A Machine Learning Ensembling Approach to Predicting Transfer Values

Ayse Elvan Aydemir[1,2*], Tugba Taskaya Temizel[2] and Alptekin Temizel[2]

[1*]Enskai Ltd., Sofia, 1000, Bulgaria.
[2] Graduate School of Informatics, METU, Ankara, 06800, Turkey.

*Corresponding author(s). E-mail(s): elvan.aydemir@ensk.ai;
Contributing authors: ttemizel@metu.edu.tr;
atemizel@metu.edu.tr;

Abstract

Predicting transfer values of association football players, despite its importance, has been studied in a limited way in the literature. The existing approaches have mainly focused on explanatory models that cannot be used in predicting future values. In this paper, we propose a method where we fuse in-game performance data, player popularity metrics from the web and actual transfer values. The method uses a model ensembling approach to capture different dynamics in transfer market. The proposed approach outperforms the state-of-the art models and commonly used benchmarks.

Keywords: Predictive Modelling, Football Analytics, Player Valuation, Data Fusion, Model Ensembling

1 Introduction

Association football is the largest spectator sport in the world, leading to an industry worth billions of dollars. Player salaries and transfer fees make a significant portion of transactions in this market. However, the player transfer market has been shown to have serious information asymmetry, significant bias and complexities coming from rules and regulations. In addition, clubs are also interested in generating revenue from marketing efforts to reach more people.

Player values are also dependent on their role, their performance, the injuries, player's agent and how influential they are. Out of these, evaluating player performance is a difficult problem on its own [1, 2]. These complex market dynamics and gaps in information make the task of determining the value of the next transfer of any given player a challenging task. Furthermore, player transfer market is highly skewed in terms of values [3]. While the majority of the transfers have low or no monetary value, there are also transfers -while limited in number- in the value range of hundreds of millions. Depending on the use-case, accuracy on one end or the other may be desirable.

Another difficulty arises from football-related data sources being segregated and hard to reach. In-game performance statistics are available from different data providers. However, the same data providers do not provide the financial data such as transfers. Furthermore, data related to players' public perception does not exist as a formal data source and must be developed. Once data is obtained, the data sources must be combined together to perform modelling and eventually valuation.

In Economics literature, player valuation has been studied from the perspective of understanding the factors that affect the value [4**?** –6]. These factors also include information such as the performance of the buying club or the properties of the competition player gets transferred into. However, in a predictive modelling context, this information cannot be used because at the time of predicting the value of next transfer for the player, the information of the buying club is unavailable.

In addition to the market complexities and modelling approaches, the topic of player valuation is also a broad one. We can consider total career value of the player as the player value. This would entail the total salary player receives, as well as their cumulative transfer fees. Player valuation problem could also be framed as predicting whether they will be transferred with or without a fee. We could model the current value of all future transfers. Finally, the topic could be framed as predicting the value of the next transfer.

We frame the problem of player valuation as the accurate prediction of the value of transfer if the transfer were to happen in the next transfer window. Existing methods in the literature typically restrict the analysis and modelling to top European leagues and high divisions. On the other hand, we do not restrict the application to specific leagues, regions or specific roles and we aim to perform transfer valuation to all active football players globally.

In this paper we take a predictive modelling approach to estimating nominal player values, where we only use information from the past, making the modelling approach ideal for real world applications. We refer to additional factors such as a player's Google Search trends information to quantify their historical popularity. We use statistically enriched player performance metrics, player ranks obtained with unsupervised learning and model stacking to extract maximum amount of information from the historical and current datasets. We fuse distinct datasets using fuzzy string and attribute matching to include detailed factors in player performance.

In addition, we introduce an extended methodology to include player performance as predictors for global comparison of player values.

Most approaches aim for a generally applicable solution which is not feasible in such a dynamic market. We explicitly address the different dynamics of transfer value skewness by ensembling predictions under different outcome variables. To the best of our knowledge, this study contains the largest player and transfer subset for modelling.

This paper first provides a detailed literature review on player valuation, then outlines the proposed methodology. We continue with providing experimental setup and results. We compare our results to other predictive player valuation methods and Transfermarkt benchmark. We conclude the paper by summarising the contributions and outlining future work.

2 Background

Player values and financial contracts in association football manifest complex transactions. In any given transfer, clubs may have to agree to transfer fees, agent fees, player salaries, performance bonuses, fees to agents from future player transfers and additional custom clauses that make every transfer a complex negotiation. Furthermore, there are restricting rules on which players can play in which regions. For example, with the UK leaving the EU, the players now have to meet eligibility criteria to be transferred into UK teams. This complexity of the landscape makes predicting even a single aspect of the whole financial spectrum difficult. This is evident by the gap between the actual transfer fees and the published player market values by journalists and websites like Transfermarkt. The financial dynamics of football are explained in detail in [7].

With the emergence of the performance data providers such as WyScout [8], Opta [9] and Instat [10], there has been an increase in academic interest in analysing football data. Some of this interest is focused on player valuation. While football has been studied from an Econometric stand-point for some time, predictive modelling approaches have been emerging only recently.

The de-facto standard for predicting transfer fees remains crowd-sourced however crowd-sourcing relies on individuals in the crowd having sufficient knowledge about the player being valued and the market itself [11]. For well-known players, this information may be assumed, however, for most players such information is not available in the crowd-wisdom.

2.1 Descriptive vs. Predictive Modelling

The econometric modelling approaches focus on identifying factors that impact the player's transfer fee or market value. Their aim is to *describe* the mechanics that drive a certain phenomenon, in this case player fee or market value. Often these models use expert judgements or information that is not available prior to transfer completion such as the buying club. These models are useful for understanding the market dynamics to help make high-level strategic decisions for clubs and policy makers. In [3], authors explore the effects of player performance, buying and selling club characteristics and time on transfer fees between 1990 and 1996 in English Premier league using linear regression methods. They use segments as a controlling factor and find that the transfer fees are volatile across segments even in a single competition. Another approach to player valuation is treating them as volatile assets [12]. The authors factor in the club's own financials and goals while valuing a player by formulating the player valuation problem as asset management. This approach values the player specifically for the buyer instead of estimating a general market value or transfer fee. A more recent study uses an expanded set of variables to estimate the player transfer fees [13], however, the authors use player ids and names as factors, which make generalisation of their results impossible.

In contrast, predictive modelling approaches focus more on accurately predicting the transfer fees than modelling the market dynamics explicitly. They are limited by this objective to use only the data that is available prior to transfers. However, unlike market dynamics modelling, they can be used for financial decision making on an individual player basis. In [14] authors use linear regression to estimate transfer fee values based on performance and popularity metrics between 2009 and 2014 for top-5 European leagues. They introduce the usage of popularity metrics into such models, however they use the total number of Wikipedia, Facebook and Google metrics at the time of data collection causing retrospective data leakage. Furthermore, they do not perform cross-validation to evaluate their model citing worries about data leakage to prior seasons ignoring out-of-time cross-validation methodologies [15]. Their approach does not perform better than Transfermarkt crowd-sourced estimations.

