ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis [Data Science & Marketing Analytics]

Assessing market values of football players using performance data

Name student: Mark Janssen

Student ID number: 545095mj

Supervisor: Michel van de Velden

Second assessor: Michiel van Crombrugge

Date final version: 01-11-2023

Contents

Assessing market values of football players using performance data

1. Introduction

Football players are the most important assets to football clubs. They are the life and soul of the club, generating money streams by means of competition prize money, ticket sales and transfer values. However unlike normal business assets, market values of football players are hard to determine. Although the public generally has an idea how much a player is worth, it can be difficult to exactly determine a player's market value. During transfer windows with limited time, negotiations can often drag on, sometimes with very insufficient results. Hence, if this process could be simplified and a more objective approach to market value could be taken, this could help ease negotiations. There is currently more and more data available in the footballing world (Asif et al., 2016). Data analysts and scouts in football make use of match statistics and performances by players to come to decisions. Therefore, there is already quite some in-house knowledge available for football clubs. Perhaps this data can be used to assess a player's market value and aid clubs during the negotiation phase.

Research on the topic has recognized that performance statistics can be useful to an extent when assessing market values of a player. Müller, Simons & Weinmann (2017) have found that taking goals and assists into account can help determining the market value of football players. They found that these attacking metrics are of positive influence on the market value of a player. Furthermore, they concluded that defensive metrics such as number of tackles, yellow cards and red card negatively influence market value. This in accordance with the finding that overall, attacking players are more highly valued than defensive players. (Felipe et al., 2020).

Furthermore, research has found that external factors are also of influence. Popularity of a player may be directly linked to their market value in a positive manner. (Müller, Simons & Weinmann, 2017) This may explain why some players are overpaid for during negotiations. In addition to this, a player's nationality (Majewski, 2017), a player's age (Müller, Simons & Weinmann, 2017) and agent fees (Monteiro et al., 2022) all play a part in the eventual market value of a player. Hence,

there are both performance related and non-performance related factors of influence when it comes to player market value.

Whilst the role of most non-performance related factors has been studied extensively, the performance measures used in current research to asses market values if often lackluster. Simple metrics such as goals, assists and number of tackles are taken into account. However, current clubs often use much more intricate data metrics, such as progressive passes per 90 minutes, in order to assess a player's quality. Given that these metrics are used to assess a player's quality, it is interesting to research whether adding these more extensive metrics proves to be useful when assessing market values of players. In addition to this, current research uses relatively simple methods such as multiple linear regression to asses market value. Whilst this is not inherently wrong, the results of current research have not been sufficient to the extent that football clubs could incorporate the models in the transfer negotiations. Adding the more extensive performance metrics in addition to a stronger machine learning model can therefore be a solution. Therefore, this paper will focus on the following research question:

To what extent can football player's market values be assessed using extensive player performance data?

Next to filling in the aforementioned gaps in current literature, the research is socially relevant as it can help technical directors of football clubs with decision making concerning buying and selling players, using data that is easily accessible to them. By assessing the value for the club in an objective manner, it should provide an objective measure of market value of the player for the club during transfer negotiations. This research seeks to find whether current market values can be assessed using extensive performance data and an intricate machine learning model. This should enhance decision making and prevent the occurrence of significantly overrated transfer fees. From a marketing perspective, this research can aid both player marketing agencies and selling clubs to push the value of players or perform damage limitation regarding their market value depending on the expected change in the respective player's value as assessed by the model.

This paper will first delve deeper into the current literature on the topic. Here, both the performance and non-performance related factors that play a part in player market valuation are assessed. It is then addressed what data is used in the analysis and what the reasons were for selecting the specific dataset. In addition to this, the paper will explain the reasoning for the use of multiple linear regression and gradient boosting in the analysis, as well as the inner workings of them. From these data and methods, the results are derived and interpreted to aid in answering the main research question. The paper will round up with both the conclusions and limitations derived from the analysis and the process.

2. Literature Review

2.1 Introduction to football transfers and market valuation

Football player market valuation is one of the most important factors when it comes to negotiations on transfer fees between clubs. A definition of player market values is given by Herm, Callsen-Bracker and Kreis (2012). They describe it “as an estimate of the amount of money a club would be willing to pay in order to make this athlete sign a contract, independent of an actual transaction.” The key word here is “estimate”, as there is no concrete or objective measure by which footballers can be valued. Football players can be considered assets to a football clubs, despite them being humans. Unlike with normal physical or financial assets, there is no intuitive or mathematical model to assess the valuation of an athlete. Footballers provide both value on the pitch as well as being able to be bought and sold for profit, to some extent similar to stocks. Therefore it is difficult to objectively state the exact value of a player. Looking back at the definition of Callsen-Bracker and Kreis (2012), the amount of money a club would be willing to pay can be perceived by the transfer fee the buying club pays for a player. Transfer fees are what both the buying and selling club consider to be an accurate valuation of the player. Although one can argue that there are occasions where a club severely overpays for players, in general the transfer fee agreed upon is a relatively objective measure. A disadvantage of this is that only players who have made transfers can be considered when analyzing this problem. Furthermore, football is often an emotional business. This can cause clubs to act irrationally during transfer negotiations. Therefore, the transfer fee is still subjective to an extent as there are occasions where clubs have significantly

