Football Player's Performance and Market Value

Miao He[1], Ricardo Cachucho[1], and Arno Knobbe[1,2]

[1] LIACS, Leiden University, the Netherlands, r.cachucho@liacs.leidenuniv.nl
[2] Amsterdam University of Applied Sciences, the Netherlands

Abstract. A lot of money is involved with the transfers of top players in the big European football leagues. For various reasons, obtaining a good economic valuation of football players throughout the year is valuable, in other words, not just when a player has just transferred. Furthermore, it is relevant to consider how the market value of a player relates to the performance of that player. Both these factors again depend on the various parameters of the player, that might be gleaned from various public sources on the web. In this paper, we demonstrate how market value and performance of La Liga (the Spanish League) players can be modeled using extensive public data sources.

1 Introduction

The transfer fees of football players are getting higher and higher each year. The UEFA Financial Fair Play Regulations [1] were recently implemented, in order to prevent professional football clubs from getting into financial problems by spending more than they earn, which might threaten their long-term survival. This will definitely affect the behaviour of clubs in the transfer market. Besides, right-valued players are not only very critical to the development of the team, but also essential to the agents and players themselves.

Consider the problem of economic valuation of football players. Most likely, the closest we can get to the real market value is the *transfer fee* of the player. This valuation is missing for most of the players, because players are not always moving from one club to another during transfer seasons. Furthermore, researching the market value individually for each football player can be quite hard work. We decided to approach this challenge by applying data-driven modeling techniques to attempt a proper valuation of football players.

Football is a team sport, thus it is quite hard to judge an individual football player's performance. Different people have different opinions on a player's performance. The responsibilities of each position are different, which leads to performance indicators also being different by position. We took the votes on "Who is the best player of this year?" by football experts [6] as the closest to represent the football players' performance. However, only top players are in the voting list. We have applied the same methods as for the real market value to get the performance indicators by positions. There is the fact that the forward players are more visible to the audience than other positions. Simply because football is goal-oriented, the forward players are overrepresented when it comes

to voting. Therefore, voting is most representative for the performance of forward players.

The goal of this research is to find the relationship between *market value* and the *performance of players*. In this paper, we develop regression models to predict the real market value and assess a player's performance. A fair market would assign a higher market value to a player with high performance. After we got the player's market value and his performance indicators, we look at the relation between the two.

2 Data Source Description

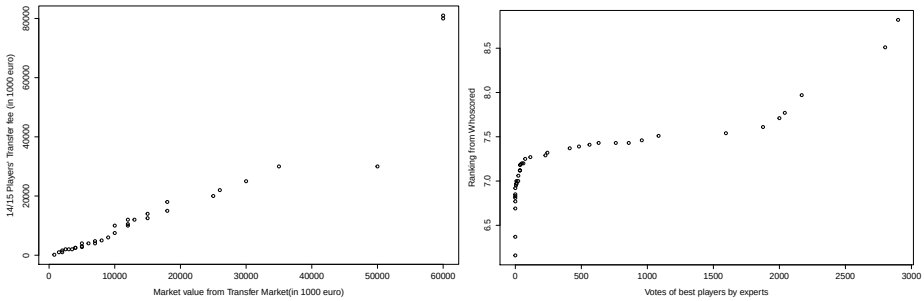
For this research, we required data containing a player's basic information (name, team, age, height, weight, ...), market information (transfer fee, former team, duration of the contract, when the player joined the team, ...) and performance information (on pitch time, actions at the ball, fouls, scores).

After an extensive online search and browsing related work [7, 8], we have found the following useful public data sources from which we got our datasets: Transfer Market [3], WhoScored [4], European Football Database [5] and Garter [6]. Due to time constraints of this study, we gathered and prepared data only for the Spanish League *La Liga*, for the first half[3] of season 2014/2015. Notice the difficulty to access and combine free football data. Firstly, due to its very high commercial value. Secondly, because combining the multiple sources is a record linkage problem. Due to space constraints, we left the record linkage task out of this paper.

The market data are from Transfer Market, which is a website to discuss and learn the latest news from the world of football. There is transfer news, rumors and also statistics on the market value, for example the length of contract, the former clubs. As for the real market value, we decide that the closest we can have to a real market value of a player is his transfer fee at the moment he is transferred from one club to another. This data was gathered from the European Football Database [5], which is a web database that presents all transfer news in tabular form, by league. It includes basic information of the transferred player and relevant clubs in the transfer.

