The Transfer Market Madness

Determinants and Predictions of Football Player Transfer Fees

Sarah Mohammad, Sophia Nagler, Simon Schölzel, Silvia Triscova

Abstract

In the recent years, excessive transfer fees for football players have turned the European transfer market upside down – a development coined as the 'transfer market madness' in this paper. Predicting those fees is assumed to be of great interest to researchers, policy makers, managers and the public alike. This paper addresses the issue by discussing the following four research questions: (1) What are the important value drivers that determine transfer fees in the European football market? (2) Is there a significant influence of a player's popularity on transfer fees? (3) Which model for estimating football transfer fees performs best in terms of predictive accuracy? (4) Where is the transfer market madness heading to?

For this purpose, five predictive modelling techniques from the regression family are proposed: linear, stepwise forward, Ridge, Lasso and polynomial regression. The models are trained on a rich data set of 2,634 transfers observed during the 2013-2019 period which is scraped from transfermarkt.de, kaggle.com and Wikipedia. The empirical results reveal numerous important predictors for the transfer fee and especially indicate that the median player transfer incorporates a €1,06 million price premium that accounts for a player's popularity. Moreover, the quadratic regression model yields the overall best predictive accuracy with the 1-standard error-rule Lasso model being the least prone to overfitting. Finally, the latter is deployed and evaluated on four currently rumoured transfers. The generated predictions are not only remarkably close to what is hypothesised by the media but also suggest that the transfer market madness is there to stay.

Eventually, the paper advocates the use of novel predictors and consideration of non-linear relationships in future research. From a practical perspective, this study develops a tool that can aid managers and agents in terms of decision-making. In addition, concerns regarding the ethical and psychological implications of the transfer market development are raised leading to the question: When will the excessive costs become unbearable for the average team as well as the players and how will it shape the competitive equilibrium in the future of football?

Keywords: football analytics, transfer fees, player popularity, predictive modelling, regression

Table of Contents

Introduction

Football is a growing sport all over Europe (FIFA, n.d.), and apart from the game itself, the football industry and the businesses built around it draw a lot of attention and motivate various interesting research topics. European football leagues have been generating increasingly higher revenues, largely due to broadcasting deals such as with Sky Sports for the Premier League in England, and other commercial deals with brands such as Adidas and Nike (Barnard, Boor, Winn, Wood & Wray, 2019; Geey, 2019). However, clubs have been allocating an increasing share of these revenues to player salaries and transfer fees to purchase them (Barnard et al., 2019). Some clubs risked long-term financial instability, because football clubs started to spend large amounts of their financial resources towards buying better players. In 2010, the Union of European Football Associations (UEFA) stepped in to take control of overspending and to restore competitive balance among European football clubs by introducing the Financial Fair Play regime (Vöpel, 2011; Preuss, Haugen & Schubert, 2014; Čeferin & Theodoris, 2018). Financial Fair Play provides a set of rules that prevents clubs from having large debt by setting a 'break-even requirement', allowing clubs to only incur as many 'relevant expenses' as they generate 'relevant income' (Vöpel, 2011).

UEFA's effort to regulate the European football transfer market and ensure financial stability was successful, but transfer prices still seemed to increase substantially. Based on data from transfermarkt.de (TM), it is visible that transfer records have almost quadrupled from 2000 onwards (Transfermarkt, 2019). TM estimates market values with the help of community members' judgements, and is heavily used by football enthusiasts to check team and player statistics, as well as to stay updated about rumours on potential football transfers (Müller, Simons & Weinmann, 2017). The increasing transfer prices despite the abovementioned regulation sparked the interest to look further into the prediction of transfer fees, which is the actual value that is paid by a club for a player during a negotiation between two clubs (Herm, Callsen-Bracker & Kreis, 2014). Thus, the goal of this paper is to explore the applicability of five predictive modelling techniques in the context of the European football transfer market. Ultimately, it aims to answer the following four research questions:

RQ1: What are the important value drivers that determine transfer fees in the European football market?
RQ2: Is there a significant influence of a player's popularity on transfer fees?
RQ3: Which model for estimating football transfer fees performs best in terms of predictive accuracy?
RQ4: Where is the transfer market madness heading to?

This paper is structured as follows. First, a review of the relevant literature on the football industry and transfers is performed, together with a review of the econometric techniques applied in this paper. Next, findings from related works are given and the methodology is discussed. Hereafter, the results are elaborated upon and detailed in the discussion. Finally, after stating the implications and limitations, the conclusion summarises this paper and provides an outlook for future research.

1. Conceptual Framework

1.1 Background of the Transfer Market Economy

In order to establish a better understanding of the topic and the related research questions, it is important to first look into existing literature related to the football industry and the player transfers. This section first focuses on the football industry itself. Next, the football transfers and the reasons for their occurrence are investigated. Finally, literature on brand value is linked to the football transfers to see if any relevance of this on transfers can be found in existing literature.

1.1.1 Football Industry

The football industry in Europe has been expanding especially in women's football, which has exponentially grown over the past years all over the world (FIFA, n.d.). However, women's football is not yet as established. Research found that the players in the English Women's Super League had an average annual salary of £26,000 in 2017, which is very low compared to the £2.6 million average salary of a male player in the Premier League (Geey, 2019). Moreover, as women's football is not as established as men's football, the available research on women's football is limited. Based on these limitations, the research of this project will focus on men's football in Europe.

Players are seen as important and valuable assets to football clubs and have therefore highly detailed contracts that contain their rights and obligations to the club. It often happens that clubs are interested in players of other teams. Clubs that would like to acquire a new player have several ways of doing so. If the player is already contracted with another team, they can wait for his contract to expire and 'receive' the player for free, which is also referred to as a Bosman transfer (Frick, 2007). Another method is to loan the player from another team via a temporary player transfer, in which the borrowing team pays the player's salary (Geey, 2019). Finally, the last option is that the buying team pays the selling team a 'transfer fee' to buy the player out of his contract prior to expiration. The transfer fee can differ from the fair market value of a player, due to, e.g., the length of the remaining contract, strategic reasons to weaken the competition, or due to the bargaining power of the selling and buying club (Frick, 2007; Bryson, Frick & Simmons, 2013; Herm et al., 2014). Overall, the contracts of football players are highly important to the clubs and are therefore filled with special clauses and all the rights of the players to ensure that no player can be released from a club easily (Geey, 2019). Thus, this paper will look closely into transfer fees as it has a large economic relevance to the football industry and is the corner stone of the present transfer market economy (Herm et al., 2014).

1.1.2 Football Transfers

Football transfers can only happen in two specific periods throughout the year, the summer and winter transfer windows. These are defined by UEFA to give every team a fair and equal chance to acquire new players as all teams have to comply with these time windows. Usually, transfer fees are not paid all at once; they are a long-term investment with multiple instalments which can be paid over several years (Oprean & Oprisor, 2014).

Buying teams usually do not approach football players directly if they are interested; they either have to ask their current football club for permission or communicate via the player's agent. Having an agent is essential nowadays as they represent the players in several ways (Herm et al., 2014). For example, if the player is not able to fully settle into his new team, the agent tries to talk about what can be changed with the management of the club. Also, this means that if another club is interested in the player, the club reaches out to the agent to negotiate potential deals. Agents are paid based on commissions and can be part of a larger agency with more financial resources (Geey, 2019).

As mentioned before, the contracts themselves contain several types of clauses that both restrict but also protect the players. Examples of these clauses are sell-on clauses in case the player transfers again in the future and clauses that guarantee more compensation, based on player and or club performance for example (Chadwick & Thwaites, 2004; Geey, 2019). In the UK, Premier League teams also usually have image rights contracts with their players in which they have the exclusive rights on the videos and photos of their players (Barnard et al., 2019; Geey, 2019).

Other clauses generally included in contracts are release clauses. These clauses regulate when a player can be purchased and come into effect when a minimum amount set in the contract is offered by a buying club (Gerrard, 2002). Often, then the buying club is allowed to speak to the player about a possible transfer, whereas some release clauses automatically allow a transfer. Overall, release clauses are set up to protect the selling clubs, so their players and the talent are not easily poached (Geey, 2019).

Apart from the release clauses, there are also buy-out clauses. The buy-out clause is similar to a release clause; however, these are set at a highly unrealistic amount. The amount in a buy-out clause does not necessarily reflect the true market value of a player but is designed to ensure that a player cannot be bought-out of his contract (Gerrard, 2002; Geey, 2019). A prime example of this is Lionel Messi's current contract which includes a buy-out clause of over €700 million; in other words, he is locked-in at FC Barcelona (Telegraph Sport, 2017). Finally, clubs can include buy-back clauses in player contracts to ensure that the club can repurchase a player if they see he developed his skills over time (Geey, 2019).

Generally, Premier League clubs pay an inflated transfer fee for a player transferring to their clubs, as they have larger budgets due to the enormous television broadcasting deals in the UK which is also referred to as the 'English premium' (Barnard et al., 2019; Geey, 2019).

The above passages emphasise the central importance of player contracts. However, there are many features other than contractual that have to be taken into consideration when transfers are negotiated. Therefore, this paper will try to answer the following research question first:

RQ1: What are the important value drivers that determine transfer fees in the European football market?

1.1.3 Player Popularity

Generally, player popularity can be used to quantify brand value, which is defined as the creation of added value to a product or service (Rohde & Breuer, 2016; Farquhar, 1989). Other research states that the financial success of a football club is partially driven by brand value (Rohde & Breuer, 2016). This paper thus assumes that popularity is a method of quantifying brand value.

Having popular individual players is very important for a club's income as with increased team popularity comes higher merchandise and ticket revenues, and better sponsorship deals (Müller et al., 2017; Bernard et al., 2019). Thus, football clubs increasingly see themselves as brands and actively make an effort to market this brand internationally (Bodet & Chanavat, 2010).

Empirically, a player's popularity has an effect on his contract and salary as well as his play time during a game (Herm et al., 2014; Müller et al., 2017). As contracts and transfer fees are differently negotiated, it might be interesting to see if player popularity also has an effect on the transfer price. Therefore, the second question this paper would like to answer is:

RQ2: Is there a significant influence of a player's popularity on transfer fees?

1.2 Econometric Techniques

This section describes the statistical methods as well as the data transformation, model selection and accuracy measurement techniques utilised in the course of this paper. Hence, it provides a theoretical understanding of the methods employed for the prediction of the football transfer fees. The set of input variables are referred to as dependent variables X, while the predicted variable referred to as the independent variable Y (Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie & Tibshirani, 2013).

1.2.1 Linear Regression

Linear regression is an econometric method which is utilised to explore the relationship between the dependent (Y) and independent variables (X). Additionally, the method is well suited and often utilised to predict quantitative outputs (James et al., 2013; Woolridge, 2009). Considering the business context, a simple linear regression is not sufficient as most cases evaluate more than one independent variable. Multiple linear regression is based on simple linear regression and defines parameter coefficients of each independent variable p within a model (James et al., 2013; Woolridge, 2009):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

The parameters of the model, i.e. the β, are unknown and can be determined with the help of training data (James et al., 2013). β_0 reflects the intercept of the linear model, while the β_p constitute the slopes (James et al., 2013; Woolridge, 2009). The intercept equals the realised value of Y, i.e. y, when all $x_p = 0$, while the slope constitutes the average increase in Y with a one-unit increase in X (James et al., 2013). The parameters must be chosen in a way that the model fits the training data as well as possible. The fit of the model can be estimated with the help of various methods, but the most common one is the ordinary least squares (OLS) method (James et al., 2013; Woolridge, 2009).

Based on the denotation of the predicted value of y as $\hat{y}$, the residual of the i-th observation of the training dataset can be expressed as the difference between this response's actual value y_i and the predicted value $\hat{y}_i$ (James et al., 2013). Essentially, the residual sum of squares (RSS) across all observations is computed. The aim is to minimise the sum of the squared residuals for every observation in the dataset in terms of the predicted value $\hat{y}_i$. The smallest possible value of RSS returns a model, which fits the data best, hence the method is called OLS (James et al., 2013; Woolridge, 2009). There might be additional influences apart from X that account for changes in Y. Therefore, a mean-zero random error term (ε) is accounted for on the top of the linear function (holding assumption that the error is independent from X) (James et al., 2013).

An issue that needs to be addressed when applying linear regression is the potential collinearity between independent variables. In case the dependent variables are correlating, it is difficult to estimate the predictive nature of these variables on the dependent variable separately. In mathematical terms this causes uncertainty concerning coefficient estimates (Woolridge, 2009; James et al., 2013). As the standard error tends to increase greatly along with the p-value, the collinearity can also result in a wrong prediction of the non-zero coefficient (James et al., 2013). Collinearity between two independent variables can be detected from the correlation matrix, however, in cases of multicollinearity (collinearity between three and more variables), advanced measures need to be taken (James et al., 2013). As the correlating variables do not provide additional value for the explanation of the variation in the dependent variable, the correlating variables can be either merged, or one of them can be dropped from the model (James et al., 2013; Woolridge, 2009).

1.2.2 Cross-Validation

Cross-validation (CV) can be utilised to assess the accuracy of a trained model. During the process of CV a high number of models is fitted and the best or average performing model is selected in order to estimate the test error. In contrast to indirect measures for the prediction test error, such as AIC, BIC, C_p or $\bar{R}^2$, this approach directly estimates the test error and does not pose restrictive assumptions on the variable distributions (James et al., 2013). Hence, this technique helps to build robust models (James et al., 2013).

CV randomly splits the training dataset into k folds. It generates numerous models by utilising $k-1$ folds for training, while the remaining fold is used for testing and estimating the error. This process is repeated k times (James et al., 2013). Model selection can then be performed based on the lowest error, or the 'one-standard error-rule' (1-se-rule*)* can be used. The latter is a technique that selects a parsimonious model which test error is just within one standard deviation from the minimum observed test error (Hastie et al., 2009; James et al., 2013). Consequently, the predictive accuracy of the 1-se-rule is sufficiently close to the optimal model's accuracy while preventing the model from excessive overfitting.

1.2.3 Best Subset and Stepwise Selection

It is often the case that the dependent variable is explained only by a selected subset of the independent variables (James et al., 2013). A multitude of techniques can be employed in order to define the subset, such as stepwise selection, best subset and similar methods (Woolridge, 2009; James et al., 2013). In this research, the best subset method and forward stepwise selection are considered, which will be explained in the following.
Best subset selection aims to select the best model from all possible combinations, meaning 2^p models have to be computed prior to the selection (James et al., 2013). Thus, it can become computationally infeasible in cases where the model is meant to choose from a large number of predictors p. Since this research utilises a dataset incorporating a large number of predictors, experiments with this method showed that it is computationally infeasible and therefore not considered going forward.

The forward stepwise selection method begins with choosing an intercept and then fitting p different linear regression models, choosing to add the variable of the model that yields the lowest RSS. The process is then repeated in order to find the additional variable until a stopping criterion is fulfilled (James et al., 2013).

1.2.4 Shrinkage Methods – The Ridge and the Lasso

Regularisation causes shrinking of the coefficients for particular variables X. Two regularisation methods are considered – the Ridge and the Lasso (James et al., 2013). The advantage of the shrinkage methods compared to the stepwise selection methods is that they do not constitute a high variability (Hastie et al., 2009).

Ridge regression is similar to the already mentioned OLS method, as it determines coefficient estimates β_j with the best possible model fit by minimising the RSS, but contains an additional shrinkage penalty term (James et al., 2013):

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad 2$$

The shrinkage term has the ability to shrink the coefficient estimates of dependent variables to nearly zero and is imposed through a hyperparameter λ, which balances out the RSS and the shrinkage penalty (James et al., 2013; Hastie et al., 2009). If $\lambda = 0$, the shrinkage term vanishes, and the model reflects the OLS model. The bigger the λ, the stronger the coefficient estimates converge to zero and in the extreme condition when $\lambda = \infty$, all coefficients shrink to zero and only the model intercept remains (Hastie et al., 2009). Therefore, when applying regularisation methods, it is crucial that the best λ parameter is found through parameter tuning (Hastie et al., 2009; James et al., 2013). For this purpose, CV can be utilised by performing a grid search to test various values of λ (James et al., 2013). The Ridge is sensitive to the scaling of independent variables; therefore, it is important that these are standardised before applying the method (Hastie et al., 2009; James et al., 2013).

The advantage of the Ridge regression compared to the OLS method is that it controls variance better in return for a slight increase in bias (James et al., 2013; Hastie et al., 2009). For this reason, Ridge regression should be a preferred choice when working with datasets, where the relationship between independent variables and the dependent variable is linear and the OLS method exhibits high variance (James et al., 2013). Another advantage of Ridge regression is that it is computationally inexpensive (nearly to the same extent as the OLS method) compared to the Best subset selection (James et al., 2013; Hastie et al., 2009).

Essentially, any fitted model exhibits a certain degree of variance and bias. In statistical terms, this is known as a bias-variance trade-off (Hastie et al., 2009; Woolridge, 2009; James et al., 2013). High variance reflects a highly inflexible estimate function that produces large prediction errors when a different dataset is used. High bias refers to the estimate function being expressed as a much simpler function than the real-life data pattern requires, which makes the function general but also introduces an error in estimates (James et al., 2013). The trade-off stems from the fact that extremely low bias estimate functions will fit the real-data pattern very well. However, when applied to a unseen dataset, will result in high variance, as it will probably not be fitting the unseen dataset equally well (James et al., 2013).