over- or underpaid for a player. However, given the large amount of transfers in the football world, the group of players who never transfer away from their club is significantly smaller than the group which does make a transfer. In addition to this, a sufficiently large dataset should filter the subjective side of transfer fees as the majority of transfer fees paid are well balanced and appropriate to the market. Therefore, the transfer fee still provides enough information when it comes to approximating overall market values. From here on, the 'transfer fee' is considered the objective market value that the analysis tries to predict in this paper. Therefore, the transfer fee is the measure by which the research question is analyzed and answered. However to answer this properly, it is interesting to know what aspects are of influence when it comes to market value of football players.

2.2 Factors of influence on market value: player performance related

Current research has found that there are many different factors involved in determining a football player's value. For example, according to Müller, Simons and Weinmann (2017), goals and assists are one of the main factors in high market values. This makes sense, as goals and assists are the performance measure most directly related to the success of a player. The research has found that on average, an additional goal increases the value of a player by 2.4%. Furthermore, an additional assist increases a player's value by 1.5% on average. Similar results were found by Majewski (2017), who also came to the conclusion that goals and assists were one of the most important factors in market value determination. This would explain why, on average, players in attacking positions are more expensive than players in defensive positions (Felipe et al., 2020).

There are more performance metrics which influence a player's valuation. On the positive side, number of minutes played, number of passes played, number of dribbles completed and aerial duels performed all influence valuation positively. On the opposite side, number of tackles performed and yellow cards conceded influence valuation negatively. (Müller, Simons and Weinmann, 2017). There could be multiple explanations for these factors influencing the market value the way they do. The variables which influence the valuation positively can all be traced to being a dominant player in a successful team. Having many minutes shows that a player is very important for their team. A high number of dribbles and passes indicate that the team and the

player are in a lot of possession, which is often the case with successful teams and players. A high number of tackles and yellow cards on the other hand can indicate a lack of possession, indicating a player plays in a team of a lower level and thus receives a lower market valuation. The other explanation may have to do with the fact that the negative factors mostly coincide with defensive players, whereas the positive factors are often related to offensive players. As defensive players are significantly lower valued than offensive players (Felipe et al., 2020), these factors could implicitly represent this value gap in the analysis performed by Müller, Simons and Weinmann (2017).

2.3 Factors of influence on market value: supplementary football related

There are also other directly football related factors outside of in-match performance statistics which are of influence. For example, it was found that the log of previous player value (Müller, Simons & Weinmann, 2017) has a positive influence on market value. Furthermore, team value of the previous season (Majewski, 2017) also positively influences player market value. Therefore, these factors also need to be taken into account. Lastly, Majewski (2017) has found that the position on the FIFA World Ranking, a ranking system for all national football teams, is also positively correlated with player market value. Players from countries performing well on the world stage are generally more highly valued. Sometimes this could have to do with the level of the domestic league influencing the national team's level, for example England and the Premier League. However, in instances such as the Brazilian league this is not the case. Therefore, it could be true that a player could genuinely be higher valued depending on their nationality. Another explanation is that countries who are highly ranked on the FIFA World Raking generally have a good youth structure. This could implicate that a higher level of national team performance is reflected by better basic skill training from the youth levels onward.

2.4 Factors of influence on market value: non-football related

There are also factors which are less directly related to on-pitch performance that play a part in a player's market value. First and foremost, age squared negatively influences player market valuation (Müller, Simons & Weinmann, 2017). This means that both relatively young and relatively old players tend to have a lower market value than players in their prime. Younger

players can often still improve more and play for a longer period of time than older players. However, they are also quite volatile. They can become world class players, but there is also a possibility of them completely flopping. Relatively old players tend to have very little future value. They are more likely to stop soon, therefore this is reflected in their current market value. With players in their prime in terms of age, you get exactly the value you expect from them. As a result, their market values are relatively higher than that of younger and older players.

