Football Player Value Assessment Using Machine Learning Techniques

Ahmet Talha Yiğit(✉), Barış Samak(✉), and Tolga Kaya

Department of Management Engineering, Istanbul Technical University, 34367 Istanbul, Turkey
{yigitahm, barissamak, kayatolga}@itu.edu.tr

Abstract. Sports analytics is a field that is growing in popularity and application throughout the world. One of the open problems in this field is the valuation of football players. The aim of this study is to establish a football player value assessment model using machine learning techniques to support transfer decisions of football clubs. The proposed models will mainly be based on the intrinsic features of the individual players which are provided in Football Manager video game. To do this, based on the individual statistics of 5316 players who are active in 11 different major leagues from Europe and South America, different value assessment models are conducted using advanced supervised learning techniques like ridge and lasso regressions, random forests and extreme gradient boosting. All the models have been built in R programming language. The performances of the models are compared based on their mean squared errors. An ensemble model with inflation is proposed as the output.

Keywords: Football analytics · Player value prediction · Machine learning · Ensemble learning · Extreme gradient boosting · Lasso regression

1 Introduction

Throughout the 20th century, a sport branch named "football" spread across the world and became the most popular form of sport. However, it was the 21st century that football expanded greatly as an entertainment sector. This expansion reflected itself on sponsorship deals, television deals and most importantly market value of football players. Nonetheless, in today's football world; values of players are still highly judgmental and fragile. Even the biggest clubs in the world with the highest level of resources are transferring players for fees that they regret later. At the same time, an unexpected player can breakout and become the savior of his small club by completing a transfer for high fees. In this competitive environment, the transfer decisions are becoming more and more important for football clubs to be successful both in financial and competitive aspects.

The aim of this study is to provide a value assessment model for football players using machine learning techniques that will provide a better guideline for the clubs in the world of football. A model that values players using their normalized abilities and relatively free from the environmental variables is expected to be built which hopefully

C. Kahraman et al. (Eds.): INFUS 2019, AISC 1029, pp. 289–297, 2020.
https://doi.org/10.1007/978-3-030-23756-1_36

will provide better results than current models that are reliant on in-field game statistics. Considering the expansion of football industry, a model using the latest developments in the area of data analytics and machine learning should address a huge problem for the industry and could be a highly valuable financial leverage for the clubs looking to expand their successes and profits.

The rest of the paper is organized as follows: the literature review of football analytics and football player valuation models, the methodology used in the study, data collection and manipulation, model building, examination of the findings, discussion and suggestions for the future studies.

2 Literature Review

The story of analytics is not an old one however; the story of analytics in sports is even younger. This story starts in 1977, when a book named "Baseball Abstract" was introduced [1]. After the first introduction years, sports analytics grew bigger and have been implemented in many different branches sports regarding different problems. Even though football is and has been the most popular sport, in the field of analytics it was always behind other branches such as baseball and basketball [2]. Today, still there are traditional metrics which are being used excessively to understand the beautiful game. New metrics are just being introduced and still have not been able to prove their worth. These new metrics have been introduced in a couple of different fields including; training, tactics, performance tracking. As Roosenburg stated, machine learning will play an important role in football analytics to find insights about football which cannot be find by human beings [3]. In the light of this vision, it can be said that football analytics with machine learning applications will be used widely in many fields of football such as player valuation as the needed data is becoming more and more available. Valuation of football players was and still is a huge problem for this growing football industry. In cooperation with the IT University of Copenhagen, scientists from the University of Liechtenstein conducted an algorithm to find an objective value for every individual football player by just regarding the data about the player instead of the opinions of experts and fans [4]. Another example is KPMG, in collaboration with OptaSports which is a leading company in the field of football analytics, developed a benchmarking model to value players [5]. These examples shows that valuing football players using objective, normalized talent-based metrics by avoiding overvaluing or undervaluing players is an area of interest and research regarding football analytics.

3 Methodology

In multiple linear regression, a line will be created by determining the coefficients.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_P + \epsilon \tag{1}$$

The ideal combination of coefficients that minimizes the residual sum of squares (*RSS*) will create the best fitted line to the dataset.

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_p x_{ip})^2 \quad (2)$$

To find the best subset of variables for the regression model, shrinkage methods will be used. The below equation belongs to ridge regression method which minimizes coefficients of lesser important variables. The lambda (λ) parameter which minimizes the mean squared errors will be selected [6].

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3)$$

Another shrinkage method that will be used is the lasso regression method. As can be seen in below equation, this method is highly similar to ridge method. In this method the coefficients of the lesser important variables will be equal to zero [6].