The Lasso is similar to the Ridge regression, however it allows for the coefficient estimates to take on zero value, if the λ parameter is large enough (James et al., 2013):

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \qquad 3$$

This way, the Lasso additionally offers variable selection, unlike the Ridge method, which shrinks the estimates very close to zero and therefore changes their magnitude, but keeps all variables in the model (Hastie et al., 2009; James et al., 2013). The Ridge method outperforms Lasso in settings, where there is no specific subset of independent variables that has a dominant influence on the dependent variable. This is because the Lasso may unnecessarily shrink coefficients to zero, via the λ parameter that is picked through CV based on the lowest test error. In these cases, Ridge regression outperforms the Lasso in terms of lower variance (James et al., 2013).

1.2.5 Polynomial Regression

The methods listed in this section so far are built on the assumption that the relationship between the independent variables and the dependent variable is linear, however in many cases this serves only as an approximation (James et al., 2013). The simplest method to be applied to deal with nonlinearity in the dataset is transforming the linear function into a polynomial function, where the polynomial degree d allows for a nonlinear model that produces a better fit (James et al., 2013):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 \ldots + \beta_d x_i^d + \varepsilon_i \quad (4)$$

1.2.6 Selected Data Transformation Techniques

As discussed in the context of the Ridge and the Lasso, both methods require independent variables to be standardised. The goal of standardisation is to transform the relevant data points to the same scale, in order to not give one variable more importance than others (Han et al., 2011). Each data point is standardised by subtracting the mean and then dividing it by the standard deviation, which results in equal variance in the data (Abu Mostafa et al., 2015).

In addition to standardising the independent variables, the log transformation is applied to the dependent variable. In cases of highly skewed data, this technique is recommended in order to approximate a normal distribution. This leads to a better fit for statistical analysis methods that assume a normal distribution (Ford, 2018). Additionally, linear modelling assumes constant variance and log transformation brings the data closer to meet this assumption (Ford, 2018). At the same time, it causes that the effect of a one-unit change in X on Y is expressed as a percentage change (Ford, 2018).

1.2.7 Accuracy Measure

A measure of accuracy is utilised within CV and to perform model selection. In this research, methods belonging to the family of scale-dependent measures (Hyndman & Koehler, 2006) are considered, namely root mean

square error ($RMSE$) derived from mean square error (MSE), where $\hat{f}(x_i)$ is the estimate that the prediction function $\hat{f}(\cdot)$ gives for the i-th observation of X (James et al., 2013):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2} \qquad 5$$

The $RMSE$ is used frequently as it is directly interpretable in terms of the measurement units. That is, the $RMSE$ is the average distance of a data point from the fitted line (James et al., 2013). It serves to assess the fit of the model during the training phase and to evaluate the accuracy of the trained model on previously unseen data (Chatfield, 1988; James et al., 2013).

One goal of the training and testing cycle is to increase the flexibility and hence the fit of the model for unseen data. If a decreasing train-$RMSE$ but an increasing test-$RMSE$ is observed, it is likely that the trained model is overfitting and capturing patterns which are caused by random chance (James et al., 2013). However, it can be observed that the train-$RMSE$ is smaller than the test-$RMSE$, regardless of overfitting, since the statistical learning models either directly or indirectly aim to minimise the train-$RMSE$ (James et al., 2013). Based on the previous section, the following research question is formed:

RQ3: Which model for estimating football transfer fees performs best in terms of predictive accuracy?

2. Related Work

Numerous research on the topic of transfer fees, market values or salaries of football players that utilised econometric techniques has already been conducted (Dobson, Gerrard & Howe, 2000; Garcia-del-Barrio & Pujol, 2007; Medcalfe, 2008; Herm et al., 2014; Drut & Duhautois, 2017; Müller et al., 2017; Ante, 2019; Serna Rodríguez, Ramírez Hassan & Coad, 2019; Singh & Lamba, 2019; also cf. appendix A.1, Tab. 8). Previous works utilise different sets of variables – the overview can be found in Tab. 1. Selected variables will be discussed more in depth in the following section.

Tab. 1 Variables utilised in previous research

Variable	Description	References
Transfer-specific variables		
Transfer fee	Player transfer fee related to the transfer	3, 6, 8
Year	Year of the player transfer	8, 9
Player-specific variables		
Player performance	Various indicators of player's performance, such as FIFA score, minutes played in last season, goals and assists	1, 2, 3, 4, 5, 6, 7, 9
Player popularity	Various indicators of player's popularity, such as Wikipedia page views or mentions on the web reported by Google's search engine	1, 2, 3, 4, 5, 7
Position	Position of the player involved in the transfer	2, 3, 4, 6, 7, 8, 9
Preferred foot	Preferred foot of the player involved in the transfer	3, 4, 5
Age	Age of the player involved in the transfer	1, 2, 3, 4, 5, 6, 7, 8, 9
Height	Height of the player involved in the transfer in meters	1, 3, 4
National player	The transferred player is a national player at the same time	4, 6, 7, 9
Transfer to home country	The transferred player transfers to his home country	3, 4, 9
Team-specific variables		
League or team 1	Name of the team or league from which the player transfers	2, 4, 8
Rank previous season team 1	Rank in previous season of the team from which the player transfers	2, 4, 8
League or team 2	Name of the team or league to which the player transfers	3, 8
Rank previous season team 2	Rank in previous season of the team to which the player transfers	8

The table gives an overview of the most important related literature. It states the numerous variables analysed by the prevalent studies, gives a brief description for these variables and lists the numerical index for the respective papers.

The indexes can be mapped to the related literature as follows: (1) Müller, Simons & Weinmann, 2017; (2) Singh & Lamba, 2019; (3) Ante, 2019; (4) Serna Rodríguez, Ramírez Hassan & Coad, 2019; (5) Herm, Callsen-Bracker & Kreis, 2014; (6) Medcalfe, 2008; (7) Garcia-del-Barrio & Pujol, 2007; (8) Dobson, Gerrard & Howe, 2000; (9) Drut & Duhautois, 2017.

The player popularity has been included as a potential determinant of transfer fees in the previous research on numerous occasions. Müller et al. (2017) introduced several variables that have the potential to reflect player popularity – Wikipedia page views, Google Trends search index, Reddit posts and YouTube videos referring to the player. They found that three of these indicators (excluding Google Trends) were significant determinants contributing to the model accuracy for market value prediction (Müller et al., 2017). Other researchers utilised the Google Search engine in order to test the popularity as a potential determinant. Garcia-del-Barrio and Pujol (2007) retrieved the number of links that reported how many sites on the web mention the player.

This was later repeated by Herm et al. (2014) and by Serna Rodríguez et al. (2019). Ante (2019) decided to use Social Media data as indicators of players' popularity and collected the number of followers per player from Instagram, Twitter and Facebook. Singh and Lamba (2019) utilised Wikipedia page views as an indicator of popularity. Most of the mentioned research confirmed positive correlation between popularity indicator and transfer fee or market value of a player (Garcia-del-Barrio & Pujol, 2007; Herm et al., 2014; Müller et al., 2017; Ante, 2019; Singh & Lamba, 2019), except for Serna Rodríguez et al. (2019).

Player performance is another group of characteristics that has received a lot of attention in the previous research. Müller et al. (2017) has grouped a number of quantifiable indicators under 'player performance' – these included variables such as number of goals, assists or number of minutes played at the national and international level. Variables such as goals and assists were used by Singh and Lamba (2019), who also employed data from the FIFA video game, which rates players based on their performance on the field. Ante (2019) also introduced goals and assists as performance indicators on the top of other variables such as number of yellow and red cards received in the past season, bad controls, offsides and others. Serna Rodríguez et al. (2019) used a variable 'performance', which constituted player appearances in the preceding season and additionally introduced variables such as the number of goals and assists respectively number of matches played in the previous season. Herm et al. (2014) aggregate goals and assists under 'scoring' and additionally use precision (% successful passes/corrected by mean at position) and assertion (won duels or save to shots ratio of goalkeeper/corrected by mean at position). However, they refer to these indicators as a 'talent' rather than performance (Herm et al., 2014). Goals and assists were utilised in Medcalfe (2008), among others. An interesting approach can be found in Garcia-del-Barrio and Pujol (2007) who carry out their research within the Spanish professional football league in order to explain Spanish football clubs not profiting from monopsony rents. In their framework, they are introducing indicators for 'worker productivity' within football and are using indexes of performance available to them – 'puntos Marca' and 'liga Fantástica' (Garcia-del-Barrio & Pujol, 2007). All of the mentioned research observed a positive association between performance indicators and market value or transfer fees (Garcia-del-Barrio & Pujol, 2007; Medcalfe, 2008; Herm et al., 2014; Drut & Duhautois, 2017; Müller et al., 2017; Ante, 2019; Serna Rodríguez et al., 2019; Singh & Lamba, 2019).

Position of the player on the field is utilised by various researchers (Ante, 2019; Serna Rodríguez et al., 2019; Singh & Lamba, 2019), as it has been previously found to have a positive effect on the transfer fee (Dobson et al., 2000; Garcia-del-Barrio & Pujol, 2007; Medcalfe, 2008). Garcia-del-Barrio and Pujol (2007) observe in their results that attackers and midfielders seem to be valued with a higher market value, while defenders seems to be undervalued.

Some researchers included information about the player's preferred foot (Herm et al., 2014; Ante, 2019; Serna Rodríguez et al., 2019). Ante (2019) found that a preferred right foot had a negative effect on the transfer fees in selected Leagues (Premier League and Bundesliga), while left-foot and both-feet preferences did not prove significant. Herm et al. (2014) found that what they denote as 'flexibility' – an ability to play

equally well with both feet – had a positive influence on the market value of a player. Serna Rodríguez et al. (2019) did not find the preferred foot to be a significant determinant of a player's market value.

Ante (2019) expresses an assumption that different spending behaviour can be observed for different leagues and their results show a positive relationship between the transfer fee and the league that the player was transferred to in case of the English Premier League. In the research carried out by Serna Rodríguez et al. (2019), a positive relation was also found between a transfer to the English Premier League and the market value. Dobson et al. (2000) utilised the same analogy and found a positive association in case that a player was transferred to particular clubs.

Some researchers have also attempted to measure the influence of the nationalities that are present in the team to which a player is transferring, more particularly what is the share of international players in the team in relation to the market value (Serna Rodríguez et al., 2019) respectively share of international players and the share of domestic players as separate variables in relation to player wage (Drut & Duhautois, 2017). Serna Rodríguez et al. (2019) were unable to prove the correlation, while Drut and Duhautois (2017), who studied the Italian league, found that the share of international players was a positive significant determinant of players' salaries. Additionally, they found a strong positive correlation between the player being a foreigner and player wage (Drut & Duhautois, 2017). Ante (2019) discovered that Spanish and Italian leagues seem to value players from South America more and attributes this to cultural similarities.

This research paper intends to expand the existing research on football player valuation by utilising statistical and predictive methods. Specifically, numerous models from the regression family are utilised in the coming section. Thus, the paper contributes to the current research by applying a sequence of regression models for the prediction of transfer fees in ten major European leagues. Additionally, novel variables are introduced, tested and used to build the models (cf. appendix A.1, Tab. 9). Finally, the fourth research question is proposed:

RQ4: Where is the transfer market madness heading to?

3. Methodology

3.1 Data Collection and Pre-Processing

This paper employs a rich dataset in order to draw inferences and make predictions about the fees of football player transfers. In total, the data consists of 45 potential variables which are collected from three different sources (transfermarkt.de, kaggle.com, wikipedia.org) and of which 24 are used for the purpose of modelling. The following passages describe how the data is collected and which technologies and software packages are utilised during the process.

First, the raw player transfers are scraped from the community-based platform *transfermarkt.de*. Today, the community-based platform is considered one of the most important websites on football transfers (Herm et al., 2014; Müller et al., 2017). Due to the initial launch of the database in Germany in 2001 (Krennhuber, 2008), the German version of the website is preferred over alternative versions due to a presumably more mature community and better maintained database. The idea of the project is to provide a platform for users to discuss the transfer market and provide crowdsourced estimates of football players' market values (Krennhuber, 2008; Herm et al., 2014; Müller et al., 2017). Beside estimates of market values the user can access a wide range of additional information, such as general news, player-, match- data, team- or league-specific data.

In order to automatically extract the relevant information, this paper employs the *rvest* and *polite* R-package for web scraping (Perepolkin, 2019; Wickham, 2019). It scrapes all remunerated players transfers that occur during the 2013-2019 period between two of the major European leagues (with respect to total market value as of 19/11/2019): Premier League – England (€9.32 billion), Primera División – Spain (€6.18 billion), Serie A – Italy (€5.31 billion), 1. Bundesliga – Germany (€4.71 billion), Ligue 1 – France (€3.59 billion), Liga NOS – Portugal (€1.13 billion), Premier Liga – Russia (€1.10 billion), Eredivisie – Netherlands (€1.02 billion), Jupiler Pro League – Belgium (€0.85 billion), Süper Lig – Turkey (€0.67 billion). To ensure economic relevance, the season 2013/2014 has been chosen as starting point since in that year the transfer fee mark of €100 million has been cracked for the first time (Rumsby, 2013). Note that the abovementioned "remuneration" condition ensures that the sample does not include any loan or redemption-free transfers. Further, it is assumed that transfers only include fixed payments and no variable components and that transfers are executed at the end of the summer transfer window, i.e. usually on the 30th of August. This approach finally yields a sample size of 2,807 transfers whereby for the 2019/2020 season only the summer transfer window is included in the data. For each transfer, TM records the player name, age, nationality, position, estimated market value, observed transfer fee as well as the involved teams. In addition, the scraper collects the following information:

- Team information, such as average team age, average team market value or team rank in the previous season (for the two involved teams),
- Permanent player metadata, such as birthdate or height,

– Transfer-specific player data, such as minutes played in the last season across team-level competitions, titles won or whether or not the player plays for the national A-team (all at the time of the transfer).

Second, this paper includes two datasets available on *kaggle.com* to also account for proxies of player performance and potential (Mathien, 2016; Leone, 2019). More precisely, these datasets cover individual football player statistics for the 2013-2019 period in the form of scores included in the yearly published video game FIFA and retrieved from *sofifa.com*. Since the publisher Electronic Arts aims to ensure data quality and homogeneity by employing a special data taskforce with more than 6,000 external contributors, so-called "Data Scouts" (Murphy, 2019), it can be assumed that the virtual values accurately approximate real-world player characteristics. Accordingly, the value of video game statistics in the context of sports analytics has already been highlighted by academia (e.g., Cotta, Vaz de Melo, Benevenuto, & Loureiro, 2013). Note that in cases where FIFA statistics were updated in the course of game updates, the datasets include several observations per player and season. In those cases, this paper relies on the first available observation which relates to the game's release date in the fall of each year and is thus as close as possible to the assumed transfer date, i.e. the end of the summer transfer window.

In order to merge the two datasets based on player names and birthdates as unique identifier tuple, a name dictionary is created on which several joint operations are performed. For 250 (~9%) player names the merger of the two datasets failed due to name inconsistencies caused by special characters. Those cases are manually checked and corrected by the authors. Further, 171 cases are removed due to missing FIFA data and two cases are removed due to non-unique identifier pairs (i.e. there are two different players with the same name and birthdate). Consequently, the final sample comprises 2634 unique football player transfers.

Third, the merged dataset is enriched by page views data from *wikipedia.org* to proxy for player popularity. In particular, the API access provided by the *wikipediatrend* R-package (Meissner, 2019a) is used to gather the page views during the previous season for the English-, German-, French-, Spanish- and Italian-speaking version of Wikipedia (if available). This approach ensures that the most important European languages are covered and thus the popularity of the player is adequately measured. However, in order to query the wikipediatrend API, the URL reference to the Wikipedia page is required (in italics: https://en.wikipedia.org/wiki/*Kylian_Mbappé*). In order to retrieve this string, this paper employs the following approach while relying on the functions provided by the *RSelenium* R-package for automated web navigation (Harrison, 2019):

1) Initialise a virtual machine that runs Google Chrome using the 'Selenium WebDriver'.
2) Run the following Google search query: "{player name} {birthdate} en.wikipedia.org".
3) Filter for search results that contain "en.wikipedia.org", select the first link and access the page.
4) Retrieve the specific part of the URL that references the player's Wikipedia page.
5) Implement a two-minute delay in order to not be detected as an automatic bot by Google.
6) Re-run step 2) to 5).

For some rare cases the search query specified under 2) does not yield the desired results but instead requests a Wikipedia page related to the birthyear of the player. Those cases are filtered for using a digit-filter and re-run with the following alternative search query: "{player name} {birthdate} football player en.wikipedia".

Finally, two additional features are derived from existing ones and missing values are imputed. In particular, missing values for a player's height are imputed with the median height respectively missing values for the number of titles won and the number of Wikipedia page views are imputed with zeros. Thereby, for the latter it is assumed that if no page views could be found for a player, no Wikipedia page for the respective exists. A description of the overall dataset can be found under appendix A.2.

3.2 Exploratory Data Analysis

Although the application of predictive modelling techniques lies at the heart of this paper, a brief exploratory analysis is inevitable to get familiar with the characteristics and pattern present in the dataset (Tukey, 1977). Fig. 1 indicates that the average number of transfers amounts to 376 per year with a of 470 transfers in 2017. Even though, the number of transfers appears to decline afterwards, Fig. 2, in fact, suggests that the average transfer fee is not, hence climaxing in 2019 with a value of ~€13 million per transfer. Considering the whole sample period from 2013 to 2019 the average transfer fee amounts to €9.11 million (for a whole range of summary statistics for the numerical variables included in the dataset see appendix A.3, Tab. 10).