In [16], authors expand the modelling data to all competitions in Europe, as opposed to top-5 and perform a similar modelling to [14] using data from the video game Football Manager. They compare multiple regression methods and report the cross-validation metrics for those, then demonstrate increased performance over the Transfermarkt predictions on the training set. In [17] authors take a particle-swarm based approach to estimating player values based on FIFA 20 dataset. The study implements a particle swarm optimisation to perform feature selection, followed by using Gaussian Kernel Support Vector Machines to perform regression on market values instead of transfer fees.

2.2 Market Values vs. Transfer Fees

The majority of the studies in the literature concern themselves with the *market values* of the players instead of transfer fees. Market values of the players are speculative values that reflect the perception of the outsiders, whereas the transfer fees are realised as a contractual obligation. As an example, Lionel Messi, one of the most valuable players in the world got transferred from FC Barcelona to Paris St. Germain in August 2021. At the time of the transfer his market valuation was 80 Million €. However, since his contract expired, the transfer was a *free* transfer. Therefore, while there is a correlation between market values of players and transfer fees, market value is not a perfect substitute for transfer fees. Despite the limitations of market values, they provide information about the value of the asset and allow the clubs to evaluate the asset value versus asset price.

In [18], authors perform a factor analysis to investigate the components of market valuation of players. They identify three orthogonal components that impact the valuation of the players: team talent, club performance and external factors.

In [19], authors investigate the effect of age on player market values. The use a linear regression to model the market value with respect to age and age squared while controlling for position, league and season. They find a significant relative age effect. The same effect was also observed in previous studies [13, 20]. Furthermore, market values are dependent on additional factors that are coupled with player age and purely linear models are not well-suited to performing this analysis.

Both [18] and [19] are exploratory studies. This study is concerned with the predictive approaches to player valuation. Therefore [21] and [22] are relevant studies as the market value counterparts to this study. In [21] authors model the dataset provided by FIFA to model player market values obtained from Transfermarkt [23]. They compare linear regression, Decision Trees and Random Forests [24] to show that the non-linear models are better suited to modelling the market values with a complex performance dataset with 70 features. In contrast, [22] use Neural Networks to perform similar modelling using the OPTA [9] dataset. They model the performance data for the Turkish Super League to predict the player market values using a shallow Neural Network (single hidden layer). The predictions on the test set using this model and actual values were found to have a negative correlation of -0.01994.

This study aims to model transfer fee transactions as opposed to market values, therefore we include the aforementioned papers for sake of completeness of literature review.

3 Methodology

In this paper we use player performance and performance of the selling team, current perceived market value, historical transfers, player visibility, player demographics, appearances and injuries to arrive at a comprehensive transfer fee estimation model. To cover as many factors as possible, we first collect data from the following sources: WyScout [8], Transfermarkt [23], Google Trends [25]. After which we perform matching between individual datasets to enrich the data. The collected data is used in feature engineering to reflect interactions between variables as well as comparison of players within a single variable. We develop multiple machine learning models to account for imbalance in the market dynamics and finally ensemble them to arrive at a well-rounded predictive

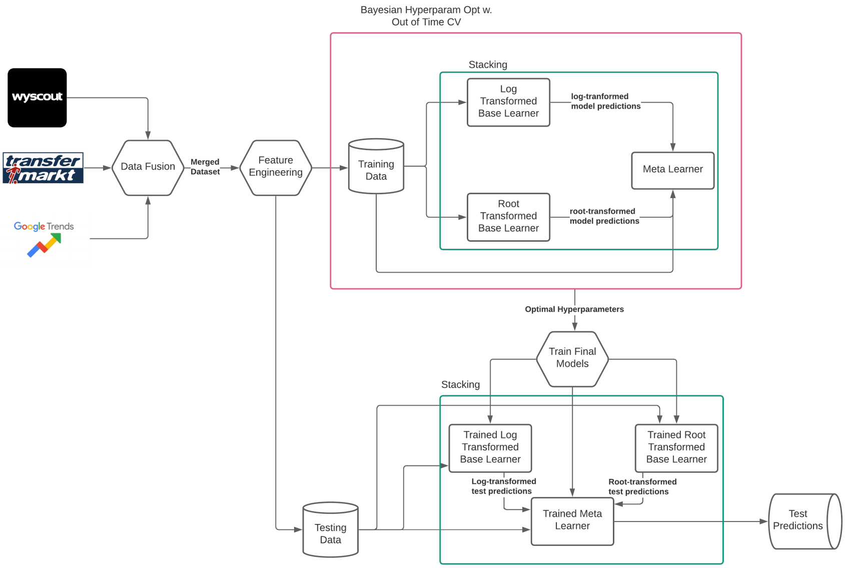

Bayesian Hyperparam Opt w. Out of Time CV
Stacking
wyscout
transfer markt
Google Trends
Data Fusion
Merged Dataset
Feature Engineering
Training Data
Log Transformed Base Learner
log-tranformed model predictions
Meta Learner
Root Transformed Base Learner
root-transformed model predictions
Optimal Hyperparameters
Train Final Models
Stacking
Trained Log Transformed Base Learner
Trained Root Transformed Base Learner
Log-transformed test predictions
Root-transformed test predictions
Trained Meta Learner
Testing Data
Test Predictions

Fig. 1 Proposed Methodology

model that outperforms existing literature and Transfermarkt benchmark by a large margin. Figure 1 shows the proposed methodology.

3.1 Data Collection

For this study, data is collected from various sources. Primarily, there are three datasets: athletic performance dataset, financial indicators dataset and player popularity dataset.

3.1.1 Athletic Performance Dataset

Athletic performance dataset is obtained from WyScout [8] under an academic use license. This dataset provides detailed log of in-game ”events” and subsequent aggregate statistics built up from the events data. For example, an ”event” would be a player attempting a pass to their teammate, which would have attributes such as the timestamp of the pass, xy location, targeted teammate, whether the pass was successful or interrupted. The events data attributes depend on the type of event. Following from events, aggregate statistics are final performance statistics per match per player (whereas events are per match, per player and *per timestamp*). These aggregate statistics provide information such as successful pass rate or number of shots on target. This data is collected from WyScout REST API using HTTP requests. WyScout also provides club and player metadata. Authors provide a sample of this dataset in [2]. Currently the dataset includes performance data for 186351 male adult football players from 613 domestic league competitions.

3.1.2 Financial Indicators Dataset

This dataset contains the details of publicly known transactions in the football market. The dataset is obtained from Transfermarkt [23] via web scraping using BeautifulSoup [26]. The website also provides public perception of how valuable a player is. Given the discrepancy of the number of fans in each country, and the number of fans per club, the public estimation of values is expected to be more accurate for popular leagues and players. Currently, there are 106773 players in the dataset, out of which 77687 have market value information. Out of this 77 thousand players, 12683 have been subject to transfers or loans with a fee. Final training dataset has 10680 paid transfers that fall into the time-frame the study is run for.