The performance of the players data were collected from the website WhoScored [4], which has detailed statistics for the top 5 leagues in Europe accumulated at different scales (powered by OPTA). Details of the offensive, defensive, and pass data have been collected from this website. We have chosen performance data accumulated by every 90 minutes, because it is a normalized version of this data, making it comparable across players. As real performance assessment indicator, we considered the votes organized by media group *the Guardian*, which gathers all the relevant information (name, team, the total votes of player, etc.). The voters consist of football experts, sports journalist and the football players themselves. There are 73 judges from 28 nations voting and the more votes a football player gets, the better performance we consider the player to have.

[3] With one transfer window passed.



(a) Q-Q plot comparing transfer fees and market value from Transfer Market.

(b) Q-Q plot comparing Garter votes and WhoScored rating.

Fig. 1. Comparison between real and proxy variables for market value and performance assessment of football players.

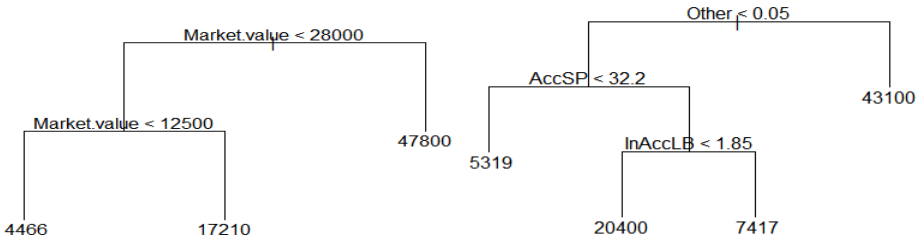
3 Market Value and Performance

In the previous section, we presented the real market value and performance indicators of players. But there is a common problem for both these variables. We have 381 players in La Liga, but not every player has been transferred this season and only some players will be on the candidate list of the voting. There are only 37 players who transferred this season and only 40 people in the voting list.

However, proxy variables for both the market value and performance have been collected. For all the players, we have a *Market Value* estimation from Transfer Market and a *Rating* from WhoScored. *Market value* is based on an algorithm built by Transfer Market to estimate the transfer fee if the players were transferred during the present season and is adjusted every year. *Ratings* are calculated based on WhoScored's algorithm, using OPTA's statistics and are updated during each game. The Rating variable is scaled from 0-10 where 10 indicates best. Both these algorithms are not public so we decided to collect and compared them the closest we could find about these variables [5, 6] (see Section 2).

The first step of our study was to find the relationship between proxy and real values. In Figure 1, we present the Q-Q plots crossing real and proxy values for market value and performance, for those cases where both values are available. The economical valuation of Transfer Market seems to match the prices paid for the transferred players. As for performance, the relation between real and proxy values appears to be non-linear but still monotonically increasing. Afterwards, we applied learning algorithms to a merged dataset containing all data sources, where proxy variables are put together with other variables as independent variables and the real variables are our target variables.

For each of our two targets, we included only those players in the training set for which the actual values were available. I.e., for *Market Value* we included the 37 players who actually transferred in that year, and for *Performance* we



(a) Regression Tree using market value from WhoScored to model transfer fee.

(b) Without using market value from WhoScored to model transfer fee.

Fig. 2. Comparison between real and proxy variables for market value and performance assessment of football players.

included the the 40 players that were nominated. The datasets used have 100 and 84 variables, respectively. This is because the performance-related variables were included to model the *Market Value*, but not the other way around. The rationale behind this decision is that performance cannot be influenced by economical variables.

Problem Statement Our main task is to find good regression models for both *Market Value* and *Performance*. More formally, we assume a wide dataset with both performance and/or economic related variables, as well as an evaluation function (R^2) that can evaluate the quality of the model, with respect to the target variables (real *Market Value* and *Performance*). The task is to find good models such that:

- The score of R^2 is high, where $0 \leq R^2 \leq 1$.
- The complexity of the model (number of variables) is low.
- The models are interpretable for further analysis.

3.1 LASSO regression

The datasets for estimating *Market Value* and *Performance* have both numeric and nominal variables. Most of the regression algorithms cannot deal with non-numeric variables, but regression trees can [15]. As an example, we trained a regression tree to estimate the real *Market value*. Both *team* and *nationality* were chosen (see Figure 2). The disadvantage of these trees is that the results are too general. Their ability to extract linear combinations of features is very poor. According to the result of the regression tree, the nominal variables do not play a large role in the result.

If we consider the subset of our dataset for which we have the real values of *Performance* and *Market Value*, there are more variables than observations. This makes it impossible to apply least squared methods [14] and avoiding over-fitting becomes a real challenge. Moreover, there are variables that are correlated.

