Towards data-driven football player assessment

Rade Stanojevic and Laszlo Gyarmati
Qatar Computing Research Institute
HBKU, Doha, Qatar
{*rstanojevic,lgyarmati*}@*qf.org.qa*

Abstract—**Understanding the value of a football player is a challenging problem. Player valuation is not only critical for scouting, bidding and negotiation processes but also attracts a large media and fan interest. Due to the complexities which arise from the fact that player pool is distributed over hundreds of different leagues and many different playing positions, many clubs hire domain experts (often retired professional players) in order to evaluate the value of potential players. We argue that such human-based scouting has several drawbacks including high cost, inability to scale to thousands of active players and inevitable subjective biases. In this paper we present a methodology for data-driven player market value estimation which tackles these drawbacks. To examine the quality of the proposed methodology and demonstrate that our data-driven valuation outperforms widely used *transfermarkt.com* market value estimates in predicting the team performance.**

Keywords-**football; data mining; performance measurement;**

I. Introduction

Association football, also known as Soccer, is one of the most popular sports globally, actively played by millions of amateurs and followed by billions of fans. A large fraction of countries have professional leagues and playing in one of the professional teams is often considered a lucrative profession. Currently, there is over a thousand professional soccer clubs and several tens of thousands professional players. Due to many factors, including relatively short career span, variable form throughout the career, risk of injuries and uncertain club budgets, both players and the clubs continuously seek to maximize their performance. From the players' perspective their goal is to find the right club where they can demonstrate their skill-set and be compensated appropriately. On the other hand clubs are on constant lookout for players which would maximize the team performance given the budget they have.

Matching the players and the clubs is a highly challenging process and almost always involves human experts (known as scouts) which use their knowledge of the game to asses the players capabilities and value. In this paper we propose a different approach to player scouting, one that purely relies on data. More specifically, our goal is to evaluate the value of the player using a non-subjective quantitative approach. Data-driven assessment of players' value is challenging in several ways. First, fine-grained data statistics for individual players is only starting to be publicly available on large scale. Secondly, most players play against only a handful of other teams which makes the comparison across different countries and different leagues very challenging. Thirdly, comparing players which play in different positions adds an additional level of complexity. And finally, there is no ground truth for testing the goodness of the assessment of individual player value. Note that it is the team nature of soccer which causes many of these issues [6]. For individual sports like tennis or snooker, it is much more straightforward to assess player's value as it is independent of the team, the competition is more global and comparing the player's value can be done directly since the players compete individually and not as a part of the team. We will discuss the team-vs-individual aspect of the problem in more detail later.

The main contribution of the present work are the following:

- We derive a methodology for assessing the player's market value using players' performance data which allows assessment and comparison of the player's value across different positions and different countries and leagues. The model we develop simultaneously ranks and valuates every active professional football player; over 12K players in our dataset.
- Using a range of available datasets we demonstrate that the team market value derived in our data-driven model is a stronger predictor of the team results compared to widely used *transfermarkt.com* market value estimates.

As we shall see in Section III there is a non-trivial discrepancy between our (data-driven) estimates of the market value and the one provided by most widely used public datasets: *transfermarkt.com* (which is in part subjective). Hence, our tool and methodology can be used by the scouts and team managers to look for the most optimal players which fit in their budget, and that search can be done globally in the pool of thousands of professional players from hundreds of different leagues.

II. Data

Our data comes from several sources.

Performance data. For 6 years, 2010-2015, sport analytics company InStat collected the performance in a bit over 100K games, from around 100K players in around 5K teams. The exact numbers can be seen in the Table I. Note that not every team has all of their games covered. Teams from the top European and South American divisions have most of their games covered, while the teams from the lower tier leagues have occasional games (e.g. cup games) which are covered. Also, the amount of data in 2015 is significantly

duration	# games	# players	# teams
1/2010-12/2015	108755	106082	5304

Table I
BASIC INFO ON PERFORMANCE DATASET.

higher than the amount of data from 2010, since InStat continuously increases its capacity to track more games. In order to collect the data they use their in-house developed technology and also employ 300 certified analysts which extract the information which is difficult to capture by the software.

