Bachelor's Studies

Quantitative Methods in Economics & Information Systems

Rafał Stępień
Student No. **82640**

Modelling football players values on the transfer market and their determinants using robust regression models

Bachelor's thesis
under scientific supervision of
dr hab. Michał Rubaszek, prof. SGH
written in Institute of Econometrics

Warsaw 2021

Table of Contents:

Abstract

This thesis aims to evaluate the key determinants of football players values on a transfer market and to describe an attempt to create models which can effectively predict a footballer's valuation using robust regression. Four separate log-linear iteratively reweighted least squares models for players on each position on the pitch are being constructed, in which the price is explained by players on-pitch statistics, physical traits and qualitative variables such as league the player plays in or participation in international club championships. Basing on models, conclusions regarding key determinants of footballer's valuation are derived and actual dependencies on a transfer market are explained. Constructed models will be further validated, checking if given specification can provide legitimate estimations.

Introduction

During the last decade we have observed the data revolution, also drastically changing the football[1] business, which is constantly growing, even despite COVID-19 pandemic. The market size in Europe has reached 28.9 billion euros in 2018/19, having risen by 32.57% in five years[2]. An especially interesting, yet very essential part of the industry is the transfer market. The prices for the best players can reach astronomical values exceeding 100 million euros, with the record beaten in 2017 by Brazilian forward Neymar Jr.'s transfer from FC Barcelona to Paris Saint-Germain with 222 million euros transfer fee[3]. The market is characterized by huge risks, especially given the nature of players professional relationships with their clubs. The contracts are usually signed for longer periods, which can reach even up to nine years, taking Atletico Madrid's Saúl Ñíguez case as an example.[4] They can also be extremely lucrative, with Lionel Messi's 4-year deal with FC Barcelona being a record in all sports disciplines combined[5]. This shows how a good use of big data and quantitative methods might prove useful in financial management of football clubs. This is actually happening, taking Argentinian midfielder Giovani Lo Celso's transfer from Paris Saint-Germain to Real Betis Balompié as an example[6], who was signed as a replacement of one of prominent players for 25 million euros and was sold a year later for 48 million euros to Tottenham Hotspur[7].

A professional football transfer market ecosystem is relatively sophisticated. The relations between players and their clubs are specified within fixed term contracts. A contract obliges the club to pay a certain amount of money to the player throughout its duration, with paychecks taking place once per month in most cases. Those contracts being signed have some clauses, bonuses and other special conditions assigned, also considering regulations imposed

[1] in this paper, the word "football" matches the definition in British English vocabulary which I am going to use throughout the paper - the discipline is called "soccer" in American English

[2] https://www.statista.com/statistics/261223/european-soccer-market-total-revenue/

[3] https://bleacherreport.com/articles/2725494-barcelona-confirm-psg-activated-neymars-eur222m-transfer-release-clause

[4] https://www.espn.com/soccer/soccer-transfers/story/3151518/saul-niguez-commits-to-long-term-atletico-madrid-contract

[5] https://www.nbclosangeles.com/news/sports/lionel-messis-contract-reportedly-worth-673-million-most-expensive-for-athlete-in-any-sport/2514971/

[6] https://eldesmarque.com/sevilla/real-betis/1159788-asi-ficho-el-big-data-a-lo-celso

[7] Lo Celso's deals were actually more sophisticated, more details: https://www.transfermarkt.de/giovani-lo-celso/profil/spieler/348795

not only by football associations, but also international and national law. As the contract terminates, the player becomes a free agent and can sign another deal with other clubs. However, until contract termination the player cannot change the team without his current club permission. Most transfers are made basing on negotiations between the clubs. If both player and his[8] current club come to an agreement with the second party, then the transfer can be realized. In most cases, the purchasing club pays a transfer fee, the value of which can reach tremendously high prices for the best and most prospective players, with ten transfers beating the gap of 100 million euros[9] in a history. There are, of course, exceptions from these rules, but they happen very rarely, and the transactions are based on more sophisticated, special conditions as well, such as payments in installments, specific clauses or players exchange.

The process drives the key question for the clubs, which I explore throughout the thesis: what is the fair value of football players, given their characteristics? In fact, that is impossible to answer this question with very high precision, especially provided that the transfer fee often depends on conditions that are not necessarily related to players characteristics, but variables such as remaining contract length or selling club's financial stability. Nevertheless, a good model can help to provide players' orientational values, based on payers skills. I can be added, that skills are not directly observed, but are proxied by players' on-pitch performance and physical characteristics. The intuition suggests that there are potential variables that can actually reflect skills, and so the players' value as well - and this is one of the key aspects of the thesis.

This is essential to mention about the importance of other qualitative players characteristics for their potential valuations as well. For a football fan, it is common knowledge that the prices vary depending on player's position or age, what will be discussed later. Each on-pitch position also puts a bigger pressure on differing skills - it is to be expected that a good defender should have more team clean sheets or successful tackles, whereas a good forward should have more goals or successful dribbles. And, at the end of the day, one should expect that Real Madrid player should be worth more on average than, for example, Luxembourgian F91 Dudelange footballer. That means that there are some more factors such as league the player plays in, which also leverages the relevance of statistics in particular competitions. It would be most likely harder for Christian Gytkjaer to score 24 goals in English Premier League rather than in Polish LOTTO Ekstraklasa, as he did in 2019/2020, becoming the top goalscorer in Polish league, playing for Lech Poznań side.

The first part of a thesis focuses on driving insights from the dataset that includes players from top five European leagues, whose characteristics and their relation to the players' valuation is analyzed. The distribution and importance of age, player's position, and more variables are further described. Dataset problematics and their impact in further analysis is discussed as well.

The second part focuses solely on modelling process, providing explanation regarding the used method and players' value determinants. Models are created depending on the players' position, with each containing different variables. Significant determinants' distribution is further analyzed. Each model is properly diagnosed, checking for models fit, specification error

[8] The paper is focusing on men's football but those rules exist in women's football as well.

[9] https://www.transfermarkt.de/statistik/transferrekorde

multicollinearity, heteroskedasticity and structural breaks. Potential omitted variables shall be discussed.

The third part consists of models' validation – both in-sample and out-of-sample validations are conducted, basing on root mean square logarithmic error and mean square logarithmic errors. Out-of-sample validation relies on real-life transfers data from 2020/21 seasons, comparing models' performance to the benchmark, which is transfermarkt.de website. Two approaches are tested and compared – one basing only on models for 2019/20 season of given player, and the second one basing on weighted means of the players' predicted value throughout 2017/18-2019/20 seasons.

In the end, a thesis contains a summary of an essay. Relevant value's determinants prove to be age, player's team performance and player's on-pitch performance – with significant aspects of the game differing depending on a position. Created models contain certain prediction error and perform worse than a benchmark out of a sample, but prove to have an utility. Interesting insights from first part of an essay are discussed as well – such as the average player's age and height depending on a position, distribution of players' nationalities or players' mean value basing on position or league they participate in.

Chapter I – Dataset overview

Dataset description

The variable I am going to explain is footballer's value, and exogenous variables are going to be footballers' statistics and characteristics. I have gathered data from transfermarkt.de, the most prominent website related to football transfers, and fbref.com, the website which provides a lot of football statistics which are available for a wide public. First of all, I have created a file containing values of selected footballers from transfermarkt.de, which are calculated by their internal algorithms, whose complete methodology remains unknown to the public. The only thing that is known is that registered users have some impact on the final value, as they are voting on it. Despite low transparency transfermarkt.de reputation is so renowned, that their calculated values are taken as a benchmark for the whole football world, including clubs, experts and journalists. For the sake of further modelling I have asserted that their valuations are significantly correct and resemble an actual footballer value. Secondly, I have gathered statistical data from fbref.com and merged it with the values from transfermarkt.de. In this part I have found a problem that some players are not named the same way at both of these sites, thus certain rows could not be merged. The main reasons were the absence of special linguistic characters in some names, different ways of writing the given footballer's full name and/or using pseudonyms. For example, the goalkeeper Wojciech Szczęsny is called "Wojciech Szczesny" on transfermarkt.de, without the "ę" mark, which appears in fbref.com records. For the 2019/20 season, which I plan to make as a representative one for the further presentation, and for chosen records in previous seasons I have edited the data on my own. Other observations which failed to match have been deleted from the dataset. In the end, I have added several important variables that were not available in data from fbref.com and that appear on transfermarkt.de, such as height, and matched them with existing records, and created some binary variables which will be used in a later part.

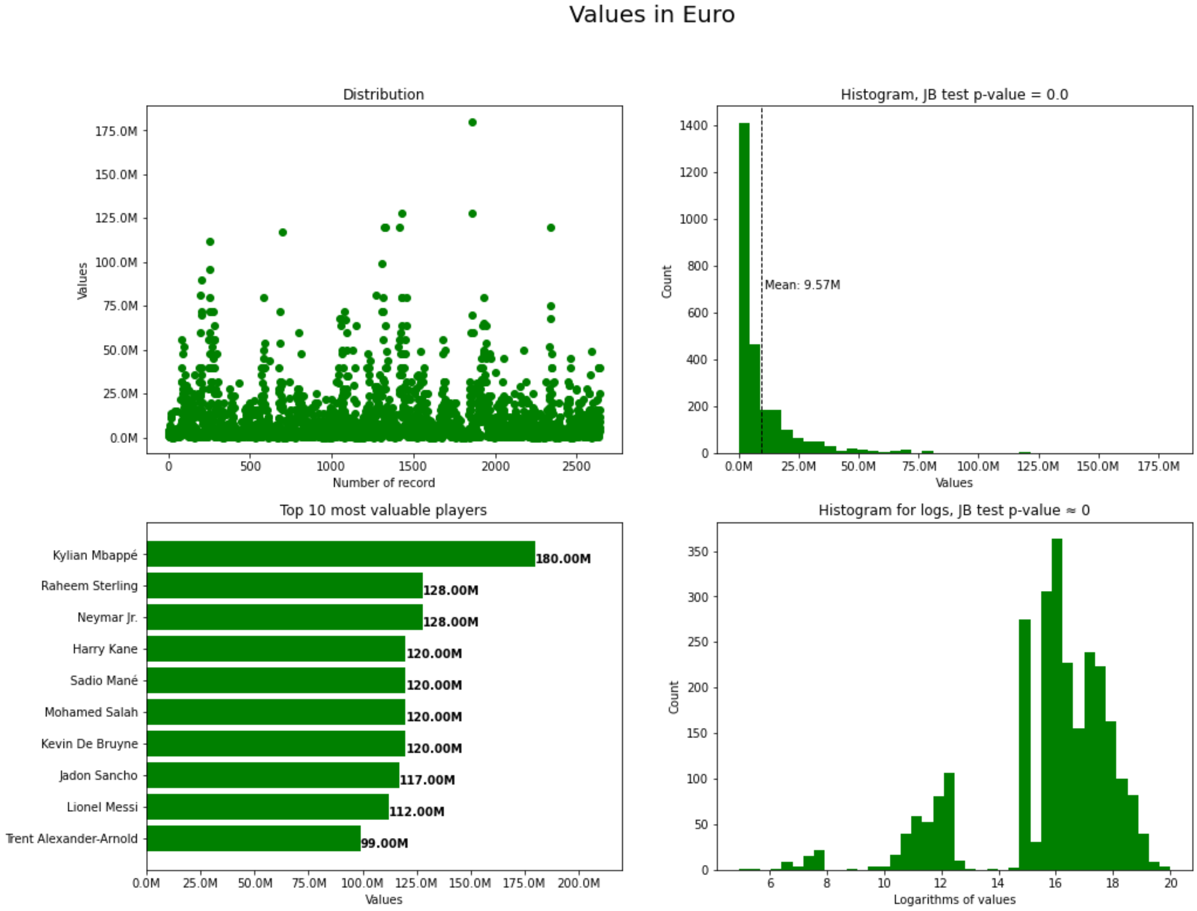
The dataset ranges from 2017/18 up to 2019/20 seasons. I selected this range as certain variables from fbref.com are not available for earlier seasons, and the 2020/21 season has not finished yet (during the process of writing this thesis). I have also decided to gather the data only from top 5 European leagues - English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French Ligue 1. I have decided for these particular leagues because there is a substantial quality gap between them and other European leagues, which means that statistical data taken from them would not have that much relevance. As mentioned in introduction, this should be on average much harder to score in a Premier League rather than in LOTTO Ekstraklasa. The other reason is that transfermarkt.de value estimates tend to be, in my opinion, not accurate enough within weaker leagues as in the five mentioned earlier, which could create the risk of biased estimations. All data gathering, wrangling, visualization and modelling was implemented in Python Anaconda and Microsoft Excel environments.

Dataset analysis

Footballers' values which I plan to explain using econometric model are gathered from transfermarkt.de website. The values there are orientational, without exact numbers, but rather a certain value that resembles some kind of an aggregate, which distinguishes more expensive from cheaper players. For example, for the 2019/20 season Raheem Sterling and Neymar Jr. have the same exact value - 128 million euros, with Kylian Mbappé topping the list alone with an insane 180 million euros. However, they are placed higher than Sadio Mané, Mohamed Salah, Kevin De Bruyne and Harry Kane, who are valued 110 million euros, without any differences between them.

There are 2644 records for the 2019/20 season and 4464 for 2017/18 and 2018/19 seasons - 2232 for each. For further presentation I am going to use the data for 2019/20, as it is the most current and complete part of the dataset. For 2019/20, an average value for players in top 5 leagues at transfermarkt.de is approximately 9.57 million euros, with standard deviation of approximately 14.09 million euros. The exact distribution of values is shown on Figure 1.1.

Figure 1.1 A dashboard presenting the distribution of transfermarkt.de values (as for the end of the 2019/20 season)



There seems to be a huge concentration of players priced on a relatively low level, which is to be expected. There is probably a certain number of outliers as well - especially of relatively unknown players, who played very little games. As we can see from the plots, there is a problem that the values are not distributed normally not only for absolute values, but for logarithms as well. The Jarque-Bera test confirms that with its p values approaching zeros. This may cause a problem in further modelling, as testing for significant variables and ordinary least squares assumptions might be misleading. However, the methods of handling these issues will be discussed in the next part – it can also be the issue related to the non-typical observations. At this moment I am going to focus on analyzing other dataset properties.

Figure 1.2 Table consisting of 10 most valuable players by transfermarkt.de as for the end of 2019/20 season and some of their key characteristics.

Player	Age	Position	League
Trent Alexander-Arnold	20	Defender - Right-Back	Premier League
Lionel Messi	32	Forward - Right Winger	La Liga
Jadon Sancho	19	Forward - Right Winger	Bundesliga
Kevin De Bruyne	28	Midfielder - Attacking Midfield	Premier League
Mohamed Salah	27	Forward - Right Winger	Premier League
Sadio Mané	27	Forward - Left Winger	Premier League
Harry Kane	26	Forward - Centre-Forward	Premier League
Neymar Jr.	27	Forward - Left Winger	Ligue 1
Raheem Sterling	24	Forward - Left Winger	Premier League
Kylian Mbappé	20	Forward - Left Winger	Ligue 1

I start by looking at ten most valuable players (Figure 1.2) and some of their characteristics. It is quite logical that the better on-pitch statistics the player has, the bigger his value should be. But there are several other factors that contribute to the overall value. For 10 most valuable players, there is one player whose age is above 30 and this is Lionel Messi, undoubtedly one of the best players in the history of football. There are three very young players (under 21), who are conditionally freed from registration to Champions League competitions. That would imply that age has a great impact on overall value and in fact it has, as it contributes to the discount factor measuring not only the actual player's skill, but also his potential. Going further, we can see that on-pitch position probably matters as well, as 8 out of 10 most valued players are forwards. In fact, 8 out of 10 most expensive transfers in history are related to forwards as well, basing on transfermarkt.de classification. The league possibly matters too, with 6 out of 10 players playing in the English Premier League. I am going to analyze this topic in the next sections, providing an in-depth description of chosen variables of players within the dataset, including age, position, league, height, nationality and preferred foot. After that, I will look for the potential determinants of players' values depending on their positions, checking for correlations.