Another of these factors not directly linked to football is media popularity. Müller, Simons and Weinmann (2017) took popularity measures into account when constructing player market value. In this research, the popularity of a player was measured by the number of hits on Google Trends, number of Wikipedia page views, number of Reddit blog posts and number of Youtube views. They found that a higher player popularity corresponded with a higher player valuation, with the latter three variables all having a significantly positive influence on a player's market value. The best performing players are often the most popular players as well. Therefore, it would make sense that these players are also more highly valued. There are also other external factors which might influence player market valuation. Monteiro et al. (2022) argues that both agent fees and player current salary could influence the market value of a player. Agent fees are more directly influencing market values, as agents often increase the transfer fee paid in order to earn commission from them. Although new FIFA approved rules limit the power that agents hold in the future (Hall, 2023), they are still very influential when it comes to player market value. Furthermore, a player's salary could also result in a higher market value. A satisfied player on a high salary may not want to leave their current club so easily. Therefore, it gives their current club a better competitive position when it comes to negotiations with other clubs. This is then reflected in a higher market value. (Monteiro et al., 2022)

2.5 Current research complications

The biggest issue that current research faces is that the models using only football related metrics often lack in accuracy. Majewski (2017) performed a very simple linear regression. Although this simple model was reasonably accurate, with an adjusted R-squared of 56% for their best model, it is not accurate enough for football clubs to base their decisions partially on the insights gained

from this model. Therefore, very strong conclusions on the factors influencing market values cannot be derived from this analysis. Müller, Simons and Weinmann (2017) used a slightly more advanced model including a large number of both football related and non-football related variables. In their analysis, they use a multiple regression model where the log of the market value is taken as the dependent variable. They then create multiple models, each time adding in additional variables. Their paper stated that a machine learning model should be better capable of determining a player's market value. They found that their model was an improvement upon crowd-based models, especially when it came to players of low, medium and slightly higher value. Their results showed that their model achieved an RMSE of 5793,474. They, however, did not solely use season match data. They also added event data, such as news and number of Reddit posts to configure a level of popularity for the player. These popularity measures are often not easily available for clubs at the player-specific level. Therefore, clubs would not have access to this when constructing a model on player transfer values. Furthermore, only very simple performance data metrics are used in current research. Metrics such as goals, assists, number of minutes played and yellow cards are very basic. Contemporary football analysis uses a wide variety of far more in-depth performance metrics. These metrics are available to most football clubs, especially those performing at the highest national level. Yet, current literature does not take into account these advanced performance metrics when assessing market value. Therefore, it still remains difficult to be certain to what extent performance data can predict or prescribe what a player's market valuation should be, as well as what factors are most influential. Although combining extensive performance data with a more advanced method may not automatically guarantee a better result for assessing what factors influence market value, current results are not strong enough for clubs to directly apply them in their organization. Therefore, creating a method using these in-depth performance data which is more suitable and valuable for football organizations is most beneficial. Furthermore, it is interesting to find out in what way certain in-depth performance measures affect market value. Additionally, there may be a difference in importance and effect depending on the position of the player on the field.

2.6 Hypotheses

From the previous research, multiple hypotheses can be derived. The main hypothesis is that using extensive performance data is a viable method to assess the market value a player represents for their club. In order to test this hypothesis, there are multiple sub-hypotheses and expectations to test. Considering the market value, we expect that goals, assists, metrics concerning attacking play (i.e. successful dribbles) have a positive influence on market value. On the other hand, we expect age squared, metrics concerning defensive play (i.e. number of tackles) and yellow cards to have a negative effect on market value. Lastly, we expect that the performance metrics' influence on market value and the model's prediction success may differ depending on player position.

3. Data

3.1 Data sources and rationale

In order to answer the research question, multiple data sources are needed. First of all, a data source containing market values is needed. Because market values are subjective, they are approximated by taking reported transfer fees. Transfer fees are what both the buying and selling club consider to be an accurate valuation of the player. Although one can argue that there are occasions where a club severely overpays for players, in general the transfer fee agreed upon is a relatively objective measure. Hence, the reported transfer fees paid for players are used to approximate market values. A disadvantage of this is that only players who have made transfers in recent year are considered in the analysis. Furthermore, the transfer fee is still subjective to an extent as there are occasions where clubs have significantly over- or underpaid for a player. However, given the large amount of transfers in the football world, the group of players who never transfer away from their club is significantly smaller than the group which does make a transfer. In addition to this, a sufficiently large dataset should filter the subjective side of transfer fees as the majority of transfer fees paid are well balanced and appropriate to the market. Therefore, the transfer fee still provides enough information when it comes to approximating overall market values. Therefore, in this paper the reported transfer fee is considered the objective market value

that the analysis tries to predict in this paper. Therefore, the transfer fee is the measure by which the research question is analyzed and answered.

Multiple data sources were considered when collecting data on player transfers. In the end, Transfermarkt is chosen as the main source of the transfer data. The main reason for this is that Transfermarkt not only provides a large database of transfers including the transfer fees, but also stores a lot of additional valuable information regarding the transfers. First and foremost, Transfermarkt provides information on both the league and club players transfer to and from. Both the league and club level is of importance when assessing the qualities and therefore the valuation of a player. Hence, having a data source which includes these pieces of information is a big advantage. Another reason why Transfermarkt is a very useful source of data, is because it has another potentially important factor to consider: crowd-based market value assessments. Transfermarkt is a website containing data on thousands of players across the world. Each of these players is assigned a market valuation. This market valuation is based on opinions of the crowd, closely moderated by the website. As such, there is an enormous external input from the crowd assessing and assigning market values for each player. Because this is closely moderated, the outcomes are often very reasonable and rational. Although this market valuation still has flaws, this valuation has been found to be a successful proxy of player value in general (Peeters, 2018). Due to this, multiple football clubs such as Olympique Lyon, Schalke 04 and FC Porto have incorporated this market value data as well (Keppel & Claessons, 2020). Additionally, this crowd-based valuation also serves as an indication of popularity. If a player is considered to be very popular or very well-known, the crowd may value two players who are similar in terms of footballing ability completely different. Hence, including the crowd-based market valuation helps finding the true market value and will thus be included in the model.

