European Journal of Operational Research 306 (2023) 389–399

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Estimating transfer fees of professional footballers using advanced performance metrics and machine learning

Ian G. McHale[a,*], Benjamin Holmes[a,b]

[a] Centre for Sports Business, University of Liverpool Management School, L69 7ZH, U.K.
[b] Department of Mathematics, University of Liverpool, L69 7SH, U.K.

ARTICLE INFO

Article history:
Received 16 June 2021
Accepted 15 June 2022
Available online 20 June 2022

Keywords:
OR in sports
Analytics
Machine learning
Moneyball
xgboost

ABSTRACT

The paper presents a model for estimating the transfer fees of professional footballers. We seek to improve on the literature in two dimensions. First, we utilise advanced player performance metrics to better capture the playing ability of footballers. Second, we adopt machine learning algorithms to improve out-of-sample prediction accuracy. The model proves to be a considerable improvement on linear regression, and the advanced performance metrics further improve the predictions. We use the model to identify value-for-money transfers, before assessing the past records of clubs in identifying value-for-money and find that, Liverpool and Atlético Madrid, for example, are successful at identifying value-for-money, whilst Manchester United and Barcelona are not.

1. Introduction

Over the last century, the money exchanging hands between clubs for the services of professional football players has grown enormously. The current transfer fee record is £199.8m, set in 2017 when Paris Saint-Germain paid Barcelona for the services of Brazilian player Neymar. Figures in excess of £10m are entirely commonplace. FIFA estimated that the global transfer market in 2020, during an unprecedented pandemic, saw 17,077 players move for fees totalling £4.28bn (FIFA, 2021).[1]

The transfer market on football players has undoubtedly become big business, and with such large sums of money being spent on a relatively small number of assets comes high risk. Mistakes are commonplace and clubs often find themselves in the predicament of having paid tens of millions of euros for a player who fails to live up to expectations. Such mistakes can have catastrophic consequences for the club and its fans including relegation and bankruptcy.

The problem of understanding and estimating transfer fees is made attractive to researchers because professional sport offers a laboratory like no other: there is no other industry where we know the identity, the physical characteristics, and the production of every worker in the labour market (Kahn, 2000). Further, the status and popularity that sport, and specifically football, attains in modern society is arguably like no other activity. Indeed, millions of fans, pundits, coaches, and scouts observe and judge each player, and each recruitment decision made by a club. Because football offers this ideal research environment, and because of the popularity of sport within society, and the large sums of money involved, the problem of estimating transfer fees has drawn the attention of economists, statisticians and operational researchers alike.

Football is not the only sport in which the labour market for players has attracted the attention of the scientific community. Indeed, one could argue that the creation of, or at least the new found prominence of, the field of sports analytics came about as a direct result of the Moneyball story in which the Oakland Athletics identified and capitalised on an inefficiency in the labour market for baseball players (Lewis, 2003). Hakes & Sauer (2006) provide an economic analysis of the Moneyball story, and find that in essence, Moneyball is a story about identifying value-for-money in the baseball labour market. Hakes & Sauer (2006) go on to show that the inefficiency was short-lived and the employment market returned to efficiency.

Assuming the market for football players is approximately efficient, in that on average, the fees paid for players do indeed reflect their value to buying clubs, the task of identifying value-for-money in the football labour market involves two important elements: (i) the search for metrics that capture playing performance and ability, and (ii) modelling transfer fees (or values) to understand what

* Corresponding author.
E-mail address: ian.mchale@liverpool.ac.uk (I.G. McHale).

[1] In 2019 the figures were £5.66bn for around 18,047 players.

https://doi.org/10.1016/j.ejor.2022.06.033

factors drive the price a club is willing to pay for a player.[2] Surprisingly, until now these two strands of literature have rarely, if ever, met. Papers estimating transfer fees have used basic performance statistics such as goals scored and minutes played (see, for example, Coates and Parshakov (2021) and Muller, Simons, & Weinmann (2017)), and/or crowd-sourced measures of playing ability, such as player ratings from websites or video games (for example, Yigit et al. (2020)). On the other hand, advanced metrics are presented in various papers and stop short of examining how the metrics explain transfer fees. It is this gap which we fill in this paper.

The paper is organised as follows. In the next section we present a literature review on estimating transfer fees. In Section 3 we describe the objective metrics we use to measure player ability and quality, together with the crowd-sourced player ratings from the sofifa.com website. In Section 4 we describe our dataset and Section 5 presents regression and machine learning models to estimate transfer fees and compare the predictive performance of our model with that of the transfermarkt.com website valuations, before using the model to identify historically good value-for-money players in Section 6. We close with some conclusions in Section 7.

2. The literature on estimating transfer fees

Valuing workers has long been a topic of interest in the literature on labour economics and the earliest work on explaining transfer fees in football originated here. Beginning with Carmichael & Thomas (1993), economists have used regression models to identify determinants of transfer fees. Since then, several papers follow in a similar vein. Linear regression is used to model the log of transfer fee as a function of a collection of player characteristics (e.g. age, height, position played), playing experience (e.g. minutes played, international appearances), performance metrics (e.g. goals scored), and level of the clubs involved in the deal (e.g. division played). For example, Dobson & Gerrard (1999) and Dobson, Gerrard, & Howe (2000) model transfer fees of professional and semi-professional footballers in England. More recently, Depken II & Globan (2021) use linear regression to identify that English clubs pay a premium in the transfer market for players relative to clubs from other European countries. Further, they find that the television rights contract that the English Premier League signed in 2012 had a causal effect on the premium that English clubs were found to be paying.

Frick (2007) provides a review of the early literature. The papers differ little in technique, with each focussing on different time frames, different leagues and marginally different sets of covariates. The types of variables used in the literature fall into five categories: (i) player characteristics (age, position played), (ii) experience (minutes played, international appearances), (iii) contract status, (iv) performance and ability, (v) popularity (e.g. number of Twitter followers).

Since the onrush of these economics papers, the topic has increasingly gained interest in the operational research literature and more recently, in the field of machine learning. An important distinction between papers in the literature is in the specifics of the data employed in studies. The data used for modelling varies in two dimensions. First, some authors choose crowd-sourced valuations of players as the dependent variable to be explained or predicted, not actual transfer fees. Second, some authors use subjective player ratings, not objective performance metrics, as the predictor variables.

[2] A third element may be in understanding the needs of the buying and selling clubs. For example, a club with a surplus of goalkeepers, may be willing to sell at a discount, whilst a club in desperate need of a central defender may be forced to pay a premium.

Crowd-sourced player valuations have most commonly been taken from the transfermarkt.com website (TM). Members of the website offer their valuations of players and a panel of experts calculates a weighted average of the values to arrive at a single transfer value for each player. The panel of experts calculate the weights based on judging how accurately each member has valued players historically. Although the TM values have proved to be highly correlated with transfer fees (see, for example, Herm, Callsen-Bracker, & Kreis 2014), the two quantities are, of course, different. One is a subjective assessment of the valuation of the player made by a large crowd, whilst the other is an actual fee paid by one club to another for the services of a player. Indeed, Coates and Parshakov (2021) found that despite a high correlation with transfer fees, the TM valuations suffer from systematic bias in that adding simple descriptive statistics (such as goals scored) as a covariate in a model in which transfer fee is regressed on TM value improves the model fit. In this paper we model actual transfer fees, not crowd-sourced valuations.

Using TM values as a proxy for transfer fees may produce spurious results. First, as Coates & Parshakov (2021) found, there exists a bias. Second, TM values are calculated for all players, whilst only a relatively small proportion are actually transferred, resulting in a selection bias. Carmichael, Forrest, & Simmons (1999) acknowledge this selection bias and, using a Heckman two-step procedure to model the probability of being transferred, identify that selection of players being transferred is not random. Modelling transfer fees directly, as we do, removes the potential for biases to be induced from selection issues. Our model provides an estimate of a player's transfer fee, conditional on being transferred.