Figure 1.3 A dashboard presenting chosen characteristics of the dataset and relations between them.

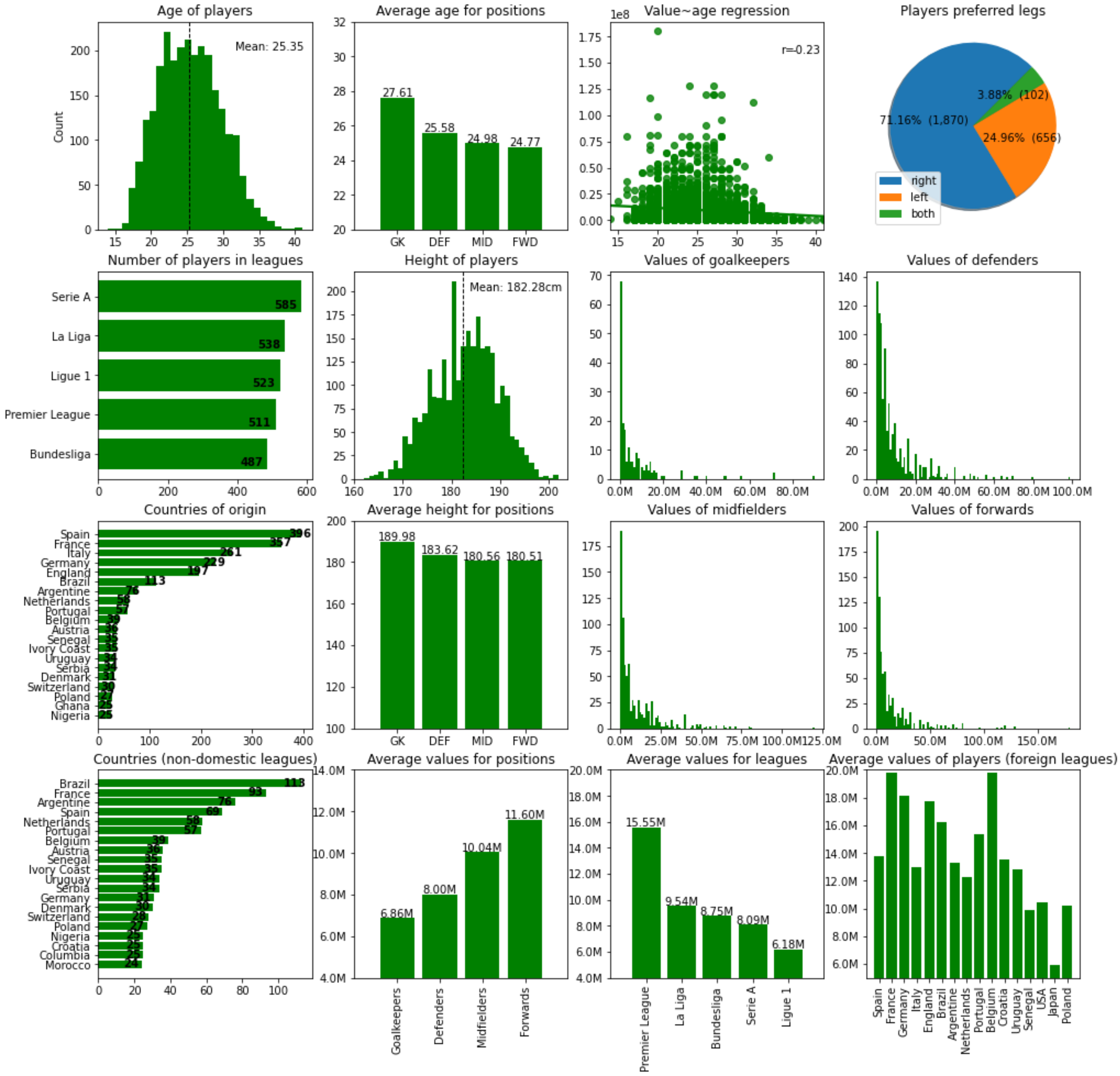
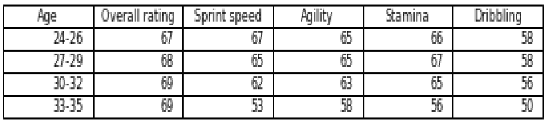


Figure 1.3 provides illustration of distribution for selected variables. Starting from age, we can see the mean age is 25.35 and that a vast majority of footballers are in their 20s (71.6% of them). The youngest footballer is 14 years old - an Argentinian midfielder Luka Romero from Spanish RCD Mallorca, while the two oldest are 41-year-old - Brazilian defender Vitorino Hilton playing in French Montpellier HSC and Italian legendary goalkeeper Gianluigi Buffon. The average age differs basing on position, with goalkeepers having undoubtedly the highest mean - 27.61. We can also see that there is a negative, relatively low correlation between value and age (r=0.23). However, the relationship might be non-linear, as for example the correlation coefficient for players above 26 years old equals to about 0.5. For that reason, for each model I will use squares of age variable as I expect potential nonlinearity within that variable. It is interesting that the Jarque-Bera test indicates that age is not normally distributed.

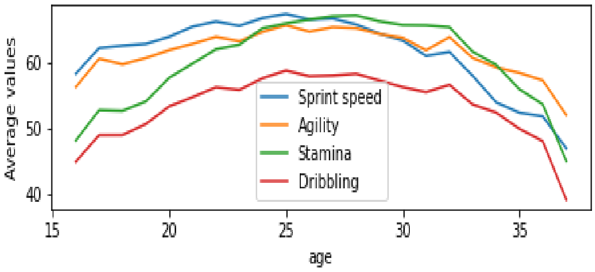
For somebody keen on football this should be a very logical outcome. First, the older the player is, the shorter is his expected remaining career length. Player value discounts his expected career length to a certain degree. That happens because age is a strong biological determinant of physical attributes, such as pace, durability or agility. For illustration, in

Figure1.4. I use the dataset with players statistics from FIFA 21, a popular football video game, and I analyze certain in-game attributes for certain age groups. Although the validity of attributes for individual players might be debatable, the aggregated values should resemble overall tendencies, as they are adjusted to the players performance.

Figure 1.4 A table and a graph presenting the relations between age and overall rating, sprint speed, agility, stamina and dribbling attributes in FIFA 21 video game.



Age	Overall rating	Sprint speed	Agility	Stamina	Dribbling
24-26	67	67	65	66	58
27-29	68	65	65	67	58
30-32	69	62	63	65	56
33-35	69	53	58	56	50



As we can see, despite overall players ratings (which are certain formulas of cumulated values of his attribute values) remaining almost the same across age groups, we can see a significant decrease of their physical abilities and dribbling skills, which are also determined by their physical form. That also explains the differences of age on positions – goalkeeper position is much more static and does not require the player to run such great distances on a pitch. That applies to the role of defender to some degree as well, whereas it is often required for modern midfielders and forwards, especially wingers, to be swift and durable.

Going back to players characteristics, we can see that the average footballer height equals to 182.28 cm. Interestingly, we also reject Jarque-Bera null hypothesis for this variable. That might be caused by height distributions for certain positions. We can see that when an average height for midfielders and forwards is almost equal to 180.50 cm, the defenders average height is 183.62 cm, and goalkeepers – 189.98 cm. For goalkeepers and central defenders tallness is a big advantage, if not requirement in professional football. Only 10.71% of center backs and 4.1% of goalkeepers are shorter than 182.28 cm. Height is crucial for them – the taller the player is, the bigger is his jumping range, which contributes to better goalkeeping range for goalkeepers and better aerial possibilities for on-pitch players. These disproportions may cause the distribution to be far from Gaussian.

A potentially interesting aspect is the role of preferred leg. This is a crucial trait especially for side backs and wingers (to some degree – the popularity of inside forward position is on a rise, where the inverted foot, opposed to the side of the pitch is in general preferred). Over 70% players are right footed and about 4% feel almost equally comfortable with both legs. Does that mean that left footed players (or both footed) are potentially worth more as they are in a minority? This is the matter I am going to look up to during modelling.

For the better understanding of football market's dynamics, I have decided to present an overview of nationalities of players as well. As we can see, most players are from nations that are home to the 5 analyzed leagues. This is what one can expect, especially as national regulations often require clubs to provide a certain number (most often 8) of homegrowns, who are most often native players. However, excluding the players playing in their domestic leagues, we have a completely different picture. Brazil is topping the list with 113 players, followed by France with 93 players and Argentine with 76 players. We will not see England on that plot, with only 11 players outside of domestic league. However, it does not necessarily imply that their players are, on average, weaker than French players. As we can see on average values for leagues plot, Premier League players are definitely most valuable compared to remaining four leagues, whereas French Ligue 1 players are the cheapest on average. One can deduct that English players are less willing to leave their native league compared to the French players. Nevertheless, if one would strongly desire to prove for player's ability depending on nationality, it would be the best to check average values of players in non-domestic leagues. However, I will not use nationalities in modelling although there are some theoretical reasons to justify that they might play a bigger role in overall player valuation – this is however the reason that, for example, certain countries have better sporting and youth facilities, what would explain potential dependencies, with prestige of domestic leagues playing its role as well.

The last insight from Figure 1.3 suggests that there is a relationship between values and players positions. This is especially important for my further modelling as I intend to construct distinct models for each position, as we can expect different variables playing important role for each. As we can see, the higher player's position on a pitch is, the bigger value we can expect, with forwards average value reaching 11.66 million euros, compared to goalkeepers 6.86 million. Values distribution for each position are relatively similar.

There are 408 columns gathered in a final dataset, including several variables created by myself, such as on-pitch statistics per minute or participation in Champions League. Dashboard explains the relationship only between few variables and neither of them is actual on-pitch statistics. There are 186 variables related only to players on-pitch performance and another 183 being results of quotients of these chosen variables by minutes played, and 16 related to their club's performance throughout the season. This is a huge number, too high for in-depth analysis for each variable to be contained in this paper. To check which variables could be significant for further modelling, I will calculate a correlation matrix between them, checking for their relations with footballer's value variable, and later describe potentially significant ones. However, before I leap to that topic, I need to discuss potential problems occurring within the dataset, including potential significant variables missing, and how to handle a problem of non-typical observations.

Dataset issues

In this point I discuss my worries about omitted variables, which might lead to endogeneity bias, and which is related to dataset's flaws. Despite a huge number of on-pitch statistics, there is a plenty of potentially significant variables missing, such as, for example, average total distance covered per match, which would measure player's work rate and in-game engagement. Some of the statistics are not available on free source websites such as fbref.com, although playing an important role in football analytics business. Other potentially important variable that comes to my mind is, for example, player's injury proneness, that could be estimated basing on his injury history from recent years. The dataset also consists only of national leagues statistics, excluding international competitions, which would probably have a strong impact on the estimations. It is also possible that appropriately weighted data from other leagues and earlier seasons could change the final estimation results. In the end, the footballer value itself is endogenous as well and there are other potential unmeasurable variables that impact the final calculations, with player's skills and talent being definitely the most important ones.

The second issue is that within the dataset we can find several outliers and other non-typical observations. There is, for example, a large number of players that play very little throughout the season or who lack data regarding certain variables such as age, height or preferred foot. Most of these players were actually deep backups and/or juniors, but several names could be known for people keen on certain national leagues. Some of the players values are also relatively low and could cause the final estimations to lose quality. To avoid potential negative impact of these variables, I have decided to include only these observations whose value is bigger than one million, the number of played games is bigger than 5 and who do not lack data regarding their age or height.

For model training, I am going to use an original dataset with data for all 3 seasons from 2017/18-2019/20, which consists only of observations meeting the requirements above. In a process, the number of observations shrunk from 7108 to 4932. I intend to create distinct models for 4 different positions – goalkeepers, defenders, midfielders and forwards, thus I split the dataset between them. The results are presented in Figure 1.5.

Figure 1.5 A table presenting the number of observations used for modelling for each positions.

Position	Number of observations
Goalkeepers	274
Defenders	1674
Midfielders	1490
Forwards	1493

The differences between numbers of observations for these positions are to be expected. Teams place 11 players on a pitch, with only one goalkeeper. Most football formations use 3-5 defenders, 2-4 midfielders and 2-4 forwards during the game. Each team can land only a limited number of substitutes for each game and make limited number of substitutions, with its limit being 3 throughout 2017/18-2019/20 (these rules are changed in 2020/21 season). In conclusion, the relatively small number of goalkeepers is to be expected and the dominance of defenders within the dataset should not be surprising.

Chapter II - Creation of models & evaluation of significant variables

Overview of methodology

Throughout the next parts of this bachelor thesis I will put emphasis on constructing regression models and drive insights regarding the impact of chosen variables on the footballers valuation, and also evaluate constructed models in terms of their validity regarding transfermarkt.de values and real-life transfer fees. There are going to be four different models for each position – goalkeepers, defenders, midfielders and forwards, as I expect different optimal specification for each.

The variables described in a previous part are only an aspect of a big picture. The major impact on the final estimations should be caused by players' on-pitch statistics. Due to the excessive amount of these they will not be analyzed in such depth as previous variables, although the influence of selected ones on the players valuation will be properly examined in the next part. In appendix one can find a table with all variables within the dataset and their correlation with players' valuation.

The crucial problems that I expect to occur – and what is confirmed later - are heteroskedasticity and multicollinearity. The first issue is a common problem for cross-sectional data, whereas the second one is strongly related to the nature of the dataset. For example, it is expected that the larger amount of assists the player has, the bigger his goal creation ratio should be, and the more passes he commits, the bigger numbers of all short, medium and long passes are. The problems of dataset might be even deeper, as there might be potential nonlinear relation, what will be suggested by RESET test results, where RESET test's null hypothesis is expected to be rejected for all models. However, the matter of fighting nonlinearity and applying specific quantitative methods will not be examined throughout the paper. To tackle these problems, I have decided to apply iteratively reweighted robust regression and turn the endogenous variable values into logarithms, creating log-linear models instead of linear. I will apply stepwise approach, inserting the variables that are both statistically significant (at significance level of 0.1) and do not enter into too strong relationship with each other.

In the later part the models will be properly diagnosed, measuring their adjusted R-squared metric, variance inflation factors and applying RESET, Breusch-Pagan and Chow tests, including their actual validation as well, running 5-fold cross validation calculating root mean square error and mean absolute error compared to transfermarkt.de valuations, but also to selected 25 real-life transfers, basing on 2019/20 statistics. All calculations have been implemented using Python StatsModels module, with Scikit-learn package to run a cross validation. Complete prints for all models, including linear and log-linear ones, find themselves in an appendix.

Robust regression

In classical ordinary least squares method there are five assumptions lying foundations for Gauss-Markov theorem, saying that OLS estimator is the best linear unbiased estimator, with one of these relating to the sphericality of errors – that means they have the same standard errors. In practice, it equals no heteroskedasticity or autocorrelation of models' residuals. When it occurs:

- ordinary least squares method is not biased, but there is a more efficient linear estimator with lower variance,
- the original formulas for ordinary least squares estimator's variance and standard errors stop being correct,
- significance tests do not have t-Student distribution and overall significance does not come from F-Snedecor or Chi-squared distribution,

With the last problem being crucial for answering the key question of the thesis – "what determines the valuation of a player on a transfer market?" – as using standard linear regression method models could have failed to distinguish actual significant variables that contribute to players' value.