The other data source needed is a data source containing performance data for the players. In addition to market values and transfer fees, Transfermarkt itself has performance data. However, this performance data only consists of the most basic forms of performance data: minutes played, goals, assists, yellow cards and red cards. Although this information is useful, it lacks the depth that current football clubs look for and need in order to make a proper assessment of a player's

quality and therefore market value. Hence, a more advanced set of performance data is needed. The leaders in the field on performance data are FBref and OPTA, data collecting agencies often used by football clubs in their current transition to more data-oriented analysis. These agencies are very thorough and collect substantial amount of data on leagues, players and clubs. However, the issue with these websites is that they require sumptuous payment to access this data. As an alternative, WhoScored will be used. WhoScored is a website specialized in data collection and data analysis. They have a large database on performance data from multiple leagues, clubs and seasons. Their performance data do not only include the more simple metrics like goals and assists, but also very detailed performance metrics. Examples of these are total successful passes per 90 minutes, successful dribbles per 90 minutes and key passes per 90 minutes. Therefore, this data allows for an in-depth analysis in terms of which football metrics are of importance when it comes to player quality and therefore player value. Furthermore, WhoScored also provides the league and club the player played in during the season that the data was collected. Therefore, it not only provides the in-depth data, but also at what footballing level these performances are. Hence, this adds additional valuable variables to assess player market valuation by.

3.2 Origins and contents of datasets

For the analysis, data from two different sources is used. First of all, a dataset containing players and transfer fees is used. This dataset consists of the top 250 transfers per season in terms of market value during the period of 2010-2018. This means that there are a total of eight seasons, which results in 2000 observations. This data was retrieved from Kaggle, which in turn was retrieved from the website Transfermarkt (Slehkyi, 2019). This data consists of the player's name, their position, age, the team and league they moved from, the team and league they moved to, the market value of the player at the time of the transfer according to Transfermarkt and the actual reported transfer fee. The second dataset was also retrieved from Kaggle, which in turn was retrieved from the Whoscored website (Dejan, 2020). This dataset consists of the player's name, along with additional information such as the club the player played at during a season. Furthermore, the dataset includes a large number of match statistics collected over a season. These statistics are corrected for the amount of minutes played, which is represented as the

average statistics per 90 minutes. These Kaggle datasets were used, because they contained nearly all necessary variables to conduct thorough analysis. The most pressing downside of the Transfermarkt dataset was that it consists only of the top 250 transfers per season. This dataset still chosen due to the fact that there were a very limited number of large Transfermarkt datasets available. Attempts of Webscraping data from Transfermarkt was unsuccessful, due to computational limitations of the equipment used for the analysis. The use of this dataset has two main downsides. First of all, there are cases in football where irrational transfer fees were paid. These are transfer fees where the player's transfer fee does not accurately reflect their quality. Transfers that suffer from this bias are often at the higher end of the transfer fee spectrum. Hence, the top 250 transfers per season also include the irrational transfers. Secondly, the analysis will not take into account the transfers at the lower end of the transfer fee spectrum. On one hand, one could argue that it is the clubs at the lower end of the footballing pyramid which benefit most from the positive sides of data analysis as they have relatively less room for financial error than large football clubs. On the other hand, these smaller clubs often do not have the resources to invest in these data focused transfer policies, opting for a more traditional hands on scouting network. Nevertheless, this dataset provides enough valuable information and is still viable to use, though these limitations must be considered when deriving any conclusions.

3.3 Data Cleaning

The two datasets were merged by the player's name. Trivial variables such as internal ID's as well as duplicate columns created due to the merger were deleted. The club the players transferred to and from were also deleted. Although these are variables containing useful information, the dataset is too limited to include these variables as most clubs only appeared once or twice in the dataset.

In addition to duplicate rows, there were also some duplicate observations in the dataset. Some of these observations were exact duplicates. Therefore, these were immediately removed. Other observations were semi-duplicates. These observations were from competitions which had both group stages and knock-out stages. The same error occurred with competitions which end with play-offs for promotion/relegation. The issue arose as there was one observation which only

included the group stage or normal competition, whereas another observation included the full details from the knock-out or playoff stages as well. In these cases, the latter observation was kept as this provides the most complete picture of the player's overall performance in the competition. Lastly, there were some observations where players participated in multiple competitions during each season. For example, player A could have statistics for both the Premier League and Europa League in the same season. However, this results in the same player having the same market value and the same transfer fee, despite having completely different numbers for each observation. Hence, for each season the observation with the most amount of minutes was kept. Although this resulted in some loss of data, the data point that was kept consisted of more minutes played than the deleted observations and thus had less room for noise.

