Soccer Transfer Efficiency Modeling

Khaled Jedoui

Stanford University

Introduction

The transfer window is the period during the year in which a football club can transfer players from other playing staff into their playing staff. Such a transfer is completed by registering the player into the new club through FIFA. This period of time represents one of the most important periods for every club, as it is the opportunity to improve their weaknesses and consolidate their strengths by deciding to sign more players or release some others.

In order to be successful in the transfer window, professional soccer teams hire a staff of professional recruiters who would work with the head managers in order to decide on which players to sign based on some budget. All recruiters mostly base their assumptions on the players' past performances with their previous teams. However, it is important to note that even though a player seems good in paper, it is hard to predict his contribution to his new team.

In this paper, we study the possibility of designing a simulation engine that is able to give us the added value of the teams' transfer activity. Such a tool, if it proves to be robust and reliable, can be very interesting in getting a rough idea of how a player can help a team achieve their goal. All code can be accessed through here.

Related Work:

In the last few years computer science and big data have played an increasingly important role in sports: players are tracked using cameras or sensors, detailed video analysis is employed, game theory is used to model in-game events, and large data bases of results and match data have become available.

Preprint submitted to Math 110: Applied Number And Field Theory *June 1, 2019*

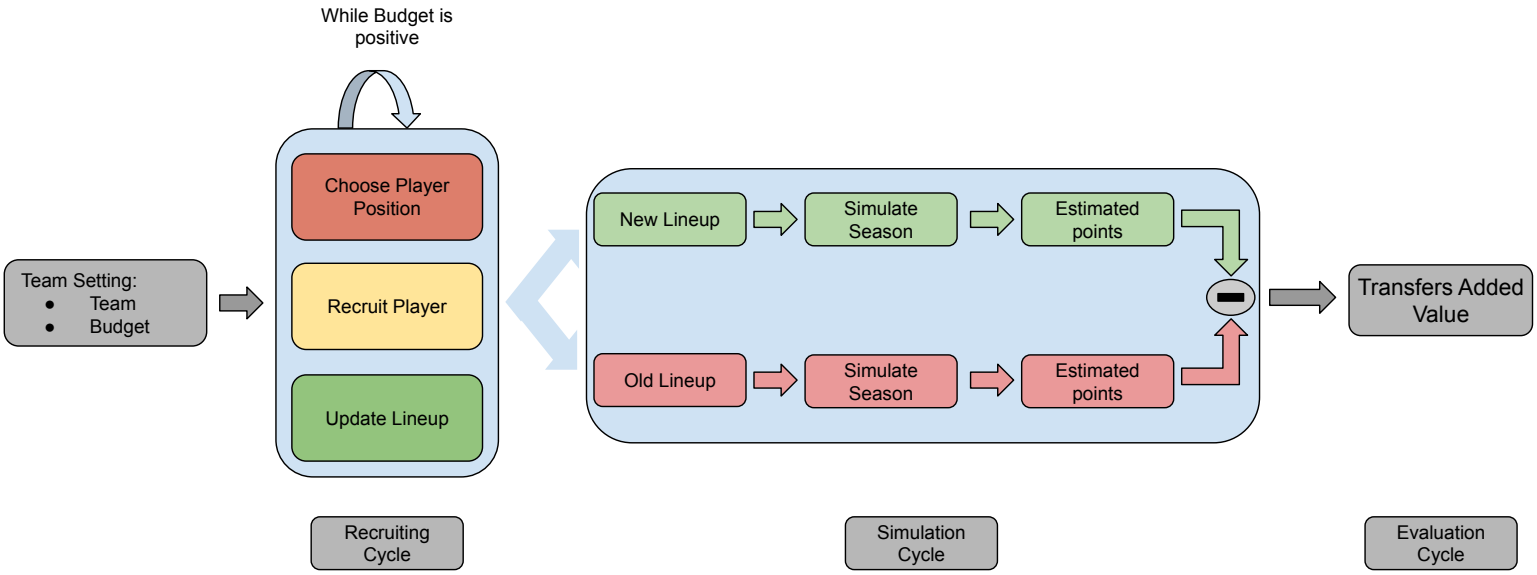

Figure 1: The Transfer Efficiency Simulation Engine architecture.