The second variation in modelling in the literature is the choice of predictor variables which measure the ability of the player. Some authors use subjective player ratings whilst others use objective performance metrics. Player ratings have most commonly been taken from the sofifa.com website. Members of the website provide ratings of players in many attributes (including passing, shooting, tackling etc), with editors reviewing these ratings before presenting single values for each attribute for each player. Using crowd-sourced player ratings has become a popular approach in the machine learning literature. For example, both Yigit, Samak, & Kaya (2020) and Behravan & Razavi (2021) model TM values using crowd-sourced sofifa player ratings.

The alternative to using crowd-sourced player ratings is to use objective performance metrics. Such metrics are typically rudimentary such as goals scored, minutes played, assists, successful dribbles, etc. (for example, Muller et al. (2017)), and are likely to fail to capture the subtleties of playing ability, and certainly struggle to capture how players in different positions contribute to a team's performances. However, it is interesting that findings from these papers suggest that the transfer market does indeed take note of such rudimentary measures of playing performance. This perhaps provides evidence of a somewhat naive labour market in which basic summary statistics prove to be strong determinants of transfer fees. It is reminiscent of the original Moneyball story in which hitting home runs (the most basic statistic in baseball) was an overvalued skill (Hakes & Sauer, 2006) relative to its contribution to team win percentage.

In this paper we use advanced performance metrics to capture the ability of players, before considering the additional contribution of the subjective, crowd-sourced player-ratings to the predictive accuracy of the model.

3. Models for rating players

The most critical component of identifying value-for-money in a transfer is accurately assessing the playing ability of a player. In team sports such as football, rating players is an extremely complex task. Players have different roles and objectives, and with 22 players moving around, interacting in continuous time, simple metrics fail to capture the true impact of a player on his team. Early papers examining determinants of transfer fees concentrated on using rudimentary metrics of performance such as goals scored. However, goals scored fails to capture the contribution of the multitude of other actions players are responsible for whilst on the pitch. As such, researchers have developed ratings systems for players. Two ratings systems are of particular interest here: the plus-minus ratings presented in Kharrat, McHale, & Pena (2020), and the action value ratings presented in Liu, Luo, Schulte, & Kharrat (2020). Both ratings attempt to capture the full contribution of a player to on-the-pitch performances. In addition to these two ratings systems, we describe the crowd-sourced ratings from the sofifa website since we experiment with including these in our models.

3.1. Plus-Minus ratings in football

Plus-minus ratings are a popular ratings system used in ice-hockey and basketball. The naive plus-minus rating answers the very simple question: “how does a team perform with a player compared to without that player?”. The measure of team performance in plus-minus ratings is known as the ‘target’. In basketball, the target is the points differential of the two teams; in ice-hockey and football, the target has tradionally been the goal difference. Rosenbaum (2004) modified the naive plus-minus rating to account for the strength of the other players on the pitch by setting up the estimation of plus-minus ratings as a regression model. In this framework, the target (goal difference per minute of play) is regressed on dummy variables representing the identity of players on the pitch. Each observation represents a period of play when the same set of players (team mates and opponents) are on the pitch. As such, a new observation occurs whenever a new match begins, a substitution is made, or a player is sent off.

To deal with collinearity issues arising from players always playing together (a problem in football, but not necessarily in basketball where players are frequently on and off the court), Macdonald (2012) used ridge regression to estimate the strengths of the players.

In football, Sæbø & Hvattum (2015) estimated plus-minus ratings and used the ratings as a covariate in a regression model to explain transfer fees. The ratings were statistically significant and proved to be a key predictor of transfer fees. However, the target, i.e. the unit of team performance, used in their plus-minus ratings was goal difference, but because football is a low scoring game, the resultant ratings are somewhat insensitive to team performance, and the vast majority of observations have a dependent variable (goal difference per minute) with a value of 0. As such, Kharrat et al. (2020) present new plus-minus ratings for football. The plus-minus rating we use here uses the difference in expected goals as the target[3], and the resultant plus-minus ratings are called xGPM and are intended to describe how a player affects the net chances a team creates.

A major advantage of using plus-minus ratings in this context is that, unlike goals scored for example, they are not position specific. Players can ‘earn’ plus-minus ratings no matter where they play on the pitch, be it striker, midfielder or defender.

[3] Expected goals, xG, is equal to the probability of a shot resulting in a goal. The main determinant of this probability is the location, relative to the goal, of the shot. For the segment of play in which the same set of players are on the pitch, the dependent variable is then the net total expected goals per minute of play.

Table 1
Top 10 players according to (a) the average 12-month GIM performance rating, and (b) the xGPM rating, and the date on which the rating occurred, and the player's position.

Player	Position	GIM	Date
(a) Top 10 players according to the average 12-month GIM performance rating			
Neymar	AMC	90.19	2018-08-12
Lionel Messi	FWR	86.98	2018-03-18
Hakim Ziyech	FWR	85.10	2019-04-30
Carlos Vela	FWR	84.08	2019-08-26
Angus Gunn	GK	84.02	2019-05-12
Lars Veldwijk	FW	83.60	2017-08-20
Tom Pettersson	DC	83.58	2017-10-19
Luuk de Jong	FW	83.32	2019-03-31
Thiago Alcántara	DMC	82.44	2017-09-19
Zlatan Ibrahimovic	FW	81.97	2016-08-27
Player	**Position**	**xGPM**	**Date**
(b) Top 10 players according to the xGPM performance rating			
Marc-André ter Stegen	GK	71.04	2018-03-01
Domenico Criscito	DC	68.92	2016-10-20
Luis Suárez	FW	68.09	2017-12-23
Neymar	AMC	67.64	2017-09-27
André Onana	GK	67.62	2017-09-09
Joshua Kimmich	DR	67.54	2019-09-28
Robert Lewandowski	FW	67.25	2019-09-28
Nick Viergever	DC	67.10	2017-04-08
Lionel Messi	FWR	66.84	2018-03-01
Joël Veltman	DR	66.80	2017-11-05

3.2. Deep reinforcement learning action value ratings

Cervone, D'Amour, Bornn, & Goldsberry (2016) presented the intuitively appealing concept of measuring the expected value of possession in basketball. Using a stochastic process model to estimate the probability of a possession (a phase of play) ending in points before and after a player's action, the authors calculate the worth of the player's contribution as the change in the expected points coming from that possession.

Following Cervone et al. (2016), others have developed similarly themed ideas for use in rating football players. Decroos, Bransen, Van Haaren, & Davis (2019) presented ratings based on whether a possession ends in a goal in the near future. Liu et al. (2020) use deep reinforcement learning to ‘learn’ the values of player actions in terms of their contributions to changing (increasing or decreasing) the probability of an episode of play ending in a goal. The authors call the resulting ratings a Goal Impact Metric (GIM) and provide evidence that the GIM performance rating is better at predicting future values of basic metrics of player performance (such as goals scored and key passes) than other similar ratings systems. It is these GIM performance ratings we use as a second family of objective player ratings in our models for determining transfer fees.

As for the plus-minus ratings, a major advantage of the GIM performance ratings is that they are not position specific and players earn ratings for defensive actions as much as offensive actions.

Table 1 shows the top ten values achieved by players in our dataset, and the date on which the rating was achieved, according to the GIM and xGPM performance ratings. The lists are populated by some very famous players such as Lionel Messi, Robert Lewandowski, Neymar and Luis Suarez. It is noteworthy that under both ratings systems, players from a variety of positions are present.

3.3. Crowd-Sourced sofifa.com player ratings

In addition to objective ratings, we also obtained crowd-sourced ratings from the sofifa.com website. The sofifa ratings provide a popular set of covariates for use in transfer fee modelling. For example, Kirschstein & Liebscher (2019), Behravan & Razavi (2021),

Table 2
Summary of the sofifa.com overall and potential ratings, as recorded in September 2020, for four positional lines.

