Rating Soccer Player:

What are Key Contributing Features? How Well They Predict?

Lasang J. Tamang

The University of Memphis

Author Note

Lasang J. Tamang, Department of Computer Science, University of Memphis.
Correspondence concerning to this paper should be addressed to Lasang J. Tamang, Department of Computer Science, The University of Memphis, Memphis, TN 38152.
Contact: ljtamang@memphis.edu

Abstract

Rating is key performance measure of player's performance. Most of the current rating system by FIFA and other official organizations use more than 50 features about players. Such large information cannot be collected in a local club making a job of rating complex. In this research, we performed statistical analysis and identified 8 features which are age, strength, long pass, shot power, heading, aggression, reaction and composure as the key contributing factor. We also showed that we can use only these 8 key contributing features and yet predict 74 percent correctly.

Keywords: Prediction, correlation, multiple regression, normality

Rating Soccer Player:

What are Key Contributing Features? How Well They Predict?

Rating is one of the most widely used indicators of player's performance measure in the soccer world. The Federation International de Football Association (FIFA) officially uses rating as the key performance measure of the players. Different stakeholder of soccer game frequently refers to the rating of players for the decision-making process in regard to their performance. For example, a fan of the player uses rating score of his/her best player in order to support his/her claim that the player he/she support is better than player chosen by another person. A team selection committee may use rating of the player to make the decision regarding the selection of the team member while the club owner wanted to know the rating of the player to decide whether to buy the player or not.

Rating players in soccer is a very complex procedure and done on the basis of various features of the players. FIFA rating, which provides one of the most reliable sources for assessment of the professional player, rates the player on the basis of more than 50 features of players which includes physical, mental and technical skills and many other factors. The professional club often collect various information on a player to rate their player for the performance measure. The initial assessment often starts with the collection of technical skills (heading, long pass, short pass, dribbling, etc.) and physical information (height, weight, stamina, vision) of player as coaches often attribute the success of the player with talent and physical characteristics (Hyllegard, Radlo, & Early, 2001). The sports phycologist suggests that physiological and mental skills also contribute to the player's success (Laguna & Ravizza, 2003; Smith, Schultz, Smoll, & Ptacek, 1995) and hence use them for rating.

Collection of data about players on such large number of feature, in order to rate them, if often infeasible for clubs that are local or college level. Having access to the limited resources and time, it is impossible both technically and economically to collect information on all such factor that affects the rating. Also, when it comes to making the assessment of players for small clubs or college team who are not listed in the FIFA, their past performance data are often unavailable. This makes the assessment procedure tedious and coaches might wonder how to evaluate players they are interested in or which features to consider while starting to collect data about their player to make the future evaluation with consideration of their limited resources and budget.

Small local or college clubs can alternatively collect data with consideration of features that most contribute to a rating. They should consider any features that are most important to measure rating, regardless it is mental or physical or technical skills. But, what are those key contributing features? How accurately can we predict the outcome of rating if we only use these key contributing feature? Which feature contributes most to rating? The goal of this research study is to answer these questions.

Spieler (2006) performed a study on deterring the physiological factor affecting the success the soccer player and found that there are 28 skills. Unlike his study, our research is not confined to finding the only physiological factor, we look any important factor regardless technical or mental or physical. Also, the purpose of our study is more focused on finding as few features as possible which still predict as good as using all those standard features. Piedmont, Hill, and Blanco (1999) studied how to use the five-factor model of personality to predict the rating. Similar to them we also attempt to make player’s rating prediction, however, we do not confine our study on personality factor only. We take into consideration of all kind of factor and allow our study to select the best factor contributing most to the rating.

Methods

Research Question

The goal of this research is to determine the key contributing factor in rating player and see how accurately we can use them to make the prediction. For this, we studied following three research questions:

Question 1: What are the key feature that contributes to the rating of players?

Question 2: How much of the variance in rating score can be explained by those features?

Question 3: Which feature is the strongest predictor of rating?

Dataset

For our research purpose, we use Kaggle(2010) dataset publicly available at their website. This dataset consists of 17588 soccer players from 18 different countries with 52 features per player. Those features with 10 categorical and rest 42 continuous scales are sourced from the EA sport's FIFA video games series, and consists of technical skills (ball control, finishing, heading, jumping, long pass, short pass, dribbling, etc.), physical abilities (stamina, running, strength, etc.), and psychological behaviors (aggression, composure, reactions, etc.), etc.

For the purpose of our study, we drop all features from the dataset that are not continuous in scale, thus we consider only 42 features that are continuous in scale. Our subject of interest i.e. rating values ranges from 0 to 100 and has maximum, minimum and mean value of 45, 94 and 66.17 respectively.