3.1.3 Player Popularity Dataset

Finally, to capture the change in public opinion, Google Search Trends [25] data is collected for the past 5 years at the time of the analysis using Google Trends REST API. Each player's search trend is obtained in conjunction with the search for their respective roles in the domain of European football. The following query provides a sample for obtaining Ronaldo's relative search popularity compared to his role "forward" in the past 5 years. The collected dataset has the daily trend for the player search, the role search and the ratio between those for 12876 players, resulting in 23.5 million trend data points for 3 features.

```
{"exploreQuery":"cat=294&q=forward,
Ronaldo&date=today 5-y,today 5-y"}
```

3.2 Data Fusion

First, we obtain in-game performance statistics from WyScout who provide a sample set of statistics for research purposes. To model the transfer fees, the transfer information must be collected in addition to performance statistics. However, this data is not reliably available from performance data providers. Instead, the industry standard for such information is Transfermarkt. The players and teams collected from WyScout and Transfermarkt are matched using the similarity between entity names and additional properties. For players, we use their demographic information to narrow down the search space. For teams, we use the rosters and players' final teams in both datasets to confirm matching. This matching is not straightforward due to difference in naming in both datasets. For instance, Paris St. Germain, one of the world's best teams, is named PSG in WyScout dataset, whereas in Transfermarkt full name is used. Therefore, simple string matching is not feasible and additional information must be used for ensuring data quality. Additional data is collected from Transfermarkt to enrich the dataset, the full list of information collected is provided in Section 4.

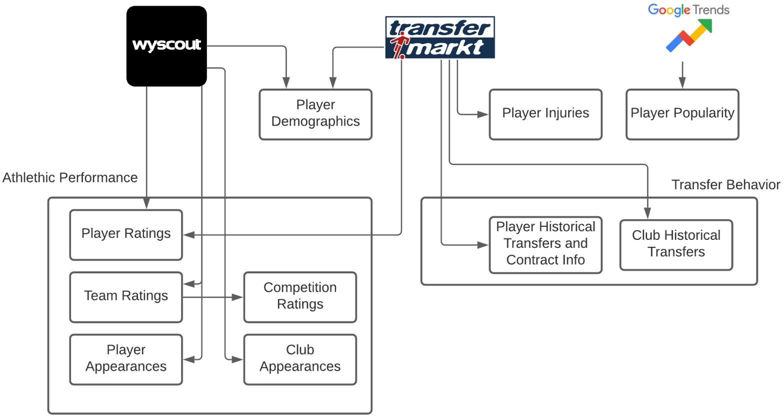


Fig. 2 High-Level Feature Engineering

3.3 Feature Engineering

To extract maximum information from the available datasets, we perform feature engineering to cover the following main factors that affect players' transfer values: Historical transfer behaviour of the club who owns the contract of the player, transfer history of the player, team quality, competition quality and team appearances in various competitions. In addition, we use player-specific properties such as their age, age squared, birth and passport areas, their prominent foot, their main role and the role distribution (i.e. frequency distribution of the matches they played in each role), player injuries, player appearances in various competitions and player popularity. The final dataset contains 220 predictive features for 80000 players in the world, for each of their paid transfers. In relevant sub-sections we provide the details of engineered features where the process includes integration of different datasets.

Figure 2 shows the high-level families of calculated features. All families are explained in the following sections except player demographics which consist of facts about player at the time of transfer such as their age or their nationality. As such, these are not engineered features but rather look-up values.

3.3.1 Athletic Performance

Typically, player performances are analysed in isolation with their competition or similar leagues. Comparing players from two different leagues could be straightforward, however, the problem becomes significantly more challenging when comparing all players from all competitions and divisions in conjunction. Similarly, player's current club and their performance are also important. To address this issue, we employ ELO ratings [27] to quantify both club and competition performance. The ELO (ξ) update for team λ facing team $\neg\lambda$ is given in Eq. 1 where K is the update parameter and S_λ is the expected score for team λ. The higher the K, the more reactive the algorithm to the new scores.In [28], authors provide a discussion on the selection of K-factor and explain the effects of different parameter values. Based on their findings, a parameter value of 64 represents higher uncertainty than average. Since the aim is to evaluate teams and players on a large scale, this uncertainty is a desired property. Hence, in this study, we use $K = 64$ which makes the algorithm reactive to new scores to reflect a particularly good or bad club performance. The ELO ratings of clubs are further transformed to arrive at player performances, therefore a less-reactive value would not yield enough separation after aggregation.

$$\xi'_\lambda = \xi_\lambda + K(S_\lambda - \frac{1}{1 + 10^{(\xi_\lambda - \xi_{\neg\lambda})/400}}) \tag{1}$$

Team quality is represented using teams' ELO [27] ratings at the beginning of the transfer season for each transfer season. As ELO is an incremental algorithm that combines historical information as well as the latest performance, there is no need to average team's ELOs within the season. Using the latest rating at the end of the season is enough to reflect the general behaviour in the season and competition athletic quality can be represented as the average ELO of the clubs in the competition given in Eq. 2 where ξ_λ is the ELO rating of team λ and N_c is the number of teams in competition c.

$$AvgCompRating_c = \frac{\sum_{\lambda=1}^{N_c} \xi_\lambda}{N_c} \tag{2}$$

Players' individual in game statistics are scaled using their play-time, their competition's average ELO, the recency of their games and the ratio between the ELO of their opponent and the ELO of their own team. The recency of the player's games and relative club strength are calculated using Eq. 3 and Eq. 4 respectively.

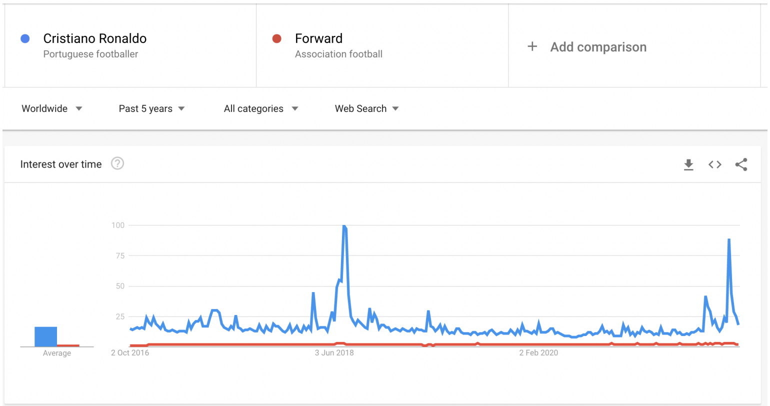
$$RecencyAdjustment = \alpha^{\Delta t} \tag{3}$$

$$RelativeClubStrength_\lambda = \frac{\xi_{\neg\lambda}}{\xi_\lambda} \tag{4}$$

where α is a real-valued number $0 < \alpha \leq 1$ and Δt is the time difference in weeks from the latest game to the date of analysis, $\xi_{\neg\lambda}$ is the opponent rating for the respective match and ξ_λ is the rating for the team λ.

In-game player statistics, adjusted to reflect these factors via scalar product are then combined into a player rating using Cosine-Kernel PCA as described in [29].

This takes major factors that affect performance into account on an individual match level for all players. Resulting player rating vector is shown to correlate with change in player values pre and post analysis. This approach maps the high-dimensional analysis factors obtained through statistical distribution of performance and contextual features into a new vector-space to rate all globally active players under various conditions.



Cristiano Ronaldo
Portuguese footballer

Forward
Association football

Worldwide
Past 5 years
All categories
Web Search

Fig. 3 Search Term Frequency for Cristiano Ronaldo and Forward in Conjunction

Finally, another factor used to quantify athletic performance is club and player historic appearances in various competitions (i.e. national team, international leagues and cups). The frequency of appearances are used as predictors, as well as players' relative performance compared to the average appearance of their teammates and to the average of players in the same role.