Fig. 1 Number of player transfers per year

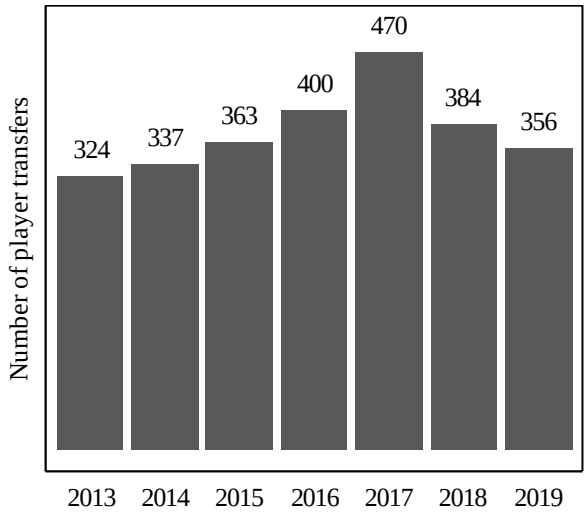
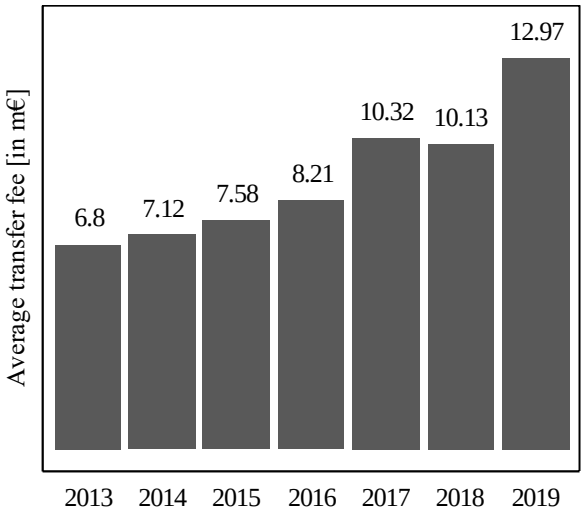
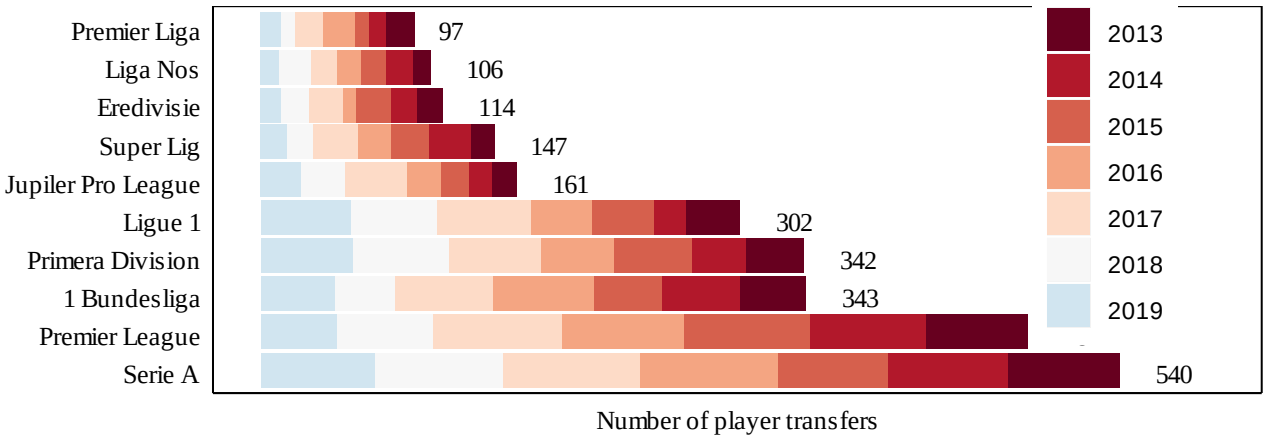


Fig. 2 Average transfer fee per year



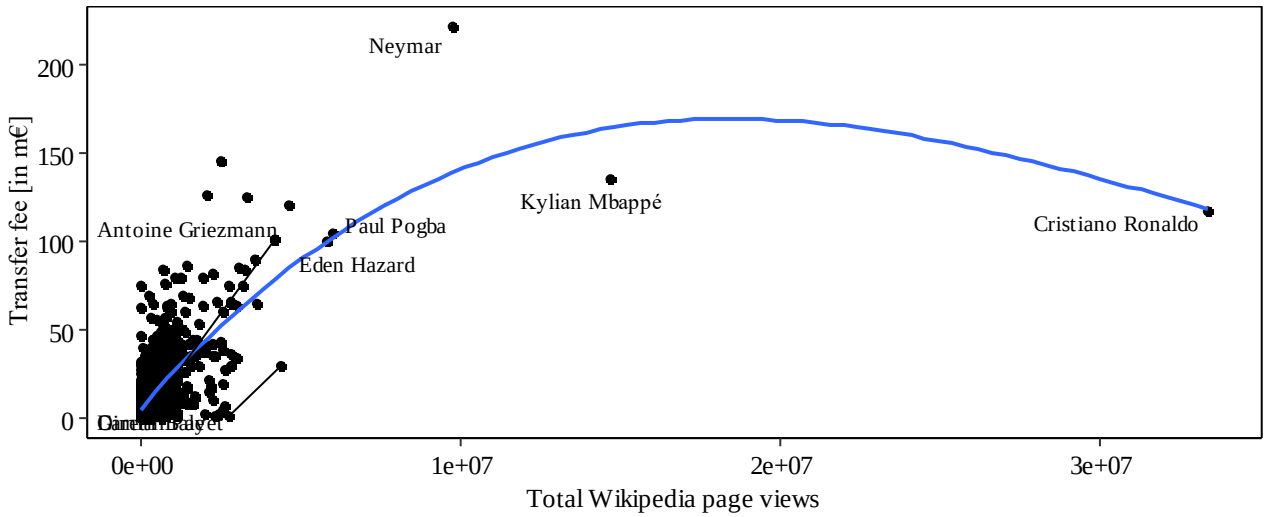
Considering the distribution of transfers across leagues in Fig. 3, a clear disbalance can be found with Serie A (Italy) accounting for 20.5% (540) and Premier Liga (Russia) accounting for only 3.7% (97) of the overall observed transfer activity. Generally, these findings are well aligned with the total estimated market values of these leagues highlighting that the more highly valued leagues are also more active on the transfer market.

Fig. 3 Distribution of player transfers per league and year



Further, considering the main variable of interest for *RQ4* – the total number of Wikipedia page views – a clear positive and non-linear relationship can be found between the total page views and the player's respective transfer fee (cf. Fig. 4). Note that the non-linear nature of the relationship is heavily distorted by the few extreme outliers, such as Ronaldo, Mbappé and Neymar. An overview of other cross-wise relationships between the transfer fee magnitude and potential value drivers can be found appendix A.3, Fig. 15, and Fig. 16.

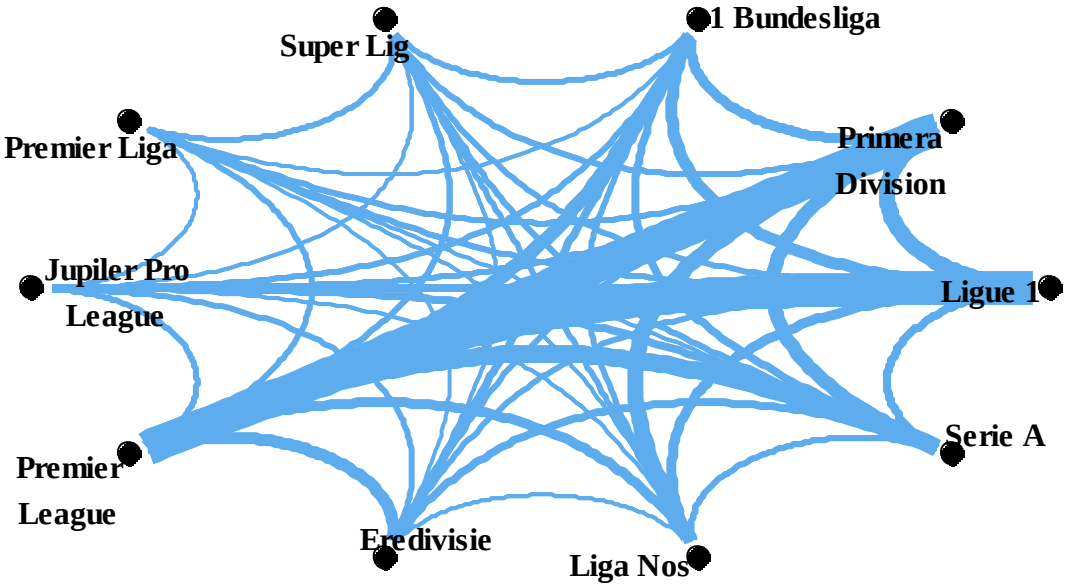
Fig. 4 Non-linear relationship between the total Wikipedia page views and the transfer fee



In order to further demonstrate the richness of the dataset, Fig. 5 illustrates a simple network analysis of the transfer market with the number of transfer mapped to the width of the network's edges. Again in line with the authors' expectations, the biggest proportion of the transfer activites are captured by the top five to six European leagues, especially the Primera División, Serie A and Premier League. Interestingly, the teams of smaller leagues, such as the Primier Liga or Süper Lig, appear to trade equally frequently with the other

European leagues whereas for example the Portuguese Liga Nos or the Italian Serie A tend to trade more with their respective geographic neighbours. Note that transfers within leagues are not caputred by the network.

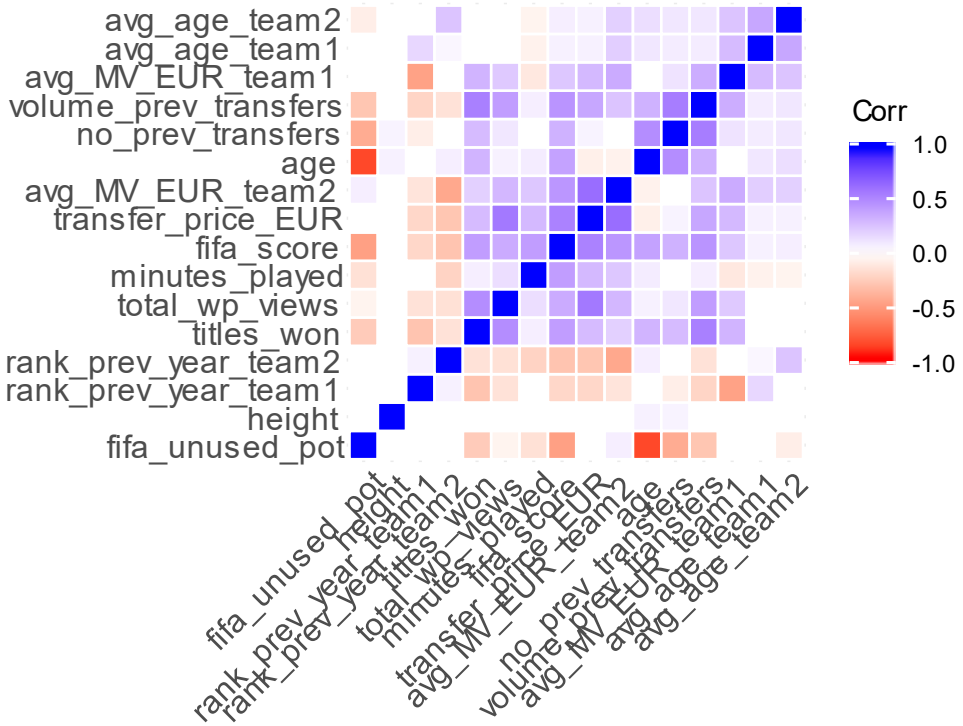
Fig. 5 European transfer network



To finally conclude the exploratory analysis and prepare for the predictive modelling section, Fig. 6 illustrates the pairwise correlations between the several numerical variables (for the non-graphical version see appendix A.3, Tab. 11). Note that blue (red) boxes denote strong positive (negative) linear correlations and that insigificant correlations (i.e. $p > 0.05$) are blanked out. The following moderate to strong relationships can be identified (for frequently applied interpretation thresholds cf. for example Cohen, 1988; Fahrmeir, Heumann, Künstler, Pigeot, & Tutz, 2016):

- For the transfer fee a positive and moderate linear relationship can be observed with the FIFA score ($\rho = 0.54$), the total number of Wikipedia page views in the season prior to the transfer ($\rho = 0.58$) and the average market value of the receiving team ($\rho = 0.63$).
- A negative and strong relationship can be observed between the player's age at the time of the transfer and his unused potential as indicated by the difference between the FIFA potential score and the FIFA overall score ($\rho =$ -0.87). Further, positive, moderate relationships exists between the number and the aggregate monetary volume of previous player transfers ($\rho = 0.56$) and the aggregate monetary volume and the amount of titles won by the player prior to the transfer ($\rho = 0.55$).

Fig. 6 Correlation matrix for the numerical variables



3.3 Modelling

This section sketches the methodological approach employed in the statistical inference as well as the predictive modelling part of this paper. In a first step, the various data manipulations are disclosed, and the baseline linear regression model is specified to address *RQ1* and *RQ2*. Subsequently, the paper discusses the implementation as well as tuning of the previously introduced prediction methods and states the criterion which selects the most accurate model to address *RQ3* and *RQ4*. To facilitate consistency, the terms 'response' and 'predictor' stand synonymously for 'dependent' and 'independent variable' throughout the rest of this paper.

3.3.1 Outlier Correction

Flaws in linear OLS models are known to stem from extreme outliers, multicollinearity or situations where the number of predictors tends to surpass the number of observations (Agresti, 2015). To address the first issue, the data is checked for outliers using a regression diagnostic. More concretely, observations with a high residual error ($\hat{\varepsilon}_i = y_i - \hat{y}_i$) as well as high leverage, i.e. extraordinarily high feature values $x_i \gg \bar{x}$ (James et al., 2013), are identified using Cook's distance (Cook & Weisberg, 1982; Williams, 1987). This approach reveals single observations that have a significant influence on the magnitude of the regression coefficients (Rousseeuw, 1987). Since one observation could be found that clearly exceeds the conventional threshold of

one (Rousseeuw, 1987; Agresti, 2015), winsorising is employed to capture extreme cases by capping numerical variables at the 0.1 respectively 99.9% quantiles (Barnett & Lewis, 1994).

3.3.2 Data Transformations

First, since the response variable (transfer fee in million €) is highly right-skewed, the variable is log-transformed (cf. section 1.2) to generate a distribution that resembles a Gaussian more closely (cf. appendix A.4, Tab. 18 for a graphical illustration of the effect). Inversely, a log-transformation of several of the skewed predictor variables cannot be justified due to several zeros produced by the data generating process which would result in a log-transformed value that converges towards negative infinity.

Second, the predictor variables are standardised (cf. section 1.2) as it is crucial for the use of the Ridge and Lasso models (cf. appendix A.4, Fig. 19 for a graphical illustration of the effect). In line with Tibshirani, (1997), this technique is simultaneously employed for the numerical as well as categorical predictors. Therefore, the latter are transformed into dummies via one-hot encoding ex ante.

Implications of the previously outlined steps for the quality of the linear model are provided in Tab. 2. When fitting the model using the whole dataset, the adjusted R-square ($\bar{R}^2$) suggests a better model performance for the models that implement all of the abovementioned transformations. Note that the model performance in terms of $\bar{R}^2$ is indepent of the standardisation step.

Tab. 2 Performance effects of the different data manipulation steps

Model type	Degrees of freedom	F-statistic	R^2	$\bar{R}^2$
Linear model	56	99.28***	0.6793	0.6724
incl. Outlier correction	56	125.37***	0.7279	0.7221
incl. Log-transformation of the response	56	135.37***	0.7428	0.7373
incl. Standardisation of the predictors	56	135.37***	0.7428	0.7373

The table presents the degrees of freedom, the F-statistic, the coefficient of determination (R^2) and the adjusted coefficient of determination ($\bar{R}^2$) for the following models: the basic linear model, the linear model with outlier correction, the linear model with outlier correction and log-transformed response as well as the linear model with outlier correction, log-transformed response and standardised predictors. Statistical significance is indicated by the asterisks as follows: *** (1%-level), ** (5%-level), *(10%-level).

For the purpose of statistical inference, i.e. to address *RQ1* and *RQ2*, the general linear model with outlier correction is tested and discussed in the course of section 0 and 0. To promote the interpretability of the model coefficients in an economically meaningful, the log-transformation and standardisations are therefore only implemented as part of the predictive modelling part of this paper.

3.3.3 Cross-Validation Approach

In order to enable the comparison of different prediction models based on the test error and prevent data leakage, i.e. the use of test data in the course of model fitting, outlier correction (Kaufman, Rosset, & Perlich, 2011), data transformation and hyperparameter tuning is implemented within a k-fold CV approach (Raschka, 2018).

The implementation of the data transformation steps within the CV approach is achieved via the *recipes* R-package (Kuhn & Wickham, 2019). The transformation steps are encoded into a formula-like object which is used to predict the transformation parameters from the test data (i.e. mean and standard deviation) and then applied to the test set. With regards to the fold parameter k, this paper follows the common practice in data science and chooses a fold size of $k = 10$ to responsibly account for the bias-variance-trade-off inherent to CV (James et al., 2013; Kohavi, 1995). Hence, the final dataset is randomly split in ten folds of which nine are then recombined in a rotating fashion to generate ten 90%-10% train-test-splits.