Error check. We checked for any error, normality (Shapiro & Wilk,1965) and missing values for those attributes and found none. Therefore, no handling of error was essential. We performed descriptive statistical analysis and accessed maximum, minimum and mean of each of those attributes confirming none of our values falls outside of the range. The same analysis also reported no missing values.

Outliers handling. Histogram of al features showed they were all normally distributed, which indicated no possible outliers. However, boxplot over few attributes showed outliers which were not extreme. Those non-extreme outliers (Barnett & Lewis, 1994) were very few in number (less than 30) and were not handled for they do not violate normality of data as our sample size is very large. When the sample size is very large, we are safe to assume that our sample is normally distributed even if it is not.

Design

We used parametric statistical techniques to answer our research question as our dataset contains no errors, outliers and passes normality test. We performed correlation analysis to answer the first research question, and multiple regression to answer the second and third research question.

Pearson correlation to check the relationship between features. In order to answer what features contribute most to a rating of players, we conducted Pearson product-moment correlation analysis (Benesty, Chen, Huang, & Cohen, 2009) to find the strength of the relationships between each of the features in our dataset after we ensured no violation of homoscedasticity.

Multiple regression for prediction of rating. To answer our second and third research about the set of features that predict rating and find the strongest predictor, we used multiple regression analysis. We used rating as outcome features and age, aggression, reactions, composure, long pass, short pass, strength and heading as predictor features. These independent variables were chosen by accessing correlation result obtained while addressing the research question. We first choose only those variables which had the correlation of above 0.3 with a rating and then we omitted one of the independent features whenever we found it had the bivariate correlation of 0.7 or above with another independent variable. The independent variable chosen to remove among two features was one which had less correlation with the rating. Preliminary analysis was conducted and ensured no violations of normality, linearity, multicollinearity, and homoscedasticity before conducting the regression analysis (Jarque & Bera, 1980).

Results

Result from Correlation Analysis

Table 1 for correlation shows the person product-moment correlation between rating and 8 features that contribute most to the prediction of player's rating. Since our original table of correlation result is too large with 42 X 42 dimension, we only present correlation between rating and 8 most contributing features to rating, which is sufficient to answer our first research question. We excluded all those features which showed small correlation (r = <0.3) (Cohen, 1988) with rating feature. If there was a bivariate correlation of more than 0.7 between any two independent features, we excluded the feature in our table whichever had less correlation with the rating. Although, some of the features which had a high correlation with the rating but strong bivariate correlation with another independent feature were eliminated and will not be considered in our interpretation of the result, the interpretation is still valid since our goal of answering research question 1 is to be able to find features which showed high correlation with the rating and be able to use them in our multiple regression analysis.

From the result of correlation analysis, we found that 8 features out of 42 which are age, aggression, reaction, composure, long pass, shot power, strength, and heading contribute to the rating of the player which is the answer to our first research question. We can see all these eight features has a positive correlation with the rating, thus more their value more is the rating.

Result from Multiple Regression Analysis

Multiple regression was applied to assess the ability of eight chosen independent variable, which we obtained after addressing the first research question, to predict rating of performance and see which feature is the strongest predictor among them. The result is shown in table 2 for coefficients, figure 1 for a normal P-P plot of regression standardized residual, figure 2 for Scatterplot plot obtained from multiple regression tests.

Preliminary analysis was conducted to ascertain that assumption of normality, linearity, multicollinearity, and homoscedasticity was not violated. From table 1, we see that each independent features have a correlation with output variable rating above 0.3 showing some relationship between them. Also, no correlations between two independent feature were found to be too high (greater than or equal to 0.70) in our original correlation table. In table 2, no any features have the tolerance of less than 0.10 and variance inflation factor (*VIF*) greater than 10, thus indicates no multicollinearity. Figure 1 of normal p-p plot shows that all of the points lie in reasonably straight line and figure 2 shows scatterplot which is roughly rectangular, which leads to conclude that no major deviation from normality (Julie, 2016).

We answer our second research question by looking the value of r^2 of the model summary. We found the value of r^2 as 0.74, which means that our 8 features (age, aggression, reactions, composure, long pass, short pass, strength, and heading) explains 74 percent of the variance in our rating feature. We say that this result is statistically significant as we see sig. = 0.00(which means *p*<0.005) in table 2.

The answer to our third research question is inferred from table 2, coefficients. If we compare (ignoring the negative sign, if any) the value of beta under standardized coefficients in the table, larger the value more is the contribution of the variable in predicting the outcome variable rating. We see that feature with largest beta coefficient is 0.654 which is reactions. This means that reactions make the strongest contribution in predicting our output variable rating, while we control all other predictor variables in our model, which gives the answer to our third research question. We can also infer that heading is making the least unique contribution with the beta coefficient of -0.047 in predicting rating. All the predictor features are making a statistically significant unique contribution in predicting since all of them have a *sig*. = 0.00 values.