3.3.2 Player Popularity

Player popularity is a major factor in determining transfer value in modern football. To include popularity measures, we collected 5-year long web search data for main football roles, and the players in comparison from Google Trends. We link this data to our two former datasets by using player names as search terms in conjuction with their high level roles. The search trend for their role, search trend for the specific player and the ratio between two time-series are used as features in modelling.The ratio between the player popularity and the role popularity quantifies the relative popularity of the player compared to the popularity of the role. Figure 3 shows the relative time-series of Ronaldo and his respective position as a player. Throughout the history, Ronaldo has been a more popular search term than his role. However, at times corresponding to his transfer, his relative popularity has sky-rocketed (July 2018 to Juventus and Aug 31st to Manchester United).

3.3.3 Player Injuries

Transfermarkt provides 343 unique injuries for all active players. These injuries cannot be used directly. Instead, we group injuries based on how frequently they occur in players, and how severe the impact is. The frequency of an injury is defined as number of unique players experiencing the injury. The severity of the injury is defined as median duration the injury causes the player to miss matches. We group these statistics into decentiles. First decentile being the least common/severe and the 10th decentile being the most common/severe injury. We opt to use decentiles instead of clustering to avoid clusters of few-values as these statistics are skewed. The most severe injuries tend to happen very infrequently and vice versa.

3.3.4 Transfer behaviour

Past transfer behaviour is also an important factor in transfers. This behaviour can be quantified on a player and on a team level. For instance, teams who bought multiple players in the previous season may be more inclined to sell players in the current one due to financial restrictions. Similarly, teams who have loaned players in the previous season may want to purchase the players they bought or may want to replace them. Furthermore, teams who consistently buy or sell players may have the financial liquidity to continue the transactions. To avoid data leakage, the transfer statistics are only calculated for player's club right before the transfer.

From a player perspective, certain players may be consistently loaned historically for their development and certain players historically could be deemed too valuable to release unless transferred for large sums.

In general, there are two main types of aggregations that quantify transfer behaviour: Frequency of type of transfers and total fee spent on type of transfers. These aggregations are performed on two main time scales ts: latest finished season and entire time scale. Similarly, the same aggregations are performed for both player's current club and the player themselves. The transfer types tr in consideration are paid transfers, free transfers, paid loans and free loans. Transfer spending is only computed for paid transfers and paid loans.

Eq. 5 and Eq. 6 respectively show frequency and spending aggregations.

$$TransferFrequency(tr, ts) = \sum_{i=1}^{N_{tr,ts}} 1 \tag{5}$$

$$TransferSpending(tr, ts) = \sum_{i=1}^{N_{tr}} Fee(tr, ts) \tag{6}$$

3.4 Modelling

Transfer market is a highly unfair market where the transfer fee distribution exhibits a long tail shown in Figure 5. There are a few cases where the transfer fee is in range of millions of Euros, and a lot of cases where transfer fee is very low. Typically, this type of data is treated by applying a normalising transformation such as log transformation to the dataset and performing modelling as such. This results in a model that on average predicts accurately. In other words, normalisation of the skewed data allows the models to avoid being skewed by outliers such as high transfer fees and predict the behaviour of the high-frequency samples better.

However, in case of player transfers, predicting high-value transfers accurately has a higher financial impact on club use-cases. Yet, majority of the transfer market operates with low fees so these data points must also be represented. Therefore, to cover different use-cases, we opt to build multiple models and then ensemble their predictions afterwards to arrive at finer-tuned estimations.

In this paper, we build three LightGBM models using different transformations of transfer fees as outcome variables. We minimise Root Mean Squared Error (RMSE) of predicted players' transfer values compared to the actual transfer fees. One property of RMSE is that it is sensitive to outliers. Typically, this would be an undesired property in an error metric, however, combining different levels of normalisation and stacking with RMSE optimisation allows us to cover various valuation cases.

As shown in Figure 5, transfer fees have a long tail. Log-transformation normalises the data and allows model to estimate the usual cases. Root transformation normalises the data slightly less and allows the model to focus on cases that happen more infrequently than usual but still mitigates the impact of outliers. Finally, in order to model the 'superstars', we apply no transformation to the target variable and use it as is. Using these three outcome variables, we build LightGBM models and used a LightGBM meta-learner to stack predictions.

Model ensembling is an advanced predictive modelling technique that aims to extract different information captured by multiple models and combine them together. There are three main types of model ensembling: Bagging [30], boosting [31] and stacking [32].

Bagging and boosting use simple aggregation and weighting methodologies to combine information coming from different predictions. In contrast, stacking uses a machine learning model on top of the predictions to combine the results in a more intelligent fashion. This model is called a meta-learner. In this study, we use this technique to extract the maximum amount of information from the base-learners to capture the different dynamics in transfer market. We elaborate on this further in Section 4.

4 Experimental Setup and Results

To model transfer values, we collect and fuse data for 70383 male football players from 85 countries and 174 competitions between 1 January 2016 and 1 June 2020, when the summer transfer season starts. From Transfermarkt we collected 141121 number of transfers involving these players, out of which 10680 were paid transfers with fee information. The rest were free transfers, or loans. For the paid transfers, we computed the features as described in Section 3. The final dataset has 224 predictive variables. We built three LightGBM regression models: without any transformation, applying root transformation and log transformation of the outcome variables (Figure 5).

To perform hyperparameter optimisation for base learners, we split into a modelling and hold-out set. To account for out of time dynamics, we use the transfer data after January 2020 as the hold-out set. To account for out

Table 1 Hyperparameter Ranges

Hyperparameter	Minimum Value	Maximum Value
Learning Rate	5e-2	1e-1
Number of Estimators	10	500
Subsample	0.9	1
Minimum Child Samples	5	15
Maximum Depth	3	9
Column Sample by Tree	0.9	1
Number of Leaves	3	1000
Maximum bins	255	255

Table 2 Runtime Statistics

Stage	Average Runtime Duration (hh:mm:ss)	Standard Deviation (hh:mm:ss)
Feature Engineering (Training)	01:10:47	00:05:10
Feature Engineering (Prediction)	00:28:36	00:00:35
Hyperparameter optimisation and Model Training (Incl. data read)	00:44:01	00:13:05
Prediction (Incl. data read)	00:06:05	00:00:14

of sample dynamics, we randomly select players using stratification on player market value as the out-of-sample hold-out set. These two hold-out sets are combined into a single set and the model metrics are reported on this set which is never used in modelling.

Using the modelling subset, we perform Bayesian hyperparameter optimisation with 4-Fold out-of-time cross-validation for three LightGBM models as shown in Figure 4. Out-of-time validation is selected as the proper cross-validation method because we assume that the transfer fees are subject to inflation, therefore a random-split would introduce data leakage. In other words, in case of traditional K-Fold split, the samples that are randomly selected for training would contain information of the inflation present at the time, therefore inflating validation results and negatively impacting model generalisation.