After all the cleaning, the dataset consists of 1613 observations and 39 variables. For the analysis, the full dataset was split into two sets: a training and test set. Both models are trained and tested using the same training and test set containing 1292 and 321 observations respectively. This means that the training and test split is 80/20.

3.4 Initial simple data analysis

Before the full analysis can be performed, the initial data must be analyzed to see whether the data set is balanced and whether there are any peculiarities. The descriptive statistics of a number of variables can be found in Table 1 below.

Table 1.

Descriptive statistics

Variable Name	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
Transfer Fee	1,613	15,746,044	14,997,702	3,000,000	135,000,000
Market Value (Transfermarkt)	1,613	13,067,070	11,615,073	100,000	120,000,000
Age	1,613	24.89	2.94	17	33
Height	1,613	182	6.49	162	203
Weight	1,613	76.61	7.17	58	101
Appearances	1,613	20.01	12.86	1	48
Minutes Played	1,613	1539	1082.23	5	4320
Goals	1,613	3.76	5.24	0	36
Assists	1,613	2.13	2.76	0	20

Note. Table 1 shows the descriptive statistics of the dataset used in this paper's analysis. The variables depicted are all continuous variables.

As can be seen in Table 1, the average transfer fee in the dataset is close to 16 million euro. However, the maximum value is 45 times larger than the minimum value. The mean value is much closer to the minimum value than the maximum value, meaning that the data is skewed to the right. The same pattern can be seen for the market values from Transfermarkt. Here, the maximum value is also much larger than the minimum value, with the mean value being much closer to the minimum value. This skewedness becomes even more evident when looking at Figure 1 and Figure 2 below.

Figure 1

Distribution of transfer fees

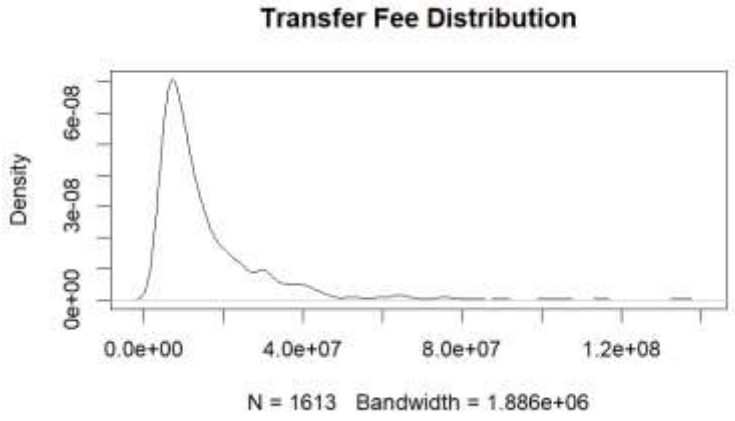
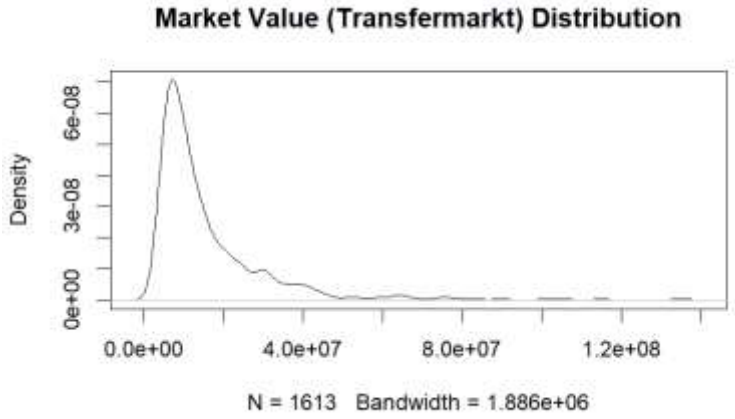


Figure 2

Distribution of market values



Here, it becomes clear that both the transfer fee and the market values from Transfermarkt must undergo a transformation before being used in the analysis as they are indeed strongly skewed to the right. In terms of the age, height and weight distribution, the dataset is relatively balanced. The number of appearances, minutes played, goals and assists are all slightly skewed to the right. However, this skewedness is not so large to the extent that a transformation of the data is needed.

Therefore, these variables remain unchanged. The other performance metrics are all relatively well distributed too.