For every tracked game InStat extracts key statistics regarding each player who played that day including: minutes played, position, #goals, #passes, #accurate passes, #challenges, #accurate challenges, #air challenges, #accurate air challenges, #key passes, #accurate key passes, #shots, #accurate shots, #dribbles, #accurate dribbles, #tackles, #accurate tackles, #crosses, #accurate crosses, etc. We use these information to build performance profile for each player; see Section III

Transfermarkt data. *trasfermarkt.com* is a large online service which follows virtually every professional and semi-professional football team in the world. For every player they report a few basic statistics for each played game such as number of minutes played, goals, assists, yellow/red cards. More interestingly, *transfermarkt.com* reports estimated market value for each player. Exact methodology used for the calculation of the market value estimate is unknown, but it involves opinions of professional scouts, as well as fans, in addition to the statistical performance indicators [8]. We crawled *transfermarkt.com* in March 2016 and extracted information from all players of the clubs with total (club) estimated market value of at least 3 million £. Overall there are 1383 such teams, and 19K players associated with them.

Actual dataset we use contains players which lie in the intersection of InStat and transfermarkt datasets. Since there is no unique player identifier which would allow us to calculate the intersection of the two sets, we use the following heuristic. Both InStat and transfermarkt datasets contain date of birth (DoB) and name for every player. We consider InStat player A to be the same as transfermarkt player B if their DoB coincide and their names have string similarity[1] of at least 0.5. We use similar names rather than the exact matches, since the two datasets have different naming convention. For example Leo Messi's name string in the InStat dataset is 'Lionel Andrés Messi Cuccittini', while it is only 'Lionel Messi' in transfermarkt. The similarity of the above two strings for Leo Messi is 0.59, though for a majority of the matched players in our dataset the similarity is 1, or very close to 1.

In the intersection of the two datasets we have 12858 players which were active (played at least one game) in 2014/15 season, which is a sizable chunk of the professional players, and it contains a huge fraction of active players from the top European and South American football leagues. Note that for many players from lower tier teams no DoB info is available, hence making it difficult to match them with their transfermarkt entries. However, among the top-tier teams matching between the two datasets is almost always successful.

III. Methodology

Our goal is to use the players' performance indicators to evaluate their market value. To that end we consider transfermarkt estimate of the market value to be equal to the true market value perturbed with a noise factor. Our approach to extract the true market value is the following. We consider the regression task of predicting player's market value given his performance feature vector. As a label, we use the *transfermarkt.com* market value estimate (TMVE) which we consider as a true label (true market value) perturbed by a noise variable. By learning (developing the ML regression model) using this noisy labels, we aim to eliminate the 'noise' and approximate the true market value for each player in the dataset. In Section IV we will evaluate the strength of TMVE and our performance-driven market value estimate (PDMVE) in predicting the team results. We would like to point the interested reader to several recent studies which analytically study the problem of learning with noisy labels [2], [3].

As we discussed above we will build a regression task, where on one hand we will have the features derived by the player performance information supervised to 'learn' the TMVE. In order to maximize overlap with the set of active players we build the features using the data from the season 2014/15 (using all the games played between 1/7/2014 and 1/7/2015). The TMVE used as a supervisory signal is from the summer of 2015[2]. In the following subsection we will detail on how are the features used for regression extracted.

A. Feature extraction

In order to pose the regression task described above we will generate a feature vector for each player with following components (features):

Performance features. The most relevant group of features are performance-related features. For each player we go through the list of games he played in, and extract the features listed in Table II. It basically contains absolute number (aggregated over the monitoring period, in our case 2014/15 season) of various events players participated into 9 important event groups: assists, tackles, passes, challenges, key passes, shots, dribbles, air challenges and crosses. Then the number and the accuracy (defined as the ratio between

[1] We use the standard metric for string similarity defined as $2.0 * M/T$, where T is the total number of elements in both strings, and M is the number of matches. Note that this is 1.0 if the strings are identical, and 0.0 if they have nothing in common.

[2] Normally, players get their transfermarkt.com market value estimate twice per year, once in summer and once in winter.

Performance features

# games
minutes played per game
# goals (total)
# assists
# passes
# accurate passes
# challenges
# accurate challenges
# air challenges
# accurate air challenges
# key passes
# accurate key passes
# shots
# accurate shots
# dribbles
# accurate dribbles
# tackles
# accurate tackles
# crosses
# accurate crosses
assists accuracy (%)
passes accuracy (%)
challenges accuracy (%)
air challenges accuracy (%)
key passes accuracy (%)
shots accuracy (%)
dribbles accuracy (%)
tackles accuracy (%)
crosses accuracy (%)
goals per minute
assists per minute
passes per minute
challenges per minute
air challenges per minute
key passes per minute
shots per minute
dribbles per minute
tackles per minute
crosses per minute

Player info features

age
position
height

Team info features

average co-players TMVE
average opponent TMVE
average points per game [3]

Table II
EXTRACTED FEATURES.