The problem of lack of sphericality of errors is expected within my dataset. Although the problem of autocorrelation for cross-sectional data is in most cases non-existent, it is expected that heteroskedasticity should occur, providing that the distribution of values is not normal and that they consist of some strong outliers. This is the reason why following models will be created basing on robust regression principals. The main idea behind robust regression is to apply weights to each observation, with better-behaved observations having a bigger weight applied, whereas the strong outliers being properly discriminated, with applied weights reaching up to zero. Specifically, I have decided to use iteratively reweighted least squares method originally proposed by Paul W. Holland and Roy E. Welsch, proposed in 1977.

Models' estimations analysis

For each position I created four separate iteratively reweighted least squares log-linear model, with its coefficients visible in a table below. They will be further analyzed in this section. Full prints from Python StatsModels module and coefficients for linear counterparts of these models used in evaluations of predictions appear in the appendix. But first, key variables described in part I will be discussed. Note that each interpretation below bases on an assumption that all other things remain equal.

Age

`Age` (or, actually, square of original age variable) has been a statistically significant determinant for each position – the older the player is, the lower his value on average. The power of the effect was relatively similar for each position.

Height

Height has been a significant variable for defenders at significance interval equal to 0.1, with positive coefficient, while tested in a simple robust regression model with height being the only coefficient – the effect was however losing its significance with more variables being added, due to what variable was not included in a final model.

Preferred foot

There is no statistical evidence that preferred foot contributes to the defenders' valuation – although variable `foot_both`, which is a binary variable describing if player feels comfortable with boot feet on the equal level, was statistically significant, its coefficient was negative, which is an illogical outcome and possibly related to the selection bias within the dataset. The effect would probably diminish with more data available, for that reason it was not included in final models. Preferred foot was also a significant factor for goalkeepers, however, due to their relatively much lower participation in game compared to other positions, I reason that this might be the case of a selection bias as well. In the end, preferred foot variables are not included in modelling.

Leagues

For all positions, players from Premier League had a higher valuation on average, on the contrary to the Ligue 1 players, where negative impact on the league was statistically significant. Binary variables for other leagues – La Liga, Serie A and Bundesliga – tended to give different conclusions regarding their significance depending on the variables added to the model (including binary variables for other leagues), for that reason they have not been included in final models. The impact of league played for prediction error will be analyzed later.

Bundesliga players might be slightly underestimated in final estimations under current specifications. The reason for that is that Bundesliga teams play 34 games throughout the season – there are 18 teams within the league, compared to 20 teams in other leagues.

Coefficients estimates

Below attached is the table with the coefficients for final models for all positions. Further description shall be provided later.

Figure 2.1 A table with coefficients for each model

	Goalkeepers	Defenders	Midfielders	Forwards
Intercept	16.7067	15.6328	16.6653	15.8348
age	-0.0022	-0.0019	-0.0018	-0.0017
CL	0.2449	0.2762	0.1735	0.2613
wins_gk	0.0781	-	-	-
draws_gk	0.0407	-	-	-
passes_pct_launched_gk	0.0128	-	-	-
psnpxg_per_shot_on_target_against	-3.8398	-	-	-
isPremierLeague	0.5557	0.5195	0.5385	0.5914
isLigue1	-0.4141	-0.2382	-0.3491	-0.3805
clean_sheetsm	62.6687	-	-	-
goals	-	0.0339	0.0532	0.0396
xg_xa_per90	-	0.9086	0.7456	0.3446
passes_ground	-	0.0006	-	-
touches_att_pen_area	-	0.0056	-	0.0029
touches_def_pen_area	-	0.0019	-	-
aerials_won_pct	-	0.0024	-	-
Pts	-	0.0103	0.0136	0.0099
xGA	-	-0.0154	-0.0142	-0.0151
xG	-	0.0078	0.0069	0.0081
passes_completed_short	-	-	0.0009	-
passes_into_final_third	-	-	0.0013	0.0054
carry_distance	-	-	0.000038	-
tackles_won	-	-	0.0048	-
gca	-	-	-	0.0137
dribbles_completed	-	-	-	0.003

As visible above, each position follows its own set of rules regarding what impact the players valuation, although there are certain things that are in common. Variables have been selected considering four main criteria: if they were statistically significant at significance level of 0.1, if there is no multicollinearity occurring, if interpretation of their coefficients made logical sense and, in the end, considering their relationship for adjusted R-squared metric – given two variables which were entering into relationship with each other, the variable which contributes better to the determination metric was chosen, provided that their estimate made logical sense and were statistically significant. Further analysis will be conducted on next pages of the thesis, including models' diagnostics and explanations of variables.

Team performance

Beyond the variables discussed earlier, the important factor that is common for each position is team's performance. It is to be expected that, on average, there is a higher probability that the better teammates the player has, the higher level of potential he can reach. The valuables which measure team's overall ability are `CL`, `Pts`, `xG` and `xGA` for defenders, midfielders and forwards, which are and `CL`, `wins_gk` and `draws_gk` for goalkeepers. Their meaning equals to:

- `CL` – binary measuring if the player's team played in Champions League following season,
- `Pts` - player's team points,
- `xG` - player's team expected goals,
- `xGA` - player's team expected goals against,
- `wins_gk` - wins of player's team when he played as a goalkeeper,
- `draws_gk` - draws of player's team when he played as a goalkeeper,

The reason for applying different variables to measure team's performance for goalkeepers is due to that the huge differences in adjusted R-squared score – it tended to raise rapidly in case of goalkeepers using player's team wins and draws when the player played as a goalkeeper rather than basing on his team overall points, whereas xG and xGA variables were insignificant under given conditions.

It is essential to explain what expected goals (xG) statistic means, as it might be an unknown term for readers less interested in football. To describe it briefly, it is a probability metric measuring if a given shot will result in goal, based on factors such as distance from the goal, angle of the shot, shooting part, passage of play and more, basing on historical data. In that case, it is an aggregated value of all expected goals throughout given season – which did not necessarily result in a goal itself. It should be quite intuitive that they should be strongly related to team's performance.

Overall, the results lead to the conclusion that player's team overall performance strongly contributes to his valuation on average – regardless of the position. Even for defenders, their team's offensive performance matters, like with team's defensive performance for attackers, what is logical, as mentioned earlier – especially in modern football, where defenders contribute to the attacking actions of the team and the popularity of so-called "defensive strikers" is on the rise.

On the next page attached is the table presenting 25 teams with the highest average value of the players within the dataset that proves the relationship between team performance, including number of points, xG, xGA and participation in UEFA Champions League with average footballer's valuation within a team, what legitimizes my conclusions regarding the observed phenomena from a model.

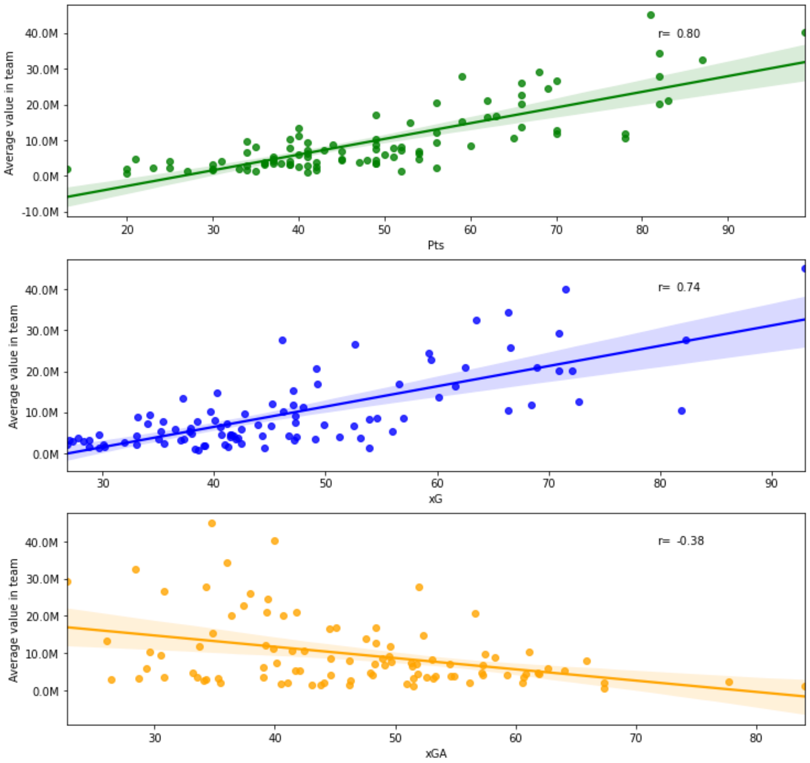
Figure 3.2.1 25 teams with highest average player's value and their team performance statistics

Team	League	League Ranking	Average Value	Pts	xG	xGA	CL
Manchester City	Premier League	2	45.09M	81	93	34.7	Yes
Liverpool	Premier League	1	40.13M	99	71.5	40	Yes
Barcelona	La Liga	2	34.39M	82	66.4	36	Yes
Real Madrid	La Liga	1	32.62M	87	63.5	28.4	Yes
Paris S-G	Ligue 1	1	29.24M	68	70.9	22.7	Yes
Bayern Munich	Bundesliga	1	27.78M	82	82.3	34.3	Yes
Tottenham	Premier League	6	27.75M	59	46.1	52	No
Atletico Madrid	La Liga	3	26.51M	70	52.6	30.8	Yes
Chelsea	Premier League	4	25.91M	66	66.6	37.9	Yes
Dortmund	Bundesliga	2	24.41M	69	59.2	39.4	Yes
Manchester Utd	Premier League	3	22.72M	66	59.4	37.4	Yes
Juventus	Serie A	1	21.07M	83	68.9	41.8	Yes
Napoli	Serie A	7	21M	62	62.5	39.3	No
Arsenal	Premier League	8	20.61M	56	49.2	56.6	No
Inter	Serie A	2	20.07M	82	72.1	40.7	Yes
RB Leipzig	Bundesliga	3	20.07M	66	70.9	36.4	Yes
Everton	Premier League	12	16.94M	49	49.3	48.4	No
Leverkusen	Bundesliga	5	16.89M	63	56.6	45.1	No
Leicester City	Premier League	5	16.46M	62	61.6	44.5	No
Wolves	Premier League	7	15.29M	59	47.1	34.8	No
Valencia	La Liga	9	14.85M	53	40.3	52.3	No
Milan	Serie A	6	13.75M	66	60.1	47.6	No
Lyon	Ligue 1	7	13.43M	40	37.2	26	No
Roma	Serie A	5	12.61M	70	72.7	48.4	No
Real Sociedad	La Liga	6	12.2M	56	45.2	39.2	No

Figure 2.2.2 Comparison of team performance statistics between top 25 teams and all teams from top 5 leagues

	Pts	xG	xGA
Average for top 25 teams:	67.84	61.008	39.788
Average for all teams:	48.520408	46.56429	46.56122

Figure 2.2.3 Distribution of team performance statistics and their relation to average value of players in a team



Goalkeepers on-pitch statistics

In case of goalkeepers, three variables that relate to his on-pitch performance are included within a model, with each of them being statistically significant. These are:

- `clean_sheetsm` - goalkeeper's clean sheets per a minute played (due to the given denominator the coefficient seems so huge),
- `psnpxg_per_shot_on_target_against` - Post-shot non-penalty expected goals per shot on target against the player playing as a goalkeeper,
- `passes_pct_launched_gk` - % of passes longer than 40 yards that were launched not including goal kicks,

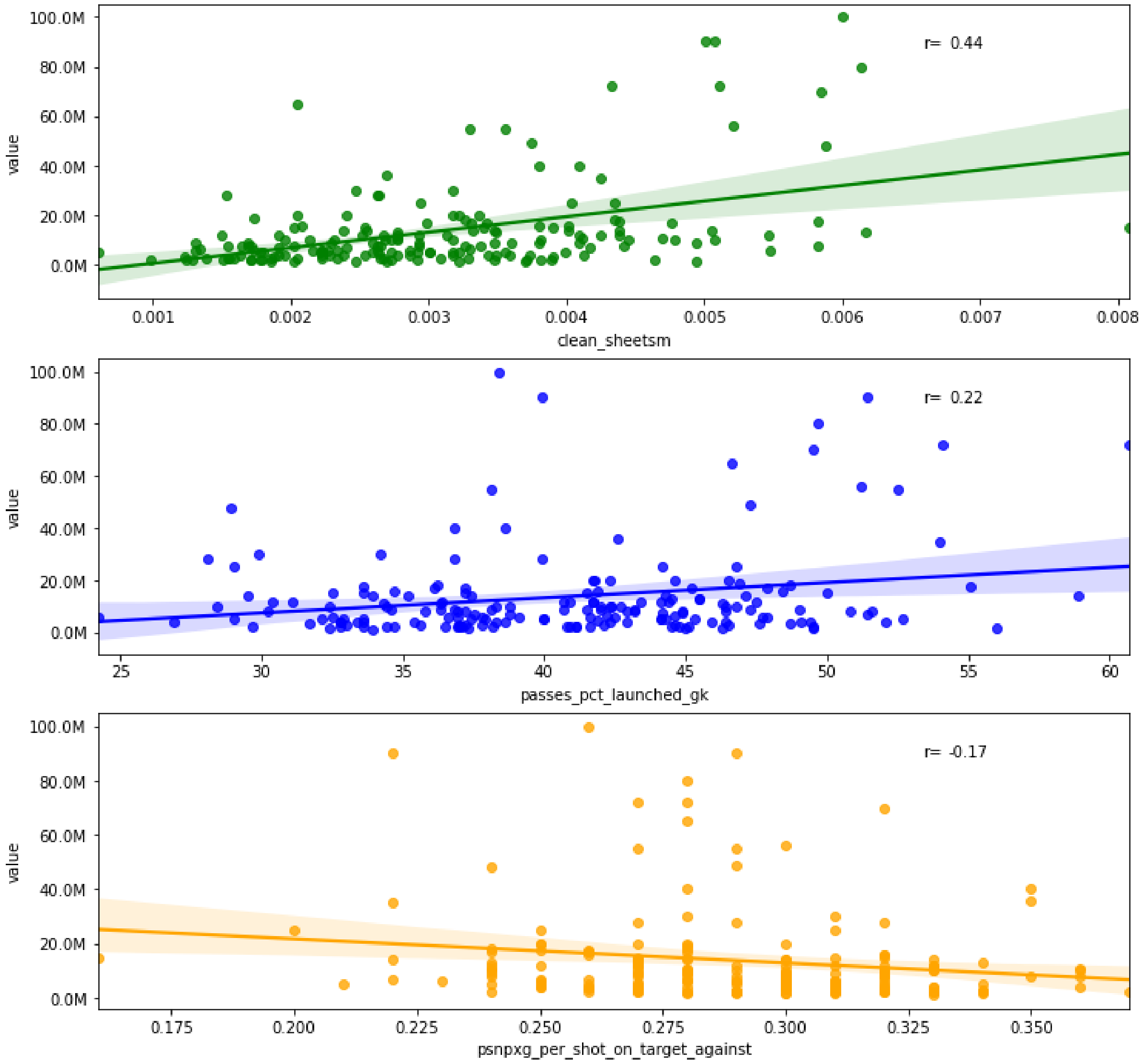
Unsurprisingly, the most important factor of goalkeepers' valuation is their goalkeeping ability. However, clean sheets per a minute played variable also possibly includes team's defensive performance as, as logic suggests, the lower the number of shots of target, the lower the chances of conceding goal by the goalkeeper. `psnpxg_per_shot_on_target_against` variable is more focused on pure goalkeeping ability and how the goalkeeper possibly can cope with the shots. Although it is possible that certain goalkeepers are simply unlucky to being shot with almost unstoppable attempts on a constant basis, what happens and could be balanced with adjusting the variable with expected goals against metric, it is most likely that those differences would be statistically insignificant and that on average it is one of the best variables that distinct world class goalkeepers. The last variable is % of passes longer than 40 yards that were launched, not including goal kicks, suggesting that in modern football goalkeepers who actively participate in playing out the ball with the long passes are higher valued on average, what is in fact a visible trend in recent years. Each effect is visible looking at top 10 most valuable goalkeepers at transfermarkt.de (basing on 2019/20 season), with `clean_sheetsm` and `passes_pct_launched_gk` variables average being higher than by the rest of the players, on the contrary to the `psnpxg_per_shot_on_target_against,` which is to be expected.