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (4)$$

In regression model building, two different variables can be in interaction with each other [6]. Interaction terms are showed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2 + \epsilon \quad (5)$$

Decision tree based models uses the below formula to minimize *RSS* value [6].

$$RSS = \sum_{j=i}^{J} \sum_{i \epsilon R_j} \left(y_i - \widehat{y}_{R_j} \right)^2 \quad (6)$$

Random forest method finds out if a certain number of random variables (m) among all variables will achieve a better accuracy. After selection of m, in every bootstrapped data set randomly selected variables are used to grow trees sequentially for each data set to create and ensemble of the trees at the end. By this way, in some of the grown trees, weaker predictors will also affect the prediction accuracy [6].

Extreme gradient boosting method has been introduced to use gradient boosting algorithm in a more scalable and computational way [7]. It is one of the most accurate ensemble learning methods which is used in many machine learning competitions.

4 Data

To replace the highly volatile in-field game statistics, we will use the data from Football Manager Simulation Game. Football Manager is a football management simulation game developed by Sports Interactive and published by Sega. Since 1992, the game was published every year [8]. Since the natural aim of a simulation is to make it as real as possible; throughout the years, the database that Football Manager possessed became the most advanced and detailed one in the football world. The news agency VICE described Football Manager as 'the world's most influential video game.' and

added that 'Sports Interactive have created something so real it has professional applications.' [9]. The data consists of individual information and attributes representing different abilities for every player. There are 49 attributes for every player under 4 main chapters which are; technical, mental, physical, and goalkeeping. This data is being created and revised every year by over 1000 professional scouts around the world employed by Football Manager. Every attribute is being judged to be somewhere between 0 and 20. The difference of this data compared to the traditional in-field statistics is that it is not completely based on the in-field performance of players and all the attributes are being judged by professionals [10]. If a player scores a lot of goals in a season, the traditional metrics like goals per game or goals per shot may not give the true estimation of how successful this player was at scoring. However, the professional scouts who are judging this player's scoring ability will judge this ability by keeping in mind a lot of other variables regarding the environmental situations that this player's currently in which makes this data a reliable alternative to value players. After the collection of data from the FM 2018, values for every player from transfermarkt.com have been collected using web scraping codes in R programming language [11]. Finally, after transfermarkt.com values have been matched with data from FM, the data that will be used in this project have become ready. However, it should be noted that sampling and manipulation will be used on this data and the most proper piece of data will be used while building models.

Analyzing this data showed us that the values of players have a right-skewed distribution. Logarithmic transformation has been applied to give it a better distribution (Fig. 1).

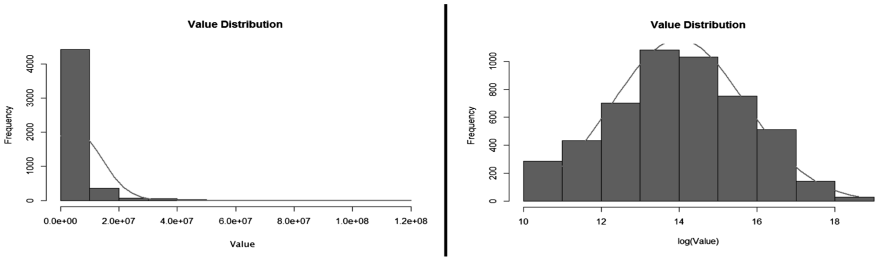

Fig. 1. Football player value distribution before and after logarithmic transformation.

On the other hand, the data at hand consists of players from a lot of different leagues regarding different levels of football. KPMG Valuation Model, which is a well-known and respected valuation model in the industry, uses the 11 highest level leagues in the world [5]. This method also will be used in this model since it will serve the aim of the project better. These leagues will be; English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, French Ligue 1, Dutch Eredivisie, Belgian Pro League, Portuguese Primeira Liga, Turkish Süper Lig, Brazilian Serie A, Argentine Premier Division. Moreover, since goalkeepers have a different ability set than the other positions, they are excluded from the data set. Moreover, feature engineering

applied on potential and current ability, *Division*, and *Club* variables. From potential ability (*PA*) and current ability (*CA*) variables, a new variable called "Ability" has been created.

$$\text{Ability} = CA + PA \times (40 - Age) \div 100 \tag{7}$$

Furthermore, *Division* and *Club* variables have been divided into different tiers. For divisions, English Premier League, Spanish La Liga, German Bundesliga, French Ligue 1 have been defined as *TIER1*, Italian Serie A, Portuguese Primeira Liga, Turkish Süper Lig have been defined as *TIER2*, and Brazilian Serie A, Argentine Premier Division,Dutch Eredivisie, Belgian Pro League have been defined as *TIER3*. Also, clubs have been separated into 5 tiers with respect to their total market values. As the result, our data consisted of 5316 players from the 11 high level leagues.