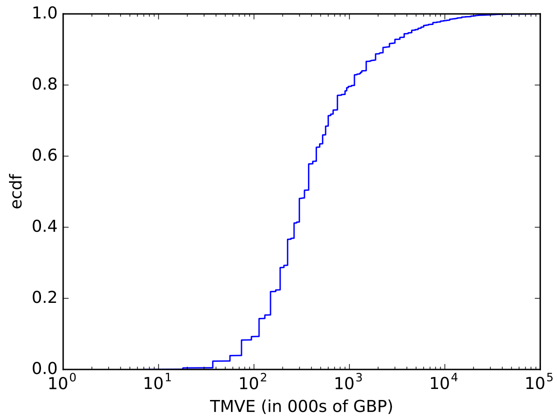

Figure 1. ECDF of TMVE. Transfermarkt.com market values estimates of professional football players span several orders of magnitude. Note log-normal distribution.

the # of accurate outcomes and the # of all trials) of successful events within each of those 9 categories. Finally we also use the frequency of the events measured in # events per minute played to measure how involved is the player within each of these 9 key categories of play.

In addition to the performance features we also extract several features related to the player himself (age, position in the team and height), and some features which capture the team he plays for: average co-player TMVE (average TMVE of all the players from the same team) represents a proxy of the strength of the team he plays for; average opponent TMVE represents a proxy of the strength of the opponents and average points per game measures how successful the team is. For TMVE calculation here we use the summer 2015 TMVE values. Note that team/opponent TMVE averages fluctuate very little in time and are primarily the proxy of the strength of the team and the league.

Apart from *player position*, all other features are numerical, suitable for a range of machine learning tools. The categorical variable *player position* we convert to a list of numerical features using the standard binary one-hot encoding: for each possible player position we have a binary variable representing that position.

B. Model building

As we mention above we use supervised learning for building player market value model. To do so, each user is represented by the described feature vector, extracted using the data from 1/7/2014 to 1/7/2015. Our supervisory variable is the TMVE of the player in the summer of 2015. Our dataset has 12858 players which had played at least one InStat-recorded game during that 2014/15 season and could be matched against the transfermarkt data. We split the data in 4 equally sized folds and use 3 folds for training and the fourth for testing. Thus, for all players within each fold we use the model trained on the other 3 folds to derive what we call performance-driven market value estimate (PDMVE). We experimented with several ensemble supervised learning methods including random forests, gradient boosting trees regression (GBT) as well as generalized linear models. Among them GBT had slightly lower errors than the others and hence we report the results obtained using them. For hyper-parameter selection we used a small validation dataset (10% of the training data) and used grid-search to select the appropriate hyper-parameters.

ECDF of relative difference between TMVE and PDMVE between all the players in our dataset is depicted in Figure 2.

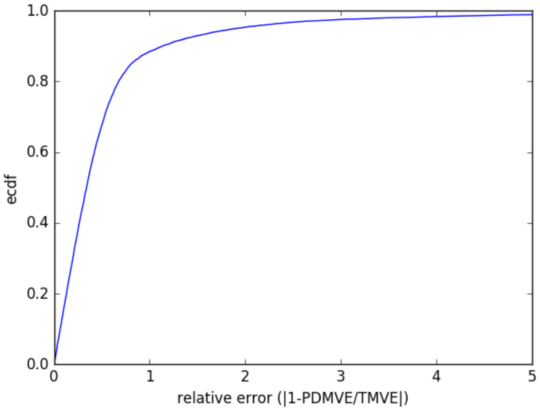

Figure 2. ECDF of relative errors between TMVE and PDMVE. Median = 33.9%, mean = 59.6%.

Player name	TMVE (M £)	PDMVE (M £)
neymar	60.0	69.55
eden hazard	52.5	57.72
cesc fabregas	37.5	49.54
sergio aguero	45	48.81
lionel messi	90	48.53
nolito	7.5	46.14
luis suarez	60.0	40.65
thomas muller	41.25	38.08
marco verratti	30.0	36.07
diego costa	37.5	35.79

Table III

Top 10 players according to PDMVE. TMVE and PDMVE are in Millions of GBP (£).

The median difference between the TMVE and PDMVE is around 34%, while the mean difference is around 60%. We would like to make several remarks here. First, as we can see from Figure 1 the TMVE spans several orders of magnitude and coming up with a single model across such a long range of market values is very challenging; especially given different styles of play in different leagues/contintent. Building more specialized models which focus on players from similar leagues/TMVEs could possibly lead to a better match between TMVE and PDMVE. Second, there are several factors regarding the market-value which our performance-driven approach omits such as commercialization capacity of injury-proneness. Finally, looking at the performance over longer time-scales (longer than one season we use here) is likely to reduce the difference between the TMVE and PDMVE. However, the median error of 34% in predicting the value which lies in the domain spanning 4 orders of magnitude is a rather satisfactory result, especially given the subjective element of TMVE which can hardly be numerically modelled.