On the next page there is a table providing an overview of distribution of given on-pitch statistics by top 10 most valuable goalkeepers and how the average of these variables for them compares to the average for all goalkeepers within a dataset, and a dashboard with subplots drawing simple regression lines between footballers' values and given variables, providing insights about the distribution of significant variables within a dataset and their relation to the endogenous variable. The results should explain that the observed phenomena explained by models are actually legitimate. The same scheme applies to the rest of the positions.

Figure 2.3.1 10 most valuable goalkeepers and their chosen statistics values, compared to the rest of the players

player	clean_sheetsm	psnpxg_per_shot_on_target_against	passes_pct_launched_gk
Jan Oblak	0.005007	0.22	39.9
Marc-Andre Ter Stegen	0.004321	0.27	60.7
Alisson	0.005112	0.28	54.1
Ederson	0.00521	0.3	51.2
Gianluigi Donnarumma	0.003749	0.29	47.3
Thibaut Courtois	0.005882	0.24	28.9
David de Gea	0.003801	0.28	38.6
Kepa Arrizabalaga	0.002694	0.35	42.6
Wojciech Szczęsny	0.004245	0.22	54
Alex Meret	0.001538	0.29	36.8
Average for top 10 players:	0.0041559	0.274	45.41
Average for all players:	0.002929	0.28396	40.319802

Figure 2.3.2 Distribution of goalkeeper's on pitch statistics and their relation to goalkeeper's value



Defenders on-pitch statistics

There are six different on-pitch variables contributing to the defenders' valuation in my model:

- `goals` – number of goals player scored,
- `xg_xa_per90` – sum of expected goals and expected assists per 90 minutes,
- `passes_ground` – player's passes that lie to the ground,
- `touches_att_pen_area` - touches in competitor's team penalty area,
- `touches_def_pen_area` - touches in own penalty area,
- `aerials_won_pct` – percentage of aerial duels won,

The key determinants of defenders' valuation from a model might be quite surprising, although in given model specification are logically explainable. Importance of offensive play of defenders is rising, including their participation in playmaking and action creating, what is proved by the impact of goals (both actual number and expected goals), expected assists and touches in the attacking area. Number of passes plays important role as well, what creates certain difficulties while considering that effect into a linear model, as there are several variables that were significant while placed within a model – however, they were entering into a strong relationship with each other, leading to multicollinearity, thus I have decided to insert the one which increased the adjusted R-squared metric at strongest, which is a number of players ground passes. This might lead the model to underestimate certain types of players, whereas the attempt to implement bigger playing style differentiations - for example, percentage of short, medium and long passes within overall number – this approach caused even more problems to the model itself and also these variables have been statistically insignificant.

Defensive abilities, naturally, play vital role as well. In this case, touches in own penalty area and percentage of aerial duels won were significant factors. First one bases more on overall participation in defensive play, including not only potential fight for a ball, but also playing out of the defence, whereas the second one can be considered as actual estimators of player's strength and aerial ability, as the percentage of aerial duels won speaks for itself – the better aerial capacities the player has, the better chances for him to win the aerial duel.

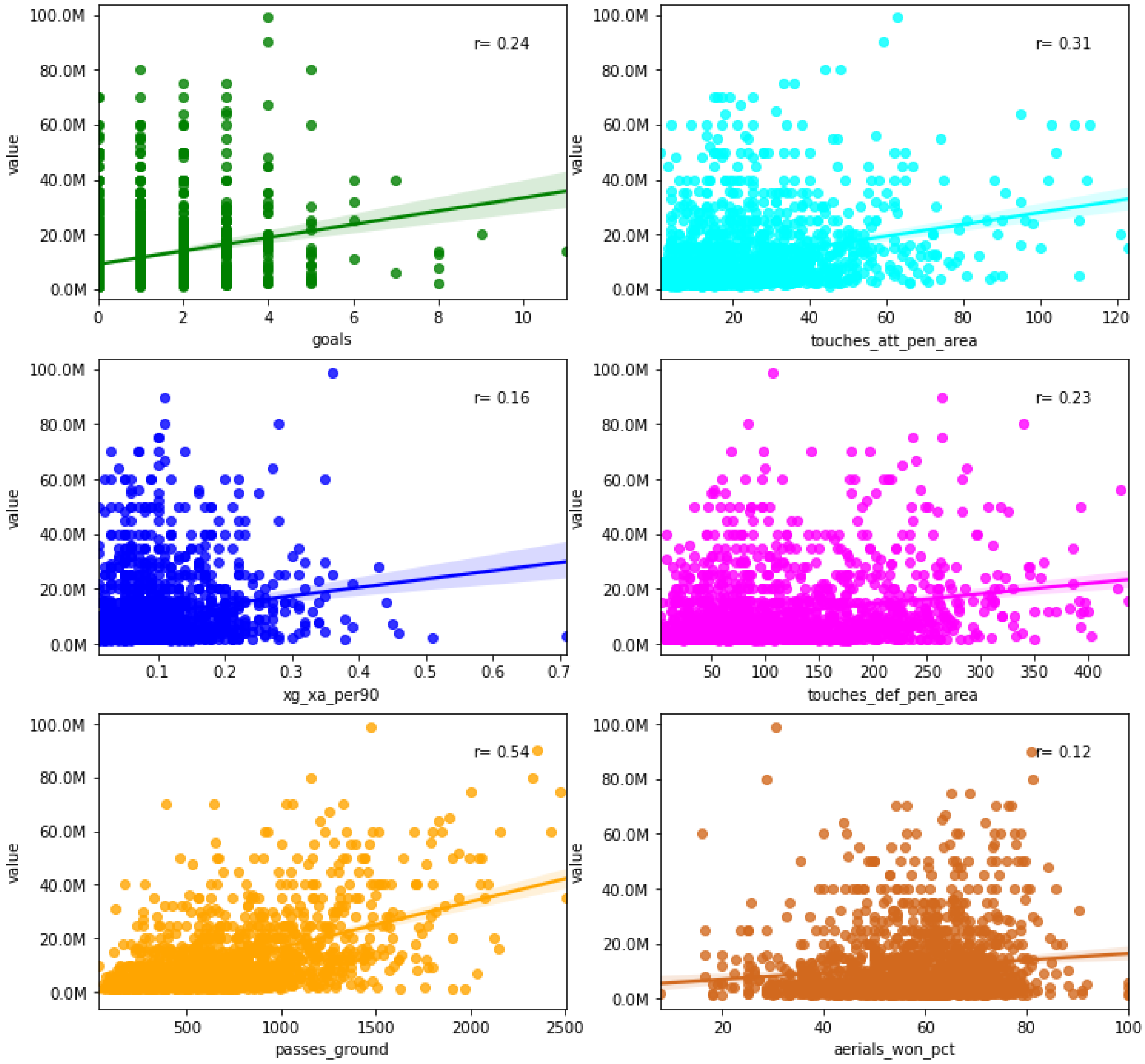
Each of the variables above had a positive effect for endogenous variable. It is however possible that creating distinct models for center backs and wing backs would cause the conclusions to be different, as playing styles for each position tend to differ in modern football, with wing backs leaning into offensive players much more often than its central counterparts.

Looking at top 10 most valuable defenders at transfermarkt.de (for 2019/20), there is an actual distinction between their statistics and for the rest of the players, although it is possible that the difference is bigger due to amount of substitute players within the dataset.

Figure 2.4.1 10 most valuable defenders and their chosen statistics values, compared to the rest of the players

player	goals	xg_xa_per90	passes_ground	touches_att_pen_area	touches_def_pen_area	aerials_won_pct
Trent Alexander-Arnold	4	0.36	1472	63	106	30.4
Virgil van Dijk	5	0.11	2330	48	340	81
Marquinhos	3	0.14	1054	17	98	56.2
Jose Maria Gimenez	0	0.03	644	15	142	76.4
Matthijs de Ligt	4	0.11	1251	22	240	63.9
Andrew Robertson	2	0.27	1834	95	100	43.9
Raphael Varane	3	0.04	1206	18	286	71.8
Aymeric Laporte	1	0.03	913	4	115	66.7
Lucas Hernandez	0	0.16	648	13	52	77.8
Harry Maguire	1	0.07	1787	57	429	73.7
Average for top 10 players:	2.3	0.132	1313.9	35.2	190.8	64.18
Average for all players:	0.916388	0.09388	667.180602	21.586957	120.09699	57.569565

Figure 2.4.2 Distribution of defender's on pitch statistics and their relation to defender's value



Midfielders on-pitch statistics

For midfielders, there are also six statistically significant on-pitch statistics:

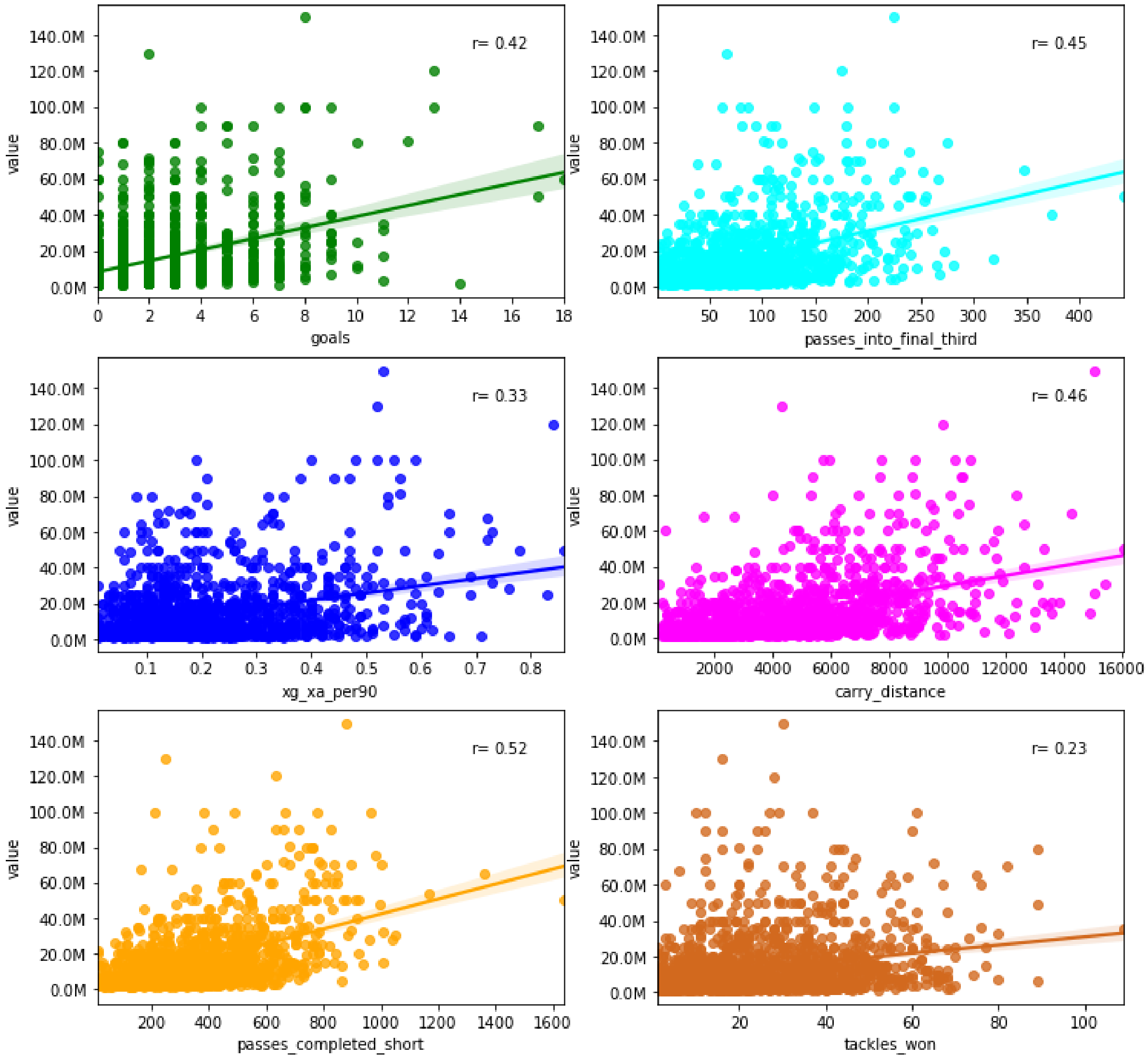
- `goals` – number of goals player scored,
- `xg_xa_per90` – sum of expected goals and expected assists per 90 minutes,
- `passes_completed_short` - passes completed between 5 and 15 yards,
- `passes_into_final_third` - passes into attacking third of football pitch,
- `carry_distance` - total distance the player moved with the ball by his feet,
- `tackles_won` - number of tackles won,

Creativity, actions creation and playmaking are essential skills for the midfielders, especially central and attacking ones, thus it is to be expected that the higher number and quality of passes overall, the more valuable the midfielder should be. However, the case is similar as by defenders' model – inserting multiple variables related to passes caused multicollinearity, and those are short passes which provide the best model fit. It is logical, as in most cases, the approach play bases on short passes, being possibly the safest way of retaining ball possession. Basing on what I written, one shall notice that importance of passes into attacking third and carries with the ball is logical, as those are strong contributors into attacking engagement of a whole team – key drivers of creating goal scoring actions, with the first variable causing direct danger for the opponent's goal, whereas the second one being crucial for pushing the ball forward in the approach play. In general, overall offensive player's contribution, including goals, expected goals and expected assists is also something that visibly differs top notch players, just like in the case of defenders and forwards as well. The last significant variable in model is the number of tackles won, what reminds of the importance of midfielders in defensive play as well. However, this might seem counterintuitive that this variable proved to be irrelevant by defenders – nevertheless I think it is explainable. This might be the case of defenders needing to be more cautious while playing near their own goal, whereas midfielders can play riskier, with potential counter attacks against the competitor being eventually more dangerous due to lower distance to the opponent's goal. On the other hand, number of tackles might be relatively high enough for each defender to not be a decisive factor for his higher valuation. In the end, the importance of defensive midfielders in certain tactics in modern football is undoubted, with ball-winning midfielders focusing on hard tackling and aggressive play finding their place in majority of teams. Regardless of my own conclusions, the matter of relation of tackles towards footballers' valuation is an interesting topic for further research.