3.3.4 Predictive Modelling

This paper implements the five different regression modelling techniques introduced in section 1.2: (1) linear regression, (2) linear regression with stepwise forward selection, (3) Ridge regression, (4) Lasso regression and (5) polynomial regression. Since the exploration of correlation coefficients in the previous section implies the presence of multicollinearity in the data, the stepwise forward and Lasso approach is applied to directly perform variable selection in the course of model fitting (James et al., 2013). More precisely, multicollinearity is addressed by omitting predictors that only possess limited additional value in explaining the remaining variation in the response (Agresti, 2015). In the same vein, the Ridge regression enables regularisation to improve the generalisability of the model, however, without explicitly performing variable selection. For each of the five techniques, the cross-validated $RMSE$ is computed on the train and test set, computed as the (weighted) average across all ten folds, to enable model selection. In the following some of the peculiarities of each technique are briefly discussed:

(1) Linear regression: The full specification of the linear model including all available predictors as well as an intercept incorporates 56 variables in total. For the year-, league-, footedness- and position-dummies the year "2013", the German "1. Bundesliga", "right" and the "offensive midfield" position act as base cases meaning that these factor levels are not represented as individual dummies. For example, if all year-dummies take on a value of zero it means that the transfer took place in the year 2013.

(2) Linear regression with stepwise forward selection: The CV approach leads to 550 models being fitted in the course of the training phase using the *leaps* R-package (Lumley, 2017). That is, due to the 55 predictors included in the modelling dataset, 55 stepwise regression models are fitted with $p \in [1, 2, \ldots, 55]$ predictors. The $RMSE$ in each step is then estimated via 10-fold CV which eventually results in 550 fitted

models. During each step, the next predictor to be added to the model is chosen based on the maximal marginal improvement of the model, measured via AIC, BIC, or $\bar{R}^2$ respectively. Eventually, the optimal number of p is determined by first identifying the model with the lowest overall CV test-$RMSE$ and then consulting the 1-se-rule. Note that this approach mitigates the need to define specific stopping criterion ex ante.

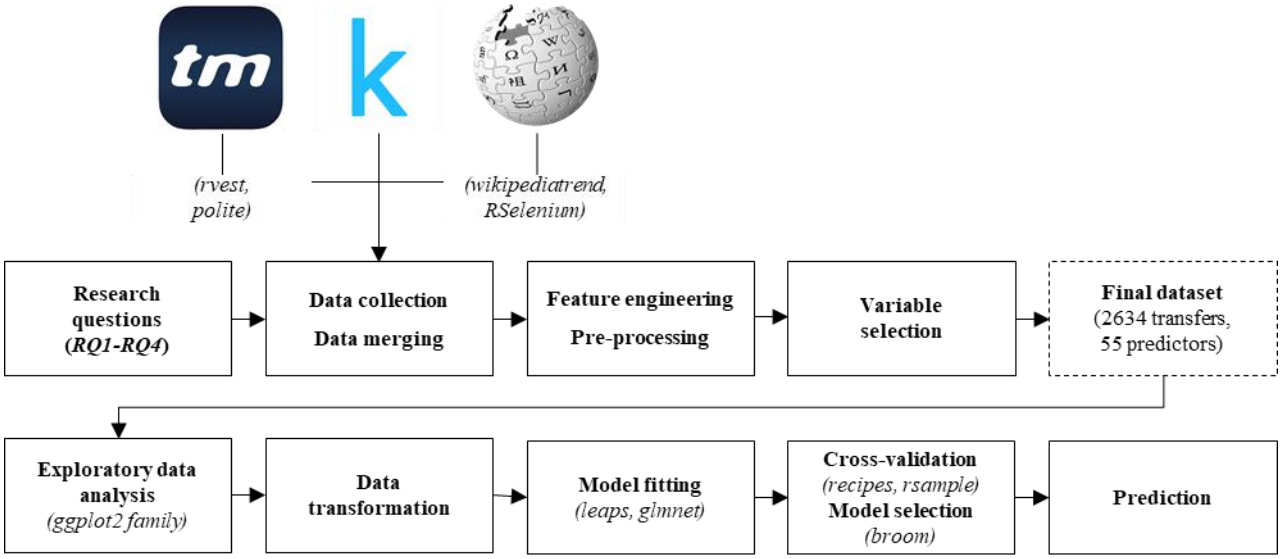
(3) Ridge regression: The CV approach leads to 1,010 Ridge regression being fitted in the course of the training phase using the *glmnet* R-package (Friedman, Hastie, & Tibshirani, 2010). For the purpose of hyperparameter tuning, Ridge regressions are fitted for a sequence of 101 values of λ with $\lambda \in [10^5, 10^{-5}]$ where the exponent is gradually decreased by 0.1. Again, the $RMSE$ is estimated across the 10-folds for each λ. Also similar to the stepwise forward approach, the optimal model is eventually chosen while taking the 1-se-rule into account.

(4) Lasso regression: The procedure for training and selecting the optimal Lasso model is identical to that of the Ridge model. The only difference is the fact that the Lasso performs not only regularisation but also variable selection which is why the optimal Lasso model will not include all the available predictors.

(5) Polynomial regression: Section 3.2 has already graphically highlighted that there exists a non-linear relationship between the total Wikipedia page views and the response, the transfer fee. Moreover, literature frequently suggests an inverse U-shape for the relationship between age and transfer fee as the player's performance usually climaxes in his mid-twenties (Herm et al., 2014; Patnaik, Praharaj, Prakash, & Samdani, 2019; see also appendix A.3, Fig. 14). Therefore, several polynomial regressions are fitted to capture such non-linear relationships. More precisely, polynomials are generated for all non-dummy predictors to fit quadratic ($d = 2$), cubic ($d = 3$) and quartic ($d = 4$) regressions in addition to the linear model ($d = 1$). Thereby, the degree $d = 4$ is chosen as the maximum degree since for $d > 4$ the model is expected to overfit (James et al., 2013). Again, the optimal d is identified by employing 10-fold CV and comparing the resulting test-$RMSE$. Note that the same d is applied to each predictor, i.e. the same non-linear relationship is assumed for each predictor of the same model of degree d.

The final data analysis process is summarised in Fig. 7. The figure illustrates the main process steps, such as data collection, pre-processing, data transformation and modelling as well as the vital employed R-packages in italics. Under the *ggplot2*-family (Wickham et al., 2019b) this paper summarises, amongst others, the R-packages *ggplot2*, *ggrepel*, *ggraph*, *Ggally* and *ggcorrplot*. Moreover, almost each step employs functions from the popular tidyverse R-packages *dplyr* and *purrr* (Henry & Wickham, 2019; Wickham, François, Henry & Müller, 2019a). Code for the whole analytics pipeline can be found under https://github.com/simonschoe/project_player_transfers.

Fig. 7 Data analysis process diagram



4. Results

4.1 Determinants of Transfer Fees

Tab. 3 presents the main findings of the general linear model with outlier correction. As stated in section 3.3, the model explains more than 70% of the variability of the response, i.e. the transfer price ($R^2 = 0.7279$, $\bar{R}^2 = 0.7221$). Overall the model finds 27 significant predictors (excl. intercept) with 17 of them being significant at the 1% level. The total number of Wikipedia page views turns out at the predictor with the lowest overall p-value ($\beta = 0.00001167$, $se(\beta) = 0.00000038$). Moreover, year fixed as well as league fixed effects are identified as statistically significant predictors. Further, only two of the twelve position dummies indicate a significant association with the observed transfer fee. Finally, it is striking that a player who transfers to or within his home country does not generate a higher transfer fee per sé. Only for national players who transfer to or within their home league the model finds explanatory evidence for the transfer fee.

Tab. 3 Determinants of transfer fees

Predictor	β	$se(\beta)$	Predictor	β	$se(\beta)$
Intercept	-37.6596***	8.9736	National player	0.7379*	0.3891
Year:			To home country	0.5389	0.4187
2014	1.9038***	0.5603	National player to home	-1.7786***	0.6520
2015	2.1566***	0.5519	country		
2016	0.9084*	0.5428	Total Wikipedia page views	0.0000***	0.0000
2017	2.7880***	0.5284	Rank previous season team 1	-0.0870**	0.0396
2018	2.3734***	0.5655	Average age team 1	-0.0279	0.1540
2019	1.3777*	0.7759	Average market value team 1	0.00124	0.0331
Age	-0.6748***	0.0934	*League team 1:*		
Height	6.9977**	2.8571	Primera División	-0.6171	0.6552
Preferred foot:			Ligue 1	-0.3036	0.6412
Left	1.0599***	0.3688	Serie A	0.7106	0.6346
Position:			Liga NOS	-0.1920	0.7730
Goalkeeper	-3.3842***	0.9197	Eredivisie	-2.2652***	0.8131
Left defense	-1.9842**	0.8272	Premier League	-0.1070	0.6617
Central defense	-0.6431	0.7185	Jupiler Pro League	-0.7696	0.8382
Right defense	-0.8404	0.7984	Premier Liga	0.7073	0.9876
Defensive midfield	-0.2029	0.7472	Süper Lig	0.1211	0.9564
Left midfield	-1.7425	1.7420	Rank previous season team 2	0.0335	0.0429
Central midfield	-0.1901	0.6917	Average age team 2	-0.2296	0.1573
Right midfield	0.5899	1.7448	Average market value team 2	0.6459***	0.0379
Secondary attack	0.2848	1.3087	*League team 2:*		
Left attack	0.2039	0.7580	Primera División	0.5892	0.6375
Central attack	0.1043	0.6810	Ligue 1	1.6656**	0.6472
Right attack	0.2187	0.7720	Serie A	1.5326**	0.6166
FIFA score	0.6142***	0.0550	Liga NOS	-0.4244	0.9441
FIFA unused potential	-0.0178	0.0854	Eredivisie	1.3094	0.9629
Play time	0.0009***	0.0002	Premier League	5.5079***	0.6089
Number of previous transfers	-0.5432***	0.1403	Jupiler Pro League	2.1767**	0.8772
Volume of previous transfers	0.0659***	0.0189	Premier Liga	2.5510**	1.0096
Titles won	-0.2419***	0.0499	Süper Lig	1.1547	0.8509

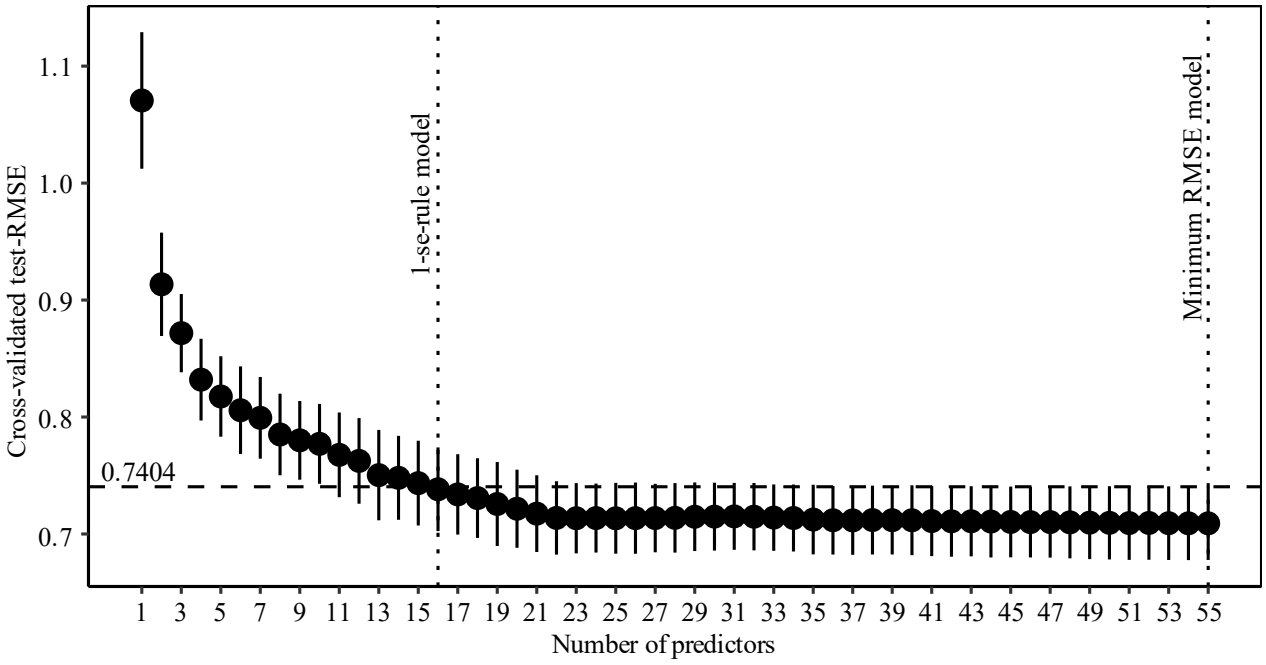
The table presents the results of the linear model with outlier correction. Depicted are the regression coefficients β as well as the standard error $se(\beta)$ for each predictor. Categorical respectively dummy variables are indicated by italics. Statistical significance is indicated by asterisks as follows: *** (1% level), ** (5% level), * (10% level).

4.2 Model Evaluation

Since the focus of the second part of this paper is more concerned with finding the optimal model for predicting football players' transfer fees, the results and discussion in this section will centre around the predictive accuracy and peculiarities of the employed prediction models. In addition, for the variable selection methods a majority vote is executed to present the most frequent predictors included in the final models.

First, the *test-RMSE* over all ten folds for the linear model amounts to 0.7091 (*train-RMSE* $= 0.6905$). Subsequently, this value serves as an accuracy benchmark for the more advanced modelling techniques. Since the linear model incorporates all 55 predictors as well as an intercept, stepwise forward selection is performed next to evaluate the potential need for variable selection. Fig. 8 illustrates the results of the CV approach with the cross-validated *test-RMSE* and *test-RMSE* standard errors being plotted against the respective number of predictors included in the model. The findings suggest that the baseline linear model discussed before ($p = 55$) produces the lowest *test-RMSE*. However, the plot also evidences that the model accuracy barely improves for $22 \leq p \leq 55$. By applying the 1-se-rule the stepwise forward selection technique proposes a model with $p = 16$ predictors plus intercept, thus suggesting only a subset of the overall set of predictors. The 1-se-rule stepwise forward model produces a *test-RMSE* of 0.7382 (*train-RMSE* $= 0.7253$). According to the majority vote presented in Tab. 4, only some of the predictors discussed in the previous section actually make the cut into the prescribed model. Hence, it can be inferred that the log-transformation and standardisations motivated in section 3.3 clearly affect the importance of the individual variables for the overall model.

Fig. 8 Cross-validated RMSE for the linear model with stepwise forward selection



Tab. 4 Majority vote – Stepwise forward selection

Predictor	Vote	Predictor	Vote
Age	10	*League team 2:* Süper Lig	10
Average market value team 2	10	*Position:* Central attack	10
FIFA score	10	Rank previous season team 2	10
League team 2: Eredivisie	10	FIFA unused potential	9
League team 2: Jupiler Pro League	10	*Year:* 2018	9
League team 2: Liga NOS	10	*Year:* 2019	9
League team 2: Premier League	10	Average market value team 1	8
League team 2: Serie A	10	*Year:* 2017	7

The table presents the majority vote for the stepwise forward selection procedure. It contains the $p = 16$ with occured most frequently in the 1-se-rule model across all ten folds. Categorical variables are indicated by italics.

Next, two regularised linear models are fitted: the Ridge and the Lasso. In both cases, CV is used to estimate the in- and out-of-sample $RMSE$ and the hyperparameter λ is tuned to identify the minimum $RMSE$ as well as the 1-se-rule model. Fig. 9 and Fig. 10 illustrate the relationship between the cross-validated train- and test-$RMSE$ for the Ridge respectively the Lasso model. The plots suggest that an optimal λ for the Ridge model of $\lambda \in [log(-10), log(-1)]$ respectively an optimal λ for the Lasso model of $\lambda \in [log(-10), log(-3)]$. In both cases, the models perform marginally better on the train than on the test set with train- and test-$RMSE$ converging to each other with λ greater than the previously defined ranges.

Fig. 9 Cross-validated RMSE for the Ridge regression

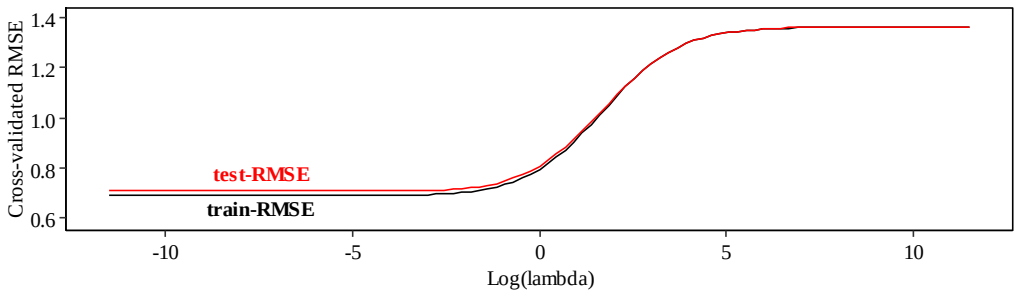


Fig. 10 Cross-validated RMSE for the Lasso regression

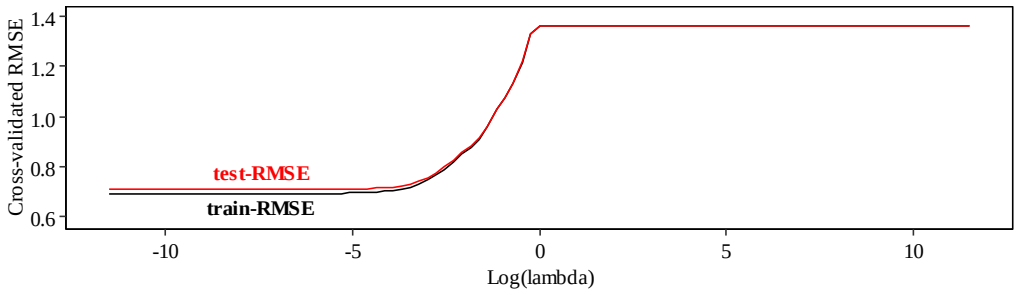


Fig. 11 and Fig. 12 picture the results of the tuning procedure. For the Ridge and Lasso model the optimal $log\ (\lambda_{min})$ amounts to -4.37 ($\lambda_{min,Ridge} = 0.0126$) and -6.68 ($\lambda_{min,Lasso} = 0.0013$) respectively. In the optimum, the results indicate that the Lasso performs marginally better than the Ridge model with test-$RMSE_{min,Lasso} = 0.7089 <$ test-$RMSE_{min,Ridge} = 0.7090$. By applying the 1-$se$-rule, the procedure again yields two models that are expected to be less prone to overfitting ($\lambda_{1-se-rule,Ridge} = 0.3162$, test-$RMSE_{1-se-rule,Ridge} = 0.7381$ and $\lambda_{1-se-rule,Lasso} = 0.0316$, test-$RMSE_{1-se-rule,Lasso} = 0.7298$). Note that for both models, the shrinkage parameter λ does not apply to the intercept (James et al., 2013).