The hyperparameters subject to optimisation and their optimal values per model are given in Table 1. Table 2 shows the average runtime of the stages of the proposed solution. All raw data and generated features are stored in a PostgreSQL 11 database with 26 GB memory and 4 vCPUs. The model training and inference are performed on a virtual machine on Google Cloud Platform with 52 GB memory and Intel(R) Xeon(R) CPU @ 2.30GHz 8 vCPU cores, running Ubuntu 18.08 OS.

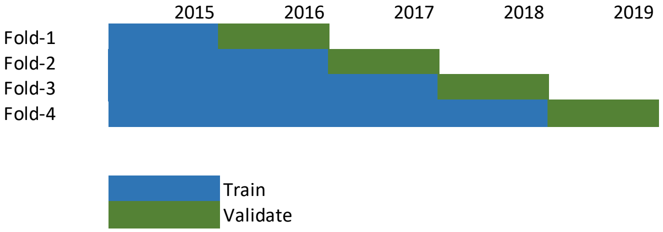


Fig. 4 Cross-Validation Folds

Each model is trained in a parallelised setting, which treats the features separately and combines them in a tree-building process. Therefore, each LightGBM model has a linear computational complexity of O(0.5 * #feature * #bin). The number of bins (#bin) refers to the discretisation of the numeric features into several bins for building the decision trees and is determined by the maximum number of bins parameter.

The outputs of the three models corresponding to three outcome transformations are then ensembled using an additional LightGBM meta-learner, which is also trained using the same hyperparmeter optimisation procedure.

The performance metrics of different models on different levels of transfer fees are reported separately in Table 3. This table shows the test set RMSE (in million) of predictions made by base models, the final stacked prediction and predictions of TransferMarkt in decentile groups of player transfer fees. For clarity we also report the number of players in each transfer fee group. Due to the magnitude of the values in different groups, the errors are not distributed normally. Given this non-normality, we use Wilcox paired two sample test to compare the errors of the proposed methodology to the baseline (Transfermarkt). Under Wilcox paired two-sample test [33] with a confidence level of 0.05, we are able to reject the following null hypothesis with $p = 0.02441$ in favour of the one-sided alternative hypothesis. The comparison of median RMSE of the proposed methodology ($RMSE_{PM}$) and median RMSE of the Transfermarkt is tested formally as follows. ($RMSE_{TM}$).

As shown in Table 3 and Table 4, base learners with different transformations capture different dynamics in various levels of transfer sizes. Ideally, base-learners should capture non-overlapping information in the dataset. In most cases, some overlap in base learner information is unavoidable, especially with a limited dataset. In our case, root transformed base-learner captures the overall transfer performance fairly well while log transformed base-learner specialises in low transfer values. And the base-learner with no transformation is affected by significant outliers in the dataset, therefore specialises in high-value transfers.

Table 3 shows that the log transform captures the transfers with low fee better while the root transformed model has overall the best predictive behaviour

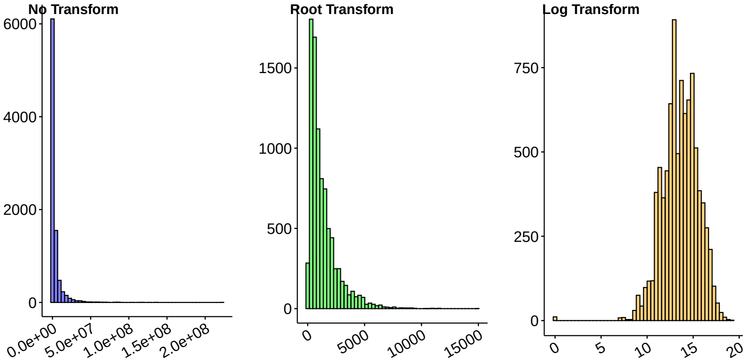


Fig. 5 Transformations Applied to Outcome Variable

Table 3 Prediction Performance on Test Set (RMSE in Million €) (N=3317)

Market Value Range	N	Base Learner No Transform	Base Learner Root Transform	Base Learner Log Transform	Stacked Model	TransferMarkt Predictions
≤ 0.07M€	345	0.43	0.25	0.15	0.06	0.26
≤ 0.17M€	338	0.62	0.44	0.38	0.70	0.55
≤ 0.3M€	382	0.80	0.52	0.41	0.29	0.52
≤ 0.5M€	362	0.86	0.63	0.47	0.24	0.96
≤ 0.85M€	276	1.12	0.84	0.66	1.11	0.97
≤ 1.5M€	412	1.70	1.27	1.05	0.58	1.62
≤ 2.3M€	278	1.67	1.48	1.30	1.62	1.40
≤ 4M€	349	2.17	1.99	2.19	2.36	2.52
≤ 8.5M€	319	3.59	3.34	3.85	3.11	4.05
≤ 220M€	317	11.12	10.79	16.59	8.65	10.38

Table 4 Learner and TransferMarkt Performance in Transfers Between 8.5 - 50 Million €(RMSE in Million, N = 300)

Learner	Error
Base Learner - No Transform	8.63
Base Learner - Root Transform	8.72
Base Learner - Log Transform	12.79
TransferMarkt Predictions	9.33

of base models. The root transformed base learner also outperforms the Transfermarkt predictions consistently. The base learner without transformation does not perform well in terms of RMSE, however as shown in Table 4, base learner with no transform outperforms the rest of the predictions when the transfer fee is between 8.5 and 50 Million €. We opted to exclude this model as a base-learner because the difference between the average error of root transformed model and model with no transform applied is statistically insignificant ($p = 0.55$).

Table 5 Top-10 Closest Predictions Above 500,000 €

Name	Transfer Date	From	To	Transfer Fee	Proposed Method Predictions	TM Predictions
S. Haller	2021-01-08	West Ham	Ajax	22.5	23.5	30
M. Busi	2020-10-05	RSC Charleroi	Parma	7.5	7.68	2.5
A. Tchouaméni	2020-01-29	G. Bordeaux	Monaco	18	17.85	14
Maicon	2020-08-01	Galatasaray	Al-Nassr	1.43	1.48	4.8
O. Maritu	2020-02-17	SX Chang'an At.	SJZ Ever Bright	3.94	3.86	0.95
N. Okafor	2020-01-31	FC Basel	RB Salzburg	11.2	11.59	8
S. Sosa	2021-02-12	River Plate	Atlanta United	4.95	4.95	7.5
A. Zeqiri	2020-10-01	Lausanne-Sport	Brighton	4	4.03	1.5
L. Bittencourt	2020-07-15	TSG Hoffenheim	Werder Bremen	7	6.87	4.8
S. Ristovski	2021-02-02	Sporting CP	Dinamo Zagreb	1	0.96	3
S. Weissman	2020-08-31	Wolfsberger AC	Real Valladolid	4	4.06	6
Tao Qianglong	2020-02-27	HB CFFC	DL Pro Res.	2.61	2.51	0.6
Gilberto	2020-08-08	Fluminense	Benfica	3	2.85	1
L. Dykes	2020-08-19	Livingston FC	QPR	2.2	2.18	0.48
C. Domínguez	2020-08-24	Independiente	Austin FC	2.27	2.16	4
E. Rigoni	2021-05-26	Elche CF	São Paulo	1.8	1.72	3.2
A. Băluță	2020-07-31	Slavia Prague	Puskás AFC	0.7	0.71	2
Breno	2020-11-12	Juventude	Palmeiras	1.5	1.54	0.3
B. Kouyaté	2020-08-25	Troyes	FC Metz	3.5	3.33	2.2
R. Centurión	2020-07-24	Racing Club	Vélez Sarsfield	1.4	1.38	2.4

To demonstrate the performance of the final model, we show the error distribution of final predictions on the test set compared to TransferMarkt predictions in Figure 6. The figure shows that the proposed methodology improves error over Transfermarkt predictions across all value groups. However, in the group with largest transfer values, the long tail of the error exceeds Transfermarkt's. This is consistent with the summary statistics such that while the proposed method predictions are better overall in all groups, the largest value transfers have higher error.