		Overall			*Potential*		
Position	N	Mean	Std. Dev.	Max	Mean	Std. Dev.	Max
Goalkeeper	2,458	63.90	7.65	91	69.73	6.27	93
Defender	7,205	65.45	6.72	90	70.85	5.80	93
Midfielder	8,466	65.51	7.06	91	71.56	6.11	92
Forward	4,478	65.52	7.21	94	71.58	6.21	95

Table 3
Top 10 players according to the sofifa overall potential ratings, as recorded in September 2020.

Name	Overall	Potential
Lionel Messi	94	94
Cristiano Ronaldo	93	93
Neymar	92	92
Jan Oblak	91	93
Eden Hazard	91	91
Kevin De Bruyne	91	91
Marc-André ter Stegen	90	93
Virgil van Dijk	90	91
Luka Modrić	90	90
Mohamed Salah	90	90

(a) Top 10 sofifa players by overall rating.

Name	Overall	Potential
Kylian Mbappé	89	95
Lionel Messi	94	94
Cristiano Ronaldo	93	93
Jan Oblak	91	93
Marc-André ter Stegen	90	93
Matthijs de Ligt	85	93
João Félix Sequeira	80	93
Neymar	92	92
Paulo Dybala	88	92
Leroy Sané	86	92

(b) Top 10 sofifa players by potential rating.

Table 4
Correlations between the three sets of player ratings and the potential rating.

	xGPM	sofifa overall	sofifa potential
GIM	0.2334	0.2698	0.1730
xGPM		0.1549	0.1496
sofifa overall			0.8787

Coates & Parshakov (2021) and Yigit et al. (2020) all use sofifa ratings to model transfer fees and/or valuations. They are constructed similarly to the crowd-sourced TM values in that members of the website offer ratings for each player for each of a wide range of skills and attributes such as heading, shooting, passing etc. In this paper we use two specific ratings from sofifa: the *overall* rating, and the *potential* rating. The overall rating is an estimate of the player's current playing ability, whilst the potential rating is an estimate of the player's future playing ability.

Whenever sofifa ratings have been used as an explanatory variable in models of transfer fees or values, they have proven to be strongly statistically significant. We note that the significance of the sofifa ratings suggests that they accurately represent skills that decision makers involved in the process of agreeing prices of players also value.

Table 2 provides summary statistics for the two sofifa ratings (overall and potential) by playing position. Goalkeepers appear to have a lower mean rating whilst the three outfield lines are roughly the same.

Table 3 provides lists of the top ten players according to (a) the overall rating, and (b) the potential rating. Compared to the xGPM and GIM performance ratings, there are some differences in the names present in the lists, though Messi and Neymar are ever-presents.

Given that these metrics are to be used as covariates/features in models explaining transfer fees, it is important to consider the correlations between them. Table 4 presents these correlations, and on the whole they are low, though the two sofifa ratings are highly correlated. As such, we include the overall rating and the difference in the overall and potential ratings as features in the models. This has the attractive interpretation as the two variables represent how good the player is at the time of the transfer, and by how much he is expected to improve. The low correlations between the other sets of performance/ability metrics shows that each of the ratings is capturing something a little different about player ability. That Lionel Messi features highly in each of the rankings is testament to just how good a player he is.

4. Data

We obtained data from three different sources for the nine seasons starting with the 2011/12 season and ending with the 2019/20 season:

1. Transfer details were scraped from transfermarkt.com. The information included: the date, clubs and leagues involved, the actual fee, and the remaining duration of the player's contract. Further information on players such as their injury history was also gathered.
2. Match 'event data' across 36 leagues[4] were obtained from Instat. *Event data* describes the actions occurring in a football match such as passes, shots, yellow and red cards, tackles, interceptions and so on. The timing, the x-y coordinates of the event, the names of the players involved, and the outcome of the event are provided. Typically, there are around 2000 events in each football match, and in total we obtained event data from 59,393 different matches. These data were used to generate the xGPM plus-minus rating, and the GIM performance rating described earlier.
3. Crowd-sourced ratings of players were scraped from sofifa.com. The two ratings we use here are the overall rating and the potential rating. In addition to these ratings, we also obtained basic information on players such as their height and weight.

The data from different sources were merged first by finding players with the same birth date and surname, then by finding players who had played for the same teams, before the remaining players were matched by hand.

[4] The leagues covered included: the top four divisions in England, the top two divisions in each of Spain, Italy, Germany, France, and the top leagues in other European countries.

4.1. Covariate/feature generation

The final sets of covariates (features) we use in the models are as follows:

1. player characteristics: age; age squared; position played[5]; height.
2. experience: percentage of minutes played in the 12 months prior to the transfer; whether the player has been loaned to another club in the last 12 months; number of senior international appearances.
3. contract status: number of weeks remaining on the player's contract.
4. ability: 12-month pre-transfer mean of the GIM performance rating; the xGPM ratings calculated using data from the 12-month period immediately prior to the transfer date; the sofifa overall rating; the difference between the sofifa potential rating and the sofifa overall rating; and several basic playing statistics as calculated using the 3-year period[6] immediately prior to the transfer date. These were: goals per 90 minutes played average, key passes per 90 minutes played average, tackles per 90 minutes played average, and aerial duels contended per 90 minutes played average.
5. financial information on the clubs involved: Bryson, Frick, & Simmons (2013) found a correlation between the buying-club characteristics and the size of transfer fee. This confirms the original contention of Rottenberg (1956) who, when examining sports labour markets suggested clubs will be prepared to pay different amounts of money depending on how successful they are, and how they can turn that fee into increased revenue (e.g. through success on the pitch or through extra shirt sales). Some authors use dummy variables for clubs and leagues but for large data sets, that requires too many variables. Instead we use the median transfer fee paid by the buying club over the previous 3 years; and the median transfer fee paid by the selling club over the previous 3 years. Our approach has the additional advantage (over using dummy variables for leagues and/or clubs) that it can deal with the changing financial situation of the transfer market and of each club. The budgets clubs have to spend on transfers changes over time for a myriad of reasons, including new owners or investment in the club, and new broadcasting rights deals. Using covariates based on summary statistics of recent transfer market activity allows the model to pick up on changing economic conditions at both the club and global levels.

Note that we do not include popularity measures such as number of Twitter followers. This is because obtaining historical data on e.g. the number of followers on Twitter a player had at the time of a transfer on a specific day is not possible as this information is not stored anywhere, and without historical data, out-of-sample predictions are not possible. However, we note that popularity of the player on social media is not a quality that can help determine on-the-pitch value-for-money. Further, Ante (2019) uses number of Twitter followers as a covariate in a series of regression models utilising backward stepwise variable selection where the dependent variables are different ranges of transfer fee. The *social media* variable (which is the sum of the logarithms of Twitter, Instagram and Facebook followers is selected in only one of the models presented. We believe that the occasional finding that the number of Twitter followers a player has is a statistically significant determinant of transfer fee must be driven by the popularity of an extremely small number of superstar players.

Our final data set comprised information on 1946 transfers between 11th August 2016 and 29th September 2020. When calculating the financial information variables used in modelling, e.g. the median fee paid by the buying club in the previous 3-years, we used a much larger database of 995,486 transfers. For various reasons, many of these transfers did not make it into the final dataset: many of the transfers are loans, free transfers in which there was no fee, or the fee was undisclosed; many transfer observations contain missing contract information, a key component of the model; or, event data for the matches played by the player being transferred was not available (and hence it was not possible to calculate the player performance ratings). The resulting 1946 transfers were all for a non-zero fee, where the player was still under contract at the parent club. The transfers involved players leaving 31 different leagues and going to 62 unique leagues - a total of 69 different leagues involved, far more than has been covered in past literature.

Table 5 displays the summary statistics for variables used in the modelling, and Fig. 1 shows a histogram of the transfer fees in our final dataset.

Note that there are a small number of missing values, as indicated by $N < 1,946$ in Table 5. A missing value could occur when, for instance, a player hadn't attempted a tackle in the match event data; or a club hadn't made a paid transfer over the last three years. Rather than discard these samples, we used a linear model to impute the missing values, where the model was estimated on the non-missing columns.