Finally, Tab. 5 presents the majority vote across all ten folds for the Lasso model, i.e. the coefficients of the 1-se-rule model that survive the variable selection procedure. In comparison to the 1-se-rule stepwise forward model the Lasso model contains 19 predictors in addition to the intercept whereby ten predictors are incorporated in both, the 1-se-rule stepwise forward and Lasso model. There is no predictor for $\lambda_{1-se-rule,Lasso} = 0.0316$ that occurs in less than seven of the fitted models. Put differently, either a predictor is shrunken to zero by the Lasso model or included in the 1-se-rule model with at least 70% probability. The threshold predictor is represented by the number of total Wikipedia page views which is, in contrast to the stepwise forward model, again considered as an important predictor of the magnitude of the transfer fee.

Fig. 11 Standardised Ridge coefficients

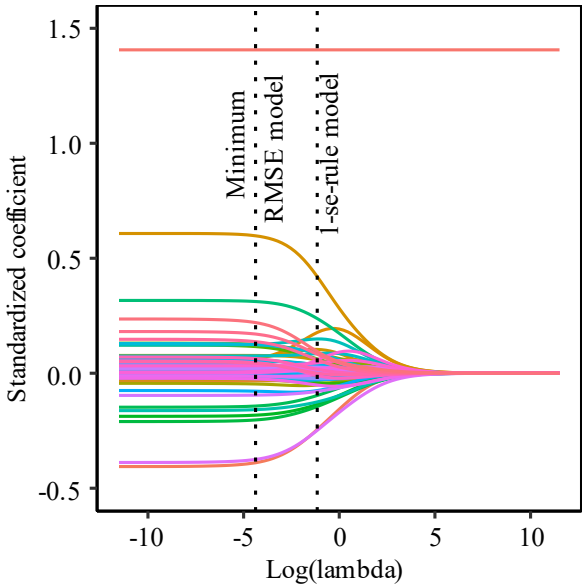
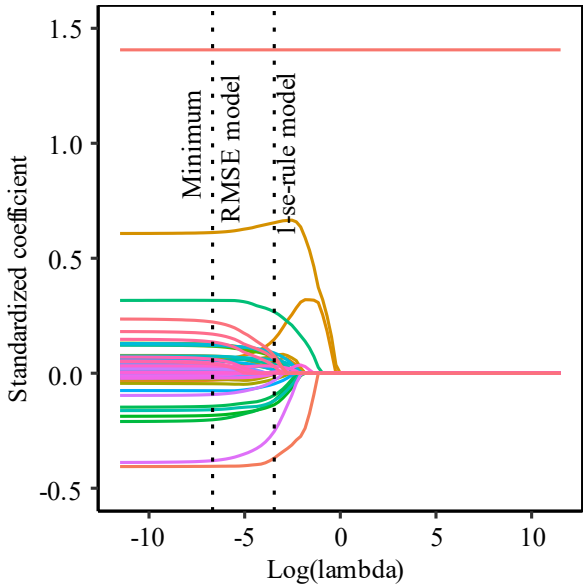


Fig. 12 Standardised Lasso coefficients



In a last step, four polynomial models are fitted with a polynomial degree of $\boldsymbol{d} \in [1, 2, 3, 4]$. Fig. 13 highlights that for $d \geq 3$ the model is prone to overfitting with the test-$RMSE_{d=4} = 32.5237 \gg$ train-$RMSE_{d=4} = 0.6451$. Correspondingly, the train-$RMSE$ decreases consistently in the degree of freedoms: train-$RMSE_{d=1} = 0.6905$

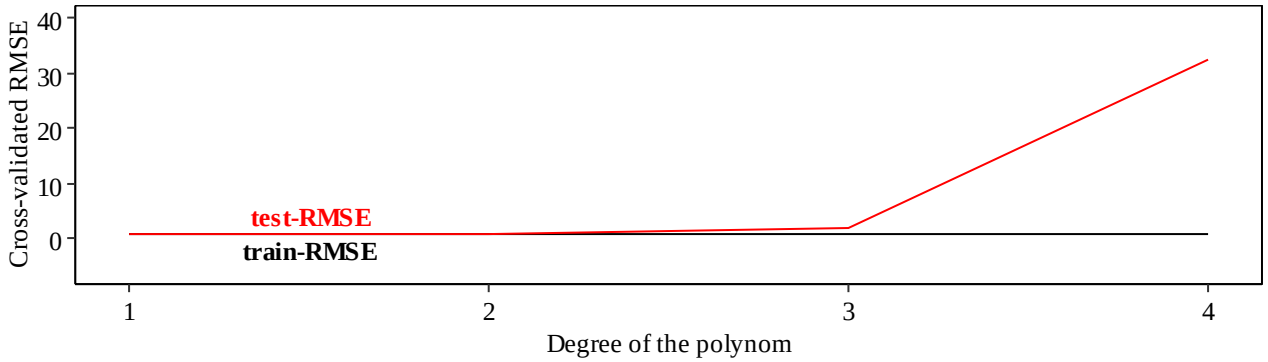
$> \cdots >$ train-$RMSE_{d=4} = 0.6451$. Although hard to infer from the plot, the minimum test-$RMSE$ can be achieved by implementing the quadratic model with test-$RMSE_{d=2} = 0.7060$ (train-$RMSE_{d=2} = 0.6661$).

Tab. 5 Majority vote – Lasso regression

Predictor	Vote	Predictor	Vote
Average market value team 1	10	*Position:* Central attack	10
Average market value team 2	10	*Position:* Central midfield	10
FIFA score	10	Volume of previous transfers	10
FIFA unused potential	10	*Year:* 2017	10
League team 1: Premier League	10	*Year:* 2018	10
League team 1: Serie A	10	*Year:* 2019	10
League team 2: Premier League	10	Height	9
League team 2: Serie A	10	League team 1: Ligue 1	9
Play time	10	Total Wikipedia page views	7
National player	10		

The table presents the majority vote for the variable selection procedure implemented by the Lasso method. It contains all predictors that enter the 1-*se*-rule model. A "vote" value of ten indicates that the predictor is included in the model independent of the CV split. Categorical respectively dummy variables are indicated by italics.

Fig. 13 Cross-validated RMSE for the polynomial regression



Tab. 6 summarises the main findings of the predictive modelling part. Overall, the quadratic model yields lowest test-$RMSE$, followed by the minimum $RMSE$ Lasso and Ridge model – all performing slightly better than the benchmark linear model in term of test-$RMSE$. Naturally, the baseline linear model, the polynomial model of degree $d = 1$ and the stepwise forward model with $p = 55$ all produce the same results. The 1-*se*-rule models are all less accurate than the full specification with the cubic and quartic model exhibiting the poorest overall fit. Further, the models confirm the theoretical property that the *train-RMSE* should be lower than the *test-RMSE* by definition (cf. section 1.2). Consequently, for all models, except the latter two, the $\Delta RMSE$, i.e. the difference between test- and train-$RMSE$, lies on the interval $\Delta RMSE \in [0.0114; 0.0399]$.

Tab. 6 Model summary and comparison

	Model	Hyper-parameter	*train-RMSE*	*test-RMSE*	$\Delta RMSE$
1	Polynomial regression (quadratic)	$d = 2$	0.6661	0.7060	0.0399
2	Lasso regression (min.)	$\lambda = 0.0013$	0.6907	0.7089	0.0182
3	Ridge regression (min.)	$\lambda = 0.0126$	0.6907	0.7090	0.0182
4	Linear regression	-	0.6905	0.7091	0.0186
5	Polynomial regression (linear)	$d = 1$	0.6905	0.7091	0.0186
6	Linear stepwise forward regression (min.)	$p = 55$	0.6905	0.7091	0.0186
7	Lasso regression (1-se-rule)	$\lambda = 0.0316$	0.7184	0.7298	0.0114
8	Ridge regression (1-se-rule)	$\lambda = 0.3162$	0.7228	0.7381	0.0153
9	Linear stepwise forward regression (1-se-rule)	$p = 16$	0.7253	0.7382	0.0130
10	Polynomial regression (cubic)	$d = 3$	0.6569	1.8104	1.1534
11	Polynomial regression (quartic)	$d = 4$	0.6451	32.5237	31.8786

The table summarises the main findings and compares the different prediction models. In total, it presents the *train-RMSE*, *test-RMSE* and $\Delta RMSE$ (defined as *test-RMSE* $-$ *train-RMSE*) for eleven models. The add-on "min." refers to the model that produces with the lowest *test-RMSE* whereas the add-on "1-se-rule" refers to the model with the lowest *test-RMSE* according to the 1-se-rule. Finally, it reports the (tuned) hyperparameter for each specification.

5. Discussion

The aim of this section is to discuss the previously presented results while addressing *RQ1-RQ4* as formulated in the introduction of this paper. Thereby, the results from section 4.1 respectively 4.2 will form the basis for evaluating the first two respectively the latter two questions.

5.1 RQ1 – Value Drivers of Transfer Fees

In order to address RQ1, the discussion focusses on selected predictors that are identified as significant at the conventional levels by the baseline linear model with outlier correction. First, the coefficient of the age predictor (β = -0.6748) confirms the results of previous research (Dobson et al., 2000; Garcia-del-Barrio & Pujol, 2007; Medcalfe, 2008; Herm et al., 2014; Drut & Duhautois, 2017; Müller et al., 2017; Ante, 2019; Serna Rodríguez et al., 2019; Singh & Lamba, 2019). As the age of the player increases the transfer fee decreases. The coefficient illustrates that the transfer fee decreases by ~ €675,000 if the age of the player increases by one year reflecting the fact that a football player's value and performance is linked to his physical shape (cf. also Fig. 15 in appendix A.3 for the juxtaposition of age and FIFA score respectively FIFA unused potential).

Second, also the height predictor shows significance ($\beta = 6.9977$). According to the results, each centimetre added to the height generates an increase of almost €70,000 in transfer fees. However, this inference

should be treated with caution as the height variable is correlated with other player-specific attributes, such as the player position (cf. the pairwise correlations in appendix A.3, Tab. 12).

Third, the coefficient of the left-foot dummy is significant and positive ($\beta = 1.0599$) which contradicts the previous research carried out by Ante (2019). Similarly, Serna Rodríguez et al. (2019) do not find foot preference to be significant. Compared to a right-footed player, left-footed players generate a transfer fee premium of €1.06 million. As can be seen in appendix A.3, Fig. 16, right-footed players are overrepresented in the dataset which suggests that left-footed players may be rarer in the population, thus generating a higher transfer fee.

Fourth, in line with prior research (Müller et al., 2017; Singh & Lamba, 2019; Ante, 2019), the number of minutes played in the season prior to the transfer across all tournaments and leagues is positive and significant ($\beta = 0.0009$). For example, if a player plays in 75% of the seasonal Bundesliga or Eredivisie matches – which would amount to 34 matches $*$ 90 minutes $*$ 75% = 2,295 minutes – the transfer fee increases by €2.12 million.

Fifth, the model produces an unexpected result for the number of titles won. The variable shows a negative coefficient of $\beta = -0.2419$. One possible explanation for this relationship could be the noise captured by this predictor. That is, the predictor incorporates a wide variety of different titles, such as international, national, personal or youth titles, and hence treats each title equivalently without accounting for the prestige of the title. Thus, making the 1st place in the 2nd division equally weighted compared to winning the world cup.

Sixth, the OLS model produces surprising results for the number ($\beta = -0.5432$) and monetary volume of previous transfers ($\beta = 0.0659$). The coefficients suggest that a player with a higher number of previous transfers is transferred for a lower price which may occur if the player is passed on from club to club. Inversely, the transfer fee increases in the cumulative monetary volume of previous transfers. Altogether, the two predictors imply that a player that transfers infrequently but for higher fees is also more likely to generate higher transfer fee payments in the future. This may reflect a scenario of a relatively loyal and highly valued player.

Seventh, the influence of the league a player transfers to aligns with previous research (Dobson et al., 2000; Ante, 2019; Serna Rodríguez et al., 2019). According to the results, a transfer to the Premier League has the strongest impact ($\beta = 5.5079$), thus providing evidence for the "English premium" (cf. section 1.1). Moreover, also the Italian, Belgium, French and Russian leagues have a positive influence on the transfer fee. Interestingly, only the dummy for the Dutch Eredivisie shows significance concerning the league a player transfers from ($\beta = -2.2652$). Hence, it can be summarised that the league a player transfers to exhibits a higher explanatory power than the league he comes from. Note that all dummies are interpreted relative to the Bundesliga.

Finally, the newly introduced dummy 'national player to home country' yields significant and negative results ($\beta = -1.7786$). Hence, the positive influence of being a national player ($\beta = 0.7379$) is evaporated by

the fact that the national player is transferred to or within his home country. Two alternative scenarios offer potential explanations for this counter-intuitive finding. First, the coefficients may reflect transferred national players that return to their home country after failing on the international stage. Second, it may be the case that returning national players are heading towards the end of their career, thus having the desire to retire in their home country. Both cases would give rise to the transfer fee for the national player being relatively lower. However, these hypotheses certainly require additional empirical analysis in the future.

5.2 RQ2 – Transfer Fees and Player Popularity

To address *RQ2*, this paper proxies a player's popularity by the total number of Wikipedia page views in the season prior to the transfer for five main European languages (cf. section X). In line with literature (Müller et al., 2017; Singh & Lamba, 2019; Ante, 2019; Herm et al., 2014; Garcia-del-Barrio & Pujol, 2007), the linear model attests the predictor a highly significant and positive influence on the transfer fee magnitude ($\beta = 0.00001167$). From the coefficient it can be inferred that one additional page view increases the player's transfer fee by €11.67. On a more meaningful scale, 91,000 additional seasonal views – reflecting the sample median – relate to a transfer fee that is roughly €1,06 million higher relative to the transfer fee of an entirely unknown player. Consequently, in the light of *RQ2* it can be stated that a player's popularity is a significant and economically meaningful driver of transfer fees in the football industry. Thereby, the estimate of €1,06 million may give a rough approximation of the brand value of player that can be attributed to his level of popularity.

5.3 RQ3 – Models for Predicting Transfer Fees

In line with *RQ3* this paper tries to address the issue of predictive accuracy – that is, which method from the regression family performs best in predicting transfer fees on the European football market? The results in section 4.2 identify the polynomial model of degree $d = 2$ as the model with the best out-of-sample accuracy. This finding evidences that there is arguably a significant proportion of non-linear relationships in the data, thus highlighting the need for considering non-linear models in the context of transfer fee prediction.

However, it should be stated that the accuracy of the quadratic model is only slightly superior to the Ridge and Lasso which, in turn, perform only marginally better than the full, linear specification. Therefore, inferences about an absolute model ranking should be viewed with caution. This is especially the case since the respective $RMSE$ are not easily interpretable, due to the various transformation steps discussed in section 3.3.

Another issue for selecting the best prediction model refers to the issue of overfitting which is assumed to be present if there is a substantial disparity between train- and test-$RMSE$ (Ghojogh & Crowley, 2019). The $\Delta RMSE$ suggest that the 1-se-rule models may be the least prone to overfitting (e.g., $\Delta RMSE_{1-se-rule,Lasso} =$

$0.0114 < \Delta RMSE_{min,Lasso} = 0.0182$). Hence, for the purpose of addressing *RQ4* in the next section, this paper refers to the 1-se-rule Lasso model as it displays the minimum overall $\Delta RMSE$ of 0.0114. In particular, it is expected that this model has the best properties for predicting transfer fees for entirely new cases.

Finally, it should be stated that the given dataset provides a data-rich environment in which the number of cases considerably exceeds the number of predictors. In this scenario, the fitting procedure tends to generate rather complex models that explain a large fraction of the variation of the response (National Research Council, 2013). Therefore, it is not surprising to find generally low values for $\Delta RMSE$ for most of the models.

5.4 RQ4 – Transfer Market Madness

Lastly, this paper's fourth research question (*RQ4*) tries to forecast what can be expected in terms of future transfer fees. For that purpose, the ten 1-se-rule Lasso models fitted during the CV procedure are deployed and applied to new data which comes in the form of four frequently rumoured player transfers:

1) The reverse transfer of Neymar from FC Paris Saint-Germain (League 1) to FC Barcelona (Primera División) (Laurens, 2019). As the highest ever paid transfer fee of €222 million in 2017, this rumour raises the question if the observed sum can even be topped.
2) The return of the FC Bayern München (1. Bundesliga) veteran Javi Martínez to his home country. That is, rumours say the defensive midfielder will return to his youth club Athletic Bilbao (Primera División) for a price of €10 million (ETB, 2019).
3) The transition of German super-talent Kai Havertz (Bayer 04 Leverkusen, 1. Bundesliga) to the Premiere League (Manchester United) for an estimated fee of close to €100 million (Metro Sport Reporter, 2019).
4) Finally, as a last rumour this paper considers a potential transfer of Jadon Sancho from Borussia Dortmund (1. Bundesliga) to Real Madrid (Primera División). In this case, news outlets speculate the required fee to amount to at least £100 million (€119 million) (Kajumba, 2019).