5 Model Building and Comparison

In the beginning of the model building process, multiple linear regression model has been built with all the variables and without adding any non-linearity or interaction. However, further investigations about finding the best subset of variables; non-linear relationships, and interaction terms should be done to improve the model to its limits.

Validation set, cross-validation, principal component regression (*PCR*), and partial least squares (*PLS*), ridge and lasso regression methods have been applied to the data to find the best subset of predictors (Table 1).

Table 1. MSE comparison of all methods.

Validation set	0.768
Cross-validation	0.771
Ridge regression	0.608
Lasso regression	0.590
Principal component regression	0.769
Partial least squares	0.594

MSE results are found using cross-validation for each model created by each method. The model created by using lasso regression method provided the lowest MSE, consequently the subset of variables obtained from it has chosen as the best. For further improvements, this model will be used.

For all possible combinations of dependent and independent variables, the scatter plots have been analyzed to find possible non-linear relationships. After the examination, it is found that only "Age" variable has a significant non-linear relationship with dependent variable (*Value*). Applying analysis of variance (*ANOVA*), "*Age^2*" variable is decided to be added to improve the accuracy of the model.

To find interaction terms, a model with all the possible double interactions has been built. After that, the significant interaction terms are found from this model and they

have been compared with the models without interaction terms by *ANOVA* method. After analysis, "*Ability* * *Age*", "*Age* *Concentration (*Cnt*)", and "Determination (*Det*) * Technique (*Tec*)" terms are found to be statistically significant and added to the model as the final modification.

After finding the best subset of variables, nonlinear relationships, and interaction terms, the final model is built by using all the found variables.

$$\ln(\text{Value}) = \text{Lasso Subset} + \text{Age}^2 + \text{Ability*Age} + Age * Cnt + Det * Tec \qquad (8)$$

After linear regression, decision tree model creating process which involves recursive binary splitting, and methods of pruning, random forests, and extreme gradient boosting methods to find the best predictions has started.

As the first step a recursive binary splitting decision tree has been fitted to the training set. To find out if recursive binary splitting model has over fitted the training set, pruning has been applied. After analysis, it is found that the pruning will not increase the performance of the model. As a result, the best MSE at this stage is calculated as 1.302.

To improve the accuracy of recursive binary splitting model, the random forests method is used. The optimum number of predictors (*m*) has been determined as 18 out of 55. After random forest application with 18 predictors to our data set, the *MSE* is calculated as 0.949.

The last method that will be used to create a better decision tree model is extreme gradient boosting method. In many cases, extreme gradient boosting is promising to have more accurate predictions.

Because of the high number of parameters in this method, firstly a parameter tuning by grid search technique has been applied to find the best value for parameters (Fig. 2).

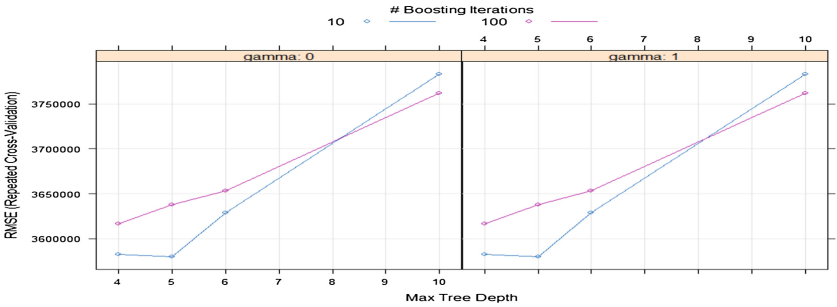

Fig. 2. RMSE value with respect to gamma, iteration number, and max depth parameters.

From this plot, it can be seen that the gamma parameter does not affect the root mean square error (*RMSE*) so it is taken as 0, number of boosting iterations as 10 and the maximum tree depth as 5 to minimize the *RMSE*. After finding the optimal values of parameters, the extreme gradient boosting method has been applied to the train set.

The *MSE* value of this model has been calculated as 0.170, consequently extreme gradient boosting model selected as the final decision tree model.

After model building, both regression and decision tree models are seen to be able to value players in accordance with FM in-game attributes. In this section, a final model will be built by using ensemble methods to combine these two reliable models. So that, the final model out of these two models will have even more accurate values than these two models separately. While decision tree model built by using extreme gradient boosting technique seems to have a much better accuracy in predicting players market values, this model have more static values. On the other hand, regression model have more dynamic values for players. For the final model the weighted average ensemble method will be used and the weights of decision tree and regression models will be selected as 0,7 and 0,3.

Since both models are built with the last year's values, the inflation for the player market in the last year should be compensated for to find accurate values for the current year of 2019. Also, FM 2019 data will be used for final results [12]. The change in values of the most valuable players will determine the overall yearly inflation for the market and this inflation rate will be added to the final model. Using the value changes of top 200 players, the inflation rate has been calculated as 0,407 for the previous year. After adding inflation rate to our ensemble model, the final model of the project has been finalized.