We also provide well-performing and badly-performing predictions of the proposed methodology in Tables 5 and 6. Predictions and fees are reported in millions of €. These predictions are further discussed in Section 5.

Finally, the relative importance of features for top-20 features used by the meta-learner is given in Figure 7. The most important features are extracted from the LightGBM meta-learner importance score which quantifies how often the feature was utilised in building the individual decision trees in the algorithm. The values show that the meta-learner utilises the predictions from the root-transformed base learner most frequently. The Transfermarkt market value and predictions of log-transformed base learner are also utilised frequently in the final model. In addition, historical club transfer spendings, player's age at the time of transfer and average player market value of the competition player is playing in before the transfer are also significant. Diminishingly, player performance metric on shot assists and player's popularity also provide information. Overall, the most impactful features seem to be the features related to the market and the transfer fees, whereas athletic performance metrics have a secondary impact on market value. This is in-line with prior

Table 6 Bottom-10 Worst Predictions Above 500,000 €

Player	Transfer Date	From	To	Transfer Fee	Proposed Method Predictions	TM Predictions
M. Pjanić	2020-09-01	Juventus	Barcelona	60	11.53	45
Arthur	2020-09-01	Barcelona	Juventus	72	24.21	56
T. Partey	2020-10-05	Atlético Madrid	Arsenal	50	11	40
V. Osimhen	2020-09-01	LOSC Lille	SSC Napoli	70	13.71	40
B. Chilwell	2020-08-26	Leicester	Chelsea	50.2	15.67	40
Álvaro Morata	2020-07-01	Chelsea	Atlético Madrid	35	11.51	36
D. van de Beek	2020-09-02	Ajax	Man Utd	39	12.79	44
Rúben Dias	2020-09-29	Benfica	Man City	68	14.42	35
J. David	2020-08-11	KAA Gent	LOSC Lille	27	9.09	25
Diogo Jota	2020-09-19	Wolves	Liverpool	44.7	12.69	28
P. Schick	2020-09-08	AS Roma	Bay. Leverkusen	26.5	11	25
Allan	2020-09-05	SSC Napoli	Everton	25	9.84	28
S. Sensi	2020-09-01	Sassuolo	Inter	20	7.9	20
A. Doucouré	2020-09-08	Watford	Everton	22.1	7.98	20
S. Dest	2020-10-01	Ajax	Barcelona	21	6.05	18
Gabriel	2020-09-01	LOSC Lille	Arsenal	26	8.08	20
D. Lovren	2020-08-01	Liverpool	Zenit S-Pb	12	0.28	12
T. Castagne	2020-09-03	Atalanta BC	Leicester	20	6.34	18
M. Doherty	2020-08-30	Wolves	Spurs	16.8	4.36	16
Lucas Paquetà	2020-09-30	AC Milan	Olympique Lyon	20	8.57	20

descriptive studies that modelled the transfer market from a dynamics perspective. The list of all features used, their definitions and families are provided as supplementary material (Feature Definitions).

4.1 Comparison With Existing Literature

We compare our results to [14], [16], [17] and Lasso Regression Baseline in Table 8. Direct comparison to methods is not possible due to authors reporting different performance metrics (if any) in their studies. Therefore, we opted to apply the published methodologies to our dataset for a fair comparison where the error metric is different. The two most similar studies [16] and [14] which model the transfer fees report comparable metrics so their results are provided as is. However since authors predict the clusters of market values through Particle Swarm SVM [17], their evaluation metric is not directly translated to the improvement in prediction error, therefore we implement the method from scratch, using the same set of parameters reported in the study.

As most studies [3, 7, 12–14] in the literature use a variant of linear or regularised regression, we also implement a so-called 'baseline' to compare the model performance to a traditional regression model. However, due to high-dimensionality of the dataset in this study, we opted for Lasso regression instead of a non-regularised regression model. For [17] and regression baseline, the methods are implemented from scratch to work with our dataset, as the datasets for these are not directly available. To compare with [17], the particle swarm optimisation, hyperparameter optimisation was initialised with parameters provided in Table 7. In this table, $c1$ and $c2$ are so-called 'trust parameters' that control the acceleration of the particles. If both are 0, particles scatter in constant speed, until they reach the end of the search space. The parameter w is the inertia parameter that further controls the movement