Altogether, these four hypothetical transfers relate to a well-established top-star (1), a highly experienced player (2) as well as two youth players with very promising prospects (3, 4). For each of the four hypothetical transfers the data is gathered as discussed in section 3.1 with only the actual transfer fee being absent.

Tab. 7 summarises the prediction results. The average 1-se-rule Lasso model predicts a transfer fee of €279.64, €9.72, €121.46 and €120.84 million for Neymar, Javi Martínez, Kai Havertz and Jadon Sancho respectively. First, these findings underline the accuracy of the prediction model with estimated fees lying in close proximity to the prices assumed by the media (if available). For example, the model almost perfectly approximates the €10 and €120 million fee hypothesised for Javi Martínez and Jadon Sancho.

Tab. 7 Transfer fee prediction of selected transfer rumours

				Predicted transfer fee ($\hat{y}$) [in million €]		
	Player Name	Team 1 (*seller*)	Team 2 (*buyer*)	$\bar{x}$	min	max
1	Neymar	FC Paris Saint-Germain	FC Barcelona	279.64	200.20	355.58
2	Javi Martínez	FC Bayern München	Athletic Bilbao	9.72	9.37	10.16
3	Kai Havertz	Bayer 04 Leverkusen	Manchester United	121.46	113.91	127.70
4	Jadon Sancho	Borussia Dortmund	Real Madrid	120.84	111.08	131.76

The table presents the predictions for the four rumoured player transfers. The table contains the name of the player involved in the transfer, the name of the team which sells (team 1) respectively buys the player (team 2). Moreover, it presents the mean ($\bar{x}$), minimum (min) and maximum (max) for the predicted transfer fee. The values are predicted using the ten 1-se-rule Lasso model fitted in the course of the CV procedure (cf. section 3.3).

Second, the predictions suggest where the transfer madness is heading to. In particular, the observation relating to Neymar indicates that transfer fees at the top-end may skyrocket even more. Yet, the minimum ($min(\hat{y}) = 200.20$) and maximum ($max(\hat{y}) = 355.58$) estimates emphasise that for these price regions predictions are extremely noisy. In contrary, transfer fees for relatively old and well-appreciated players appear to barely participate in the inflated development of fees with an average predicted price for Javi Martínez of €9.72 million. Finally, it could be argued that it is especially the highly talented players that are caught in the vortex of the transfer madness. Where there are only ten transfers to date that cracked the €100 million mark (as of 16/12/2019), rumour 3 and 4 illustrate that there may be more cases in the future – likely also occurring at a faster pace.

6. Limitations

Even though the data collection, the merging of the different datasets as well as the predictive modelling is been conducted with utmost diligence, it cannot be ruled out that this paper is subject to important limitations.

6.1 Dataset Limitations

First, in section 1.1 it has been highlighted that the player's agent, the remaining contract term as well as specific contractual clauses, such as a pre-installed buy-out clause, play a central role in determining the transfer fee. For example, it may be expected that star agents with a prestigious portfolio of widely-known players frequently leverage their negotiating power to bargain higher transfer fees, as they receive a share of the overall deal value. Yet, it should be highlighted that Herm et al., 2014 did not find empirical evidence confirming this hypothesis. Moreover, a long time-to-maturity of the player's contract at the time of transfer can imply higher transfer fees as the seller holds a superior position in the price negotiation (Geey, 2019). Finally, a fixed buy-out clause may cause players to be transferred for significantly more or less than their fair value (Gerrard,

2002; Geey, 2019). However, data on those cases is extremely cumbersome to collect, especially with regards to historical transfers for which the data may not even be available at all.

Second, transfer data from *transfermarkt.de* could not be scraped on a daily but only on a seasonal basis. That is, for each player the web scraper returns only the season in which the player has been transferred and not the exact date. That leads to a central simplifying assumption when reconciling the three datasets: It is assumed that transfer took place at the last day of the summer transfer window which is frequently the last day in August. This also implies that the data does not differentiate between transfers executed in the summer respectively the winter transfer window. Yet, it may not be surprising when players regularly trade at a premium during the winter window as teams may be desperate to replace injured players or reinforce the team after a poor performance in the first half of the season.

Third, the predictive models only account for fixed effects on the year and league level. As has been illustrated in section 1.1, English teams tend to pay an "English premium" due to extensive internal funds stemming from relatively large TV licensing deals paid on the British market. Therefore, the models may be enhanced by including a predictor that takes this very fact into account and thus isolates the influence of licensing deals. Again, getting access to team or even league level data of licensing deals depicts a major challenge which is why this paper ultimately introduces league fixed effects to account for the lack of licensing data.

Fourth, the total Wikipedia page views as one of the central predictors discussed in this work is subject to potential limitations. Although, the web scraper has been designed as robust as possible it might be the case that the Google-searches do not identify the actual Wikipedia page of the player amongst the search results. Still, a manual and randomised check by the authors confirms that the approach yields highly satisfactory results. Moreover, this approach relies on the data that is accessed via the API offered by the *wikipediatrend* R-package. As the author of the R-package states: "One issue is article coverage" (Meissner, 2019b, p. 1). Thus, missing or noisy page views data recorded by the database underlying the R-package may distort the predictor. An alternative to proxying for a player's popularity constitutes the analysis of Facebook or Instagram data related to each player. Unfortunately, since the Cambridge Analytica incident has caused a change of heart among the major social media platform operators, this data is not accessible to the paper's authors as well.

Fifth, it has been emphasised that two-footedness is a highly attractive player characteristic as it allows players to be more flexibly deployable on the field (Patnaik et al., 2019; Herm et al., 2014). Yet, the Kaggle data only includes the preferred foot of the player which does thus not allow for an investigation of this hypothesis.

Sixth and last, the analysis is restricted to the ten major European leagues and the 2013-2019 period. Thus, the findings do not necessarily generalise to other football leagues or former time periods.

6.2 Methodological Limitations

Aside from the shortcomings of the dataset itself, future research should also consider the implementation of alternative or more advanced prediction techniques.

First, principal component regression (PCR) comes forward as a technique to account for potential multicollinearity (Wicklin, 2017). When multicollinearity occurs, OLS estimates are unbiased, but their variances tend to be large (Jackson, 1991). PCR generates linear combinations of existing predictors, thus eventually yielding more reliable estimates. The not used variables give insights in which linear combinations of variables are responsible for the collinearities (Jackson, 1991).

Second, an Elastic Net could be trained on the underlying data. As a hybrid of the Ridge and the Lasso it employs shrinkage techniques to generate penalised coefficients for redundant predictors (Zou & Hastie, 2005). In this context, empirical studies have suggested that the Elastic Net can outperform the Lasso on data with highly correlated predictors (Zou & Hastie, 2005; Schams, 2019), thus predestining the technique for an application in the domain of transfer fee prediction.

Third, a regression tree learner can be implemented as a non-parametric model that splits the data into smaller subsamples in which the response is predicted based on the partition's average response (Breiman, Friedman, Olshen & Stone, 1984). Since the regression tree does not impose any assumptions on the data, it is well-suited to handle non-linear relations and simultaneously performs variable selection (Gupta, 2017). Moreover, the model can be optimised by using bagging and boosting techniques and by generating a more robust random forest (Boehmke, 2019).

Lastly, it should be stated that the time trend in the recent development of transfer fees has been identified as a potential issue when applying of k-fold CV (Müller et al., 2017). Consequently, the CV approach discussed in section 3.3 could be refined by relying on stratified sampling (Thompson, 2012) to draw a balanced split that maintains the underlying structure in the dataset with respect to distribution of transfers across years.

7. Implications for Future Research and Practice

The methodology and findings presented by this paper bear several implications for future research. First, the quadratic model is identified as the most accurate model. Hence, future research should consider potential non-linear relationships when building a prediction model for transfer fees. Second, the comparison of the predictions of the deployed 1-se-rule Lasso model with the rumoured transfer fees (cf. section 5.4) suggests that the given dataset can compensate for some of the omitted predictors discussed in section 6. Hence, future research should not be discouraged by the absence of theoretically meaningful predictors. Third, the paper proposes numerous predictors that are statistically significant and also survive variable selection. As such, the results may guide future research in picking the best predictors for football transfer fees. Fourth, the paper advocates

the use of novel predictors, such as the number and monetary volume of previous transfers, titles won, or whether or not a national player transfer to a team of his home country, which have not yet been considered by prior research. Fifth, the approach and data employed in this paper could also be projected to the domain of market value prediction. Although section 1.1 outlines that there is a conceptual difference between the transfer fee and a player's market value, both are found to be highly correlated (He, Cachucho, & Knobbe, 2015). Accordingly, the data indicates a strong positive relationship between transfer fee and market value with the realised transfer fees being slightly below the market value at the time of the transfer (cf. appendix A.3, Fig. 17).

Moreover, the findings also yield several implications from the practitioner's point of view. First, the use of data-driven analytical methods – as proposed by this paper in the form various regression techniques – can be expected to aid managers in terms of decision-making (Geey, 2019). That is, a team's management could complement its subjective judgement with the quantitative output of a prediction model in the course of a transfer negotiation. Naturally, the same rational also applies to a player's agent who might want to validate the worth of his client. Second, section 5.4 once more gives rise to the perception that the transfer madness has not yet reached its climax. The presented predictions do not appear to point at a slowing down of the money inflow to the transfer market raising the question: When will the excessive costs become unbearable for the average team and how will it shape the competitive equilibrium in the future of football? The answer to this question can be expected to have important implications for the policy makers in designing effective regulations, such as the UEFA's financial fair play rules (Čeferin & Theodoris, 2018). In addition, the question also shifts the view towards the ones who are shouldering the burden of this development – the players. Not only are transfer fees in the tens and hundreds of millions questionable from an ethical perspective (McLean, 2017), but also can they be criticised from a psychological perspective. For example, Dohmen (2008) emphasises the burdensome effects of monetary incentives on a player's performance in the light of failing expectations. Further, the team's management is required to earn a return on the initial investment in the player thus amplifying the pressure put on the player (Smith & Stewart, 2010).

Eventually, it should be stated that the techniques advocated in this paper are not limited to the domain of football. By accounting for the nature of other sports and the peculiarities in the underlying data, the methodology could be easily adapted by researchers or practitioners from other fields of sports analytics.

8. Conclusion

The goal of this paper is to explore the applicability of five predictive modelling techniques in the context of the European football transfer market. By doing so, the following four research questions are addressed in the course of the study:

RQ1: What are the important value drivers that determine transfer fees in the European football market?
RQ2: Is there a significant influence of a player's popularity on transfer fees?
RQ3: Which model for estimating football transfer fees performs best in terms of predictive accuracy?
RQ4: Where is the transfer market madness heading to?

Regarding the first question, there are several variables that are significant, positive drivers of football player transfer fees: height, left-footedness, play time, monetary volume of previous transfers, the league a player transfers to, and the year of the transfer. There are also several variables that negatively influence the fee, such as: age, the goalkeeper position, the number of titles won, the number of previous transfers, if a player transferred from the Eredivisie, and lastly if a national player transfers to his home country. In particular, findings for the second question confirm that player popularity has a positive significant effect on transfer prices. This is proxied by looking if the number of Wikipedia page views per player in the season prior to the transfer is positively associated with the transfer fee of the player. Thereby, the median transfer is found to incorporate a €1,06 million price premium due to player popularity.

Further, this paper finds that a polynomial model of degree two performs best in terms of predictive accuracy. All the discussed models perform rather similar as there is not a large variance in the test-$RMSE$. Moreover, it is expected that the 1-se-rule Lasso model is the least prone to overfitting. Finally, by deploying this model and testing it on four frequently rumoured transfers, this paper finds that the predictions are quite close to the hypothesised prices proposed by the media. The findings underline the accuracy of the fitted model and simultaneously suggest that there is no indicator that the transfer market madness will calm down anytime soon.

To put these findings into perspective, several limitations are addressed. On the one hand, various potential value drivers motivated by the prevalent literature are not publicly available and thus not accounted for in the dataset. Moreover, the analysis could be refined by taking the exact transfer date into account as well as by extending it to more time periods and leagues. On the other hand, the predictive modelling toolset could be complemented by using alternative modelling techniques, such as PCR, Elastic Net or regression trees.

Regarding implications for future research, this paper can guide researchers in choosing the best possible transfer fees predictors and the results show that non-linear relationships in the data should be accounted for. Furthermore, this paper proposes that data-driven methods can aid managers and agents in their transfer fee negotiations and that this methodology can also easily be adapted for analytics in other sports.

In the end, the predictions stress that the transfer market will not slow down any time soon. Not only should the excessively high transfer fees be questioned from an ethical perspective, but they can also put a large psychological burden on the oftentimes young football players. Therefore, an interesting and highly relevant question for the future arises: At what point do these transfer costs become unbearable for the average team as well as the players and how will it shape the competitive equilibrium in the future of football?

References

Agresti, A. (2015). Foundations of linear and generalized linear models. John Wiley & Sons.

Ante, L. (2019). Determinants of transfers fees: Evidence from the five major European football leagues. *University of Hamburg*. DOI: 10.13140/RG.2.2.18356.91526/1

Barnard, M., Boor, S., Winn, C., Wood, C., & Wray, I. (2019). World in motion. In *Deloitte Annual Review of Football Finance 2019*. Retrieved from https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: John Wiley. Retrieved from http://www.loc.gov/catdir/description/wiley032/93029289.html

Bodet, G., & Chanavat, N. (2010). Building global football brand equity: Lessons from the Chinese market. *Asia Pacific Journal of Marketing and Logistics*, *22*(1), pp. 55-66. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/13555851011013155/full/html

Boehmke, B. (2019). Regression Trees · UC Business Analytics R Programming Guide. Retrieved 10 December 2019, from http://uc-r.github.io/regression_trees

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *Wadsworth Int. Group*, *37*(15), pp. 237-251.

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), pp. 606-628.

Čeferin, A., & Theodoris, T. (2018). UEFA club licensing and financial fair play regulations: Edition 2018 (PDF File). In *UEFA*. https://www.uefa.com/MultimediaFiles/Download/Tech/uefaorg/General/02/56/20/15/2562015_DOWNLOAD.pdf

Chadwick, S., & Thwaites, D. (2004). Advances in the management of sport sponsorship: fact or fiction? Evidence from English professional soccer. *Journal of General Management*, *30*(1), pp. 39-60. Retrieved from https://pdfs.semanticscholar.org/cdd9/83f1d007c57173949928f7f5f5d08f8d469c.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). ISBN: 978-0805802832

Cook, D. R., & Weisberg, S. (1982). *Residuals and Influence in Regression*. ISBN: 978-0412242809

Cotta, L., Vaz de Melo, P. O. S., Benevenuto, F., & Loureiro, A. A. F. (2013). Using FIFA Soccer video game data for soccer analytics. In *ACM Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. Retrieved from https://homepages.dcc.ufmg.br/~fabricio/download/lssa_fifa_CR.pdf*

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

Dohmen, T. J. (2008). Do professionals choke under pressure? *Journal of Economic Behavior & Organization*, *65*(3), pp. 636–653. DOI: 10.1016/j.jebo.2005.12.004

Drut, B., & Duhautois, R. (2017). Assortative matching using soccer data: Evidence of mobility bias. *Journal of Sports Economics*, *18*(5), pp. 431–447. DOI: 10.1177/1527002515588134

ETB. (2019). *Athletic reportedly negotiating €10 million January transfer deal for Javi Martinez.* In *ETB*. Retrieved 13 December 2019, from https://www.eitb.eus/es/deportes/futbol/athletic/detalle/6830275/el-athletic-trabaja-fichaje-javi-martinez-mercado-invierno/

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der Weg zur Datenanalysen* (8th ed.). Berlin, Heidelberg: Springer Spektrum. DOI: 10.1007/978-3-662-50372-0

Farquhar, P. H. (1989). Managing brand equity. *Journal of Marketing Research, 1*, pp. 24-33.

FIFA (n.d.). Women's football strategy. *In FIFA Women's Football*. Retrieved 11 December 2019, from https://www.fifa.com/womens-football/strategy/

Ford, C. (2018). Interpreting log transformations in a linear model. Retrieved 13 December 2019, from https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

Frick, B. (2007). The football players' labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy*, *54*(3), pp. 422-446. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9485.2007.00423.x

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1). DOI: 10.18637/jss.v033.i01

Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets - sport and economic contribution of Spanish soccer players. *Managerial and Decision Economics, 28*, pp. 57–70. DOI: 10.1002/mde.1313

Geey, D. (2019). *Done deal: An insider's guide to football contracts, multi-million pound transfers and premier league big business*. Bloomsbury Sport.

Gerrard, B. (2002) The muscle drain, coubertobin-type taxes and the international transfer system in association football. *European Sport Management Quarterly*, *2*(1), pp. 47-56. DOI: 10.1080/16184740208721911

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. Retrieved from http://arxiv.org/pdf/1905.12787v1

Gupta, P. (2017). Regression trees - Uc business analytics r programming guide. Retrieved 13 December 2019, from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Harrison, J. (2019). RSelenium: R bindings for 'Selenium WebDriver' (v1.7.5). Retrieved from https://rdrr.io/cran/RSelenium/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Dordrecht: Springer.