3.5 Data transformations

Based on the literature, we expect age to have a parabolic relationship with player market value. As such, the age squared variable is added as this would provide a more accurate representation of the average market valuation curve a player experiences. Therefore, the multiple regression and gradient boosting methods incorporate both age and age squared to account for this non-linear relationship. Furthermore, both the dependent variable transfer fee and the independent variable market value from Transfermarkt are log-transformed. This is done in order to conform to the normality assumption for the multiple regression method, as well as decreasing the influence of outliers to which the gradient boosting method is prone. In order to enhance the performance of both models, the log-transformation is a necessity.

4. Methodology

4.1 General notes on the dependent and independent variables

The dependent variable in both models is the log of the transfer fee paid for each player. The independent variables are mostly performance metrics. These range from simple metrics such as goals and assists per season, to more advanced metric such as interceptions, aerial duels and successful dribbles per 90 minutes. Furthermore, there are other football related variables which correspond to the leagues that player's transferred to and from, as well as the position of each player. Lastly, some independent variables contain more personal information on the player, including the age, height and weight at the time of the transfer. A full index of the variables used in the analysis can be found in Table 6 under Appendix A.

4.2 Rationale for models used in the analysis

In order to construct the model to assess market value based on performance data, two main methods are used. The main focus is a gradient boosting model. This is then compared to a multiple linear regression model as a baseline method. The gradient boosting model was chosen due to the preferable ratio between general effectiveness of the model and the model's intricacy.

Boosting models generally tend to give a decent prediction accuracy. It allows for tuning certain parameters to increase the accuracy of the prediction model. At the same time, the gradient boosting model is not as intricate as even more advanced models such as neural networks. Although these more advanced models might increase the general prediction accuracy to an extent, the increased difficulty level in terms of model construction comes with two major disadvantages. First of all, the construction of a boosting model is a lot less time and resource consuming than the more advanced models. With the quickly adapting football transfer market, the chosen machine learning model will most likely need to be adapted regularly. Having to adapt a more advanced model to function to a sufficient extent multiple times would cost football clubs too much time and resources. Adapting the boosting model on the other hand would still provide the club with a sufficient prediction accuracy whilst also being relatively adaptable. The second reason to choose the gradient boosting model over a more advanced model is because the aim of football clubs is not to construct a machine learning model that can be blindly followed. The aim is rather to have a general idea what a player should be worth before going into transfer negotiations. This is done to prevent large overpayments to the extent that a player's transfer fee significantly larger than their actual market value. The gradient boosting model aids this goal sufficiently. The additional time and resources spent on more advanced models would hence only increase the barrier for football clubs to include data guidance during their transfer negotiations and scouting.

The multiple linear regression was chosen as a baseline method for multiple reasons. First of all, it is one of the most simple models to use for any data analysis. Therefore, the entry level for the football clubs to start including this type of data analysis in the process is very low. Training the members from the data, scouting and transfer departments to perform and interpret such analyses would not cost too much time and resources. Furthermore, the multiple linear regression model adds an additional value: interpretation. This model can help football clubs find what variables makes for more valuable players. It can help the directors of football and the scouting team to assess what qualities they need to look for in order to find valuable players for the right price. Although this type of model does not directly show cause and effect, it does show

the correlation between the player's market value and the dependent variables. Hence, the multiple linear regression model can not only be used as a baseline method for prediction accuracy, but also as a more deep dive for football clubs to assess which footballing qualities are important.

4.3 **Comparison metric**

The accuracy comparison between the models is done based on the Root Mean Squared Log Error (RMSLE) of both models. This RMSLE is constructed as follows:

[1]

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2}$$

Where

n = The total number of observations

i = The i'th observation in the calculation

$\hat{y}_i$= The predicted value of the dependent variable for the i'th observation

y_i= The actual value of the dependent variable for the i'th observation

We assess based on RMSLE and not RMSE due to the log-transformation of the dependent variable. RMSLE measures the ratio between predicted and actual values in a logarithmic scale, whereas the RSME measures the absolute difference. The benefit of this method is that it decreases the issue of the large variance within the dependent variable which would occur without the log transformation, as RMSLE only assesses relative errors instead of absolute errors. Because of this, the outliers in the data set regarding the transfer fee no longer pose a problem. The drawback of this assessment method is that there is no direct explanation as to what the absolute value of the error is. Therefore, we cannot directly say how far away a certain prediction is from the actual value in absolute terms. However, the benefits of the RMSLE method outweigh the disadvantages as the scale difference is the most pressing issue in the dataset. RMSLE takes a

value between 0 and 1, where being closer to 0 indicates a more accurate prediction by the model.

4.4 Multiple Linear regression explained

The multiple linear regression analysis is a relatively simple form of analysis, which this research uses as the baseline method. The model is as follows:

[2]

$$Log(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

Here we see that the log of the dependent variable Y, transfer fees in this case, is constructed by a constant β_0, the variables X_i with their respective coefficients β_i ranging from i=1 to n and an overall error term ε. The variables X_i are all as they are presented in the dataset, with exception of age squared where it follows X_{age}^2 and the log of the market value from Transfermarkt where it follows $\log(X_{\text{Market Value TM}})$. This model entails that a change in an independent variable X_i causes a change in the dependent variable log(Y) with respect to their specific coefficient β_i. For the market value from Transfermarkt, a change in $\log(X_{\text{Market Value TM}})$ changes log(Y) with respect to $\beta_{\text{Market Value TM}}$.