Discussion

It is not sufficient to look psychological or technical or physical attribute alone to measure player performance (Humara, 2000; Niednagel, 2004). The outcome of our result also suggests the same as we find 8 key contributing factor that requires all three psychological, psychical and technical skills. These 8 key contributing factors are age, strength, long pass, shot power, heading, aggression, reaction, and composure; the first two are physical abilities, following other three are technical skills and the last three are mental skills. This finding is intuitive in the soccer game for we know that to be a good player it is not only sufficient to have good physical abilities (e.g. run quickly to get the ball control), one should have good technical skills (e.g. dribble the ball skillfully through opponent) as well as good mental skills (e.g. react to a situation quickly and decide when to pass the ball to a friend or shoot to make the goal). We select only 8 out of all 52 features using statistical procedure and surprisingly it still contains features that span over those three skill set.

Using Cohen (1988) guidelines for interpretation of correlation score, we see age, aggression, long pass, short pass, strength, and heading has a medium correlation with rating while rest of the 2, composure and reaction, have large correlation suggesting former 6 contribute moderate and later 2 features high individually to the rating. Heading has least contribution (correlation, $r = 0.343$) and reaction (correlation, $r = 0.828$) has the highest contribution for rating. In football, the heading is least use skill and therefore its contribution being less to rating is obvious. It is not surprising that reactions alone have such high correlation and contribute uniquely 65 percent to rating because this single feature implicitly takes into consideration of many physical, mental and psychological factor. For example, while we talk about quick reaction, we also implicitly consider physical ability (run quickly, turn quickly), mental skills (quickly decide to turn or not, pass the ball or not), and technical skills (quickly dribbling).

Iyer and Sharda (2009) have successfully predicted the rating of the player using neural network approach with the accuracy of 80 percent. Our multiple regression model uses less knowledge of users and predicts with the accuracy of 74 percent. Our prediction in soccer, which is 6 percent less than the neural model, is very good given the different nature of the game. In cricket, the performance of the player is more dependent on individual effort and thus past data on the player can be used to predict the future performance of the player more accurately. But, in a soccer game, the performance of the player is dependent on the effort of an individual as well as the performance of other players in the team. Performance of other team member accounts more in soccer than in cricket; this makes the prediction of rating comparatively harder in soccer compared to a cricket game.

In our study, we consider only features that are continuous in scale. We did this to simplify our statistical model, but, it could be possible that consideration of features that we dropped can impact our model to behave differently. So, in future, it would be interesting to conduct another study without dropping those features. Also, our dataset is sourced from EA sport's FIFA video games series, which does not reflect the performance of the real player. So, it would be also interesting to replicate this study with data of real players.

References

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3, No. 1). New York: Wiley.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer Berlin Heidelberg.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences . Hilsdale. *NJ: Lawrence Earlbaum Associates*, *2*.

Frey, M., Laguna, P., & Ravizza, K. (2003). Collegiate athletes' mental skill use and perceptions of success: An exploration of the practice and competition settings. *Journal of Applied Sport Psychology*, *15*(2), 115-128.

Humara, M. (2000). Personnel selection in athletic programs. *Athletic insight*, *2*(2).

Hyllegard, R., Radlo, S. J., & Early, D. (2001). Attribution of athletic expertise by college coaches. *Perceptual and mot Niednagel, J. (2004). Jon Niednagel appears again on ESPN TV, this time questioned on his perspective of the top quarterbacks in the 2005 NFL Draft. Jon highlights Utah's Alex Smith and Aaron Rogers of Cal Berkeley.*

or skills, *92*(1), 193-207.

Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications*, *36*(3), 5510-5522.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, *6*(3), 255-259.

Kaggle. (2010). Complete FIFA 2017 Player dataset (Global)[FullData.csv]. *Available from Kaggle Web Site: https://www.kaggle.com/datasets.*

Pallant, Julie. (2016). SPSS survival manual : a step by step guide to data analysis using SPSS. Maidenhead :Open University Press/McGraw-Hil.

Piedmont, R. L., Hill, D. C., & Blanco, S. (1999). Predicting athletic performance using the five-factormodel of personality. *Personality and Individual Differences*, *27*(4), 769-777.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

Smith, R. E., Schutz, R. W., Smoll, F. L., & Ptacek, J. T. (1995). Development and validation of a multidimensional measure of sport-specific psychological skills: The Athletic Coping Skills Inventory-28. *Journal of sport and exercise psychology*, *17*(4), 379-398.

Spieler, M. J. (2006). Predicting starting status: factors contributing to the success of collegiate football players.