This gave engineers and sports scientists the opportunity to exploit the collected data to gain further insight into the mechanics of the game, as well as to analyze information efficiency in related markets[1]. This led to several published methods for rating and ranking teams in soccer [2, 1, 3]. Other player rating methods based on a regression model capturing the performance of players relative to their team mates and the opposition[4].

The interest of using data in the transfer market started with a financial goal. Teams were interested in maximizing their profit and to help assess their investments, it helped to know the market value of the players and to identify inefficiencies in the pricing of players. Several past studies have used ordinary least squares regression to model transfer fees based on a range of independent variables[5], some of which describe the talent of the players, while others describe external factors related to the buying or selling club.

In this work, we focus on the performance side of transfers. Instead of prioritizing the team's financial gain, we develop a system that gives the coaching staff an idea of future team performance depending on their recruitment strategy.

Transfer Efficiency Engine Design

Our system architecture is represented in Figure 1. We design a pipeline that is composed of 4 parts:

Team Settings:

In this section, the user (recruiter/coach) chooses their respective team and decides on the total budget for a specific transfer window.

Recruiting Cycle:

The system contains a database of players with their favorite playing position, market value and their FIFA Index attributes. The FIFA index attributes are a set of scores ranging from 0 to 100 given to every player defining their skills offensively and defensively. In total, we have 7 categories (Ball Skills, Defence, Mental, Passing, Physical, Shooting and Goalkeeper) for a total of 37 skills.

Once the team and budget are chosen, the system loads the players dataset and the user starts recruiting players. While the budget is positive and the user still needs players, the recruiting cycle goes as follows:

1. Choose Player's Position of interest.
2. The System, at this point, displays the set of players the user is able to recruit with the remaining budget. The players are sorted by their overall rating.
3. The user recruits a player.
4. The team lineup and budget are both updated.

Simulation Cycle:

When the user is done recruiting, we get both the old and the resulting team and simulate a season for both line-ups in their corresponding league. From each simulation, we get the resulting number of points.

Evaluation Cycle:

We define the added value of a transfer window, E_T, as the change in the total number of points at the end of the season. It is calculated as follows:

$$E_T = P_{new} - P_{old}$$

where P_{old} is the total number of points if we continue with the same team and P_{new} is the total number of points with the new line-up. When $E_T \leq 0$, we assume that the recruitment is not very efficient. Otherwise, the recruitment is efficient. Of course, the higher the E_T, the better the transfer strategy.

Prediction Model Description

0.1. Problem Formulation:

The soccer game prediction task is a 3-class classification problem, where the input is the two teams line-ups matrix X and the output is a label $y \in \{0, 1, 2\}$ indicating the home team win, away team win and draw respectively. We train our model using the cross entropy loss:

$$L(X, y) = -\sum_{c=1}^{3} y_{o,c} log(p_{o,c}),$$

where $y_{o,c}$ is the correct classification for observation o and $p_{o,c}$ is the predicted probability of the observation o is of class c.

0.2. Model Architecture and Training

We train a number of models on our dataset. The model that achieves the best validation results is a 4-blocks convolutional neural network[6], with batchnorm[7] and dropout[8]. The network is trained end-to-end using Adam[9] optimizer. We train the model using minibatches of size 64. We use an initial learning rate of 0.001, and pick the model with the best validation performance.

After fine-tuning our model, we achieve a train accuracy of around 70% and a validation accuracy of a round 50%.

0.3. Model Averaging: Using Dropout MC in Season Simulation:

In a situation where really strong team plays a weaker side, an expert can be confident about which side will win. Yet, soccer is a sport that is very unpredictable and upsets happen regularly. Factors such as morale of a team (or a player), skills, coaching strategy, equipment or luck have a impact in the results for a sport match. Therefore, when simulating a season or a game, it is important to take into account what can happen.