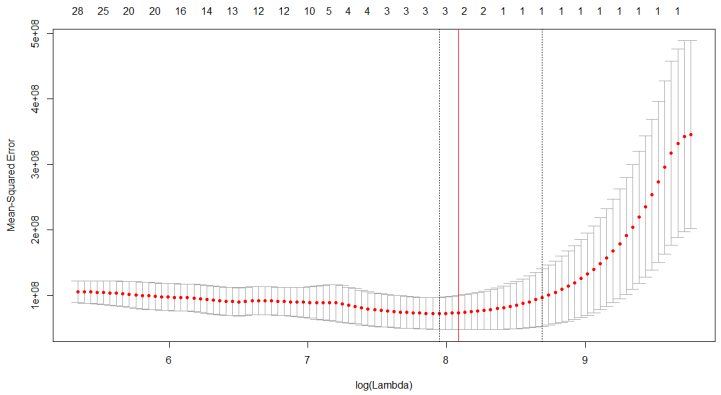


Fig. 3. LASSO regression for real market value prediction.

LASSO is a well-known regression technique for these cases [9]. It is able to perform variable selection in the linear model and it can have better accuracy than linear regression in a variety of scenarios, depending on the choice of *lambda* (λ). As λ increases, more coefficients will be zero which means fewer variables are selected and more shrinkage is employed among the non-zero coefficients. With a bound on the sum of the absolute values of the coefficients, it minimizes the usual sum of squared errors.

In the R package Glmnet [10], the algorithm uses cyclical coordinate descent in a path-wise fashion. Using cross-validation (CV), a suitable value for λ can be chosen. Glmnet proposes two significant λs. The λ_{min} option refers to value of λ at the lowest CV error. Sometimes λ_{min} might cause over-fitting, because the error at this value is the average of the errors over the k folds. The second option offered by Glmnet is to use λ_{1se}. This λ ensures the largest pruning of variables while keeping the minimum standard error, thus creating simpler models. The most suitable threshold is normally between λ_{min} and λ_{1se}. After choosing the right λ, the coefficients can be obtained with that λ and the unknown observations calculated.

4 Results and Discussion

4.1 Real Market Value

There are negative values for the predicted real market value when we using λ_{min} as threshold. Clearly, it goes against common sense when you pay money for selling your players. However, λ_{1se} is too restrictive by only introducing one variable.

According to Figure 3.1, the mean-squared error will be smaller when lambda is bigger. By taking all these issues into consideration, the criterion for λ is:

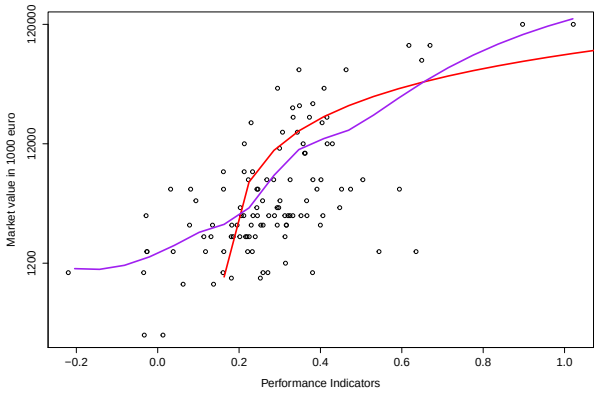


Fig. 4. Relation between market value and performance assessments.

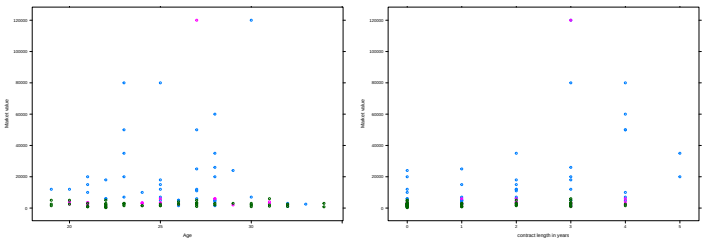
$min(\lambda_{min} - \lambda)$ provided that all the predicted market values are non-negative. Based on the above criterion, we decided for $\lambda = 3247.0$. It is the closest model to the best which would never cause negative market values. It is indicated by the vertical red line in Figure 3.1. It is in the interval of the best model and simple model. The model is:

$$\hat{M} = 231.26 + 0.89 \cdot Market.value + 2723.26 \cdot Assists$$

4.2 Performance Indicators for Forward Players

There are significant differences in performance between positions. It has been suggested [13], that there are key characteristics needed to play in certain positions within soccer. The data on market value has classified positions into 12 categories, which is too specific. Especially, most players have played in more than one position. In addition, the whole pool of our data is very small. If we made 12 subgroups, it would be too small for each subgroup. Therefore, we have used the categories suggested by [12]. They undertook a technical analysis of playing positions within elite level international soccer at the European Championships 2004. Players were classified by position into *goalkeepers*, *defenders*, *midfielders* or *strikers*.