Table

Table 1

Correlation

Scale	Age	Aggression	Reactions	Composure	Long Pass	Strength	Heading	Shot Power
Rating	0.458	0.405	0.828	0.614	0.483	0.369	0.343	0.442

Table 2

Coefficients[a]

Model	Standardized t Coefficients	Sig.	Collinearity Statistics	
	Beta		Tolerance	VIF
Age	.060	.000	.712	1.405
Aggression	-.056	.000	.416	2.406
Reactions	.654	.000	.560	1.784
Composure	.119	.000	.364	2.749
Long Pass	.118	.000	.390	2.565
Strength	.155	.000	.623	1.604
Heading	-.047	.000	.379	2.635
Shot Power	.038	.000	.384	2.602

Note: a. Dependent Variable: Rating

Figures

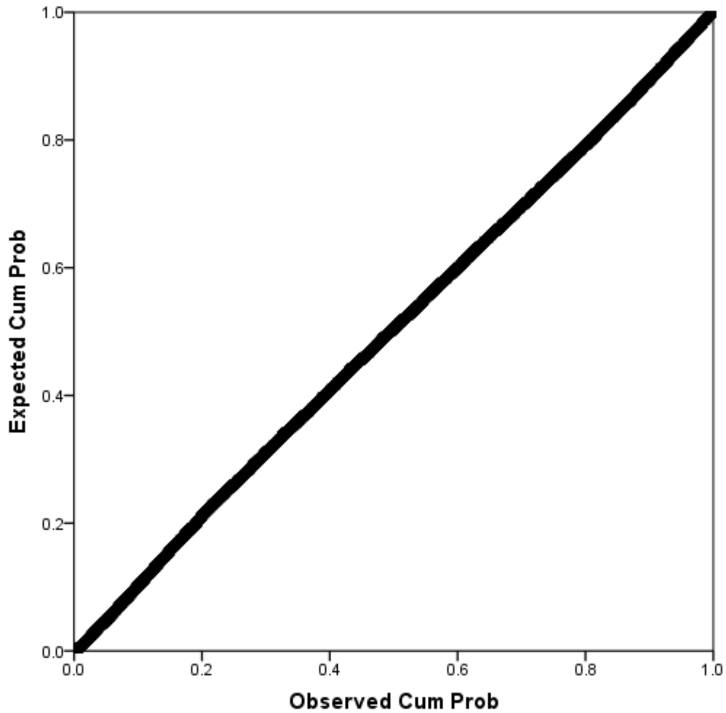

Figure 1. Normal P-P plot of regression standardized residual

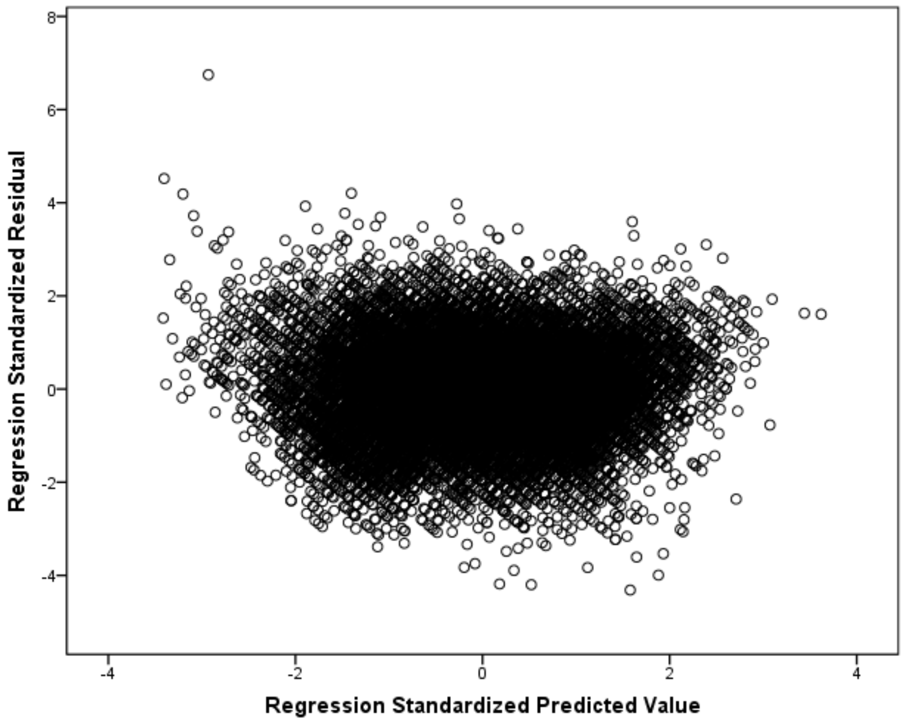

Figure 2. Scatterplot plot obtained from multiple regression test