Figure 2.4.1 10 most valuable midfielders and their chosen statistics values, compared to the rest of the players

player	goals	xg_xa_per90	passes_completed_short	passes_into_final_third	carry_distance	tackles_won
Kevin De Bruyne	13	0.84	634	175	9868	28
Kai Havertz	12	0.56	681	101	8880	20
N'Golo Kante	3	0.19	438	105	5309	43
Paul Pogba	1	0.35	374	118	4009	16
Frenkie de Jong	2	0.14	785	154	9514	22
Saul Niguez	6	0.17	660	135	6343	65
Christian Eriksen	3	0.48	433	96	4271	18
Casemiro	4	0.12	753	183	6120	75
Sergej Milinković-Savić	7	0.34	562	131	8260	59
Joshua Kimmich	4	0.31	846	252	12639	40
Average for top 10 players:	5.5	0.349622222	616.6	145	7521.3	38.6
Average for all players:	1.771325	0.204918	309.177858	67.738657	4130.255898	24.188748

Figure 2.5.2 Distribution of defender's on pitch statistics and their relation to midfielder's value



Forwards on-pitch statistics

And, in the case of forwards, there are also six on-pitch statistics included in final models:

- `goals` – number of goals player scored,
- `xg_xa_per90` – sum of expected goals and expected assists per 90 minutes,
- `passes_into_final_third` - passes into attacking third of football pitch,
- `touches_att_pen_area` - touches in competitor's team penalty area,
- `gca` - player's goal creating actions,
- `dribbles_completed` – number of player’s successful dribbles,

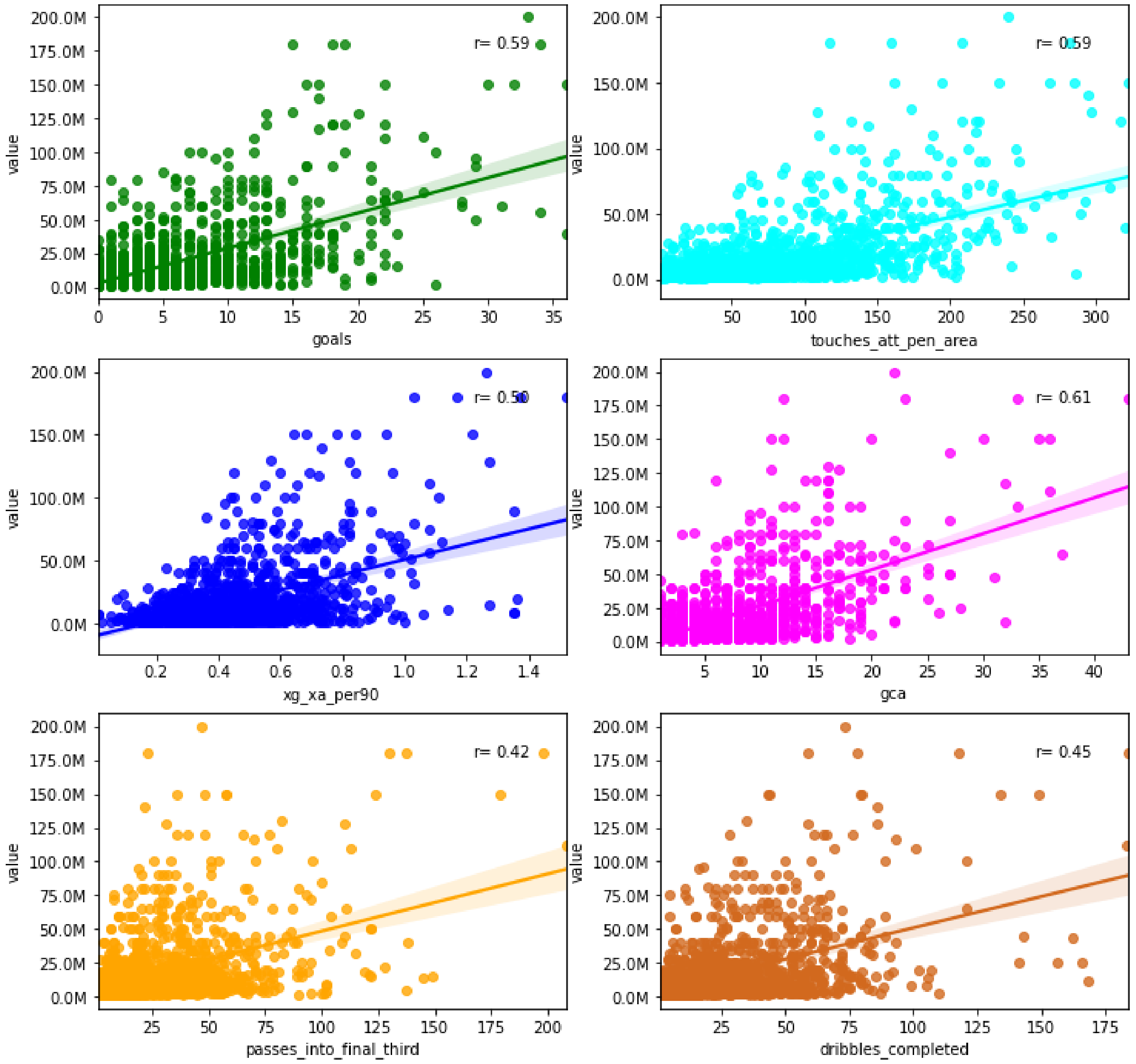
It is to be expected that the main contributor to the forwards valuation is goal threat he creates, and all overall variables relate to that. Count of goals, expected goals and expected assists naturally matter for forwards – intuitively even more than in the case of defenders and midfielders. Creating direct goal threat with passes relates with number of touches in opponent’s penalty area, whereas passes into final third and player’s goal creating actions relate to providing good chances to score for teammates, pushing the game forward and putting pressure on opponent’s defence. Number of player’s successful dribbles shall, on average, reflect the ability to the player to pass the opponent, which, logically, can create new opportunities for player’s team in approach play or in a final third.

Compared to defenders and midfielders, the number of non-specific passes (excluding passes into final third) is not significant for footballers’ valuation, ceteris paribus. Given models suggest that for midfielders, the ability to create goal threat is much more important than participation in playmaking and approach play itself, although this could be changed providing different model specification. This and other variables that could be significant under other circumstances will be discussed in the next part.

Figure 2.6.1 10 most valuable midfielders and their chosen statistics values, compared to the rest of the players

player	goals	xg_xa_per90	passes_into_final_third	touches_att_pen_area	gca	dribbles_completed
Kylian Mbappe	18	1.52	23	208	12	59
Raheem Sterling	20	0.82	31	296	11	59
Neymar	13	1.27	110	109	17	86
Mohamed Salah	19	0.84	41	317	16	66
Sadio Mane	18	0.69	65	219	16	76
Harry Kane	18	0.45	48	132	6	28
Jadon Sancho	17	0.72	70	144	32	93
Lionel Messi	25	1.08	208	217	36	183
Antoine Griezmann	9	0.42	51	161	10	18
Joao Felix	6	0.51	41	85	4	20
Average for top 10 players:	16.3	0.832	68.8	188.8	16	68.8
Average for all players:	5.294677	0.446293	23.726236	75.519011	5.92	26.754753

Figure 2.5.2 Distribution of defender's on pitch statistics and their relation to defender's value



Diagnostics of models

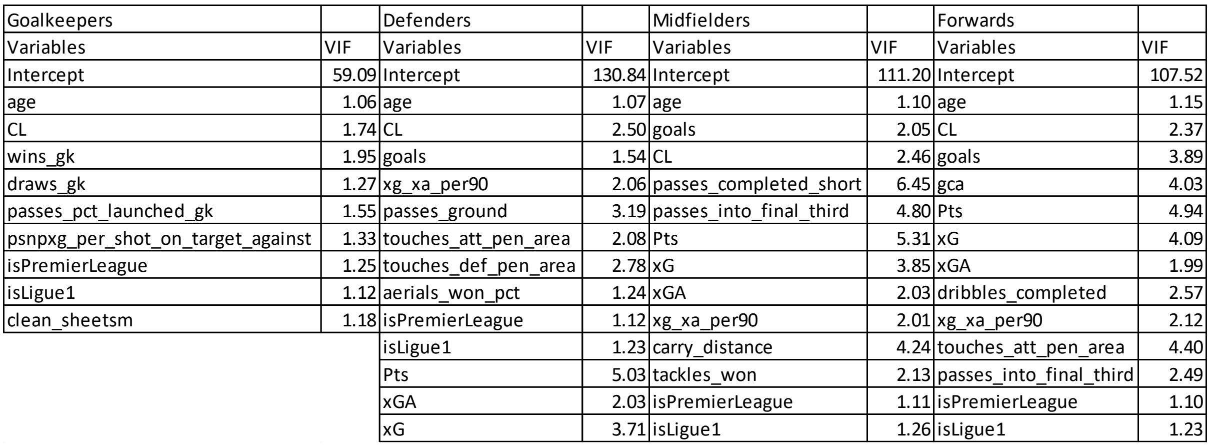
Figure 2.6.1 Table with chosen metrics and test statistics for each model

	Goalkeepers	Defenders	Midfielders	Forwards
Adjusted R-squared:	0.701	0.635	0.642	0.735
F-Statistic p-value:	~0	~0	~0	~0
Breusch-Pagan p-value:	0.14	0.11	~0	~0
Chow p-value:	0.73	0.99	0.85	0.91

The adjusted R-squared metrics are relatively high, suggesting that created models provide good fits to the observed data. However, the F-statistics from RESET test are very high, with probability values reaching approximately zeros, suggesting that there is a strong specification error, caused by either omitted variable bias or multicollinearity, assuring that further research might be vital. Despite using heteroskedasticity consistent method and applying logarithms of endogenous variable, there is still statistical evidence that heteroskedasticity occurs at midfielders and forwards models – however, iteratively reweighted least squares method includes the impact of heteroskedasticity, thus the conclusions regarding significance of variables should be considered legitimate – although it is not excluded that other heteroskedasticity resistant method would work better. The models pass Chow test for structural instability, basing on samples from 2019/20 and 2018/19 seasons.

Variance Inflation Factors

Figure 2.6.2 Table with variance inflation factors for each variable within each model



Goalkeepers		Defenders		Midfielders		Forwards	
Variables	VIF	Variables	VIF	Variables	VIF	Variables	VIF
Intercept	59.09	Intercept	130.84	Intercept	111.20	Intercept	107.52
age	1.06	age	1.07	age	1.10	age	1.15
CL	1.74	CL	2.50	goals	2.05	CL	2.37
wins_gk	1.95	goals	1.54	CL	2.46	goals	3.89
draws_gk	1.27	xg_xa_per90	2.06	passes_completed_short	6.45	gca	4.03
passes_pct_launched_gk	1.55	passes_ground	3.19	passes_into_final_third	4.80	Pts	4.94
psnpxg_per_shot_on_target_against	1.33	touches_att_pen_area	2.08	Pts	5.31	xG	4.09
isPremierLeague	1.25	touches_def_pen_area	2.78	xG	3.85	xGA	1.99
isLigue1	1.12	aerials_won_pct	1.24	xGA	2.03	dribbles_completed	2.57
clean_sheetsm	1.18	isPremierLeague	1.12	xg_xa_per90	2.01	xg_xa_per90	2.12
		isLigue1	1.23	carry_distance	4.24	touches_att_pen_area	4.40
		Pts	5.03	tackles_won	2.13	passes_into_final_third	2.49
		xGA	2.03	isPremierLeague	1.11	isPremierLeague	1.10
		xG	3.71	isLigue1	1.26	isLigue1	1.23

As visible above, there is no multicollinearity between variables themselves (considering that when VIF is higher than 10 there is a multicollinearity), however there is a huge variance inflation factor value for intercept within the models. A large VIF in the constant indicates that the explanatory variables have also a large constant component. Nevertheless, VIF value for intercept is irrelevant for models' diagnostics.

Other potential factors not included in models

Although models provide information about key determinants impacting players valuation depending on a position, they surely are not perfectly illustrating the reality. As proved later, there are statistical arguments to assume that omitted variable bias might occur. Despite dataset's broadness certain variables are not included, such as, for example, total distance run by a player throughout the season, which is strongly related to player's work rate and overall participation in on-pitch events. Dataset includes only these on-pitch statistics that were available on fbref.com website. Certain statistics are not available to the wide public, although widely used within a football analytics business, and example of total distance run is most likely one of them.

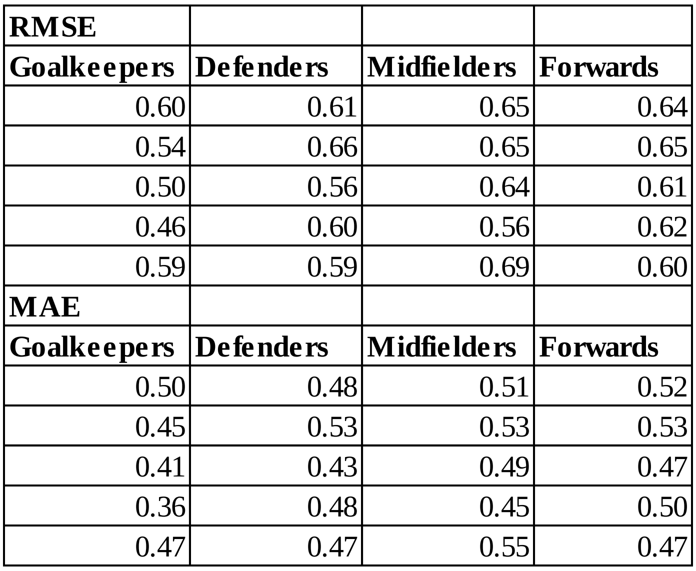
However, there are other potential candidates for significant variables within the dataset itself. A football fan's intuition would suggest that, for example, the number of tackles should contribute to the defender's valuation, as suggested earlier, or that number of long passes might be a statistically significant variable for midfielders just like a number of short passes, or that for forwards, the number of shots on target contributes to the final valuation just like the number of goals scored. Note that conclusions regarding variables significance could change under other conditions, using different model specifications. Some of the variables that would be most likely significant were not inserted in final models due to causing multicollinearity. It is of course possible as well that final conclusions would change using different estimation methods or appending more observations to the final dataset. The table with correlations of all variables from the dataset with transfermarkt.de values for each position is attached in the appendix, providing suggestions for further research.

Chapter III - Evaluation of models

In-sample evaluation

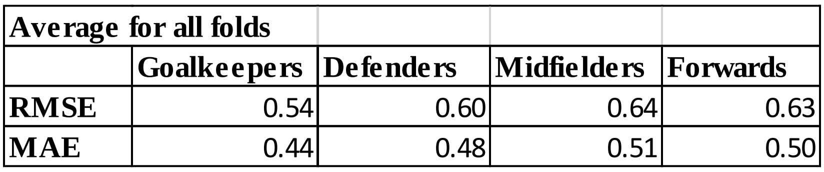
In this part created models will be evaluated checking for their utility to predict actual footballer price. First, an in-sample evaluation will be conducted, with logarithmic root mean squared error and mean absolute error being calculated and discussed, where I will also check if models could create a decent “transfer market hierarchy”, that means if they could successfully distinguish more valuable and less valuable players and rank them accordingly. For out-of-sample evaluation I will compare predicted valuations with 50 real-life transfers from 2020/21 season, where I will also attempt to include the impact of players valuations throughout all three seasons in real time.

Figure 4.1.1 A table with 5-fold cross validation for each model, measuring root mean square (logarithmic) error and mean absolute (logarithmic) error)



RMSE			
Goalkeepers	**Defenders**	**Midfielders**	**Forwards**
0.60	0.61	0.65	0.64
0.54	0.66	0.65	0.65
0.50	0.56	0.64	0.61
0.46	0.60	0.56	0.62
0.59	0.59	0.69	0.60
MAE			
Goalkeepers	**Defenders**	**Midfielders**	**Forwards**
0.50	0.48	0.51	0.52
0.45	0.53	0.53	0.53
0.41	0.43	0.49	0.47
0.36	0.48	0.45	0.50
0.47	0.47	0.55	0.47

Figure 3.1.2 A table with average root mean square (logarithmic) error and mean square (logarithmic error) for each model from all folds



Average for all folds				
	Goalkeepers	**Defenders**	**Midfielders**	**Forwards**
RMSE	0.54	0.60	0.64	0.63
MAE	0.44	0.48	0.51	0.50

Presented above is 5-fold cross validation measuring errors for errors between predicted logarithms of values and logarithms of true values. I have applied two metrics – root mean squared error (RMSE) and mean absolute error (MAE) – in order to draw conclusions, but first I would like to shortly elaborate what those metrics actually are and how to interpret them in case of logarithms use.