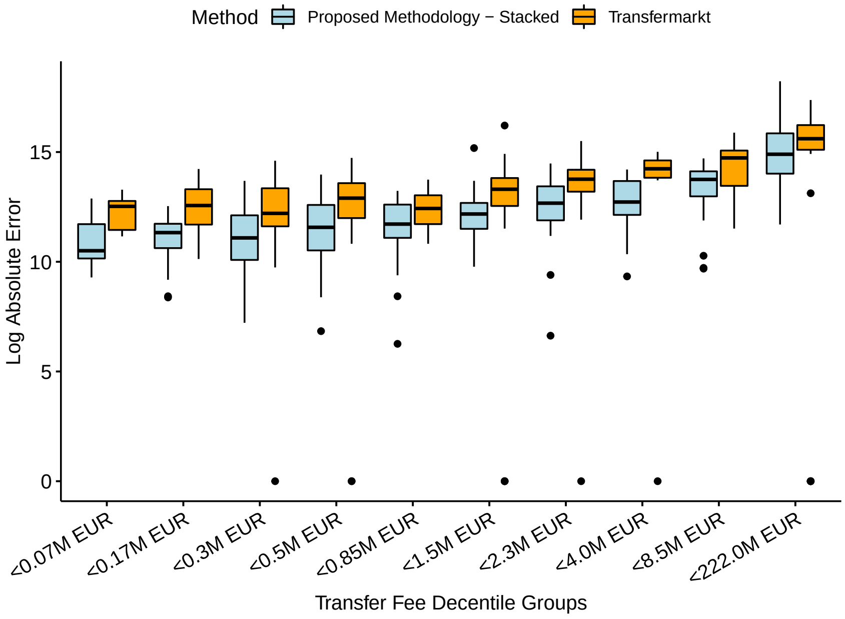


Fig. 6 Error Distributions on Test Set Per Transfer Fee Buckets

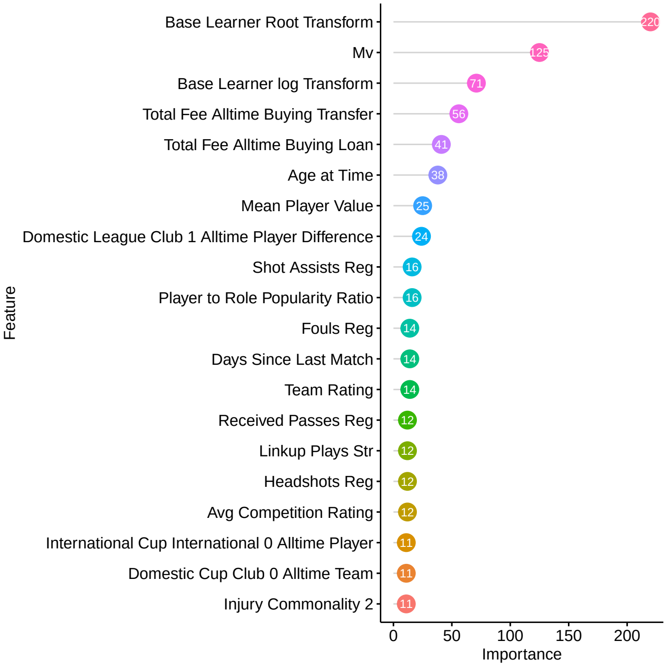


Fig. 7 Most Important 20 Features

Table 7 POS Initial Parameter Values

Parameter	Value
c1	0.5
c2	0.5
w	0.9
k	30
p	2

Table 8 Comparison to Existing Literature

Method	% Error Change
Proposed Methodology	-6.12%
Lasso Regression Baseline	+6.89%
Muller et al. [14]	+3.6%
Yigit et. al [16]	-5.08%
Behravan et. al [17]	+23.85%

behaviour of the particle. The parameter p is the Minkowski norm that specifies which distance metric will be used. If p=1 the algorithm uses $L1 - norm$, if $p = 2$ the algorithm uses $L2 - norm$ (Euclidean distance). The number of particles to launch is held constant at 30, and L2-Norm is used as distance metric during hyperparameter optimisation.

To perform comparison, we compute the RMSE of Transfermarkt market value predictions compared to the actual transfer fees, and compute the RMSE of predictions of all methods and report the improvement over TM error (i.e. Error Change) in Table 8. In case of improvement over TM predictions, the error change is negative. Therefore the lower the error change, the better. Equation 4.1 shows the Error Change (EC) between RMSE of the proposed methodology ($RMSE_{PM}$) and RMSE of the Transfermarkt ($RMSE_{TM}$).

$$EC = \frac{100 * (RMSE_{PM}(y, \hat{y}) - RMSE_{TM}(y, \hat{y}))}{RMSE_{TM}(y, \hat{y})} \tag{7}$$

The results of [17] applied to the dataset of this study appear to be an outlier. Their dataset is small with fewer features. The kernel is used to map the data into a linearly-separable hyper-space, however, this is not applicable to such a complex dataset as in this work, therefore the model diverges.

4.2 Discussions

Table 6 shows the transfers where the methodology does not perform as well as Transfermarkt predictions. This table is dominated by the transfers in English Premier League (EPL), which is known to have high transfer fees and arguably is the biggest spectator country when it comes to football. This finding is an

Table 9 Winter 21/22 Transfer Predictions

Name	Predicted Value (mil. €)	Actual Value (mil. €)	Selling Club	Buying Club
Dusan Vlahovic	51.8	78	Fiorentina	Juventus
Ferran Torres	60.4	55	Manchester City	Barcelona
Luis Diaz	24.2	43	Porto	Liverpool
Bruno Guimares	34.4	40.3	Olympique Lyon	Newcastle
Lucas Digne	25.4	30	Everton	Aston Villa
Chris Wood	**0.6**	**30**	**Burnley**	**Newcastle**
Vitaliy Mykolenko	14.8	22.5	Dinamo Kyiv	Everton
Rodrigo Bentancur	28.4	20.2	Juventus	Tottenham Hotspurs
Yuri Alberto	7.2	19	Internacional	Zenit St. Petersburg
Julian Alvarez	16.7	16.3	River Plate	Manchester City

unexpected outcome of the study. This systematic error, which can be corrected by heuristic post-processing, raises the question of whether the transfers into EPL are overpriced.

In addition to comparing and contrasting the methodology with the existing methods using historical data for overlapping time-frames, we applied the proposed methodology and predicted the fees in the 21/22 Winter transfer window, to demonstrate its applicability to real-world scenarios. Table 9 shows the predictions and the real-world transfer fees for the Winter transfer window of 21/22 transfer season for major transfers. This table highlights an interesting phenomenon. The transfer of Chris Wood from Burnley to Newcastle is widely regarded as a strategic move by Newcastle's new owners to weaken their direct competitor Burnley [34, 35]. At the time of the transfer, both clubs were in the delegation zone. By transferring a staple player from Burnley, Newcastle both weakened their opponent and strengthened their lineup. However, purely based on the player's performance, Chris Wood is not regarded as valuable as the 30 million € Newcastle paid for the player. While being a completely legal transaction, this transfer is a financial anomaly that is driven by additional club objectives. In addition to the obvious use-cases of the proposed methodology for budget and profit planning, it also has applications in detecting anomalous transactions, allowing the governing bodies to examine these transactions further for financial fair play purposes.

5 Conclusion

In this study we propose an ensembling approach to estimating the transfer fees in association football. In contrast to the existing literature, the proposed approach is able to perform transfer fee estimation across the globe, for all active players. In addition, the proposed methodology uses real-world in-game statistics, as opposed to data obtained from games, transfer data, player popularity metrics obtained from Google Trends and performs data fusion and enrichment to arrive at a comprehensive dataset.

One important contribution of the proposed methodology is using the stacking approach in modelling transfer fees to arrive at more finely-tuned individual predictions. Furthermore, we show that different mathematical transformations affect predictions in different ways and even predictions that are not particularly accurate can be used in an informative fashion through advanced machine learning techniques. The proposed methodology outperforms both the defacto industry standard TransferMarkt predictions, as well as the existing methodology in the literature.

Another contribution of the methodology compared to the existing exploratory analyses in the literature is that it does not require knowledge from buying clubs. This allows the approach to work in a predictive setting instead of performing factor analysis retrospectively, however, it also comes at the price of omitting an important factor of buying club financial details, which play an important role in determining the final transaction fee. To address this situation, the proposed methodology could be expanded to include the buying club characteristics and to maintain the predictive application of the proposed solution, it could be incorporated in a simulation that bootstrap samples buying club characteristics and estimates the transfer fee in different situations.

In conclusion, the proposed methodology could be used to estimate an appropriate transfer fee for players. The methodology can guide the club professionals to make financially informed decisions on whether a prospective player is worth their fee, or to decide on the appropriate fee for selling a player who has drawn interest from other clubs. The proposed methodology also demonstrates the impact of different data transformations on the predictive capabilities of the models. The general consensus in the machine learning community is transforming the output variable to be as close to Normal distribution as possible, however we demonstrate that for some use-cases, long-tailed distributions are more suitable and the choice of data transformation must depend on the use-case. Further formal analysis is needed to determine the best-practices in data transformation in a practical setting, however this is out of the scope of this study.

6 Future Work

There are limitations of the proposed methodology. As illustrated in Tables 5 and 6 model predictions fall short in case of high-profile transfers. As future work, we aim to study this problem and improve upon the predictive limitations of the proposed approach.