In the Table III we report the top-10 players according to PDMVE which includes many high-profile players but also a non-prolific (at the time: end of 2014/15 season) Nolito from a mediocre La Liga team of Celta Vigo. Meanwhile his career surged to the level where is a starter in the Spanish national team and most recently signed for Manchester City F.C. powerhouse. Nolito also tops the list of most undervalued players according to the difference between the PDMVE and TMVE, see Table IV. In the list there are several non-established players (as of summer 2015), Nolito, Bellarabi, San Jose and Gaspar whose TMVE during 2015/16 season approximately doubled for *each* one of them representing market correction, which our data-driven approach picked up by looking only at the data from 2014/15 season. The list contains also several established players who had a very successful 2014/15 season (Fabregas, Willian, Mertens and Pelle) as well as two veterans of the game Alves and Robben who in spite of having high-performance indicators are valued low due to their age. We would like to stress here that even though our model uses age as one of the features we are not satisfied with the ability to infer the decay of market value which inevitably comes with aging players. We discuss this and several other directions for improving the model in Section V.

Player name	TMVE (M £)	PDMVE (M £)
nolito	7.5	46.14
dani alves	7.5	25.39
karim bellarabi	9.0	23.39
dries mertens	13.5	26.73
cesc fabregas	37.5	49.54
willian	22.5	34.32
graziano pelle	8.25	19.90
arjen robben	21	32.00
mikel san jose	3.75	14.56
mario gaspar	4.5	14.49

Table IV

Top 10 most undervalued players according to the difference between PDMVE and TMVE.

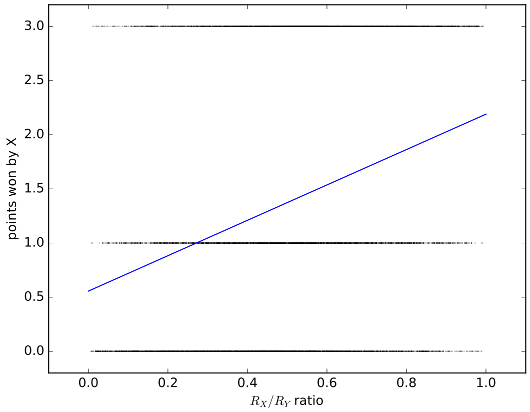

Figure 3. Points won by team X as a function of R_X/R_Y ratio; linear regression line. Ratings R = team-TMVE were used here, for the games in season 2014/15.

IV. Market value as performance predictor

In single-player sports such as tennis, snooker or chess, there are several methods for measuring the quality of value of the player such as ELO rating [4] or Glicko rating [5]. Normally they strive to not only rank players but also provide likelihood of the win/loss for any game played between the two players. However, when it comes to players of team sports it is much more challenging to evaluate the value or rating of the player. Moreover, comparing two different player ratings (e.g. TMVE and PDMVE) is very challenging since no two players compete face-to-face solo, but always as a part of the team. Note that on a team level applying Elo or Glicko rating is rather straightforward (see [6]) yet such ratings say little about the ratings of individual players.

In this section we use the following approach to evaluate the two player ratings: TMVE and PDMVE. Since the games are played and scores are reported on a team level, for each team we consider two metrics: team-TMVE and team-PDMVE, defined as sums across all the players of the team of TMVE and PDMVE, respectively. We ask the following question: **which one of the two metrics is a better predictor of the game outcome?**

If the team X with ratings (e.g. team-TMVE or team-PDMVE) R_X plays team Y with rating R_Y, what is the expected number of points[4] s team X will win? We can pose this question as a linear regression problem: $s = \alpha + \beta \cdot R_X/R_Y$. For example, using our InStat dataset for all the games in 2014/15 season the linear regression parameters for team-TMVE ratings are $\alpha = 0.55$ and $\beta = 1.63$; see Figure 3.

To slightly improve the regression we add an extra binary feature h which encodes whether the team X plays at home $h = 1$ or away $h = 0$. Thus for the two different ratings R (team-TMVE and team-PDMVE) we define the regression problem:

$$s = \alpha_0 + \alpha_1 \cdot R_X/R_Y + \alpha_2 \cdot h. \qquad (1)$$

Parameters $\alpha_0, \alpha_1, \alpha_2$ were in both cases derived using the data from 2014/15 season as a training set.