Root mean squared error is actually a standard deviation of model's residuals, describing how far from the regression line the data points are, providing information about how concentrated actual values around the models fit are, being one of the key metrics in case of regression model evaluation. Mean absolute error is a measure of an average absolute value of difference between predictions and actual values of endogenous variable, saying how big is the prediction error on average. While using logarithms, both metrics calculate a percentage error rather than absolute value error, suggesting how much do the predicted values deviate from the geometric mean.

Basing on models, we can see that RMSLE and MALE values vary between 0.46-0.69 and 0.36-0.55 accordingly, basing on the fold and position we take under consideration. That would equal that models yield between 46% and 69% or 36% and 55% prediction error, depending on metric. Basing on the table on the previous page we can see that model for goalkeepers provides possibly the best fits, whereas midfielders and forwards models are slightly worse than defenders. This might be caused by outliers impacting estimations, because as the reader might remember from the previous part of the essay heteroskedasticity is still a problem for midfielders and forwards models. The outliers, especially among the most valuable players might severely impact the average error values, causing it to go up – although for each fold within cross validation the errors estimates are relatively stable.

As we can see from the scatter plots on the next page, residuals have similar distribution for each position, with the prediction error being higher for more valuable players, steadily decreasing for lower observations. This is not surprising, however it suggests that value estimations remain stable independently from positions, what is a good news. However, there are some severe outliers, especially in case in forwards, what might suggest that models could be better specified, although they are expected to appear in case of heteroskedasticity.

Figure 3.2.1 Graphs presenting distribution of transfermarkt.de values and predictions from robust regression models and differences between them for each observation (residuals)

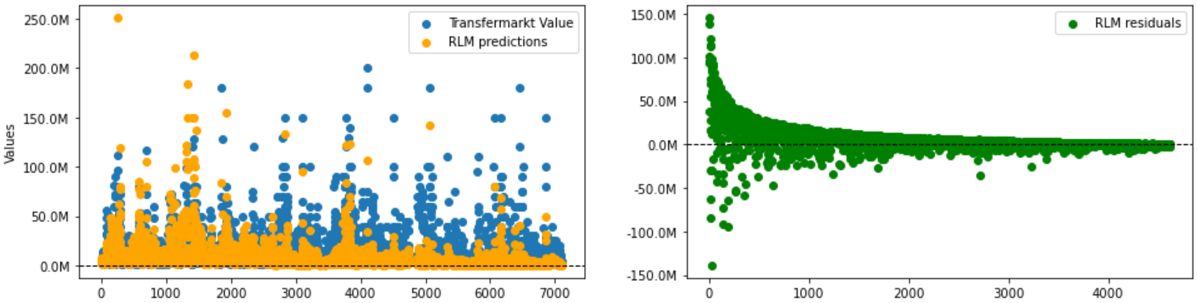
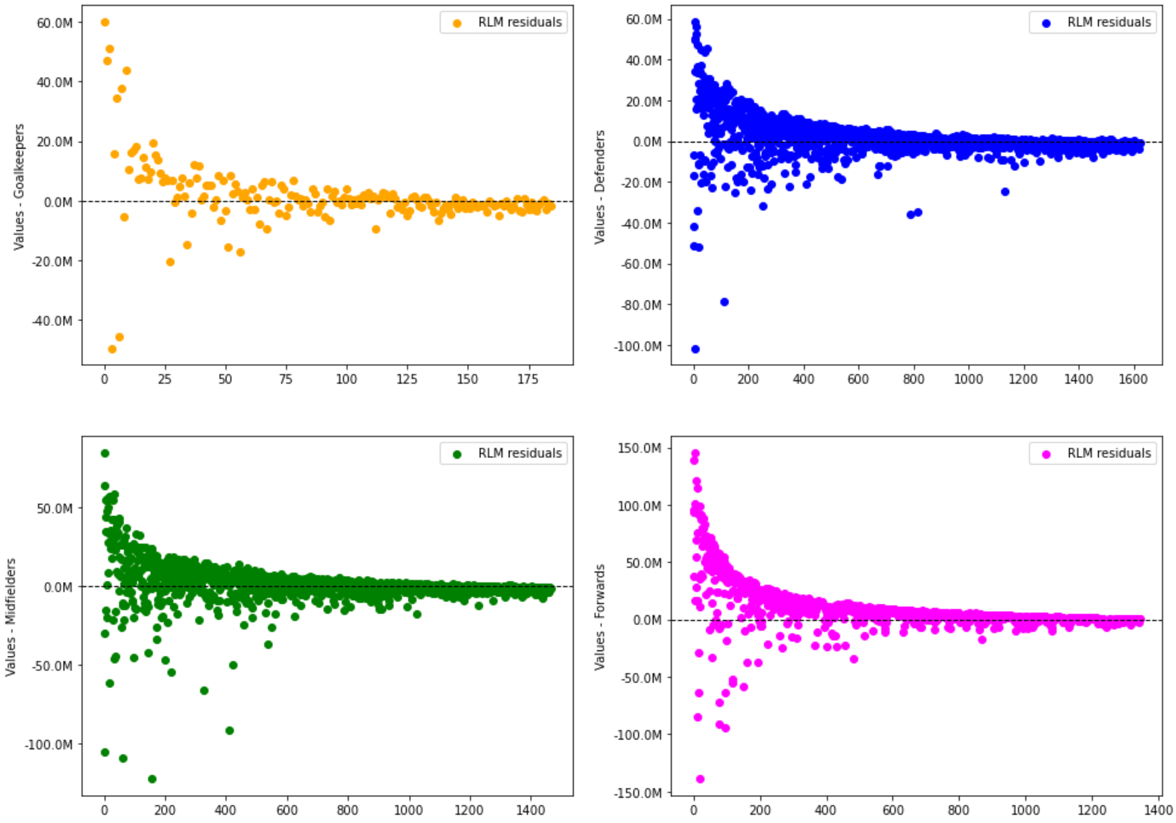
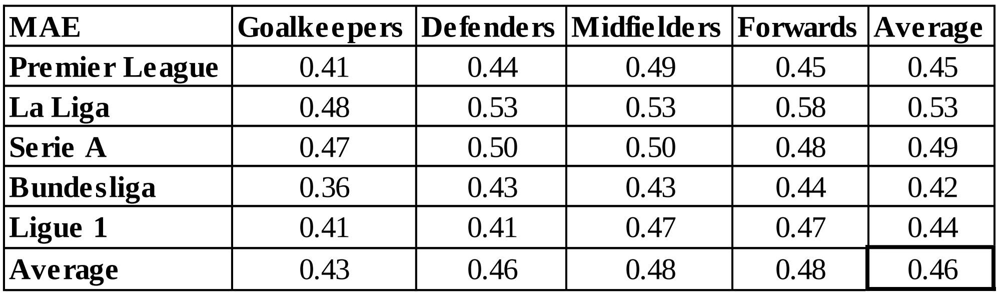


Figure 3.2.2 Distribution of residuals for each model



Below compared are average mean absolute logarithmic errors for each league, depending on positions:

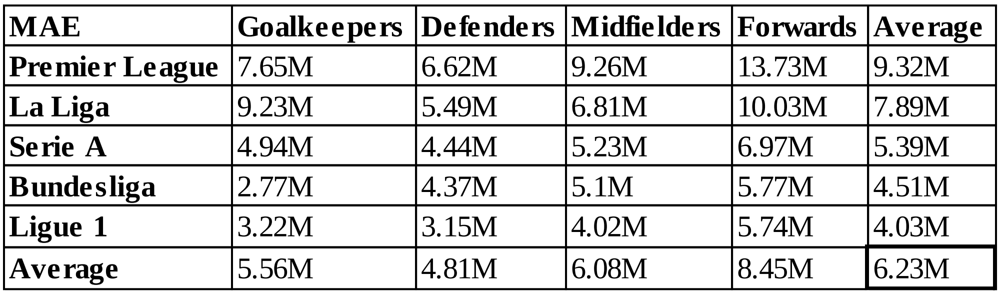
Figure 3.3 A table presenting mean absolute logarithmic error per each league and position



MAE	Goalkeepers	Defenders	Midfielders	Forwards	Average
Premier League	0.41	0.44	0.49	0.45	0.45
La Liga	0.48	0.53	0.53	0.58	0.53
Serie A	0.47	0.50	0.50	0.48	0.49
Bundesliga	0.36	0.43	0.43	0.44	0.42
Ligue 1	0.41	0.41	0.47	0.47	0.44
Average	0.43	0.46	0.48	0.48	0.46

As we can see, the mean absolute logarithmic error for each league ranges from 0.42 up to 0.53. The predictions for La Liga seem to have the biggest error on average, reaching up to 58% for forwards, with Serie A players having a slight difference in average error as well compared to Bundesliga, Ligue 1 and Premier League leagues. Possibly, a more comprehensive specification of models – including inserting binary variables for other leagues – would solve the issue, causing the predictions to be more stable structurally. In the end, we can see that predictions for all positions and leagues combined differ by 46% from a geometric mean on average. Below attached are results for exponents of values, giving more concrete insights about what exact values do the predictions differ – we can see that average mean absolute error for all positions equals 6.23 million euros:

Figure 3.4 A table presenting mean absolute error per each league and position



MAE	Goalkeepers	Defenders	Midfielders	Forwards	Average
Premier League	7.65M	6.62M	9.26M	13.73M	9.32M
La Liga	9.23M	5.49M	6.81M	10.03M	7.89M
Serie A	4.94M	4.44M	5.23M	6.97M	5.39M
Bundesliga	2.77M	4.37M	5.1M	5.77M	4.51M
Ligue 1	3.22M	3.15M	4.02M	5.74M	4.03M
Average	5.56M	4.81M	6.08M	8.45M	6.23M

The last aspect of in-sample evaluation I plan to conduct is to evaluate if models rank the most valuable players in a similar as transfermarkt.de values do. For that I have selected ten most valuable players from predictions and compared them with actual transfermarkt.de valuations.

Figure 3.5 A table including 10 most valuable players for each position basing on my models, comparing their valuation to transfermarkt.de

Goalkeepers

Rank	Name	Position	Team	Predicted value	Transfermarkt.de value	Difference
1	Ederson	Goalkeeper	Manchester City	61.3M	56M	5.3M
2	Alisson	Goalkeeper	Liverpool	56.32M	72M	-15.68M
3	Thomas Strakosha	Goalkeeper	Lazio	40.36M	20M	20.36M
4	Jan Oblak	Goalkeeper	Atletico Madrid	38.73M	90M	-51.27M
5	Marc-Andre ter Stegen	Goalkeeper	Barcelona	37.71M	72M	-34.29M
6	Gianluigi Donnarumma	Goalkeeper	Milan	32.85M	49M	-16.15M
7	David de Gea	Goalkeeper	Manchester Utd	32.63M	40M	-7.37M
8	Thibaut Courtois	Goalkeeper	Real Madrid	31.2M	48M	-16.8M
9	Pierluigi Gollini	Goalkeeper	Atalanta	29.98M	13M	16.98M
10	Dean Henderson	Goalkeeper	Sheffield Utd	29.36M	14M	15.36M

Defenders

Rank	Name	Position	Team	Predicted value	Transfermarkt.de value	Difference
1	Virgil van Dijk	Defender - Centre-Back	Liverpool	122.08M	80M	42.08M
2	Trent Alexander-Arnold	Defender - Right-Back	Liverpool	115.58M	99M	16.58M
3	Andrew Robertson	Defender - Left-Back	Liverpool	98.16M	64M	34.16M
4	Harry Maguire	Defender - Centre-Back	Manchester Utd	76.31M	56M	20.31M
5	Alphonso Davies	Defender - Left-Back	Bayern Munich	62.07M	45M	17.07M
6	Joe Gomez	Defender - Centre-Back	Liverpool	55.28M	33M	22.28M
7	Benjamin Pavard	Defender - Right-Back	Bayern Munich	53.15M	28M	25.15M
8	Cesar Azpilicueta	Defender - Right-Back	Chelsea	47.85M	24M	23.85M
9	Kurt Zouma	Defender - Centre-Back	Chelsea	47.72M	28M	19.72M
10	Matthijs de Ligt	Defender - Centre-Back	Juventus	46.56M	67M	-20.44M

Midfielders

Rank	Name	Position	Team	Predicted value	Transfermarkt.de value	Difference
1	Kevin De Bruyne	Midfielder - Attacking Midfield	Manchester City	150.09M	120M	30.09M
2	Rodri	Midfielder - Defensive Midfield	Manchester City	108.24M	64M	44.24M
3	Mason Mount	Midfielder - Attacking Midfield	Chelsea	85.47M	40M	45.47M
4	Joshua Kimmich	Midfielder - Defensive Midfield	Bayern Munich	80.14M	64M	16.14M
5	Achraf Hakimi	Midfielder - Right Midfield	Dortmund	79.98M	54M	25.98M
6	Phil Foden	Midfielder - Central Midfield	Manchester City	73.76M	27M	46.76M
7	Luis Alberto	Midfielder - Attacking Midfield	Lazio	62.17M	50M	12.17M
8	Ilkay Gundogan	Midfielder - Central Midfield	Manchester City	54.53M	40M	14.53M
9	Jorginho	Midfielder - Defensive Midfield	Chelsea	53.64M	50M	3.64M
10	Georginio Wijnaldum	Midfielder - Central Midfield	Liverpool	53.54M	40M	13.54M

Forwards

Rank	Name	Position	Team	Predicted value	Transfermarkt.de value	Difference
1	Lionel Messi	Forward - Right Winger	Barcelona	255.47M	112M	143.47M
2	Raheem Sterling	Forward - Left Winger	Manchester City	214.98M	128M	86.98M
3	Mohamed Salah	Forward - Right Winger	Liverpool	183.58M	120M	63.58M
4	Timo Werner	Forward - Centre-Forward	RB Leipzig	159.7M	64M	95.7M
5	Gabriel Jesus	Forward - Centre-Forward	Manchester City	152.66M	56M	96.66M
6	Sadio Mane	Forward - Left Winger	Liverpool	148.19M	120M	28.19M
7	Marcus Rashford	Forward - Left Winger	Manchester Utd	138.1M	64M	74.1M
8	Robert Lewandowski	Forward - Centre-Forward	Bayern Munich	120.19M	56M	64.19M
9	Jadon Sancho	Forward - Right Winger	Dortmund	107.94M	117M	-9.06M
10	Roberto Firmino	Forward - Centre-Forward	Liverpool	104.4M	72M	32.4M

First of all I would like to remind once again how transfermarkt.de valuations works. The website provides rather approximate valuation, dividing players into segments – multiple players can get the same valuation. This fundamentally makes evaluating the ranking of all players a bit ill-advised, nevertheless the method should give insights if models can distinguish the best players accordingly, although it should not be strictly followed in evaluation, providing no more than the overview and information for reader’s subjective opinion.

Nevertheless – looking at top 10 most valuable players for prediction models, I can conclude that models can appropriately resemble the hierarchy on a transfer market, although surely, they have certain flaws. Goalkeepers' valuations seem relatively lower compared to the rest of positions, which is to be expected, but possibly not to such a degree – the highest valuated player is Ederson with over 61 million euro of valuation, whereas originally Jan Oblak has the highest valuation of 90 million euro, which is a much higher price. We also see a strong dominance of Premier League players, which is to be expected providing a model specification and distribution of values per league – the question is if the model does not favor Premier League players too strongly. On the other hand, there are no Ligue 1 players in a table – including Kylian Mbappe and Neymar Jr. being the most valuable players by transfermarkt.de, what rises question if a penalty set on Ligue 1 players is too strong.