He, M., Cachucho, R., & Knobbe, A. (2015). Football player's performance and market value (PDF File). Retrieved from https://dtai.cs.kuleuven.be/events/MLSA15/papers/mlsa15_submission_8.pdf

Henry, L., & Wickham, H. (2019). purrr: Functional programming tools (v0.3.3). Retrieved 16 October 2019, from https://cran.r-project.org/web/packages/purrr/purrr.pdf

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), pp. 484-492. DOI: 10.1016/j.smr.2013.12.006. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S144135231300096X

Jackson, J. E. (1991). *A user's guide to principal components*. New York: John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer. DOI: 10.1007/978-1-4614-7138-7

Kajumba, A. (2019). Real Madrid back in for Jadon Sancho in blow to Manchester United's hopes of landing England star with Dortmund ready to sell him for more than £100m. In *Daily Mail. Retrieved 11 December 2019, from* https://www.dailymail.co.uk/sport/football/article-7648919/Real-Madrid-step-pursuit-Borussia-Dortmund-star-Jadon-Sancho.html

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(4).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1137–1143.

Krennhuber, R. (2008). Die Statistik-Paten. Retrieved 01 November 2019, from http://legacy.ballesterer.at/heft/thema/die-statistik-paten.html

Kuhn, M., & Wickham, H. (2019). Recipes: Preprocessing tools to create design matrices (v0.1.7). Retrieved 01 November, from https://cran.r-project.org/web/packages/recipes/recipes.pdf

Laurens, J. (2019). Neymar's failed PSG to Barcelona move: The definitive story of the biggest transfer that didn't happen. Retrieved 02 November 2019, from https://www.espn.com/soccer/paris-saint-germain/story/3934704/neymars-failed-psg-to-barcelona-move-the-definitive-story-of-the-biggest-transfer-that-didnt-happen

Leone, S. (2019). FIFA 20 complete player dataset: 18k+ players, 100+ attributes extracted from the latest edition of FIFA. Retrieved 10 November 2019, from https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

Lewis, J. (2013). Gareth Bale £105 million transfer to Real Madrid 'agreed'. In *Evening Standard*. Retrieved from https://www.standard.co.uk/sport/football/gareth-bale-105-million-transfer-to-real-madrid-agreed-8743140.html

Lumley, T. (2017). Leaps: Regression subset selection (v3.0). Retrieved 10 November 2019, from https://cran.r-project.org/web/packages/leaps/leaps.pdf

Mathien, H. (2016). European soccer database: 25k+ matches, players & teams attributes for European professional football. Retrieved 02 November 2019, from https://www.kaggle.com/hugomathien/soccer

McLean, M. (2017). What does a professor of business ethics make of football's relationship with money? In *Independent IE*. Retrieved 18 December 2019, from https://www.independent.ie/world-news/and-finally/what-does-a-professor-of-business-ethics-make-of-footballs-relationship-with-money-36087738.html

Medcalfe, S. (2008). English league transfer prices: Is there a racial dimension? A re-examination with new data. *Applied Economics Letters*, *15*(11), pp. 865–867. DOI: 10.1080/13504850600949178

Meissner, P. (2019a). Wikipediatrend: Public subject attention via wikipedia page view Statistics (v2.1.4). Retrieved from https://cran.r-project.org/web/packages/wikipediatrend/wikipediatrend.pdf

Meissner, P. (2019b). Wikipedia page view statistics late 2007 and beyond: The {wikipediatrend} package. Retrieved 13 November 2019, from https://petermeissner.de/blog/2019/10/09/wikipediatrend-v2.1.4/

Metro Sport Reporter. (2019). Kai Havertz opens door to £80m Manchester United transfer. Retrieved 15 November 2019, from https://metro.co.uk/2019/11/06/kai-havertz-opens-door-80m-manchester-united-transfer-11052378/

Müller, 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), pp. 611-624. Retrieved from https://www.sciencedirect.com/science/article/pii/S0377221717304332

Murphy, R. (2019). FIFA player ratings explained: How are the card number & stats decided? Retrieved 12 November 2019, from https://www.goal.com/en-us/news/fifa-player-ratings-explained-how-are-the-card-number-stats/1hszd2fgr7wgf1n2b2yjdpgynu

National Research Council (2013). *Frontiers in massive data analysis*. National Academies Press.

Oprean, V. B., & Oprisor, T. (2014). Accounting for soccer players: Capitalization paradigm vs. expenditure. *Procedia Economics and Finance*, *15*, pp. 1647-1654. Retrieved from https://www.researchgate.net/profile/Tudor_Oprisor/publication/258178537_Accounting_for_Soccer_Players_Capitalization_Paradigm_vs_Expenditure/links/5486112d0cf268d28f044bd4.pdf

Patnaik, D., Praharaj, H., Prakash, K., & Samdani, K. (2019). A study of prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market. In 2019 *IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1-7.

Perepolkin, D. (2019). *polite: Be Nice on the Web* [v0.1.0]. GitHub. Retrieved from https://github.com/dmi3kno/polite

Preuss, H., Haugen, K., & Schubert, M. (2014). UEFA financial fair play: The curse of regulation. *European Journal of Sport Studies*, *2*(1), pp. 33-51. Retrieved from http://www.ejss-journal.com/index.php/uefa-financial-fair-play-the-curse-of-regulation

Raschka, S. (2019). *Model evaluation, model selection, and algorithm selection in machine learning*. Retrieved from http://arxiv.org/pdf/1811.12808v2

Rohde, M., & Breuer, C. (2016). Europe's elite football: Financial growth, sporting success, transfer investment, and private majority investors. *International Journal of Financial Studies*, *4*(2), p. 1-20. DOI: 10.3390/ijfs4020012

Rousseeuw, P. J. (1987). *Robust regression and outlier detection*. John Wiley & Sons.

Rumsby, B. (2013). *Gareth Bale's proposed transfer to Real Madrid for £100 million is a 'joke' and 'crazy', says Arsene Wenger*. In *The Telegraph*. Retrieved from https://www.telegraph.co.uk/sport/football/teams/arsenal/10218901/Gareth-Bales-proposed-transfer-to-Real-Madrid-for-100-million-is-a-joke-and-crazy-says-Arsene-Wenger.html

Schams, A. (2019). Bias, variance, and regularization in linear regression: Lasso, ridge, and elastic net - differences and uses. Retrieved 10 December 2019, from https://towardsdatascience.com/bias-variance-and-regularization-in-linear-regression-lasso-ridge-and-elastic-net-8bf81991d0c5

Serna Rodríguez, M., Ramírez Hassan, A., & Coad, A. (2019). Uncovering value drivers of high performance soccer players. *Journal of Sports Economics*, *20*(6), pp. 819–849. DOI: 10.1177/1527002518808344

Singh, P., & Lamba, P. S. (2019). Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. *Journal of Discrete Mathematical Sciences and Cryptography*, *22*(2), pp. 113-126.

Smith, A. C. T., & Stewart, B. (2010). The special features of sport: A critical revisit. *Sport Management Review*, 13, pp. 1-13.

Telegraph Sport (2017). Lionel Messi signs new Barcelona contract to 2021 with a new buyout clause set at £626million. In *The Telegraph*. Retrieved 11 December 2019, from https://www.telegraph.co.uk/football/2017/11/25/lionel-messi-signs-new-barcelona-contract-2021-new-buyout-clause/

Thompson, S. K. (2012). Simple random sampling (3rd ed.). *Sampling Wiley Series in Probability and Statistics*.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), pp. 385-395.

Transfermarkt (2019). Historic transfer fee records. In *Transfermarkt*. Retrieved 11 December 2019, from https://www.transfermarkt.de/statistik/transferrekordehistorie

Tukey, J. W. (1978). *Exploratory data analysis*. Addison-Wesley.

Vöpel, H. (2011). Do we really need financial fair play in european club football? An economic analysis. *CESifo DICE Report*, *9*(3), pp. 54-59. Retrieved from https://www.econstor.eu/bitstream/10419/167048/1/ifo-dice-report-v09-y2011-i3-p54-59.pdf

Wickham, H. (2019). rvest: Easily harvest (scrape) web pages (v0.3.4). Retrieved 02 November 2019, from https://rvest.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2019a). dplyr: A grammar of data manipulation [v0.8.3]. Retrieved 02 November 2019 from https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H. (2019b). R Package 'ggplot2': Create elegant data visualisations using the grammar of graphics, CRAN v3.2.1. Retrieved 18 December 2019, from https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

Wicklin, R. (2017). Should you use principal component regression? - The do loop. Retrieved 10 December 2019, from https://blogs.sas.com/content/iml/2017/10/25/principal-component-regression-draw-backs.html

Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 36*(2), pp. 181-191.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Canada: CENGAGE Learning.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), pp. 301-320.

Appendices

A.1 Literature Tables and Novel Variables

Tab. 8 Related literature on the determinants of players' market values, transfer fees and salaries

Authors	Title	Dependent variable	Method
Dobson, Gerrard & Howe, 2000	The determination of transfer fees in English nonleague football	Transfer fee	General-to-specific modeling (backward elimination)
Garcia-del-Barrio & Pujol, 2007	Hidden Monopsony Rents in Winner-take-all Markets: Sport and Economic Contribution of Spanish Soccer Players	Market value	Regression with white-corrected standard errors
Medcalfe, 2008	English league transfer prices: Is there a racial dimension: A re-examination with new data	Transfer fee	Regression (OLS)
Herm, Callsen-Bracker & Kreis, 2014	When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community	Market value	Regression (OLS) with heteroskedasticity-consistent standard errors, 95% quantile regression
Drut & Duhautois, 2017	Assortative Matching Using Soccer Data: Evidence of Mobility Bias	Salary	Pooled Regression (OLS), Fixed effect Regression (OLS)
Müller, Simons & Weinmann, 2017	Beyond crowd judgments: Data-driven estimation of market value in association football	Market value	Multi-level regression analysis
Ante, 2019	Determinants of Transfers Fees: Evidence from the Five Major European Football Leagues	Transfer fee	Stepwise regression selection (Backward)
Serna Rodríguez, Ramírez Hassan & Coad, 2019	Uncovering Value Drivers of High Performance Soccer Players	Market value	Bayesian model averaging, Markov chain Monte Carlo model composition, instrumental variable Bayesian model averaging
Singh & Lamba, 2019	Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players	Market value	Decision Tree, Random Forest, Gradient Boost, Linear Regression, Ridge Regression

The table presents a brief overview of the most important related literature. For each of the nine studies it states the authors, the title, the main dependent variable examined in the paper as well as the econometric models applied. Due to the conceptual similarities of quantifying the monetary value attached to a football player, this paper includes not only studies concerned with estimating transfer fees but also market values and player salaries.

Tab. 9 New variables utilised in this research

Variables	Description
Number of previous transfers	Number of remunerated transfers prior to the current transfer
Volume of previous transfers	Cumulated monetary volume of remunerated transfers prior to the current transfer
Titles won	Number of titles won
National player to home country	The transferred player is a national player and transfers to his home country
Average age team 1	Average age of the team from which the player transfers
Average MV team 1	Average market value of the team from which the player transfers
Average age team 2	Average age of the team to which the player transfers
Average MV team 2	Average market value of the team to which the player transfers

The table gives an overview of the variables that have not yet been investigated by the related literature before. It states the variable names and provides a brief description for these variables.

A.2 Summary of the Overall Dataset as Well as Modelling Variables

Variable	Description	Example	Data source	Used in 3.3
Technical variables and identifier				
Transfer ID	Unique identifier for each player transfer (sequential number)	2033	4	
Transfermarkt ID	Unique identifier for each player assigned and used by *transfermarkt.de*	342229	1	
FIFA ID	Unique identifier for each player assigned and used by *sofifa.com*	231747	2	
Transfermarkt href	Part of the URL that references the profile of a player on *transfermarkt.de*	/kylian-mbappe/ profil/spieler/ 342229	1	
Wikipedia href	Part of the URL that references the page of a player on *en.wikipedia.org*	/Kylian_Mbappé	4	
Transfer-specific variables				
Transferred player	Name of the player involved in the transfer	Kylian Mbappé	1	
Transfer fee	Player transfer fee related to the transfer in million €	135	1	✗
Market value	Player market value related to the transfer in million € (as estimated by *transfermarkt.de*)	120	1	
Season	Season related to the transfer	2018/2019	1	
Previous season	Season previous to the transfer	2017/2018	4	
Year	Year of the player transfer ranging between 2013 and 2019 as proxy for potential year fixed effects	2018	1	✗
Player-specific variables				
Birthplace	Birthplace of the player involved in the transfer	Paris	1	
Birthdate	Birthdate of the player involved in the transfer	1998-12-20	1	
Age	Age of the player involved in the transfer in years	19	1	✗
Nationality	Nationality of the player involved in the transfer	France	1	
Height	Height of the player involved in the transfer in meter	1.78	1	✗
Position	Position of the player involved in the transfer (one of: goalkeeper, left defense, central defense, right defense, defensive midfield, left midfield, central midfield, right midfield, offensive midfield, left attack, secondary attack, right attack, central attack)	Center forward	1	✗

Preferred foot	Preferred foot of the player involved in the transfer (i.e. left- or right-footed)	Right	2	✗
FIFA score	FIFA total score of the player involved in the transfer with $x \in [1; 100]$	87	2	✗
FIFA potential	FIFA potential score of the player involved in the transfer with $x \in [1; 100]$	95	2	
FIFA unused potential	FIFA unused potential of the player involved in the transfer as difference between the FIFA potential (upper bound) and FIFA score	8	4	✗
Play time	Total amount of minutes played across all tournaments and leagues (national, international) in the season prior to the transfer	3554	1	✗
A-team debut	Date of the debut for the national A-team (i.e. not on the junior level)	25/03/2019	1	
Number of previous transfers	Number of remunerated transfers prior to the current transfer	1	1	✗
Volume of previous transfers	Cumulated volume of remunerated transfers prior to the current transfer in million €	45	1	✗
Titles won	Number of titles won (individual titles, team titles, national team titles, youth titles, etc.)	8	1	✗
National player	Dummy variable that takes on 1 if the player plays for his home country's national A-team (0 otherwise)	1	1	✗
To home country	Dummy variable that takes on 1 if the player transfers to his home country (0 otherwise)	1	1	✗
National player to home country	Dummy variable that takes on 1 if a national player transfers to his home country (0 otherwise)	1	4	✗
Average Wikipedia page views	Daily average Wikipedia page views across the English, German, French, Spanish and Italian version of Wikipedia in the season previous to the transfer	8,069.836	3	
Total Wikipedia page views	Total average Wikipedia page views across the English, German, French, Spanish and Italian version of Wikipedia in the season previous to the transfer	14,727,451	3	✗
Team-specific variables				
Team 1	Name of the team from which the player transfers	AS Monaco	1	
League team 1	Name of the league from which the player transfers as proxy for potential league fixed effects (one of:	Ligue 1	1	✗

	Premier League, Primera División, Serie A, 1. Bundesliga, Ligue 1, Liga NOS, Premier Liga, Eredivisie, Jupiler Pro League, Süper Lig)			
Country team 1	Name of the country from which the player transfers	France	1	
Rank previous season team 1	Rank in previous season of the team from which the player transfers	2	1	✗
Average age team 1	Average age of the team from which the player transfers in years	22.4	1	✗
Total market value team 1	Total market value of the team from which the player transfers in million €	385.58	1	
Average market value team 1	Average market value of the team from which the player transfers in million €	6.76	1	✗
Team 2	Name of the team to which the player transfers	FC Paris Saint-Germain	1	
League team 2	Name of the league to which the player transfers as proxy for potential league fixed effects (one of: Premier League, Primera División, Serie A, 1. Bundesliga, Ligue 1, Liga NOS, Premier Liga, Eredivisie, Jupiler Pro League, Süper Lig)	Ligue 1	1	✗
Country team 2	Name of the country to which the player transfers	France	1	
Rank previous season team 2	Rank in previous season of the team to which the player transfers	1	1	✗
Average age team 2	Average age of the team to which the player transfers in years	23.9	1	✗
Total market value team 2	Total market value of the team to which the player transfers in million €	842.4	1	
Average market value team 2	Average market value of the team to which the player transfers in million €	21.6	1	✗

The table presents the variables included in the overall dataset. The first column contains the variable name followed by a description of the variable in column 2. If not of permanent nature, each variable value refers to the point in time of the player transfer. Column 3, "Example", contains a real-world example for a player-transfer as included in the overall dataset (Transfer ID 2033). "Data Source" contains an indicator for the origin of the data. Thereby, indicators are mapped to the employed data sources as follows: 1 → *transfermarkt.de*, 2 → *kaggle.com/sofifa.com*, 3 → *wikipediatrend* R-package API, 4 → derived feature via feature engineering.