5. Models for estimating transfer fees

In this section we first present the results of fitting linear regression models to explain the transfer fees, and compare our findings with those in the literature. Following that we use machine learning algorithms to explain transfer fees. Our objective in fitting these models is to explore the improvement in model fit by allowing for highly nonlinear relationships between the transfer fees and the predictor variables, and to minimise the chances of overfitting by first training the models before calculating measures of goodness-of-fit on a testing dataset of unseen data.

5.1. Linear regression in-sample fitting

To compare our data and findings with the findings presented in the literature, we fit four models to the entire sample of 1946 transfers. The dependent variable is the natural logarithm of transfer fee. Each model varies in how the ability and performance of a player is accounted for. Model 1 includes only basic metrics of player performance; model 2 includes the advanced metrics; model 3 includes the sofifa metrics; and model 4 includes all types of performance metrics. Table 6 shows model fit summaries for the four models. Note that, with exception to the position dummies, each variable is normalised.

The best fitting model is model 4 which includes both the sofifa ratings and the objective performance metrics. We performed ANOVA tests on the nested models (models 2 and 3) with the full model (model 4), testing to see if adding the extra sets of variables significantly improved the fit. For models 2 and 4 we found a p-value of < 0.001, indicating a significant improvement when adding the analytical ratings. Further, we found a p-value of < 0.001 between models 3 and 4, indicating that adding the sofifa ratings significantly improves the fit also.

Comparing the model fit with the literature, other authors have found R^2 values typically in the range of 0.3 to 0.7. For example,

[5] Positions are given as full-back, centre-back, centre midfield, central defensive midfield, side midfield, winger, and striker.

[6] We experimented with 1-year and 2-year summaries but found that 3-year provided the better model fit, though the improvement was limited.

Table 5
Summary statistics for a number of the variables used within our models.

Statistic	*N*	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Transfer fee (millions)	1946	8.22	13.25	0.01	1.35	10.35	199.80
Age	1946	25.67	3.35	17.53	23.20	27.95	36.47
Overall rating (sofifa)	1946	74.10	4.89	59.00	71.00	77.00	94.00
Potential rating (sofifa)	1946	78.41	4.82	61.00	75.00	82.00	94.00
Height (cm)	1946	181.90	6.39	163.00	177.00	187.00	201.00
xGPM	1930	50.30	2.99	40.06	48.43	51.90	63.58
Goals per 90, 3-yrs	1946	0.17	0.17	0.00	0.04	0.28	1.11
Key passes per 90, 3-yrs	1946	0.84	0.54	0.00	0.45	1.15	3.28
Tackle accuracy, 3-yrs	1942	0.66	0.11	0.19	0.59	0.74	1.00
Aerial accuracy, 3-yrs	1945	0.45	0.15	0.00	0.34	0.57	1.00
Purchasing club median buying fee, 3-yrs	1935	4.52	5.85	0.01	1.08	5.40	42.84

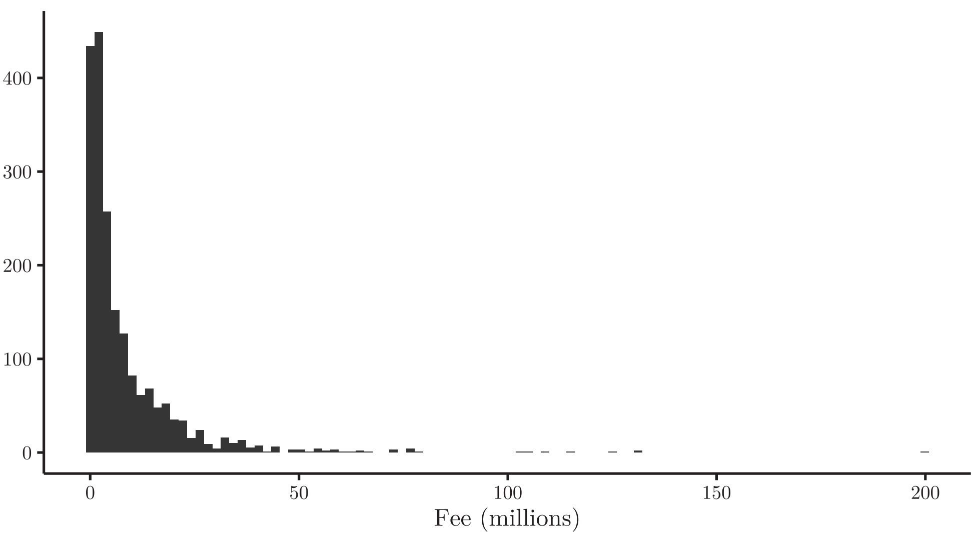

Fig. 1. Histogram of fees contained in the final dataset used for modelling and testing.

Table 6
Summary of four in-sample regression models. The dependent variable is the natural logarithm of the transfer fee paid. Standard errors are shown in parentheses. Note that the reference category for playing position is *attacking centre midfield*.

	(1)	(2)	(3)	(4)
age	-0.325*** (0.026)	-0.313*** (0.026)	-0.210*** (0.057)	-0.185*** (0.057)
age^2	-0.038** (0.018)	-0.038** (0.018)	-0.070*** (0.020)	-0.075*** (0.020)
contract remaining (weeks)	0.430*** (0.023)	0.426*** (0.023)	0.293*** (0.020)	0.297*** (0.020)
on-loan in last 12 months	-0.060*** (0.023)	-0.045** (0.022)	-0.059*** (0.020)	-0.051*** (0.020)
international caps in last 3 years	0.126*** (0.024)	0.116*** (0.024)	0.039* (0.021)	0.039* (0.021)
height	-0.012 (0.030)	-0.009 (0.029)	0.063** (0.026)	0.062** (0.026)
minutes played (3 yrs)	0.393*** (0.026)	0.327*** (0.026)	0.198*** (0.023)	0.175*** (0.024)
goals per 90 mins (3 yrs)	0.157*** (0.036)	0.082** (0.036)	0.027 (0.031)	0.001 (0.032)
key passes per 90 (3 yrs)	0.034 (0.032)	-0.038 (0.032)	-0.013 (0.028)	-0.043 (0.028)
aerial duels won per 90 (3 yrs)	0.049 (0.034)	0.00004 (0.034)	0.023 (0.030)	0.002 (0.030)
tackles won per 90 (3 yrs)	-0.034 (0.028)	-0.057** (0.027)	-0.034 (0.024)	-0.043* (0.024)
median purchase price club to (3 yrs)	0.497*** (0.025)	0.459*** (0.025)	0.270*** (0.024)	0.263*** (0.024)
median purchase price club from (3 yrs)	0.340*** (0.024)	0.347*** (0.024)	0.159*** (0.022)	0.171*** (0.022)
centreback	0.127 (0.143)	-0.002 (0.141)	-0.044 (0.123)	-0.098 (0.123)
central defensive midfield	0.102 (0.136)	0.043 (0.134)	-0.056 (0.117)	-0.080 (0.117)
centre midfield	0.182 (0.115)	0.157 (0.113)	0.150 (0.099)	0.138 (0.099)
fullback	-0.096 (0.126)	-0.167 (0.124)	0.040 (0.109)	0.002 (0.109)
side midfield	-0.054 (0.106)	-0.044 (0.104)	0.039 (0.092)	0.040 (0.091)
striker	0.194* (0.113)	0.268** (0.111)	0.387*** (0.098)	0.410*** (0.097)
winger	0.181 (0.123)	0.155 (0.121)	0.268** (0.106)	0.254** (0.106)
GIM performance		0.218*** (0.027)		0.105*** (0.024)
xGPM		0.035 (0.023)		-0.002 (0.020)
overall rating			0.799*** (0.031)	0.769*** (0.032)
potential rating - overall rating			0.377*** (0.057)	0.389*** (0.057)
Constant	1.138*** (0.099)	1.163*** (0.098)	1.139*** (0.087)	1.157*** (0.087)
Observations	1,946	1,946	1,946	1,946
R^2	0.605	0.620	0.707	0.710
Adjusted R^2	0.601	0.616	0.704	0.707
Residual Std. Error	0.944 (df = 1925)	0.926 (df = 1923)	0.813 (df = 1923)	0.809 (df = 1921)
F Statistic	147.368*** (df = 20; 1925)	142.649*** (df = 22; 1923)	211.277*** (df = 22; 1923)	196.262*** (df = 24; 1921)