For multiple linear regression analysis, there are multiple assumptions that have to hold. First of all, the relationship between the dependent and independent variable needs to be linear. In order to check for this, scatterplots of the variables were assessed to find any non-linear relationships. For most variables, the linearity assumption held. Age was the only variable which did not have a linear relationship with transfer fees. This is understandable, as both young players and old players are relatively lowly valued. For young players, this is due to the uncertainty that comes with lack of experience. With older players, this is due to the limited amount of years of peak physical performance left. Players in their prime age, however, are generally worth the most. To account for this, the squared age of the players was added in the analysis. The second assumption that has to hold is there should be no multicollinearity. This means that the independent variables should all be independent of each other and must not be correlated. This was checked using a correlation plot between the independent variables. For most variables the no multicollinearity assumption held. The only variable that was problematic, was the 'Minutes Played' variable.

There are multiple metrics in the data set that are corrected for game time already, i.e. tackles per 90 minutes. Because this is directly related to the number of minutes played, there were some cases of high correlation. However, the 'Minutes Played' variable was still kept in, as the information loss during the prediction was apparent. Therefore, the variable was kept in despite violating the independence assumption.

The third assumption of multiple linear regression is the assumption of normality. This means that the data should be normally distributed. This was checked using distribution plots of each of the independent variables. For the transfer fee and to a lesser extent the market values assigned by Transfermarkt, this did not hold. However, after performing the log transformation for both variables, the distribution became mostly normally distributed again. The other continuous variables did follow a normal distribution. Lastly, the independence assumption has to hold. This means that the observations should all be independent of each other. Nearly all observations were indeed independent of each other. There were a select few observations where there was a slight form of interdependence. This was due to the fact that there were instances where players from the same team within the same season made a transfer. One could argue that the players' performance data are influenced by each other, as they shared the same pitch all season. However, this effect should be limited. Although the players affect each other metrics to an extent, there are 9 more outfield players in the team. Hence, the direct impact one player has on another player's metrics should be negligible over the course of a season. Therefore, these observations were not excluded.

4.5 Gradient boosting explained

For the main analysis, gradient boosting is used. Gradient boosting was first proposed by Friedman (2001). Gradient boosting is a relatively strong prediction method, whilst at the same time being very intuitive. Gradient boosting is an ensemble method. Ensemble methods combine multiple weak learners, in this case small decision trees, to create a strong prediction model. It starts by fitting an initial model to the data, after which it adds a second model that mostly focuses on the mistakes that the first model makes. Combining these two models should then yield a

better result than the first model alone. This is then repeated multiple times in order to achieve the best model possible. With gradient boosting, the weak learners are combined sequentially. This means that each new tree builds directly on their predecessors, taking into account mistakes from the earlier models and learning to create more accurate predictions in the process.

How is the best model determined using gradient boosting? Gradient boosting works by taking a look at a loss function. The idea behind this is that the next model that you add by combining the previous models, minimizes the overall prediction error. The idea is to set target outcomes for the next model based on the mistakes that the previous model made. Therefore, it looks at the residuals between the old model and the new one. When a change in prediction causes a large drop in error, we make a big change for the target outcome. When a change in prediction does not cause a large drop in error, we change the target outcome very little. How strongly the target outcome adjusts is also determined by the shrinkage parameter. This is a weight with a value smaller than one that states how strongly we adjust the target outcome. For example, say that the target outcome at a certain point is 10. The next weak learner tree predicts an outcome of 20. If the shrinkage parameter is 0.1, the new target outcome is adjusted by $(20 - 10) * 0.1 = 1$. If the shrinkage parameter is 0.01, this would become $20 - 10 * 0.01 = 0.1$. If we continue shifting and shaping the target outcome until we have found the lowest mean squared error loss, we have optimized the model. This is the basic idea behind the model.

