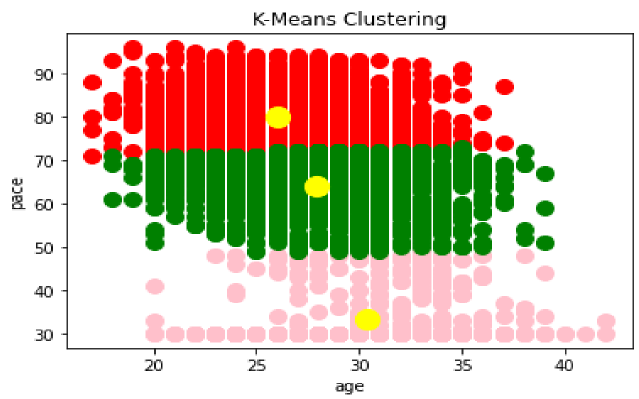
Algorithm Used	Accuracy Score
Multiple Regression	0.90
Polynomial Regression	0.973
SVR	0.971
Decision Tree	0.912

Random Forest	0.966

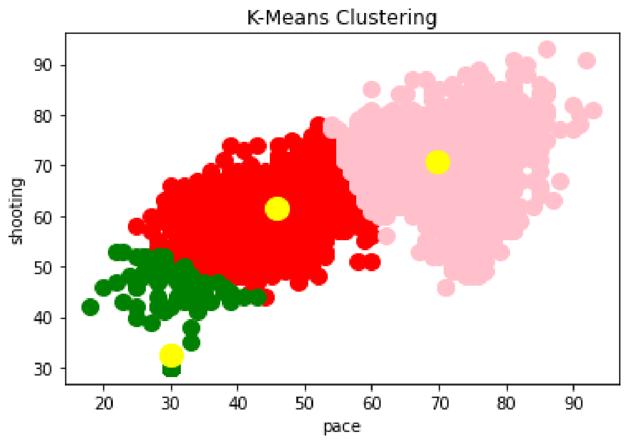
^Dependent Variable – 'Overall' (regression)

Algorithm Used	Accuracy Score
Multiple Regression	0.896
Polynomial Regression	0.973
SVR	0.971
Decision Tree	0.912
Random Forest	0.966

^Dependent Variable – 'Potential' (regression)



< AGE vs PACE



< PACE vs SHOOTING

Conclusion

From the observations we can conclude that SVC as a classification and Polynomial Regression as regression algorithm are selected due to their accuracy score. In case of classification due to there being about 300 plus classes the accuracy was less than 0.5 but when compared few samples from the result it was clear that the model gave good results as left midfielder being placed as left wing back and left back do not make very much difference in fact, it provides more versatility for a player. Also, the data can be used to form various clusters like we can get which age group of player has better pace or estimate the players that has both good pace and passing.

The researches that uses the EA rating dataset is mostly used in video game engine but with our model the same can be applied to real world challenges by giving various scouts and sport related communities an opportunity to choose wisely among a huge array of non-professional players available.

This model can be used to find how much potential does a young player carries and which roles in a team he can fulfil. The model once trained will be applied to the dataset consisting of non-professional players and by virtue of which various features of the player can be calculated like the position or role of the player in a real or a fantasy team.

Another feature of the model is its flexibility as instead of football we can train it on other dataset that belong to other sport.

References

Research Papers: -

- PlayerRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach (25 Jan 2019/Italy)
- Influences of player nationality, playing position and height on relative age effects at women's under-17 FIFA World Cup (August 2012)
- A football player rating system (October 2020)
- On the Development of a Soccer Player Performance Rating System for the English Premier League (August 2012)
- A methodology for football players selection problem based on multi-measurements criteria analysis (December 2017)

Other: -

- https://arxiv.org/pdf/1802.04987.pdf
- https://rstudio-pubs-static.s3.amazonaws.com/524786_4d3da4afc6bd406c8f720572b460cc3e.html
- https://towardsdatascience.com/predicting-premier-league-odds-from-ea-player-bfdb52597392
- https://www.kaggle.com/
- https://medium.com/analytics-vidhya/are-footballers-playing-in-their-best-positons-using-football-manager-dataset-a5346605541