*p<0.1; **p<0.05; ***p<0.01

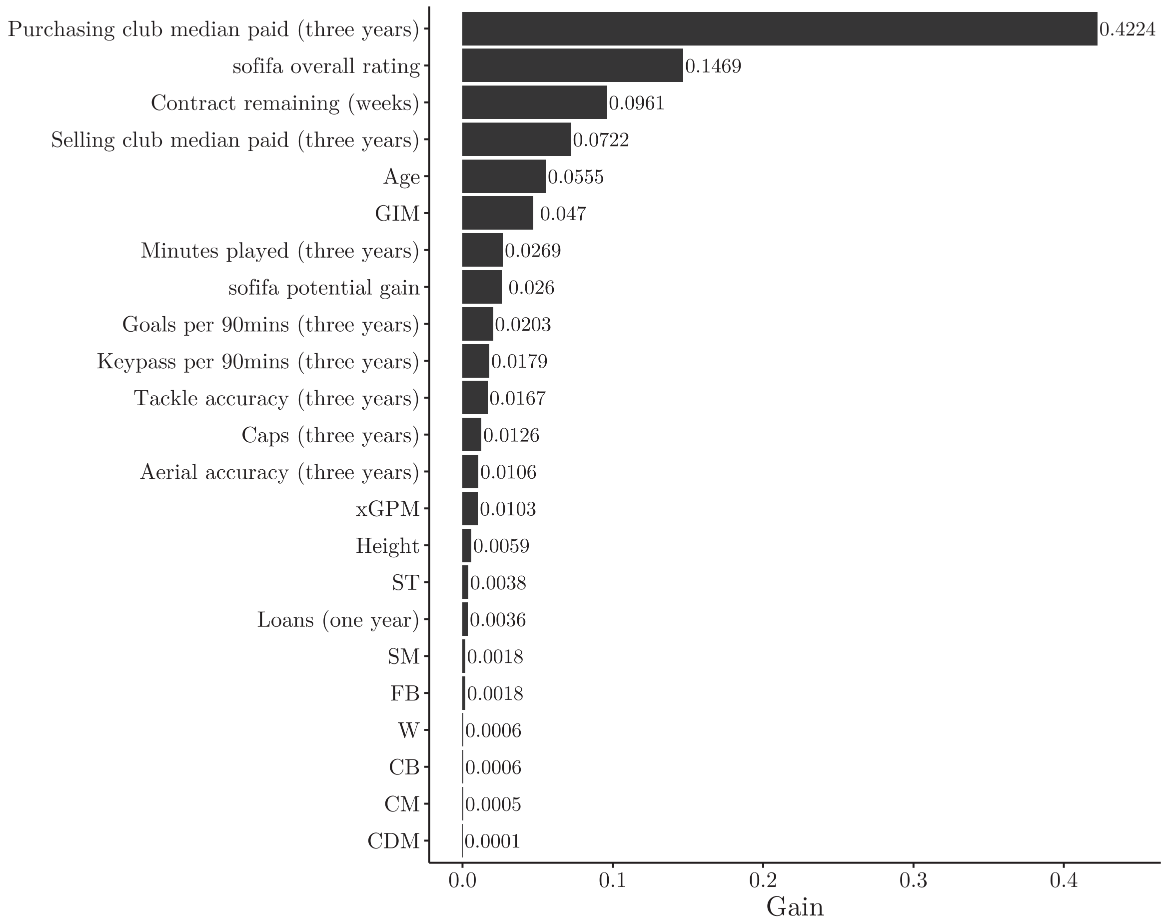

Fig. 2. Importance of each variable in the final xgbTree model as measured by the gain.

Sæbø & Hvattum (2015) achieve R^2 values of 0.331 (several European leagues) and 0.534 (English market only), whilst the model employed by Coates & Parshakov (2021) which includes the crowd-sourced TM valuations as a covariate achieves an R^2 value of 0.840. However, we note that the R^2 is lower when the authors fit their model to transfers between lower league clubs. As our dataset includes transfers between clubs from over 30 leagues, it is unsurprising that the values of R^2 our models attain (between 0.601 and 0.707) are a little lower.

In addition to the goodness-of-fit of the model being of the expected order of magnitude, the statistical significance, and signs of the estimated coefficients, is entirely unsurprising. That said, it is noteworthy that in the presence of either the sofifa ratings, or the advanced performance metrics, the basic metrics all fail to achieve statistical significance.

5.2. Machine learning

The linear regression models presented in the previous section were fitted to the entire sample of data. We now present the results of using machine learning algorithms to ‘predict’ transfer fees. To avoid data leakage, we split our data by time, using the first 80% of the transfers for training (1,557) and the remainder for testing (389). To be clear, we are now performing true out-of-sample prediction modelling. This has the advantage that a user of the model can gauge what level of accuracy/agreement with the market the model’s transfer estimate is likely to give with transfer fees typically agreed upon.

As well as ordinary least squares (OLS) linear regression, we estimate a mixed effects linear model (*glmer*) in which random effects are used for the buying and selling clubs. For this model, we drop the club median price paid variables. One advantage of the random effects model (as opposed to a fixed effects model in which the buying and selling clubs are represented by dummy variables) is that the number of parameters estimated remains reasonable. Further, because the club effects are assumed to come from an underlying distribution, those clubs with few observations will have their effect on the transfer fee paid (their intercept) shrunk towards the population mean effect. However, the random effects model has the disadvantage over using the ‘median price paid in the last three-years’ variables since it fails to allow for changing club effects.

We also use models based on three machine learning algorithms. Elastic-net regression, through *glmnet*, combines lasso and ridge regression. Lasso aims to find a sparse model with borderline variables receiving estimated coefficients of zero, whilst ridge introduces a shrinkage parameter to reduce the impact of collinearity; both reduce the risk of over-fitting. Whilst clearly solving several problems associated with OLS, interactions are not implicitly handled.

Extreme gradient boosting (xgboost) is one of the most popular ML algorithms and has often won predictive competitions (Github, 2021). The “basic” version, *xgbTree*, builds a model by iteratively learning an ensemble of decision trees which are only weakly correlated to the dependent variable but collectively achieve a strong correlation. Since the trees estimated at the start of the procedure often influence predictions to a greater degree, a third machine learning algorithm is used in which “dropout” is applied, through *xgbDART*. In this algorithm, trees are randomly removed from the ensemble, in an effort to reduce over-fitting. An advantage of the

Table 7
Description of the hyper-parameters tuned in the elastic-net regression and xgboost models.

Model	Parameter	Description
glmnet	alpha	Mixing parameter which determines the type of penalisation.
	lambda	Determines the strength of penalisation.
xgbTree	max_depth	The maximum number of nodes within an individual tree. Higher values make the model more complex.
	eta	The learning rate of the model. Lower values increase the computation time but increases the likelihood of reaching the optimal solution.
	subsample	Determines the ratio of training data observations to be randomly selected to train a tree. Higher values can lead to over-fitting.
	colsample_bytree	Determines the ratio of features to be randomly selected to train a tree.
	min_child_weight	The minimum weight necessary before a further node can be added to a tree. Smaller values allow nodes based on fewer observations, thus leading to more complex models and possible over-fitting.
	gamma	The minimum reduction in the loss function required before splitting a leaf node on a tree. Smaller values lead to more splits and possible over-fitting
	nrounds	The number of decision trees in the full model.
xgbDART	rate_drop	The dropout rate: what proportion of trees should be dropped during the dropout phase.
	skip_drop	The probability of skipping the dropout procedure for a tree.