For each position – possibly excluding forwards – I personally have an impression that the distribution of values might be too flat comparing to the original transfermarkt.de valuations. For example, only 7 defenders valuations reach over 50 million euros and only 7 midfielders are predicted to have more than a 60 million euro price. It is debatable if such a difference in distribution of values between forwards and the rest of the positions is actually strong, although it was suggested throughout the paper that it would occur. One can also consider potential bias among observations, which is possible as discussed in previous part of the essay. Let us consider Manchester City players – there are overall 7 players within a top 10 for each position. Manchester City playing style is very characteristic, with Pep Guardiola as a coach, a pioneer of "tiki-taka", a tactical style extremely focusing on short passes, approach play and ball possession. It is to be expected that average Manchester City player will pursue more (especially short) passes than the average player of the other average team. Considering that our models strongly promote players that commit a lot of passes, it is to be expected that my models will be biased for Manchester City players in certain degree – especially midfielders, where one of six on-pitch variables is number of completed short passes, what is possibly resulting in four out of top ten predicted midfielders' valuations are those of Manchester City players. What is even more interesting, players such as Gabriel Jesus and Phil Foden, who have a very big prediction error, are not Manchester City's first team key players as for 2019/20 season, starting in respectively 21 and 9 games in given season.

In the end, one should also remember that transfermarkt.de values are not perfect. For a football fan, certain results from my model might be even more accurate than transfermarkt.de evaluations – it is up to a subjective opinion. Let me take Robert Lewandowski as an example – by the end of 2020 he was awarded with one of the most prestigious awards FIFA Best Men's Player (of the year), even though he was worth "only" 56 million euros according to the transfermarkt.de, not even making it in top 10 most valuable forwards. Despite that and being older than the average top 5 footballer (being 32 years old in 1988), he is 8th most valuable forward according to my model, being valued more than two times higher than on transfermarkt.de – the difference there is, in my opinion, justified. Another interesting outlier is Lionel Messi, debatably one of the best footballers in the history, being the most valuable player according to my models despite being 33 years old. His statistics in 2019/20 have been tremendously great, what is visible in a table in a previous part of the essay, despite FC Barcelona's worse overall disposition. The final comparison of performance of original transfermarkt.de values and those of my own model will be conducted in out-of-sample evaluation.

Out-of-sample evaluation – comparison to real-life transfers

In order to conduct an out-of-sample evaluation, I have gathered 116 real-life transfers for 2020/21 that found a match with observations within a dataset and calculated an average mean absolute error for them.

Figure 3.6 Tables with mean absolute error results for out of sample evaluation for both transfermarkt.de and created models, measured by position and league

MAE Transfermarkt:	4.7M
MAE model:	7.8M

	Goalkeeper	Defender	Midfielder	Forward
Count:	9	36	40	39
MAE Transfermarkt:	5.78M	3M	1.16M	1.1M
MAE model:	3.95M	3.03M	6.98M	4.11M

	Premier League	La Liga	Serie A	Bundesliga	Ligue 1
Count:	17	19	43	20	28
MAE Transfermarkt:	3.81M	0.63M	0.55M	2M	4.04M
MAE model:	6.41M	5.46M	4.49M	4.22M	6.73M

Despite my best efforts, the models perform visibly worse than transfermarkt.de valuations – excluding the performance for goalkeepers. For each league transfermarkt.de benchmark performs better, although the conclusions are not impossible to change providing bigger amount of data as the sample is not the biggest.

One of the key reasons for which my models could perform worse is the fact they do only include the statistics for one selected season, not including performance throughout recent years. Transfermarkt.de valuations are more static, where the aspect of player's reputation is considered, and that bases on his performance in previous seasons as well. My models in current form are vulnerable for potential obstacles that lead to player underperforming, or not performing at all, such as injuries or worse form. If one wants to include it, one should at best change the specification of given models. However, the overhaul of models would require too much space as for this thesis, for that I will apply the alternative to consider player's performance over previous seasons. The method I will use is calculation of weighted mean of predicted value, using predictions for 2017/18 and 2018/19 seasons as well (providing they exist for given player), appropriately weighted. The weights applied are based on the correlation coefficients between 2019/20 season value and, respectively, 2017/18 and 2018/19 – although it is possible that other combination would create the most optimal results. The weights equal, accordingly:

Figure 3.7 Table with applied weights for each season in weighted means approach

	Weight
Season 2019/20	1
Season 2018/19	0.7846
Season 2017/18	0.6648

Figure 3.8 Tables with mean absolute error results for out of sample evaluation for both transfermarkt.de and created models using weighted means, measured by position and league

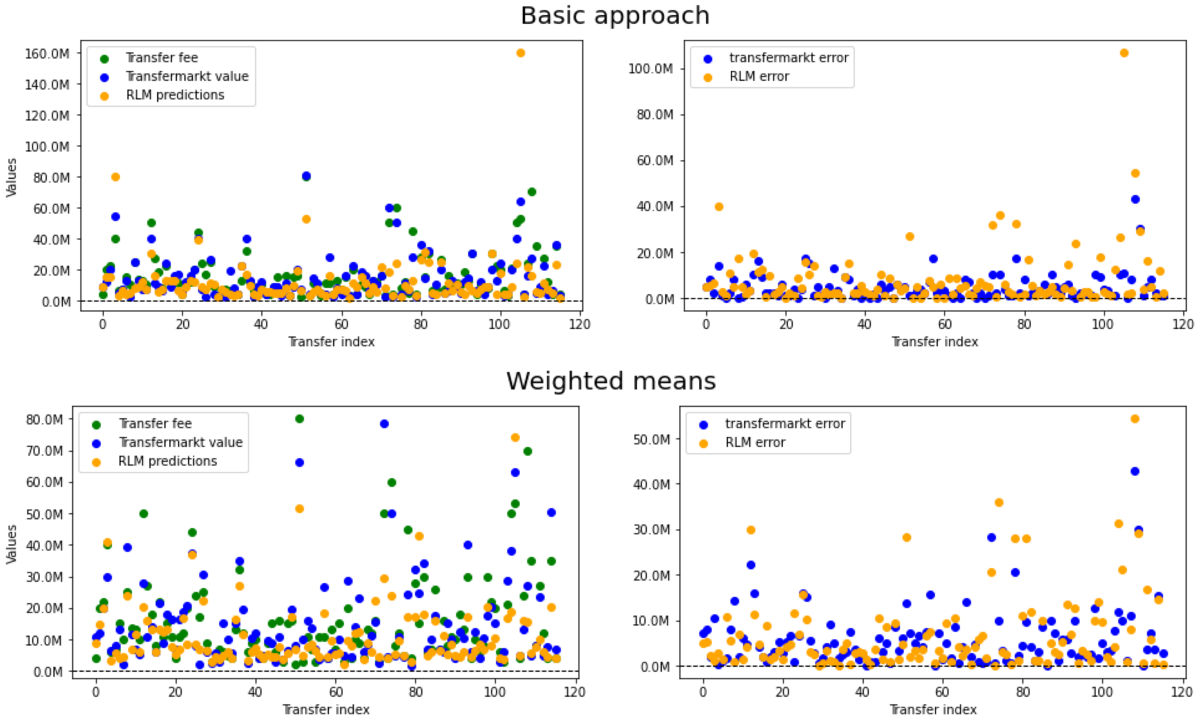
MAE Transfermarkt:	4.7M
MAE model:	6.7M

	Goalkeeper	Defender	Midfielder	Forward
Count:	9	36	40	39
MAE Transfermarkt:	5.78M	3M	1.16M	1.1M
MAE model:	2.65M	3.03M	0.83M	4.59M

	Premier League	La Liga	Serie A	Bundesliga	Ligue 1
Count:	17	19	43	20	28
MAE Transfermarkt:	3.81M	0.63M	0.55M	2M	4.04M
MAE model:	0.77M	7.79M	4.01M	1.45M	3.58M

Indeed, application of weighted means lowered the average mean absolute error, however models combined still perform worse than transfermarkt.de valuations – although, in this case, goalkeeper's model is still more effective - and midfielder's model began to provide lower errors than transfermarkt.de. Let me notice that using weighted means, model performs better for Premier League, Bundesliga and Ligue 1 players, but still worse for Serie A and La Liga players – what was actually suggested in in-sample evaluation. However, it is possible that conclusions would differ providing more observations.

Figure 3.9 Graphs comparing distributions of transfer fees, transfermarkt.de valuations, models' predictions for each player and prediction errors for each approach



Summary

Two key purposes of the paper were 1. To find out what determines the value of the football players, 2. To attempt to create a linear model that would value footballers with certain quality. These goals have been fulfilled partially – although evaluating certain key variables that contribute to the footballers valuation, there are more questions to be answered – more research is needed to discover more specific determinants, and whereas created models provide acceptable prediction errors and distinguish better performing players accordingly, they perform on average worse than transfermarkt.de value basing on out-of-sample evaluation on 116 observations. It is possible that using other model specification or estimation method, such as multicollinearity consistent method – for example, ridge regression, that would consider other significant variables that were not inserted into created models due to multicollinearity issues, possibly eliminating a bias towards, for example, players from teams focusing on ball possession and short passes. There are also other variables out of the dataset that might prove significant, such as player's injury proneness, contract length, or diverse on-pitch statistics that were not included, such as total distance ran throughout the season. Other improvements for my analysis would be to append more data from previous or from other leagues, or to consider the impact of past seasons valuation at the very beginning of model creation process. It is also possible that paper would provide more accurate and detailed information regarding determinants of footballer values having analyzed it basing on more concrete positions – distinguishing center backs from wing backs, offensive midfielders from defensive midfielders etc. rather than splitting them into given four segments: goalkeepers, defenders, midfielders and forwards.

The estimation method I have used for modelling purposes is robust regression basing on iteratively reweighted least squares. This is a logical choice due to occurring heteroskedasticity within a dataset, and although applying heteroskedasticity consistent method did not eliminate heteroskedasticity among midfielders and forwards models, it allowed me to appropriately evaluate important determinants of footballers ‘values. Each model did not pass RESET test, what would imply specification error, possibly related to omitted variable bias or nonlinearity, although they provided satisfying adjusted R-squared scores and did not suffer from structural breaks.

This thesis gives statistical evidence for significance of certain determinants for footballers' valuation. For all positions age and team performance – both offensive and defensive – were strongly related to the footballers' transfermarkt.de value, what has logical fundamentals. For goalkeepers, the goalkeeping ability is the key for their value, with the importance of participation in playing from the goal being evident as well. What distinguishes the most valuable defenders is their ability to contribute to the offensive and approach play of their team, although their defensive abilities, including aerial capacities, remain naturally important. For midfielders, the offensive participation is important as well, with bigger pressure on playmaking, creativity and ball control, but also the tackling abilities were crucial for their overall valuation – what surprisingly was not the case for defenders, although it is most likely logically explainable. For forwards, the conclusions are quite straight-forward, with their offensive abilities being the most important, including dribbling abilities, goal efficiency, but also participation in action's creation – in short, the overall goal threat they create is important.

In the end, there is another endogenous variable which contributes to footballers' valuation, and it is the reputation, which, in most cases, remains for longer than one season – in simpler words, the performance from past seasons has its impact for footballers valuation in given season as well.

At the beginning, paper presents a distribution of footballer's values, which is far from normal, and although most players are valued up to 20 million euros, there are some strong outliers, especially in case of the best players. Paper also provides an overview of characteristics of players in top 5 leagues, including information about age, height or preferred legs distribution. The average age of a player in top 5 leagues is over 25 years old and the distribution is close to normal whereas the average height is over 182 cm – higher than the average human height, although each characteristic mean differs basing on position one takes under consideration. More than 25% of players in top 5 leagues are left- or both-footed, whereas most players – excluding domestic players for each league – come from Brazil, France, Spain, Argentine and the Netherlands/Holland. On average, forwards are the most valuable, followed by midfielders, defenders and goalkeepers, whereas the average value of players is highest in English Premier League, followed by La Liga, Serie A, Bundesliga and Ligue 1, with valuation in La Liga, Serie A and Bundesliga being relatively similar.

Measuring appropriate footballer's valuation is impossible, as the actual player's value is unknown – as the classic says, the players is worth as much as somebody is willing to pay for him. However, such models can help in decision making process, giving an approximate valuation suggesting how high should be an appropriate price. The problem of endogeneity of player's valuation and his skill is one of the aspects why data analytics are revolutionizing the football world, as each team wants to maximize its potential squad quality, what contributes to overall performance. Although numbers do not play on a pitch, they definitely play its role behind the scenes, providing support for final success of a team.

Bibliography:

1. J. VanderPlas. (2016), “Python Data Science Handbook”
2. R.C. Hill, W.E. Griffiths, G.C. Lim (2018), “Principles of Econometrics”
3. I. Hendriks (2017), “Modelling the transfer prices of football players”
4. P.W. Holland, R.E. Welsch (1977), “Robustness regression using iteratively reweighted least-squares”
5. www.statista.com/statistics/261223/european-soccer-market-total-revenue/
6. https://bleacherreport.com/articles/2725494-barcelona-confirm-psg-activated-neymars-eur222m-transfer-release-clause
7. https://www.espn.com/soccer/soccer-transfers/story/3151518/saul-niguez-commits-to-long-term-atletico-madrid-contract
8. https://www.nbclosangeles.com/news/sports/lionel-messis-contract-reportedly-worth-673-million-most-expensive-for-athlete-in-any-sport/2514971/
9. https://eldesmarque.com/sevilla/real-betis/1159788-asi-ficho-el-big-data-a-lo-celso
10. https://anfieldindex.com/44942/reds-refuse-to-pay-timo-werner-release-clause-fee.html
11. https://www.transfermarkt.de/
12. https://www.statsmodels.org/

Data sources:

1. www.transfermarkt.de
2. www.fbref.com
3. www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset

Appendix:

Figure 4.1.1 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
aerials_lost	aerial duels lost	-0.04	0.14	0.09	-0.01
aerials_won	aerial duels won	-0.04	0.21	0.08	-0.03
aerials_won_pct	% of aerial duels won.	0.09	0.11	-0.01	-0.01
age	Player's age.	-0.25	-0.16	-0.12	-0.06
assisted_shots	number of assisted shots	-0.02	0.14	0.39	0.47
assists	Player's assists.	0.14	0.22	0.43	0.51
assists_per90	Assists per 90 minutes.	0.08	0.13	0.29	0.30
avg_distance_def_actions_gk	Average distance of defensive actions as goalkeeper.	0.15	-0.01		
ball_recoveries	Number of ball recoveries.	0.22	0.28	0.22	0.28
blocked_passes	number of times player blocking passes	-0.05	0.06	0.23	0.14
blocked_shots	number of times player blocking shots	-0.04	0.11	0.06	0.12
blocked_shots_saves	number of blocked shots that were on target		0.09	0.02	-0.02
blocks	number of player's blocks	-0.05	0.10	0.21	0.15
cards_red	Red cards	0.01	-0.01	-0.02	0.07
cards_yellow	yellow cards	0.00	0.03	0.07	0.09
cards_yellow_red	red & yellow cards		-0.01	-0.01	0.03
carries	Number of times the player controlled the ball with the feet.	0.28	0.50	0.49	0.52
carry_distance	total distance the player moved with the ball by his feet	0.18	0.45	0.46	0.48
carry_progressive_distance	Number of times the player controlled the ball with the feet towards enemy's goal.	0.16	0.44	0.46	0.47
CL	Binary measuring if the player's team played in Champions League following season.	0.45	0.52	0.48	0.53
clean_sheets	Player's clean sheets.	0.59			
clean_sheets_pct	Percentage of matches that result in clean sheet.	0.45			
clearances	number of ball clearances	-0.03	0.13	0.03	0.01
corner_kick_goals_against_gk	number of goals from corners against as goalkeeper	-0.03			
corner_kicks	number of corner kicks	-0.03	0.08	0.20	0.21
corner_kicks_in	corner kicks in a penalty goals	0.00	0.08	0.11	0.10
corner_kicks_out	number of outswinging corner kicks		0.08	0.13	0.07
corner_kicks_straight	number of straight corner kicks		0.03	0.14	0.07
crosses	number of crosses	-0.04	0.05	0.21	0.29
crosses_gk	Crosses as goalkeeper	0.17	-0.01		
crosses_into_penalty_area	number of crosses into penalty area	-0.05	0.01	0.15	0.18
crosses_stopped_gk	Crosses stopped as goalkeeper	0.16	-0.01		
crosses_stopped_pct_gk	% of stopped crosses as a goalkeeper.	0.06	-0.01		
D	Player's team draws	-0.10	-0.19	-0.18	-0.21
def_actions_outside_pen_area_gk	Defensive actions outside the penalty area as a goalkeeper.	0.25	-0.01		