We do this by enforcing dropout in our model during inference and simulating a season n times. We get our league ranking by averaging the number of points collected over the n simulations. Intuitively, we can think of this as perturbing the model n times to leave room for some luck in guessing the outcome of games. This result has been presented in the literature before as model averaging.

Data Description

The transfer efficiency engine requires 2 datasets:

Game Prediction Dataset: For each game i, we build a matrix $X_i \in R^{n \times m}$, where $\frac{n}{2}$ represents the number of rows for each team. Each row in X represents a specific player at a specific position and each column gives the player's FIFA Index attributes ($m = 37$ attributes in total).
We let our dataset D be as follows:

$$D = \{(X_1, y_1), ..., (X_k, y_k)\}$$

where y_n is the game result and k is the total number of games in our dataset. For the sake of simplifying our problem, we only take into consideration the starting line-ups, so $n = 22$, with the first 11 rows representing the home team and the next 11 representing the away team.

For the game results and starting players, we use the European Soccer Data (Click here) from kaggle which covers games from 2008-2009 to 2015-2016. We collect all division I games from England, Spain, Italy and France. In total, we have around 11280 games with 5231 home team wins, 3124 away team wins and 2925 draws. We split our data to 90%/10% for the train and validation sets.

We get the player attributes for each season by scraping the website https://www.fifaindex.com/players/ which contains players' attribute data of all FIFA recognized leagues. We combine these 2 datasets together in order to build our D.

Simulation Dataset: We follow the same format as the game prediction dataset. We collect the 2018-2019 Premier League season starting line-ups, as well as the FIFA Index attributes for all active players at the start of the 2018-2019 season. In total, we have 380 games.

Experiments and Analysis:

Season Simulation: We use our best model in order to simulate the 2018-2019 Premier League season. As discussed previously, we simulate the season $n = 500$ times with dropout. We get the following ranking:

Rank	Team	Number of Points	W	L	D
1	Manchester City	99	31	1	6
2	Tottenham Hotspur	93	28	1	9
3	Chelsea	92	29	4	5
4	Everton	89	28	5	5
5	Liverpool	81	23	3	12
6	Manchester United	74	21	6	11
7	Arsenal	73	22	9	7
8	Leicester City	51	15	17	6
9	Bournemouth	51	16	19	3
10	Wolverhampton Wanderers	45	12	17	9
11	West Ham United	43	12	19	7
12	Southampton	42	11	18	9
13	Brighton & Hove Albion	40	11	20	7
14	Watford	40	11	20	7
15	Burnley	35	10	23	5
16	Fulham	32	7	20	11
17	Newcastle United	32	9	24	5
18	Huddersfield Town	30	9	26	3
19	Crystal Palace	17	3	27	8
20	Cardiff City	12	3	32	3

It is interesting to note that the model correctly predicts the 2018-2019 season winner correctly. We also correctly predict 2/3 relegated teams and 3/4 of the teams qualified to the champions league (1^{st} four) and 1/2 of the teams qualified to the Europa League. This shows that our model is reliable and can be used in modeling transfer efficiency.

Transfer Engine Simulation:

In this section, we answer a few interesting questions using our transfer efficiency engine. In all studies, we assume a classic 4-3-3 formation for the sake of simplicity.

Study 1: Players with the highest average added value

In this experiment, we get the list of the 50 Top rated players. For each team in the premier league, we add one of the selected top players and simulate a season $n = 20$ times (we choose 20 as 500 is very computationally

demanding and slow). We then get the average added value of each players. The following table shows the best 5 players.

Rank	Name	Added Value
1	Neymar Jr.	+2.39
2	L. Insigne	+1.65
3	Coutinho	+1.29
4	T. Kroos	+0.98
5	M. Ter Stegen	+0.94

Study 2: Highest contribution by player position

We get the list of the top 5 rated players per position and redo the same experiment from the previous section. We show the best and worst player in terms of contribution.