Table 8
Summary statistics of the predictions of our models on the out-of-sample test set.

Model	MAE	MAPE	R^2
xgbTree	3.60	67.47	0.77
xgbDART	3.64	68.90	0.76
glmer	4.11	69.05	0.74
glmnet	9.93	90.57	0.50
OLS	10.34	91.81	0.50

xgboost algorithms is the implicit regularization and interactions of features. An introduction to many machine learning models, including xgboost, can be found in Hastie, Tibshirani, & Friedman (2009).

In the case of the glmnet, xgbDART and xgbTree models, there are hyper-parameters to be tuned. A description of each hyper-parameter in each model is given in Table 7. Note that the xgbDART model includes all of the hyper-parameters described for xgbTree.

To tune the hyper-parameters, we estimate 2500 models with each model randomly selecting a set of hyper-parameters from a grid. An exhaustive grid-search is unfeasible and impractical in most cases, particularly in the case of xgboost models which have numerous parameters to tune. Indeed Bergstra & Bengio (2012) and Mantovani, Rossi, Vanschoren, Bischl, & de Carvalho (2015) found that randomly selected grids, as we have used, perform comparably to manual searching and exhaustive grids, with the obvious efficiency benefits. We find the best fitting model through 10-fold cross validation, according to the average MAE achieved over the different folds. Fitting in this way means the performance measures are calculated over many different validation sets, resulting in more realistic and robust estimates of goodness-of-fit. We ensured all models were estimated using the same folds throughout fitting.

Table 8 shows out-of-sample measures of prediction accuracy. According to the R^2, the xgbDART and xgBoost tree models are front-runners for the best fitting model. It is striking that even out-of-sample, they achieve R^2 values higher than the in-sample linear regression models presented in Table 6. The random effects model (glmer) performs well, whilst the linear regression models perform considerably less well in comparison to the other models demonstrating the caution that must be exercised when using a model to make predictions on new data. Even in the presence of lasso and shrinkage (the glmnet model), the regression approach fails to get near to the xgbDART and xgBoost tree models. This suggests the non-linearities in the data are complex and poorly modelled by regression.

In practice, the user's specific needs should determine which goodness-of-fit metric they optimise over. Clubs at the top of the football pyramid who spend tens of millions on transfers may not be overly concerned with over-paying by a few million, in which case they might choose to use the model that minimises the MAPE. On the other hand, a club with less wealth will probably be less willing to risk over-paying in the order of millions of pounds, and would be better to use the model that minimises the MAE.

The variable importances from our best fitting model (xgbTree) is shown in Fig. 2. The most important variable for determining transfer fee is the median price paid for players by the purchasing club in the last 3 years. This variable must incorporate many elements of the factors contributing to the transfer fee including: the ability of the player since a club that is used to spending a lot of money is likely to be targeting top players, and of course, the financial strength of the club and league. The importance of this variable in the machine learning models explains the strong performance of the random effects model.

The sofifa overall rating is the next most ‘important’ variable in the model and features higher in the list than the two objective player ratings (GIM performance and xGPM). It is noteworthy that the GIM rating is identified as more important than the basic performance measures (such as goals per 90, and key passes per 90). This suggests that the decision makers who sign cheques for players are taking into account information over and above the traditionally used metrics. Conversely, one could argue that in the presence of the set of advanced performance metrics, the basic statistics should cease to be important at all. That this is the case may be due to either: the performance metrics not fully capturing the ability of the player; or, that the market reads too much into the relevance of goals scored when pricing players. This would be an interesting area for future research.

5.3. Comparison with transfermarkt values

Although they are not explicitly meant to be used for estimating transfer fees, the player valuations on the TM website provide a tough benchmark for models predicting transfer fees. In some part this is because of the circular nature of how the valuations are nowadays used by clubs themselves. We have anecdotal experience that clubs refer to TM valuations in negotiations with agents and clubs over transfer fees.

Fig. 3 shows the predicted fees (out-of-sample) according to the xgbTree model versus the actual fees (left), and the TM valuations versus the actual fees (right). Visual inspection of the scatter plots demonstrates a tighter relationship for the model predictions compared to the TM valuations. The xgbTree model has a slightly higher correlation, 0.77 against 0.75. On the other hand, the MAEs

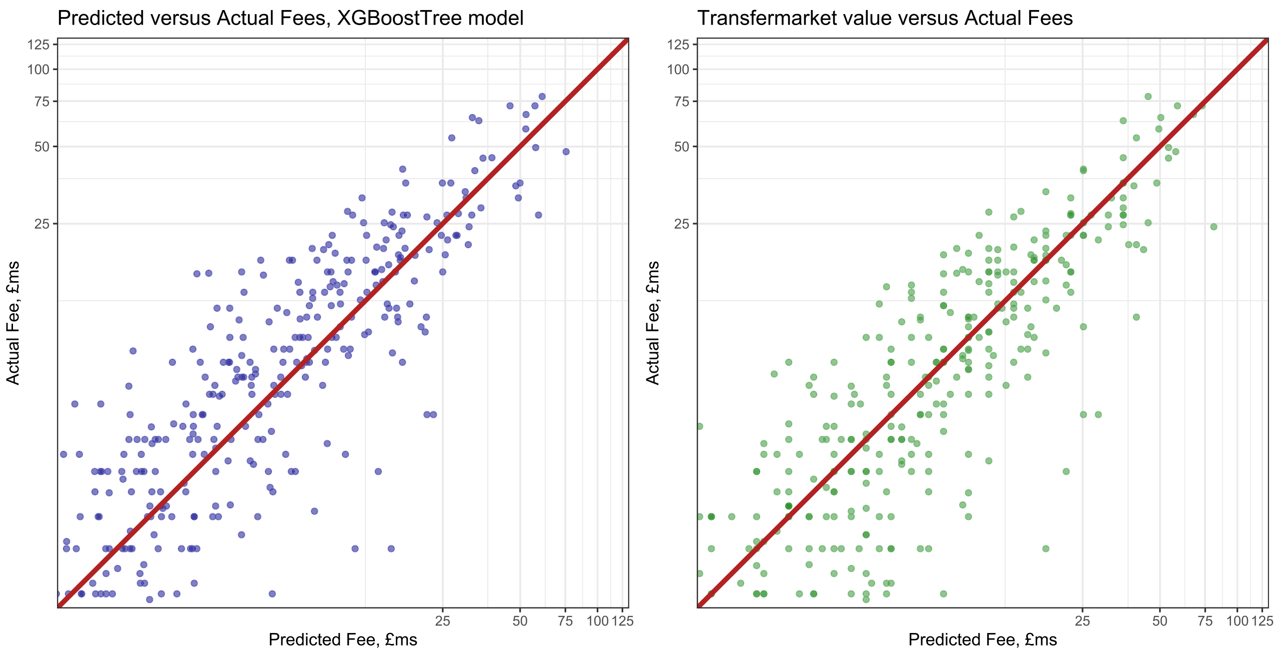

Fig. 3. Scatter-plot comparing the predicted transfer fees according to the xgBoostTree model and the actual transfer fees (left) and the transfermarkt valuations and the actual transfer fees (right). All axes are shown on a log scale.

Table 9
Comparison of the MAE of predicted transfer fees for the xgbTree model and the TM market values for different ranges of actual fees.

Fee bounds (£m)	Average fee (£m)	MAE (xgbTree)	MAE (TM)	N
(0,1]	0.52	0.58	1.05	80
(1,2.5]	1.67	1.22	1.66	65
(2.5,5]	3.52	2.26	2.56	60
(5,10]	7.33	2.70	3.12	64
(10,20]	14.76	5.70	5.20	67
(20,35]	25.30	8.57	8.79	34
(35,78.3]	51.40	15.11	10.15	19

Table 10
Comparison of the MAE of predicted transfer fees by the xgbTree model and the TM market values for different lengths of remaining contract (measured in weeks).