6 Findings

The results of the final model are in accordance with the aim of the project. Valuing players using their normalized abilities instead of in-field game statistics have been achieved (Fig. 3).

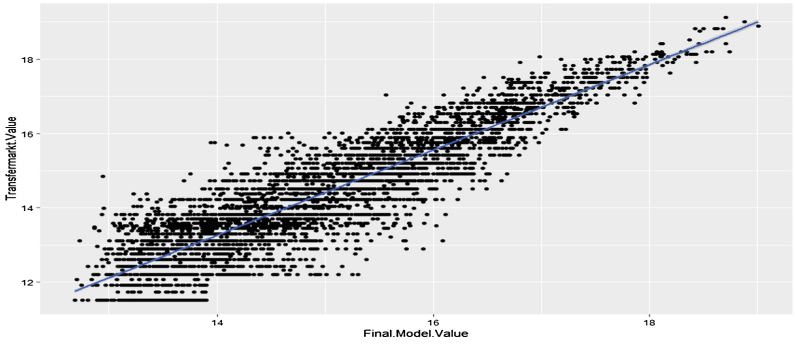

Fig. 3. Transfermarkt (y) and final model (x) values after logarithmic transformation for 2019.

Most of the values that have been found are in accordance with the current market values which prove the reliability and stability of the model. However, since the problem that has been described consisted of current models; it can be seen as a good

sign that some of the results are different than the current market values at certain points. Data points that are under the trend line represents players that the model found to be undervalued and data points that are over the trend line represents players that the model found to be overvalued (Table 2).

Table 2. Some transfers from January 2019.

Name	Market value	Transfer fee	Our valuation
Giorgian De Arrascaeta	9 M €	15 M €	*15.858.061,61 €*
Manolo Gabbiadini	10 M €	12 M €	*13.466.272,50 €*
Garry Rodrigues	7,5 M €	9 M €	*10.437.560,07 €*
Lucas Paqueta	15 M €	35 M €	*33.178.330,00 €*
Lisandro Magallan	6 M €	9 M €	*11.183.626,06 €*
Rodrigo Caio	9 M €	5 M €	*5.343.258,95 €*
Juraj Kucka	2,5 M €	5 M €	*4.803.908,11 €*
Jonny Castro	12 M €	20,5 M €	*19.109.354,14 €*

Above table shows transfers from January 2019, where the model found different values that the current transfer market. This comparison allows the model to make claims about the accuracy of undervalued and overvalued players that has been found. The actual transfer values of players has found to be closer to the model's valuations.

7 Conclusion and Discussion

The model proposed as the output of this study has the potential to be updated with future data to become even more precise and reliable. Many different approaches could be applied in the data collection and model building stages of the study. For example, a division or position based model could be built for better value predictions in a specific division or position. On the other hand, the data used in this study can be useful to create many other kinds of machine learning based football analytics models. Also, more advanced approaches can be used during ensemble and adding inflation steps. Using this idea and data, deep learning techniques such as artificial neural networks can be applied for further improvements. Another possible improvement can be provided by using the lightGBM technique which is gaining huge popularity in the data science world [13].

References

1. James, B.: Baseball Abstract, 1st edn. Ballantine Books, New York (1982)
2. The Unprofessionals article. https://unprofession.com/why-arent-soccer-analytics-a-bigger-deal-706670ab8685. Accessed 16 Feb 2019
3. SAS SciSports. https://www.sas.com/tr_tr/customers/scisports.html. Accessed 16 Feb 2019

4. Müller, O., Simons, A., Weinmann, M.: Beyond crowd judgments: data-driven estimation of market value in association football. Eur. J. Oper. Res. **263**, 611–624 (2017)
5. KPMG Football Benchmark. https://www.footballbenchmark.com/methodology_player_valuation. Accessed 16 Feb 2019
6. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R, 1st edn. Springer Science + Business Media, New York (2013)
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2016)
8. Football Manager: More than just a game. https://www.fmscout.com/i-1005-Football-Manager-More-Than-Just-A-Game.html. Accessed 16 Feb 2019
9. Blickenstaff, B.: Football Manager, the world's most influential video game. VICE SPORTS. https://sports.vice.com/en_us/article/gv74p3/football-manager-the-worlds-most-influential-video-game. Accessed 16 Feb 2019
10. "Football Manager 2018". Sports Interactive, 2017
11. Transfermarkt. https://www.transfermarkt.com/. Accessed 16 Feb 2019
12. "Football Manager 2019". Sports Interactive, 2018
13. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.: LightGBM: a highly efficient gradient boosting decision tree: In: Advances in Neural Information Processing Systems, vol. 30 (2017)