Poistion	Best	Worst
GoalKeeper	M. Ter Stegen (+0.94)	D. De Gea (-0.28)
Center Back	D. Godn (+0.80)	L. Bonucci (-0.67)
Lateral back	K. Koulibaly (+0.86)	F. Lus (-1.20)
Defensive Midfield	N. Kant, (+0.10)	Fabinho (-0.63)
Center Midfield	N. Keta (+0.12)	S. Milinkovic-Savic (-0.69)
Lateral Midfield	S. Man (+0.67)	K. De Bruyne (-1.43)
Attacking Midfield	Coutinho (+1.28)	D. Valeri (-1.07)
Winger	Neymar Jr. (+2.39)	A. Di Mara (-0.41)
Center Forward	S. Agero (+0.83)	L. Surez (-1.04)

It is interesting to note that some players with the highest ratings are not the ones always with the highest contribution. Also, it is important to notice that even though some players are incredibly talented and would normally have a great impact on a team in a real life setting, they prove to be inefficient according to the simulation. This might be explained by the 4-3-3 formation we use that might be not very compatible with their style of play. Indeed, Leonardo Bonucci, for example, was most successful in a 3-5-2 setting alongside Georgio Chiellini and Andrea Barzagli (even though he played in 4-3-3 this season in Juventus F.C).

Study 3: Impact of player absence: Eden Hazard and Paul Pogba

The 2019-2020 season summer transfer window opened a few days ago and mega transfer rumours have been the interest of the media. Probably,

the most awaited transfers are the one of Eden Hazard and Paul Pogba to Real Madrid. Even though the latter still has not confirmed he wanted to leave Manchester United, Hazard has been very clear about his willingness to leave Chelsea for Real Madrid. In the last few years, Eden Hazard proved to be one of the best players in the world and his presence in the Chelsea line-up has been primordial in helping his team win titles. For this study, instead of looking at a player's contribution to a team, we look at what would happen if the team replaces him with an average player when he leaves or gets injured. We define an average player as the average of all player attributes from FIFA Index. We study the impact of Hazard and Pogba leaving Chelsea and Man U, respectively. With this setting, we simulate our model $n = 500$ times. We get the following results.

Name	Absence Impact
E. Hazard	-4.00
P. Pogba	-7.29

Notice that the exit/unavailability of Hazard and Pogba has a great impact on the performance of both Chelsea and Man U. This proves that both teams need high quality replacements in order to ensure staying competitive at the top stages.
Compared to study 2, study 3 has results that are more intuitive.

Limitations:

Even though this system shows good promise, it is not reliable enough to be used in a professional setting. As discussed in our experiments, the model only takes into consideration one team formation (4-3-3) . Besides, the game prediction model is trained on unrealistic/incomplete data. Indeed, our data represents only 11 players per-team per-game, so no substitutions and no tactics involved. Furthermore, even though dropout MC simulation does a great job in making our predictions more natural by adding luck to the equation, it is still complex to model human factors in the game.

This shows that the model still has a lot of problems that can easily be fixed and others that are more complex. However, we can think of this system as a successful proof of concept showing the importance of such model for recruiters/coaches.

Conclusion:

In this paper, we provide a simplified transfer efficiency evaluation system based on season simulation. We show that such a model has a great potential in giving recruiters or coaching staff important insight on the impact of potential incoming or leaving players.

[1] L. M. Hvattum, et al., Analyzing information efficiency in the betting market for association football league winners, The Journal of Prediction Markets 7 (2013) 55–70.

[2] A. C. Constantinou, N. E. Fenton, Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries, Journal of Quantitative Analysis in Sports 9 (2013) 37–50.

[3] J. Lasek, Z. Szlávik, S. Bhulai, The predictive power of ranking systems in association football, International Journal of Applied Pattern Recognition 1 (2013) 27–46.

[4] O. D. Sæbø, L. M. Hvattum, Evaluating the efficiency of the association football transfer market using regression based player ratings., in: NIK.

[5] B. Frick, The football players'labor market: Empirical evidence from the major european leagues, Scottish Journal of Political Economy 54 (2007) 422–446.

[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[7] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (2014) 1929–1958.

[9] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).