Contract bounds (weeks)	Average contract (weeks)	MAE (xgbTree)	MAE (TM)	N
(0,25]	22.20	1.682	4.924	19
(25,50]	44.42	2.152	2.577	110
(50,75]	63.11	1.732	2.241	22
(75,100]	93.50	3.540	3.342	107
(100,150]	128.54	3.571	3.499	59
(150,200]	164.27	6.725	4.799	57
(200,255]	211.84	7.653	8.112	15

are in favour of the TM valuations: 3.60 versus 3.56 for the model and TM valuations respectively. Finally, the MAPEs are 67.47 and 98.00 for the model and the TM valuations respectively. Overall, the model is performing at least comparably to the TM valuations; it massively outperforms the TM valuations in terms of MAPE, and achieves similar correlations and MAEs.

In Tables 9 and 10 we disaggregate the predictive performance of both the xgbTree model and the TM valuations based on the final fee paid and the contract remaining respectively. Table 9 reveals that the model out-performs the TM valuations for the majority of ranges of fees (especially so for small fees), but is worse than the TM valuations in the top range of fees (though there is only a small number of observations in this band). Table 10 shows a similar pattern in that the model tends to outperform the TM valuations for observations with shorter time remaining on the player's contract. This lack of predictive accuracy of the TM valuations (relative to the xgbTree model) for small fees, and players with less time remaining on their contracts is likely a consequence of the TM valuations not taking into account contract length. Indeed, as said in the opening lines of this section, the purpose of the TM valuations is not explicitly stated to be to predict transfer fees. As such, it is not surprising that as contracts run down the predictive accuracy lessens for the TM valuations. We believe it is important to have established this fact as practitioners (who do refer to the TM valuations) should know that the valuations become less useful as a player's contract nears expiry.

6. Estimating value for money

The models presented here can be used by fans and the media to obtain expected transfer fees for players. Perhaps the most valuable use of the model, certainly in monetary terms, is for clubs to identify a benchmark price for players. Looking retrospectively at transfers that have happened, and comparing the predicted transfer fee with the actual transfer fee can identify 'good' and 'bad' transfers. Table 11 provides the best and worst value-for-money transfers, as measured by the excess fee paid. Football fans will be unsurprised that players like Luis Suarez and Bruno Fernandes have been identified as good value for money transfers, whilst the fees paid for Neymar, Coutinho and Harry Maguire are known high fee transfers.

By estimating 100 bootstrapped models, we calculated the 5% and 95% quantiles of the predicted fee for each player and use these to form a prediction interval. This can be of use to clubs when negotiating: the buyer/seller would likely try to aim for the lowest/highest reasonable price, perhaps within these bounds.

In addition to looking at individual transfers, one can examine the performance of clubs in the transfer market by calculating, for example, the median value-for-money for all transfers by each club. Table 12 presents a selection of the top performing and worst performing clubs. It is intriguing that some clubs known for using

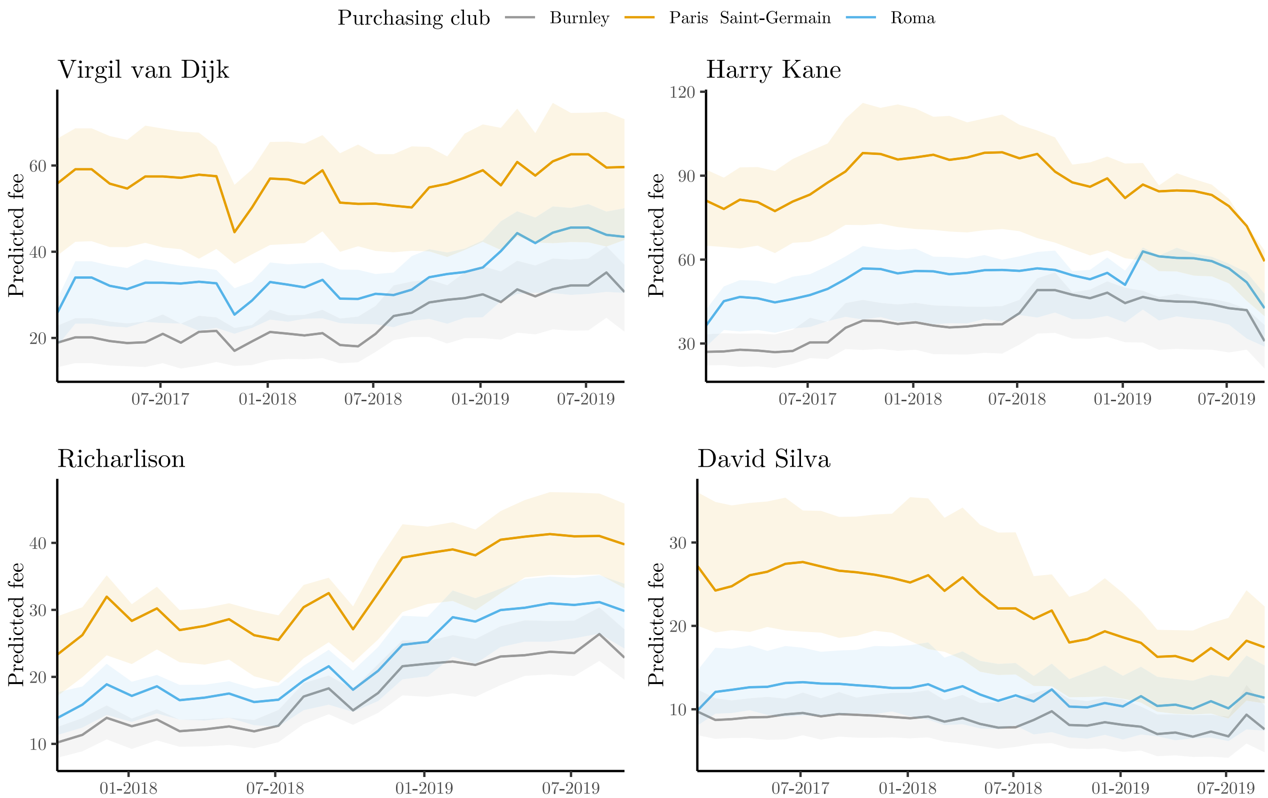

Fig. 4. Predicted transfer fees over time for different purchasing clubs.

Table 11
Value-for-money, found as the difference between the estimated fee and reality, for several interesting transfers.

Player	Club from	Club to	Date	Estimated price pctl (5%)	Estimated price	Estimated price pctl (95%)	Actual price	Value for money
Timo Werner	RB Leipzig	Chelsea	07/01/20	60.7	76.1	92.0	47.7	28.4
Fabinho	Monaco	Liverpool	07/01/18	49.3	60.5	69.1	40.5	20.0
Luis Suárez	FC Barcelona	Atlético Madrid	09/25/20	15.1	23.8	31.2	4.5	19.3
Gelson Martins	Sporting CP	Atlético Madrid	07/25/18	30.6	36.8	47.3	19.8	17.0
Hakim Ziyech	Ajax	Chelsea	07/01/20	38.6	50.8	59.8	36.0	14.8
Bruno Fernandes	Sporting CP	Man Utd	01/29/20	44.7	58.7	69.1	49.5	9.2
Xherdan Shaqiri	Stoke City	Liverpool	07/13/18	17.8	22.1	29.7	13.2	8.9
Harry Maguire	Leicester	Man Utd	05/08/19	40.6	50.9	60.4	78.3	-27.4
Philippe Coutinho	Liverpool	FC Barcelona	08/01/18	55.0	83.0	111.3	130.5	-47.5
Cristiano Ronaldo	Real Madrid	Juventus	10/07/18	23.4	43.9	73.7	105.3	-61.4
João Félix	Benfica	Atlético Madrid	03/07/19	28.2	51.8	78.5	114.5	-62.7
Neymar	FC Barcelona	Paris SG	03/08/17	63.4	103.8	137.7	199.8	-96.0

Table 12
Value-for-money, found as the median of the difference between the estimated fees and actual fees, for selected clubs.