Evaluating error of the prediction for individual games brings out very large errors which makes it hard to compare different ratings. Hence we look at the teams which played at least 20 games in the first half of 2015/16 games (our dataset has only games played before January 2016) and for each team we observe the actual number of points won per game and the expected number of points per game according to the regression model (1). The median error, root mean square error (RMSE) between the actual and expected points per game are reported in Table V together with Pearson correlation index, for both predictions made by team-TMVE as well as team-PDMVE rating. Across all three metrics predictions made by team-PDMVE performed 4-7% better: smaller median errors and RMSE and larger Pearson correlation.

ratings	Pearson corr.	median error	RMSE
team-TMVE	0.631	0.117	0.200
team-PDMVE	0.669	0.109	0.192

Table V
Strength of team-TMVE and team-PDMVE rating in predicting the average points per game. Team-PDMVE has 4-7% better strength across the three metrics.

The improvements in prediction power of PDMVE compared to TMVE are not very large, but they nevertheless demonstrate that even very simple data-driven models can have similar if not better prediction power compared to the ones derived by the human experts. An important point we would like to make here is that our analysis could allow clubs to identify players whose performance (PDMVE) outweighes their market value (TMVE) and sign them without burning budget constraints. Since our analysis is not conclusive regarding whether PDMVE is stronger or weaker predictor compared to TMVE we plan to extend the analysis presented above to capture more generic predictive models (rather than the linear regression used here) or compare the team performance estimators using the league simulators [10].

V. Discussion

The model we described above is rather simple and leaves a significant room for improvement. Here we will hint several directions we plan to pursue in the near future which could improve the overall quality of the player value estimation.

Longer data. The performance features used in our model are extracted using only one season. We believe that profiling players over a longer timespan can lead to better performance model.

Deeper data. The performance data we study here is relatively simple and hence collected at scale of thousands of players simultaneously. However, new automatic technologies emerge [9] for tracking players to much deeper detail which would also allow more accurate profiling of the player performance.

External factors. Our model currently does not take into account external factors such as injuries or conflicts with manager which can affect the valuation of the player by, for example, low number of played games. Proneness to injuries or bad-tampered character are also important characteristics which affect the player value which our model currently does not utilize.

Team strategy. The performance of a player (say number of tackles, or passes per game) largely depends on the team strategy set by the manager. Incorporating team strategy into individual player performance model is highly nontrivial. The difficulties of assessing the individual players in team sports have been recognized in the literature [7].

[4] Note that in soccer team wins 3, 1 or 0 points if it wins, draws or loses the game.

Specialized models. One of the goals of this work is building of generic model which would work across different leagues, team styles and player positions. However specialized models which could be trained on smaller corpus of players (e.g. Forwards, or Top-European division players) may offer more insights and stronger learning capabilities.

VI. Conclusion

In this paper we developed a simple yet general model for assessing the player market value using the performance data. While our model uses *transfermarkt.com* market value estimates in the training phase it only uses the performance data to evaluate the value of the player virtually eliminating the noise of the subjective biases which are common in the traditional market value estimates. We show that our value estimates are a stronger predictor (albeit the improvement is relatively small) for the team performance than estimates from *transfermarkt.com* allowing team managers to better manage their budget by looking for and signing undervalued players. Finally, the simplicity of our model allows for significant improvements which we plan to pursue in the future.

References

[1] *transfermarkt.com* online service.

[2] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., Tewari, A. (2013). Learning with noisy labels. In Advances in neural information processing systems.

[3] Ghosh, A., Manwani, N., Sastry, P. S. (2015). Making risk minimization tolerant to label noise. Neurocomputing, vol. 160.

[4] Elo, A. (1978). The Rating of Chessplayers, Past and Present. Arco. ISBN 0-668-04721-6.

[5] Glickman, M. E. The Glicko system. *http://www.glicko.net/glicko/glicko.pdf*

[6] Hvattum, L. M., Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. International Journal of forecasting, 26(3), 460-470.

[7] Gerrard, B. (2007). Is the Moneyball approach transferable to complex invasion team sports?. International Journal of Sport Finance, 2(4), 214.

[8] He Y. Predicting Market Value of Soccer Players Using Linear Modeling Techniques. Technical Report. University of California, Berkeley. 2015

[9] Gyarmati, L., Hefeeda, M. (2015). Estimating the maximal speed of soccer players on scale. In Proc. Machine Learning and Data Mining for Sports Analytics Workshop.

[10] Cintia, P. , et al. ”The harsh rule of the goals: data-driven performance indicators for football teams.” Data Science and Advanced Analytics (DSAA), 2015. IEEE International Conference on. IEEE 2015.