A.3 Descriptive Statistics

Tab. 10 Descriptive statistics of the numerical variables

Variable	$\bar{x}$	sd	min	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	max
Transfer fee [in million €]	9.11	13.98	0.00	1.60	4.00	11.00	222.00
Age	24.83	3.45	16.00	22.00	25.00	27.00	36.00
Height [in meter]	1.82	0.06	1.63	1.78	1.83	1.87	2.04
FIFA score	75.04	5.03	55.00	72.00	75.00	78.00	94.00
FIFA unused potential	4.26	3.82	0.00	1.00	4.00	6.00	22.00
Play time [in thousand min]	2.32	1.00	0.00	1.57	2.41	3.09	5.22
Number of previous transfers	1.39	1.43	0.00	0.00	1.00	2.00	8.00
Volume of previous transfers [in million €]	6.17	11.59	0.00	0.00	1.50	7.10	138.56
Titles won	2.69	4.05	0.00	0.00	1.00	3.00	75.00
Total Wikipedia page views [in thousand]	267.68	861.89	0.00	31.29	91.00	253.83	33,436.82
Rank previous season team 1	7.37	5.08	1.00	3.00	6.00	11.00	20.00
Average age team 1 [in years]	24.24	1.33	21.10	23.40	24.10	25.00	29.90
Average market value team 1 [in million €]	5.79	7.10	0.23	1.69	3.20	7.10	53.52
Rank previous season team 2	7.98	5.58	1.00	3.00	7.00	12.00	20.00
Average age team 2 [in years]	24.46	1.30	21.10	23.60	24.40	25.30	29.90
Average market value team 2 [in million €]	5.64	6.32	0.32	1.97	3.43	6.76	53.52

The table presents summary statistics for the actual numeric variables used in the modelling part of this paper. Presented is the mean ($\bar{x}$), standard deviation (sd), minimum (min), 25%-quartile ($Q_{0.25}$), median ($Q_{0.5}$), 75%-quartile ($Q_{0.75}$) and maximum (max) for each variable.

Fig. 14 **Relationship between player age and player transfer fee**

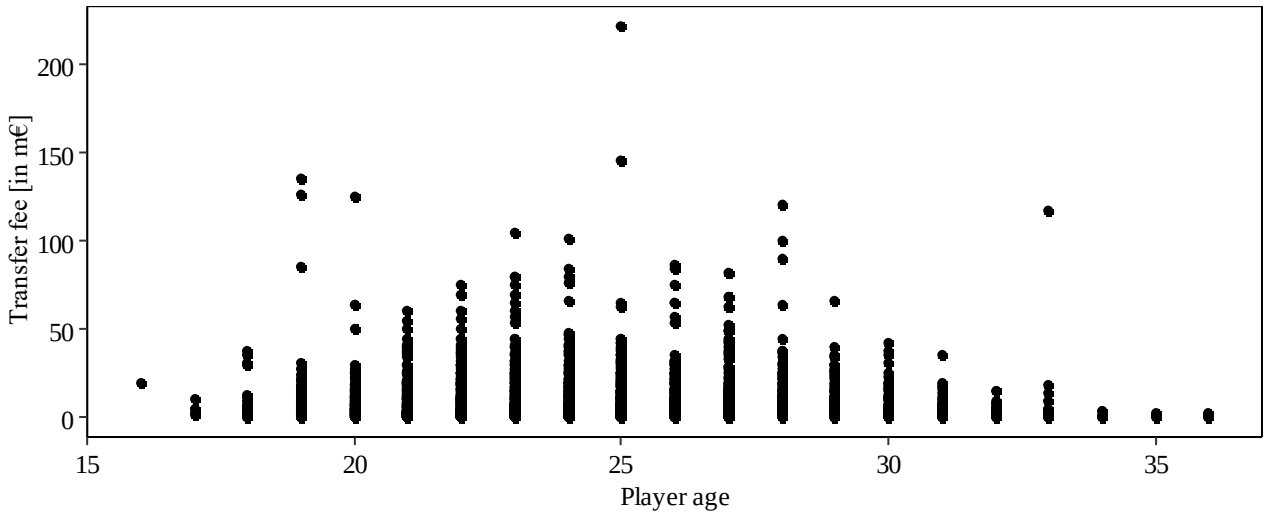


Fig. 15 **Plot matrix for numerical predictors**

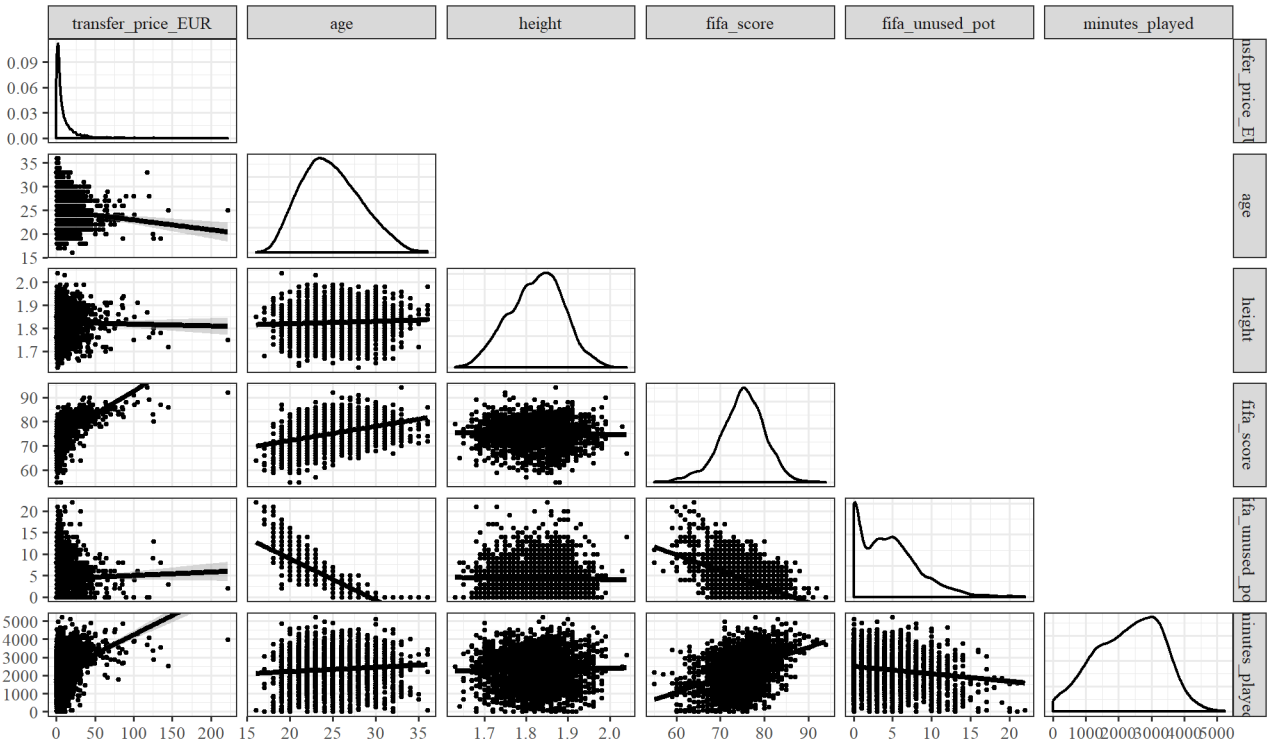
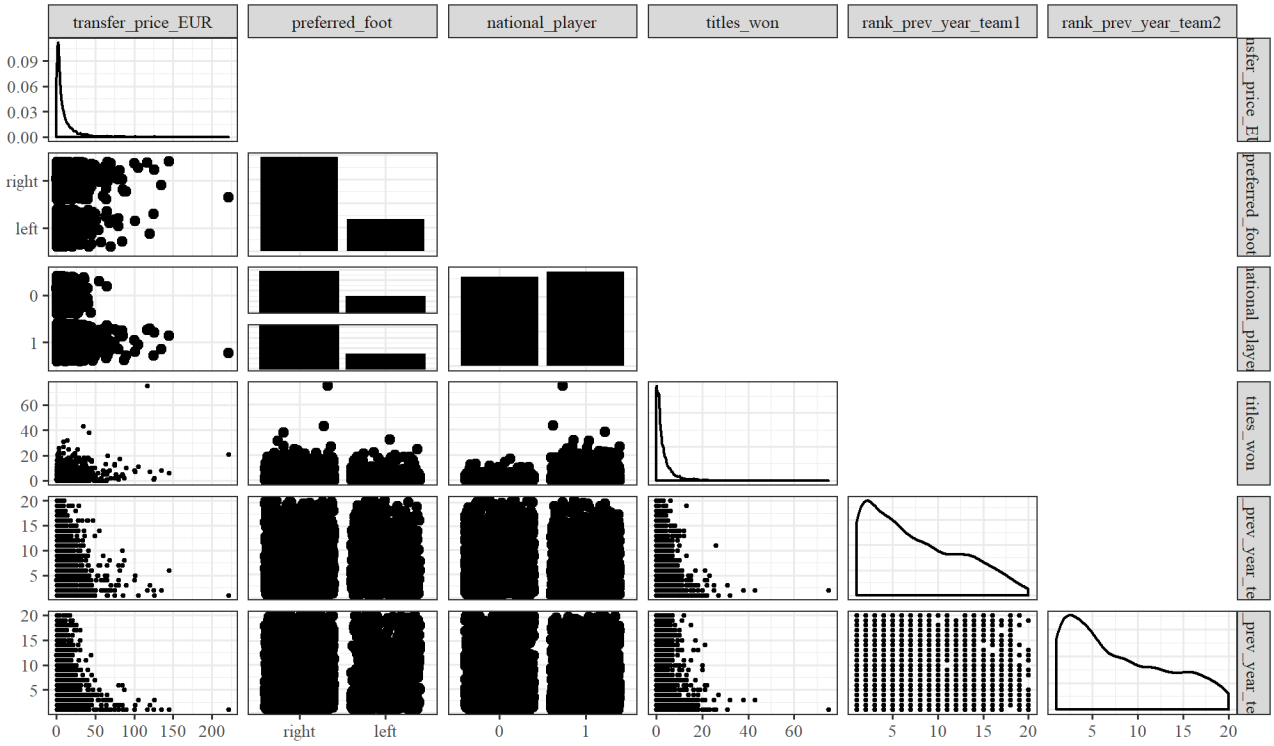


Fig. 16 **Plot matrix for categorical predictors**



Tab. 11 Correlation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1) Transfer fee	-	***		***		***	**	***	***	***	***	***	***	***	***	***
(2) Age	-0.08	-	***	***	***	***	***	***	***	***		***		***	***	***
(3) Height	-0.02	0.06	-				***						*			
(4) FIFA score	0.54	0.40	-0.03	-	***	***	***	***	***	***	***	***	***	***	***	***
(5) FIFA unused potential	0.03	-0.87	-0.02	-0.49	-	***	***	***	***	***		*		*	***	***
(6) Play time	0.30	0.08	0.03	0.42	-0.16	-		***	***	***		***	***	***	***	***
(7) Number of previous transfers	0.05	0.49	0.05	0.33	-0.43	-0.01	-	***	***	***	***	***	***		***	
(8) Volume of previous transfers	0.38	0.33	-0.01	0.46	-0.30	0.07	0.56	-	***	***	***	***	***	***	***	***
(9) Titles won	0.29	0.32	-0.03	0.42	-0.27	0.07	0.29	0.55	-	***	***		***	***		***
(10) Total Wikipedia page views	0.58	0.06	-0.01	0.36	-0.06	0.14	0.10	0.42	0.49	-	***		***	***		***
(11) Rank previous season team 1	-0.21	0.02	0.02	-0.21	-0.01	0.01	-0.09	-0.22	-0.31	-0.15	-	***	***	***		***
(12) Average age team 1	0.06	0.10	0.01	0.06	-0.03	-0.07	0.08	0.08	0.03	0.02	0.17	-	***	**	***	***
(13) Average market value team 1	0.30	0.03	-0.03	0.24	0.03	-0.12	0.12	0.35	0.32	0.23	-0.48	0.29	-		***	***
(14) Rank previous season team 2	-0.30	0.07	0.01	-0.31	-0.03	-0.23	0.02	-0.15	-0.15	-0.16	0.06	0.04	-0.03	-	***	***
(15) Average age team 2	0.06	0.14	-0.02	0.07	-0.09	-0.06	0.10	0.10	0.03	0.02	-0.02	0.38	0.25	0.25	-	***
(16) Average market value team 2	0.63	-0.07	-0.01	0.45	0.07	0.24	0.02	0.25	0.20	0.31	-0.14	0.21	0.36	-0.45	0.20	-

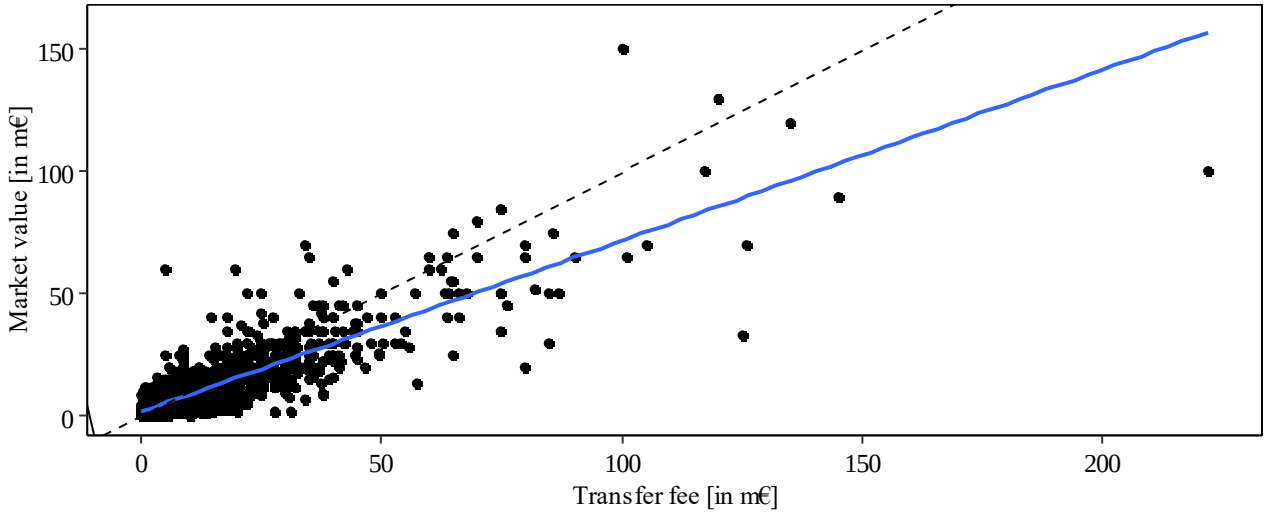
The table presents the pairwise Pearson correlation coefficients for the numerical variables included in the final modelling dataset (below the diagonal) as well as the corresponding p-values (above the diagonal). Statistical significance is indicated by the asterisks as follows: *** (1%-level), ** (5%-level), *(10%-level).

Tab. 12 Correlations between player height and position

	Position					
	Goalkeeper	Left defense	Central defense	Right defense	Defensive midfield	Left midfield
Height	0,2788	-0,1335	0,3763	-0,0924	0,0180	-0,0677
	Central midfield	Right midfield	Secondary attack	Left attack	Central attack	Right attack
Height	-0,1176	-0,0396	-0,0346	-0,2430	0,1704	-0,2190

The table presents the Pearson correlation coefficients for the height variable and the respective position dummies.

Fig. 17 Relationship between player transfer fee and market value



A.4 Graphical Illustration of the Data Transformation Steps

Fig. 18 Illustration of the effect of a log-transformation

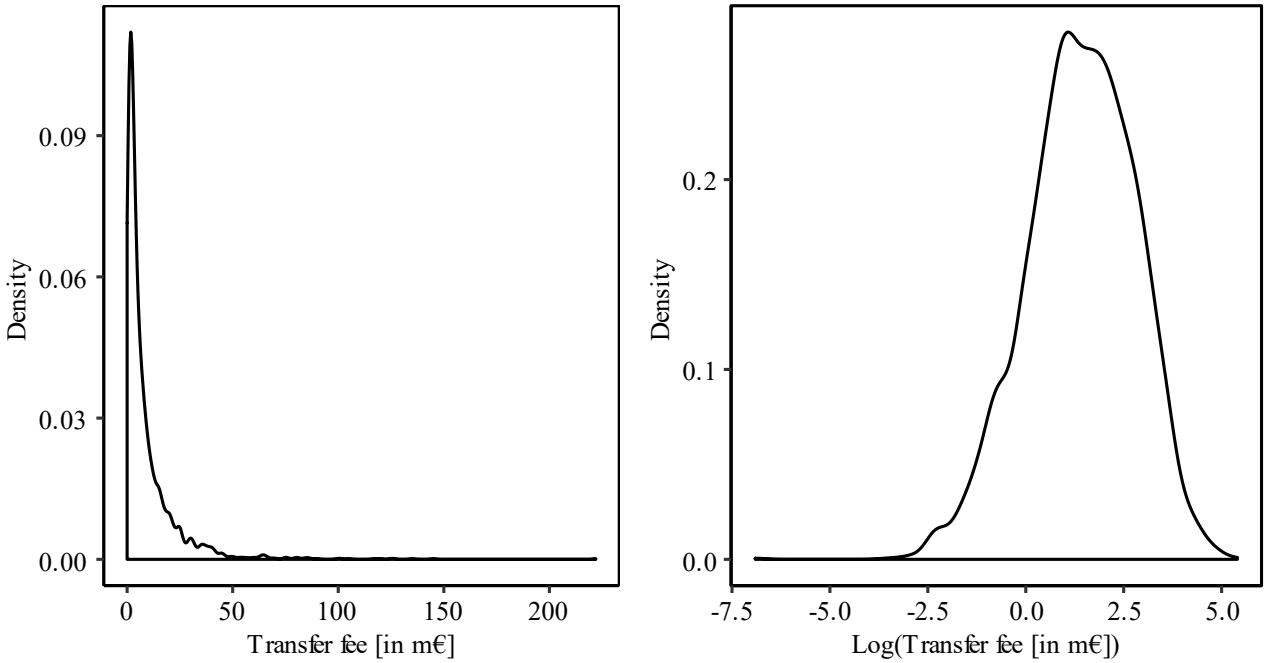


Fig. 19 Illustration of the effect of a standardisation