Purchasing club	N	Spent	Worth	Total value	Median value
Arsenal	11	361.00	373.78	12.78	5.07
Bayern Munich	9	206.55	203.27	-3.28	4.07
Southampton	9	150.71	165.74	15.03	3.22
Atlético Madrid	14	442.62	407.65	-34.97	2.52
Liverpool	11	333.41	321.79	-11.62	1.40
Brentford	6	16.02	20.98	4.95	1.12
RB Leipzig	12	170.73	145.60	-25.13	-1.30
Zenit St. Petersburg	7	104.85	62.97	-41.88	-2.92
West Ham	12	224.55	169.43	-55.12	-3.10
FC Barcelona	14	685.44	523.52	-161.92	-3.94
Man Utd	10	426.76	391.66	-35.10	-4.14
CSKA Moscow	3	27.00	11.81	-15.19	-5.70

analytics as part of their recruitment departments appear as clubs doing well in identifying value-for-money: notably Liverpool and Brentford. Whilst other clubs, who have in recent times been generally accepted to have over-paid for players, are identified by the model: notably, Manchester United and Barcelona. The presence of CSKA Moscow is interesting as for several years Russian clubs were believed to be overpaying for older players, and it appears CSKA Moscow were particularly guilty of this (though the sample size is very small). Perhaps club owners can use information like this to improve their recruitment strategies.

Our final experiment with the model is to examine predictions of transfer fee for players over time. As the covariates change, due to changing age, performance metrics, and market conditions (e.g. inflation in fees), the predicted transfer fee evolves. Fig. 4 shows the evolution of the predicted transfer fee for four players: Virgil van Dijk, Harry Kane, Richarlison, and David Silva.

For each player we plot three predicted transfer fees, depending on which club is buying. This demonstrates how instrumental the identity of the purchasing club is in determining the transfer fee. In the case of Virgil van Dijk for example, the expected purchase price for one of the richest clubs in the world, Paris Saint Germain, is nearly £60m, but that falls to around £30m if the buying club is Burnley. For David Silva, one can see how his fee falls as he nears retirement and as his contract at Manchester City neared expiry.

7. Conclusions

The paper presents a model for predicting the transfer fees in the labour market for footballers. The fees paid are functions of player and club characteristics, and player performance and ability. Until now, player performance has been measured using either rudimentary statistics such as goals scored, or player ratings from websites such as sofifa.com. Unlike previous work, we include advanced player ratings systems based on analytics which are shown to improve the predictive accuracy of the model. In addition to presenting results of linear regression models, we use algorithms from machine learning and demonstrate a remarkable gain in predictive accuracy of the models.

Models of this type can be used by fans, the media and clubs. Despite the vast sums of money exchanged between clubs for the services of players, decision making on transfer fees remains rather rudimentary in the football industry. Our model could be used by clubs as a benchmark to be used in negotiations, or as a check that a price is reasonable. Here we look at value-for-money transfers and find that some clubs have overpaid considerably for players, whilst others appear to be able to identify, and capitalise on, value-for-money in the transfer market.

Our model is used semi-retrospectively to look at transfer fees and value-for-money and answers the question “given past performance, what is a player’s expected fee?”. Future work however could look at the problem fully retrospectively and ask “given the performance of a player after a transfer, was the price paid reasonable?”.

Acknowledgements

We would like to thank Instat for allowing us access to their event data and Tarak Kharrat for allowing us access to the GIM performance ratings.

References

Ante, L. (2019). Determinants of transfers fees: Evidence from the five major european football leagues,.

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players’ value in the transfer market. *Soft Computing, 25*(3), 2499–2511. https://doi.org/10.1007/s00500-020-05319-3.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research : JMLR, 13*, 281–305.

Bryson, A., Frick, B., & Simmons, R. (2013). The returns to scarce talent: Footedness and player remuneration in european soccer. *Journal of Sports Economics, 14*(6), 606–628.

Carmichael, F., Forrest, D., & Simmons, R. (1999). The labour market in association football: Who gets transferred and for how much? *Bulletin of Economic Research, 51*(2), 125–150. https://doi.org/10.1111/1467-8586.00075.

Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: Theory and evidence. *Applied Economics, 25*(12), 1467–1476. https://doi.org/10.1080/00036849300000150.

Cervone, D., D’Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association, 111*(514), 585–599. https://doi.org/10.1080/01621459.2016.1141685.

Coates, D., & Parshakov, P. (2021). The wisdom of crowds and transfer market values. *European Journal of Operational Research*.

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDDInternational conference on knowledge discovery and data mining*. In *KDD ’19* (pp. 1851–1861). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3292500.3330758.

Depken II, C. A., & Globan, T. (2021). Football transfer fee premiums and Europe’s big five. *Southern Economic Journal, 87*(3), 889–908. https://doi.org/10.1002/soej.12471.

Dobson, S., & Gerrard, B. (1999). The determination of player transfer fees in english professional soccer. *Journal of Sport Management, 13*(4), 259–279. https://doi.org/10.1123/jsm.13.4.259.

Dobson, S., Gerrard, B., & Howe, S. (2000). The determination of transfer fees in english nonleague football. *Applied Economics, 32*(9), 1145–1152. https://doi.org/10.1080/000368400404281.

FIFA (2021). Global transfer market report 2020: A review of international football transfers worldwide. online, https://img.fifa.com/image/upload/ijiz9rtpkfnbhxwbqr70.pdf.

Frick, B. (2007). The football players’ labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy, 54*(3), 422–446. https://doi.org/10.1111/j.1467-9485.2007.00423.x.

Github (2021). Xgboost: Machine learning challenge winning solutions. online, https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the moneyball hypothesis. *Journal of Economic Perspectives, 20*(3), 173–186. https://doi.org/10.1257/jep.20.3.173.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. *Springer Series in Statistics*. New York, NY, USA: Springer New York Inc..

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players’ market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006.

Kahn, L. M. (2000). The sports business as a labor market laboratory. *Journal of Economic Perspectives, 14*(3), 75–94. https://doi.org/10.1257/jep.14.3.75.

Kharrat, T., McHale, I. G., & Pena, J. L. (2020). Plus-minus player ratings for soccer. *European Journal of Operational Research, 283*(2), 726–736. https://doi.org/10.1016/j.ejor.2019.11.026.

Kirschstein, T., & Liebscher, S. (2019). Assessing the market values of soccer players – Arobust analysis of data from german 1 and 2 bundesliga. *Journal of applied statistics*, 1336.

Lewis, M. (2003). Moneyball: The art of winning an unfair game. *Norton paperback*. W.W. Norton.

Liu, G., Luo, Y., Schulte, O., & Kharrat, T. (2020). Deep soccer analytics: Learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery, 34*, 1531–1559.

Macdonald, B. (2012). Adjusted Plus-Minus for NHL players using ridge regression with goals, shots, Fenwick, and Corsi. *Journal of Quantitative Analysis in Sports, 8*(3), 1–24.

Mantovani, R., Rossi, A., Vanschoren, J., Bischl, B., & de Carvalho, A. (2015). Effectiveness of random search in SVM hyper-parameter tuning. https://doi.org/10.1109/IJCNN.2015.7280664.

Muller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005.

Rosenbaum, D. T. (2004). Measuring how NBA players help their teams win. *82Games.com*, 4–30. http://www.82games.com/comm30.htm

Rottenberg, S. (1956). The baseball players’ labor market. *Journal of Political Economy, 64*(3), 242–258.

Sæbø, O. D., & Hvattum, L. M. (2015). Evaluating the efficiency of the association football transfer market using regression based player ratings. In *Nik*.

Yigit, A. T., Samak, B., & Kaya, T. (2020). An XGBoost-lasso ensemble modeling approach to football player value assessment. *Journal of Intelligent & Fuzzy Systems, 39*(5), 6303–6314. https://doi.org/10.3233/JIFS-189098. Publisher: IOS Press