JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Player popularity as a substitute for player ability in Premier league football using machine learning

Fahad Hilal[1*], Mohamad Saalim Wani[2]

[1,2]Department of Computer Science & Engineering, NIT Srinagar, J&K, India

Corresponding author: fahad.hilal11@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.647649 | Available online at: www.ijcseonline.org

Accepted: 15/Mar/2019, Published: 31/Mar/2019

Abstract- The effectiveness of player popularity as a proxy for ability, and the predictive power it would have in a model estimating a player's market value is examined.

Keywords**:** Multiple Linear Regression, k-means

I. INTRODUCTION

The Premier League is an English professional league for men's association football clubs. At the top of the English football league system, it is the country's primary football competition. Contested by 20 clubs, it operates on a system of promotion and relegation with the Football League. Besides English clubs, the Welsh clubs that compete in the English football league system can also qualify to play. The Premier League is a corporation in which the 20 member clubs act as shareholders. Seasons run from August to May, with teams playing 38 matches each (playing each team in the League twice, home and away). So, 380 matches per season in total. It is commonly known to as English Premier League mostly outside the UK.

For most football fans, May - July represents a lull period due to the lack of club football. What makes up for it, is the intense transfer speculation that surrounds all major player transfers. Their market valuations also lead to a few raised eyebrows, lately more than ever.

Fantasy premier league (FPL) is a game in which participants assemble an imaginary team of real life footballers and score points based on those players' actual statistical performance or their perceived contribution on the field of play. It has over 5 million registered players competing over a period of 10 months.

We aim to see how effective a proxy popularity could be for ability and the predictive power it would have in a model estimating a player's market value. The data-set obtained featured 461 observations. Categorical variables were encoded and transformed into dummy variables. The parameters used to predict market values include the club (reputation), player age, position (attacker, midfielder, defender, goalkeeper), social media page views in a specified time-frame, FPL value, points, region.

II. PROPOSED SYSTEM

The training data set is used to train the model to predict the market value (in million pounds) of the player based on actual transfer prices from the previous season. We use the multiple linear regression algorithm.

In addition, k-means clustering has been applied to draw inferences between the parameters and the market value of the players.

1.Algorithms

Multiple Linear Regression

Multiple linear regression [1] is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummy coded as appropriate). Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables x1, x2, ... , xp is defined to be y = 0 + 1x1 + 2x2 + ... + pxp. This line describes how the mean response y changes with the explanatory variables. The observed values for y vary about their means y and are assumed to have the same standard deviation. The fitted values b0, b1, ..., bp estimate the parameters 0, 1, ..., p of the population regression line.

K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms[2] that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because different location causes a different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point, we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop, we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm aims at minimizing an objective function know as the squared error function given by:

where,

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\left\|x_i - v_j\right\|\right)^2$$

‘||xi - vj||’ is the Euclidean distance between xi and vj.

‘ci’ is the number of data points in the ith cluster.

‘c’ is the number of cluster centres.

2.Libraries

Pandas: It is an open source library that provides easy to use data structures and tools for data analysis. It makes available the tools needed for reading and writing data. It provides the flexibility of insertion and deletion of columns. Handling of missing data is also made possible by this library. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. It has the broader goal of becoming the most powerful and flexible open source data analysis or manipulation tool available in any language.

Scikit-learn: It is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is largely written in Python, with some core algorithms written in Cython to achieve performance.

Matplotlib: It is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. It allows for the generation of plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

III. RESULTS AND DISCUSSION

➢ From the analysis of the associated data-set, we were able to prove that we can predict the market value of a player using just their popularity (which was extracted using theparameters mentioned. An accuracy of 71% was achieved, within a margin of 5 million (This was chosen as the threshold/margin because players rarely cost that little)).

- The k-means clustering algorithm was able to cluster the data-set into two groups, based on age. Players in their 20's commanded a higher market value than those in their 30's. This tells us that younger players have a higher market demand than their older counterparts, being fitter and having, potentially, a larger playing-tenure.

- The Fantasy Premier League values also seem to directly correlate to the players' market values. Players which had the FPL values in the range 1-5 commanded a market value of 10 million max. Those that had values in the range 5-9 had the highest value of 20 million. Those over 9 had values over 40 million.

- Clustering page view (social media) and market value also yielded substantial results. Players having over 2000 views cost more than 25 million, in general. Players with less than 800 views cost less than 8 million, for the most part.

The application of multiple linear regression proves that there is a relation between a player's market value and popularity. If teams were to consider player popularity parameters, instead of just going by the opinions of a small group of people (even though they may be experts) or the claims of the selling club, they would benefit immensely in the long run.

Also, the clustering algorithm was able to identify a more or less direct relationship between a player's market value and their age, FPL values, and social media page views.

This shows that we can use a player's popularity as a substitute for their ability, which is especially useful to pick up good up-and-coming talent for which we don't have much ability data to go on.

IV. CONCLUSION

During the player transfer window that opens from the month of May and lasts till July, and the one in January, teams often end up overpaying for player buyouts. Doing this results in teams not being able to fill up all their desired positions with suitable players. In addition, this also results in an increase in ticket prices for the home matches of the club.(Arsenal ticket prices for the season 2016 were almost twice as much as those for the second most expensive club).

REFERENCES

[1]H. W. Altland, "Regression analysis: statistical modeling of a response variable," Technometrics, vol. 41, no. 4, pp. 367–368, 1999.
[2]Jigui Sun, Jie Liu, Lianyu Zhao, "Clustering algorithms Research", *Journal of Software*, vol. 19, no. 1, pp. 48-61, January 2008.