Furthermore, paid transfers make up a very small portion of the entire transfer market. As a prior to estimating expected value of transfers, we must have an idea about the likelihood of the paid transfer as opposed to free transfers or loans. Lack of this probabilistic information limits the usefulness of predictions of fees. Without this information, it is virtually impossible to estimate the amount of financial investment required, as well as the return on investment reliably. The use of similar advanced machine learning techniques might provide value in classifying the most likely type of future transfer, as well as the most likely time of the future transfer for players.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] McHale, I.G., Scarf, P.A., Folker, D.E.: On the Development of a Soccer Player Performance Rating System for the English Premier League. Interfaces **42**(4), 339–351 (2012). https://doi.org/10.1287/inte.1110.0589

[2] Pappalardo, L., Cintia, P.: Quantifying the relation between performance and success in soccer. Advances in Complex Systems **21**(03n04), 1750014 (2018)

[3] Dobson, S., Gerrard, B.: The Determination of Player Transfer Fees in English Professional Soccer. Journal of Sport Management **13**(4), 259–279 (1999). https://doi.org/10.1123/jsm.13.4.259. Publisher: Human Kinetics, Inc. Section: Journal of Sport Management. Accessed 2021-01-21

[4] Gerrard, B.: Analysing Sporting Efficiency Using Standardised Win Cost: Evidence from the FA Premier League, 1995 – 2007:. International Journal of Sports Science & Coaching (2010)

[5] Lucifora, C., Simmons, R.: Superstar Effects in Sport: Evidence From Italian Soccer. Journal of Sports Economics **4**(1), 35–55 (2003). https://doi.org/10.1177/1527002502239657

[6] Torgler, B., Schmidt, S.L.: What shapes player performance in soccer? Empirical findings from a panel analysis. Applied Economics **39**(18), 2355–2369 (2007)

[7] Berg, E.W.A.v.d.: The valuation of human capital in the football player transfer market. Master's thesis (August 2011). http://hdl.handle.net/2105/9763

[8] Wyscout: Wyscout. https://wyscout.com/ Accessed 2020-11-12

[9] Opta: Opta Sports. https://www.optasports.com/sports/football/ Accessed 2020-11-12

[10] InStat: InStat. https://football.instatscout.com/ Accessed 2020-11-12

[11] Herm, S., Callsen-Bracker, H.-M., Kreis, H.: When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an

online community. Sport Management Review **17**(4), 484–492 (2014). https://doi.org/10.1016/j.smr.2013.12.006. Accessed 2021-06-10

[12] Kedar-Levy, H., Bar-Eli, M.: The Valuation of Athletes as Risky Investments: A Theoretical Model. Journal of Sport Management **22**(1), 50–81 (2008). https://doi.org/10.1123/jsm.22.1.50. Publisher: Human Kinetics, Inc. Section: Journal of Sport Management. Accessed 2021-01-21

[13] Nsolo, E., Lambrix, P., Carlsson, N.: Player valuation in European football. In: International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 42–54. Springer, Ghent, Belgium (2018)

[14] Müller, O., Simons, A., Weinmann, M.: Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research **263**(2), 611–624 (2017). https://doi.org/10.1016/j.ejor.2017.05.005. Accessed 2021-06-10

[15] Browne, M.W.: Cross-validation methods. Journal of Mathematical Psychology **44**(1), 108–132 (2000)

[16] Yiğit, A.T., Samak, B., Kaya, T.: Football Player Value Assessment Using Machine Learning Techniques. In: Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making. Advances in Intelligent Systems and Computing, pp. 289–297. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-23756-1_36

[17] Behravan, I., Razavi, S.M.: A novel machine learning method for estimating football players' value in the transfer market. Soft Computing **25**(3), 2499–2511 (2021). https://doi.org/10.1007/s00500-020-05319-3. Accessed 2021-06-10

[18] Rodríguez, M.S.: Factor analysis of the market value of high-performance players for three major european association football leagues. Managing Sport and Leisure **26**(6), 484–507 (2021) https://doi.org/10.1080/23750472.2020.1771197. https://doi.org/10.1080/23750472.2020.1771197

[19] Gyimesi, A., Kehl, D.: Relative age effect on the market value of elite european football players: a balanced sample approach. European Sport Management Quarterly **0**(0), 1–17 (2021) https://doi.org/10.1080/16184742.2021.1894206. https://doi.org/10.1080/16184742.2021.1894206

[20] Müller, O., Simons, A., Weinmann, M.: Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research **263**(2), 611–624 (2017)

[21] AL-ASADI, M.A., Tasdemir, S.: Predict the value of football players using FIFA video game data and machine learning techniques, 1–1. https://doi.org/10.1109/ACCESS.2022.3154767. Conference Name: IEEE Access

[22] Inan, T., Cavas, L.: Estimation of market values of football players through artificial neural network: A model study from the turkish super league. Applied Artificial Intelligence **35**(13), 1022–1042 (2021) https://doi.org/10.1080/08839514.2021.1966884. https://doi.org/10.1080/08839514.2021.1966884

[23] Transfermarkt: Transfermarkt. https://www.transfermarkt.com/ Accessed 2020-11-12

[24] Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)

[25] Google: Google Trends. https://trends.google.com/trends/ Accessed 2021-07-13

[26] Richardson, L.: Beautiful soup documentation. April (2007)

[27] Elo, A.E.: The Rating of Chessplayers, Past and Present. Arco Pub., Arco (1978)

[28] Langville, A.N., Meyer, C.D.: Who's #1? Princeton University Press, Princeton (2012). https://doi.org/10.2307/j.ctt7rwdt

[29] Aydemir, A.E., Temizel, T.T., Temizel, A., Preshlenov, K., Strahinov, D.M.: A dimension reduction approach to player rankings in european football. IEEE Access **9**, 119503–119519 (2021). https://doi.org/10.1109/ACCESS.2021.3107585

[30] Hothorn, T., Lausen, B., Benner, A., Radespiel-Tröger, M.: Bagging survival trees. Statistics in Medicine **23**(1), 77–91 (2004)

[31] Ridgeway, G., Madigan, D., Richardson, T.S.: Boosting methodology for regression problems. In: Seventh International Workshop on Artificial Intelligence and Statistics (1999). PMLR

[32] Smyth, P., Wolpert, D.: Linearly combining density estimators via stacking. Machine Learning **36**(1), 59–83 (1999)

[33] Bauer, D.F.: Constructing confidence sets using rank statistics. Journal of the American Statistical Association **67**(339), 687–690 (1972) https://www.tandfonline.com/doi/pdf/10.1080/01621459.1972.10481279. https://doi.org/10.1080/01621459.1972.10481279

[34] Blow, T.: Newcastle's Chris Wood Transfer Explained After Gabby Agbonlahor 'joke' Claim. Section: News. https://www.mirror.co.uk/

sport/football/news/newcastle-sign-wood-burnley-transfer-25933545 Accessed 2022-03-02

[35] Waugh, C., Ornstein, D.: Newcastle Sign Chris Wood from Burnley. https://theathletic.com/news/newcastle-sign-chris-wood-from-burnley/wQF0WcWlwWwz/ Accessed 2022-03-02