Figure 4.1.2 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
def_actions_outside_pen_area_per90_gk	Defensive actions outside the penalty area per 90 minutes as a goalkeeper.	0.07	-0.01		
dispossessed	number of dispossessions	-0.03	0.06	0.30	0.36
draws_gk	Draws of player's team when he played as a goalkeeper.	0.14	-0.01		
dribble_tackles	Number of dribbles tackled	-0.01	0.07	0.16	0.06
dribble_tackles_pct	% of dribbles successfully tackled by a player.	0.21	0.04	-0.03	-0.04
dribbled_past	players dribbled past	-0.03	0.06	0.18	0.13
dribbles	number of dribbles	-0.02	0.08	0.32	0.47
dribbles_completed	number of successful dribbles	-0.01	0.10	0.33	0.47
dribbles_completed_pct	% of dribbles completed.	0.16	0.04	0.04	0.07
dribbles_vs	number of dribbles vs given player	-0.03	0.07	0.18	0.12
errors	Player's errors on-pitch.	0.18	0.05	0.03	0.02
foot_both	Binary measuring if player prefers both feet equally.		-0.05	0.10	0.07
foot_left	Binary measuring if player prefers left foot.	0.12	0.00	-0.03	0.07
foot_right	Binary measuring if player prefers right foot.	-0.12	0.02	-0.02	-0.10
fouled	number of players being fouled	0.00	0.08	0.23	0.28
fouls	number of fouls committed	-0.04	0.09	0.13	0.07
free_kick_goals_against_gk	free kick against the player as goalkeeper	0.00			
GA	player's team goals against	-0.44	-0.40	-0.36	-0.39
games	Games played.	0.35	0.23	0.29	0.30
games_gk	Games played as a goalkeeper.	0.35	-0.01		
games_starts	Games played - the player as a starter.	0.35	0.24	0.30	0.38
games_starts_gk	Games played as a goalkeeper - the player as a starter.	0.35	-0.01		
gca	Player's goal creating actions.	0.13	0.27	0.50	0.62
gca_dribbles	goal creating actions from dribbles		0.06	0.18	0.45
gca_fouled	goal creating actions - being fouled, leading to the goal		0.09	0.21	0.34
gca_og_for	goal creating actions leading to the opponent's own goal		0.09	0.03	0.11
gca_passes_dead	Goal creation from dead balls (including free kicks etc)	0.18	0.07	0.13	0.16
gca_passes_live	Goal creation actions from live-ball	0.05	0.26	0.53	0.59
gca_per90	Goal creation actions per 90 minutes.	0.08	0.17	0.36	0.39
gca_shots	goal creating actions with shots	-0.04	0.14	0.16	0.33
GDiff	Goals vs goals lost difference.	0.48	0.55	0.52	0.58
GF	Goals for a team.	0.40	0.53	0.51	0.57
goal_kick_length_avg	average length of goal kicks	-0.11	-0.01		
goal_kicks	Number of goal kicks.	0.13	-0.01		
goals	Player's goals.	-0.04	0.24	0.42	0.60
goals_against_gk	goals against as a goalkeeper	-0.03	-0.01		
goals_against_per90_gk	goals against per 90 minutes as a goalkeeper	-0.38	-0.01		
goals_assists_pens_per90	Goals and assists related to penalties per 90 minutes.	0.03	0.18	0.38	0.50
goals_assists_per90	Goals and assists per 90 minutes	0.02	0.17	0.38	0.51
goals_pens_per90	goals from penalties per 90 minutes	-0.04	0.13	0.31	0.43

Figure 4.1.3 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
Height	Player's height.	0.03	0.10	0.00	-0.08
interceptions	number of interceptions	-0.04	0.12	0.14	0.14
isBundesliga	Binary checking if player plays in Bundesliga.	-0.13	0.00	0.01	-0.05
isLaLiga	Binary checking if player plays in La Liga.	0.11	-0.01	0.02	-0.03
isLigue1	Binary checking if player plays in Ligue 1.	-0.15	-0.13	-0.12	-0.08
isPremierLeague	Binary checking if player plays in Premier League.	0.16	0.17	0.17	0.17
isSerieA	Binary checking if player plays in SerieA.	-0.03	-0.05	-0.08	-0.02
L	Player's team losses.	-0.45	-0.47	-0.45	-0.47
LgRk	player's team league ranking	-0.46	-0.47	-0.48	-0.48
losses_gk	team's losses while player was playing as a goalkeeper	-0.18	-0.01		
minutes	Minutes played.	0.35	0.24	0.30	0.39
minutes_90s	Minutes played divided by 90.	0.35	0.24	0.30	0.39
minutes_90s_gk	Minutes played as a goalkeeper divided by 90.	0.35	-0.01		
minutes_gk	Minutes played as a goalkeeper.	0.35	-0.01		
miscontrols	number of ball miscontrols	-0.02	0.06	0.30	0.32
MP	Teams matches played.	0.15	0.06	0.06	0.07
npxg	non penalty expected goals	-0.05	0.29	0.41	0.58
npxg_net	non-penalties expected goals - goals allowed	0.02	0.06	0.21	0.33
npxg_per_shot	non-penalty expected goals per a shot	-0.02	0.10	0.12	0.16
npxg_per90	non penalty expected goals per 90 minutes	-0.05	0.17	0.26	0.35
npxg_xa_per90	non penalty expected goals and assists per 90 minutes	-0.04	0.17	0.34	0.47
nutmegs	number of nutmegs - rushing the ball between opponent's legs	-0.04	0.07	0.19	0.36
offsides	number of offsides	-0.04	0.13	0.16	0.29
own_goals	number of own goals	-0.04	0.02	0.02	0.01
own_goals_against_gk	Own goals against the player as a goalkeeper.	0.06			
pass_targets	Number of pass targets.	0.28	0.51	0.51	0.51
passes	Number of passes.	0.24	0.46	0.47	0.48
passes_blocked	number of passes blocked	-0.04	0.08	0.31	0.39
passes_completed	Number of successful passes.	0.32	0.51	0.48	0.51
passes_completed_launched_gk	Passes completed launched by goalkeeper.	0.08	-0.01		
passes_completed_long	Long passes completed	0.18	0.39	0.36	0.36
passes_completed_medium	Medium length passes completed.	0.36	0.50	0.44	0.46
passes_completed_short	passes completed between 5 and 15 yards	0.27	0.42	0.52	0.54
passes_dead	Passes from dead balls (including free kicks etc)	0.09	0.02	0.19	0.24
passes_free_kicks	passes from free kicks	-0.02	0.19	0.18	0.19
passes_gk	Passes as a goalkeeper.	0.27	-0.01		
passes_ground	Player's ground passes.	0.36	0.54	0.49	0.54
passes_head	passes with head	-0.03	0.20	0.11	0.05
passes_high	passes above shoulder length	0.01	0.12	0.24	0.24
passes_intercepted	Opponent's passes intercepted	0.16	0.17	0.42	0.47

Figure 4.1.4 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
passes_launched_gk	passes launched as a goalkeeper	0.02	-0.01		
passes_left_foot	Passes with left foot.	0.16	0.22	0.17	0.35
passes_length_avg_gk	average length of passes for goalkeeper	-0.15	-0.01		
passes_live	Passes in a game - excluding corners, goal kicks etc.	0.28	0.49	0.46	0.49
passes_long	Long passes, above 40 yards.	0.07	0.31	0.32	0.29
passes_low	Passes that do not stick to the ground but stay under the shoulder length.	0.20	0.19	0.32	0.31
passes_medium	Medium length passes, between 15 to 30 yards.	0.36	0.48	0.43	0.44
passes_offsides	passes leading to offsides	-0.01	0.13	0.31	0.34
passes_oob	Out of bonds passes.	0.06	0.11	0.18	0.21
passes_other_body	Passes attempted using other body parts than head or feet.	0.33	0.27	0.18	0.11
passes_pct	% of passes completion.	0.25	0.34	0.22	0.26
passes_pct_launched_gk	% of passes longer than 40 yards that were launched not including goal kicks.	0.20	-0.01		
passes_pct_long	% of long passes completion.	0.27	0.32	0.18	0.14
passes_pct_medium	% of medium passes.	0.06	0.27	0.19	0.21
passes_pct_short	% of successful short passes	0.09	0.26	0.21	0.21
passes_pressure	Passes under pressure.	0.26	0.36	0.45	0.43
passes_progressive_distance	Total distance, in yards, that completed passes have traveled towards enemy's goal.	0.19	0.39	0.40	0.40
passes_received	Number of received passes.	0.29	0.52	0.51	0.53
passes_received_pct	% Number of received successful passes.	0.03	0.16	0.11	0.16
passes_right_foot	Passes with right foot.	0.10	0.33	0.41	0.33
passes_short	Number of short passes.	0.24	0.40	0.51	0.53
passes_switches	Passes that travel more than 40 yards of the width of the pitch	0.18	0.31	0.29	0.28
passes_throws_gk	Throws attempted as a goalkeeper.	0.33	-0.01		
passes_total_distance	Total distance of passes.	0.25	0.49	0.44	0.47
pct_goal_kicks_launched	% of goal kicks launched successfully	-0.17	-0.01		
pct_passes_launched_gk	% of launched passes as a goalkeeper	-0.21	-0.01		
pens_allowed	Penalties allowed	0.02	-0.01		
pens_att	penalties while attacking	-0.04	0.01	0.10	0.29
pens_att_gk	Penalty kicks attempted	0.02	-0.01		
pens_conceded	Penalties conceded	0.04	-0.02	0.00	-0.03
pens_made	penalties made	-0.04	0.01	0.10	0.28
pens_missed_gk	Penalties missed against goalkeeper.	0.04	-0.01		
pens_saved	penalties saved	-0.01			
pens_won	penalties won	-0.04	0.10	0.11	0.27
players_dribbled_past	number of players assisted past	-0.02	0.10	0.33	0.47
pressure_regain_pct	% of successful pressures regains	-0.02	0.17	0.13	0.13
pressure_regains	number of possession regains	-0.04	0.15	0.30	0.29
pressures	number of pressures against the opponent	-0.04	0.10	0.26	0.24
pressures_att_3rd	pressures in the attacking 3rd	-0.04	0.11	0.37	0.33
pressures_def_3rd	number of pressures in defensive 3rd	-0.03	0.06	0.14	0.06
pressures_mid_3rd	number of set pressures in middle 3rd	-0.04	0.13	0.23	0.18

Figure 4.1.5 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
psnpxg_per_shot_on_target_against	Post-shot non-penalty expected goals per shot on target against the player.	-0.11	-0.01		
psxg_gk	Post-shot expected goals measuring how likely is a player to save a shot.	0.04	-0.01		
psxg_net_gk	Post-shot expected goals minus goals allowed.	0.25	0.01		
psxg_net_per90_gk	Post-shot expected goals minus goals allowed per 90 minutes.	0.20	0.01		
Pts	Player's team points.	0.54	0.55	0.53	0.57
save_pct	% of successful saves	0.20	-0.01		
saves	Number of saves.	0.15	-0.01		
sca	Shot creation actions	0.01	0.20	0.45	0.54
sca_dribbles	shot creating actions with dribbles	-0.04	0.08	0.28	0.50
sca_fouled	shot creating actions - fouled during the process	-0.04	0.05	0.25	0.41
sca_passes_dead	Shot creation actions from dead balls (including free kicks etc)	0.07	0.03	0.16	0.13
sca_passes_live	Shot creation actions from live balls	0.01	0.23	0.49	0.54
sca_per90	shot creating actions per 90 minutes	-0.05	0.09	0.33	0.35
sca_shots	shot creating actions with shots	-0.05	0.18	0.27	0.44
shots_free_kicks	number of shots from free kicks	-0.04	0.07	0.16	0.35
shots_on_target	Player's number of shots on target.	-0.04	0.27	0.43	0.60
shots_on_target_against	Shots on target against the player.	0.10	-0.01		
shots_on_target_pct	% of shots per target	0.04	0.06	0.12	0.18
shots_on_target_per90	shots on target per 90 minutes	-0.03	0.13	0.27	0.40
shots_total	number of total shots	-0.04	0.26	0.41	0.54
shots_total_per90	total shots per 90 minutes	-0.04	0.14	0.25	0.32
tackles	number of tackles	-0.03	0.12	0.21	0.12
tackles_att_3rd	tackles in the attacking 3rd	-0.04	0.10	0.31	0.23
tackles_def_3rd	tackles in a defensive area	-0.02	0.09	0.13	0.02
tackles_mid_3rd	number of tackles in a middle 3rd	-0.04	0.13	0.20	0.12
tackles_won	number of tackles won	-0.03	0.14	0.23	0.14
through_balls	The balls sent between the defenders into open space for an attacker.	0.21	0.29	0.47	0.47
throw_ins	Player's throw-ins.	0.04	-0.03	0.04	0.09
touches	Number of touches.	0.24	0.43	0.44	0.48
touches_att_3rd	touches in the attacking 3rd	-0.04	0.17	0.49	0.56
touches_att_pen_area	Touches in competitor's team penalty area.	-0.04	0.31	0.40	0.59
touches_def_3rd	Touches in defensive area.	0.25	0.32	0.22	0.14
touches_def_pen_area	Touches in own penalty area.	0.27	0.24	0.10	0.04
touches_live_ball	Number of live-ball touches.	0.27	0.46	0.45	0.48
touches_mid_3rd	number of touches in middle 3rd	-0.04	0.48	0.42	0.41
W	Player's team wins.	0.54	0.55	0.53	0.58
WinCL	Binary explaining if the player's team won the Champions League following season.	0.19	0.29	0.15	0.20
wins_gk	Wins of player's team when he played as a goalkeeper.	0.63	-0.01		
xa	expected assists	-0.01	0.19	0.43	0.55
xa_net	Expected assists minus assists allowed.	0.19	0.11	0.16	0.13

Figure 4.1.6 A table presenting the correlation between given variables and transfermarkt.de values.

Variables	Descriptions	GK	DEF	MID	FWD
xg	Player's team expected goals.	-0.04	0.26	0.38	0.57
xg_net	xG - goals allowed	0.01	0.06	0.21	0.34
xg_per90	expected goals per 90 minutes	-0.05	0.16	0.26	0.38
xg_xa_per90	Expected goals and expected assists per 90 minutes.	-0.04	0.17	0.33	0.48
xGA	A measure of expected goals against the player's team.	-0.38	-0.40	-0.35	-0.38
xGDiff	xG vs xG against difference.	0.44	0.54	0.50	0.56
xG	Player's team expected goals.	0.37	0.51	0.49	0.54
goals_per90	goals per 90s	-0.04	0.13	0.31	0.45
passes_into_final_third	passes into attacking third of football pitch	-0.05	0.46	0.46	0.44
passes_into_penalty_area	number of passes into opponent's penalty area	-0.02	0.14	0.50	0.54
goals_per_shot	number of goals per a shot	-0.04	0.08	0.17	0.24
goals_per_shot_on_target	number of goals per shot on target	-0.04	0.12	0.16	0.18
progressive_passes	number of passes towards enemy's goal	-0.04	0.35	0.46	0.50
xa_per90	expected assists per 90 minutes	-0.02	0.11	0.30	0.37