In addition, we have used the t-test to test whether there are significant differences between positions when it comes to market value. When considering the market value between the four categories, $p = 0.001$ which indicates a very significant difference between the four positions. Furthermore, when comparing the market value between specific sub-categories within each group (e.g. comparing various types of defenders amongst each other), we get p-values above 0.8, which suggest it makes sense to group such very similar sub-categories. Hence, also from a market value perspective, the four categories are justified.



(a) Market value and age. (b) Market value and contract duration.

Fig. 5. Relation between market value and players characteristics and career.

We attempted modelling the performance over the entire set of players, but failed to find satisfactory results, due to the variance of performance per positions. Focusing on the forward players specifically, it becomes possible to model their performance. Forward players play an important role in the valuation of players. Although they only represent 30% of all players, they appear as winners of the FIFA World Player of the Year (since a few years called the Ballon d'Or) in 17 out of 24 years. No goalkeeper has ever won the prize.

In this case, we use LASSO to train the forward players. Five-fold cross validation was used. The threshold is λ_{min}. The KPIs for forward players have been selected as follows. A good player should have few Fouls (F). Shots and Goals in Penalty area ($SP\&GP$) are a big plus for player performance. Shots on Target (ST), Goals from out of Box (GB), Dribble successfully (D), Assists total (A) also contribute to the final results. The model for forward players now is:

$$\hat{P} = 0.28 - 0.073 \cdot F + 0.06 \cdot SP + 0.04 \cdot ST + 0.02 \cdot GP + 0.05 \cdot GB + 0.02 \cdot D + 0.08 \cdot A$$

4.3 Market Value vs Performance

Since we have predictions for the market value and performance assessments of players, the relation between the two can be studied. The over-all trend of market value follows the trend of performance. The better performance, the higher the market value will be (Fig. 4). There seems to be a ceiling for market value, where the top performing players have similar market values and very different performance ratings.

For the transferred players, we calculated the difference between the real market value and the estimated market value based on performance, $\Delta = ln(M) - ln(\hat{M})$. The smaller Δ is, the more proper market value for the players is according to his performance. We also considered that if $\Delta > 0.3$ the player is over-valued and if $\Delta < -0.3$, the player is under-valued. In general, the majority of over-valued players are also high-performance. This might be due to the marketing value of high-performance players. Normally, high-performance players also bring revenues to the clubs in terms of publicity and merchandise sales. This market variables are not incorporated in our model.

5 Conclusions and Future Work

We have built a model to value economically all the players of La Liga. Furthermore, the method could be applied to other leagues. As for the performance, the operational model for now applies only to forward players. We believe this could be extended to other positions if a unique model is created for performance across different leagues. As part of future work we would like to scale up the project to all European leagues.

Even considering covering all leagues, this project will keep on dealing with incomplete data because not all players are valued every year (by being transferred), neither are all players' performance evaluated by the voting system. In the future, we would like to consider semi-supervised methods to solve the tasks of *Performance* and *Market Value* estimation.

Finally, the voting system to access performance is biased towards forward players and good players. We would like to explore other data mining techniques that account for this problem, such that we could create an Elo Rating model alike for performance of football players, across positions and leagues.

References

1. www.uefa.org/protecting-the-game/club-licensing-and-financial-fair-play
2. http://www.optasports.com/en/about/who-we-are/about-opta.aspx
3. http://www.transfermarkt.co.uk/wettbewerbe/national
4. http://www.whoscored.com/AboutUs
5. http://www.footballdatabase.eu
6. http://www.theguardian.com/football/2014/dec/21/how-the-guardian-ranked-the-2014-worlds-top-100-footballers
7. Pavlović, V., Milačić, S., Ljumović, Controversies about the Accounting Treatment of Transfer Fee in the Football Industry. Management, 70, 2014
8. Kumar, G., Machine Learning for Soccer Analytics, 2013
9. Robert,T., Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, 267–288, 1996
10. Simon, N., Friedman, J., Hastie, T., Tibshirani, R., Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software,1-13, 39(5), 2011
11. Mehryar, M., Afshin, R., Ameet, T., Foundations of machine learning, MIT press, 2012
12. Hughes, M., Caudrelier,T., James,N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C., Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position, Universidad de Alicante. Área de Educación Física y Deporte, 2012,
13. Di Salvo, V., Baron, R., Tschan, H., Calderon Montero, F.J., Bachl, N., Pigozzi, F., Performance characteristics according to playing position in elite soccer, International journal of sports medicine, 222, 2007
14. Stodden, V., Model selection when the number of variables exceeds the number of observations, PhD Thesis, 2006
15. Ripley, B., Classification and regression trees, R package version, 2005