The first mathematical step of the model is as follows:

[3]

$$F_0(x) = argmin_\theta \sum_{i=1}^{n} L(y_i, \theta)$$

Where $L = (y_i - \theta)^2$

Formula [3] describes setting the first target outcome for the model to reach. Here, $F_0(x)$ describes the initial target outcome we want to reach. L describes the loss function. In the analysis, we look at squared loss, therefore the loss function is squared. Y_i is the actual value of the dependent variable. The $argmin_\theta$ indicates that we wish to find the value of θ for which the loss function is

minimized. Once this value is decided upon, we find the initial target outcome $F_0(x)$. Having found this target outcome, the iterative tree process can start. This process is repeated from m to M times. The m indicates the iteration of the tree that is present. The first tree created by the model is m=1, the second tree m=2 et cetera. M indicates the total number of iterations that are performed. The first step of this process we want to calculate the residuals using the target outcome we have found. This is done by

[4]

$$r_{im} = y_i - F_{m-1}$$

Y_i still represents the actual value of the dependent variable. F_{m-1} represents the target outcome from the last model in the iterative process. For the initial iteration, we use the F_0. With this formula, we can thus calculate what the residuals are for each observation using the target outcome from the previous iteration. After calculating the residuals, we want to create the simple regression tree with features x on the residuals r. The features x are the independent variables in the analysis. The regression tree has a collection of terminal nodes of R_{jm}, where m again represents the iteration and j represents the number of leaves there are until the tree reaches the terminal node. After the tree is created, we want to compute the following:

[5]

$$\theta_{jm} = argmin_{\theta} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \theta)$$

Here the formula computes given the regression tree with what θ_{jm} the target outcome of the previous iteration has to be adjusted in order to minimize the loss function. If we complete this $argmin_{\theta}$, we find that the optimal θ that minimizes the loss function is the average of all residuals in each terminal node. Hence, we adjust the target outcome of the new iteration with the average of all residuals in each terminal node coming from the regression tree. The last step is update the model and target outcome for the next iteration, also taking into account the chosen shrinkage parameter. This is done by:

[6]

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J} \theta_{jm}(x \in R_{jm})$$

With this formula, the target outcome for the next iteration can be updated. $F_{m-1}(x)$ represents the target outcome from the previous iteration. The j represents the leaf and J represents the total number of leaves in the tree. θ_{jm} ($x \in R_{jm}$) indicates that we take into account the calculated θ_{jm} only when all features x fall within one of the terminal nodes. As all independent variables are taken into account at the collection of terminal nodes, this holds. Therefore, we add each θ_{jm} in the summation. As such, we find the θ_{jm} factor with which the target outcome has to be adjusted. The v in front of the summation represents the shrinkage parameter. This parameter lies between 0 and 1 and controls how strongly we adjust the target outcome based on the regression tree iteration. This prevents a single bad tree iteration from completely shifting the target outcome out of proportions.

4.6 Cross-Validation

In order to optimize our model even further, we want to tune a certain set of parameters. These parameters are tuned using five times repeated, 10-fold cross-validation. According to Refaeilzadeh, Tang & Liu (2018): "Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against." After the cross-validation, the parameter set which gives the optimal results is chosen. There are three parameters that are tuned. The first parameter is the interaction depth. This variable indicates the maximum size of each of the weak learner trees. This is tuned in order to prevent overfitting, as the weak learner trees would otherwise already become too large and intricate. This would defeat the purpose of the boosting model. The second parameter we tune is the number of trees. This indicates how many weak learner trees are used in the final model. More trees generally result in a more accurate prediction, but there is again a risk of overfitting here. This is why the number of trees must also be tuned. The third and final parameter we tune is the shrinkage parameter. This parameter indicates the learning rate for our algorithm. This essentially shows with what magnitude each additional weak learner tree modifies the overall model. If this learning rate is very high, there is a risk of overfitting the final model. By using a

lower learning rate, we slow down the process and prevent the model from overfitting. On the other hand, this does mean that we need more trees to gather an accurate model. It therefore comes at the cost of computation time. The cross-validation window for each parameter was as follows:

- The interaction depth was tested between 1 and 10, with intermittent steps of 1. An interaction depth of 1 meant that there was only one split for each regression tree. After this split, the end of the regression tree is reached. 10 means that the end of the tree is reached after 10 splits.
- The number of trees was tested between 100 and 1000, with intermittent steps of 50.
- The shrinkage parameter was tested between 0.01 and 0.1, with intermittent steps of 0.01.

These parameters were chosen in such a way that many rational options were cross-validated, whilst also keeping the computation time at a reasonable level. The high starting value for the number of trees is due to the use of the shrinkage parameter. Any value below 100 would not large enough to arrive at a sensible conclusion. Furthermore, this would only add to the already exhaustive computation time. Due to computational limitations of the equipment used for the analysis, taking smaller intermittent steps was not possible.

5. Results

5.1 Multiple linear regression

The multiple linear regression was trained on the training set containing 1292 observations. In Table 2 the results of the multiple linear regression can be found. The table does not contain all variables, but rather a number of variables with interesting results. All variables from Table 6 under Appendix A were used in the model.

Table 2.

Multiple Linear Regression Results

	Coefficient	Standard Error
Age^2	0.0007	0.0013
Rating	0.6195***	0.0984
Minutes Played	0.0000	0.0002
Goals	0.0035	0.0065
Assists	0.0234*	0.0096
Total passes per 90	0.0091***	0.0025
Passing success percentage	0.0076*	0.0032
Times dispossessed per 90	-0.0711**	0.0237
Tackles per 90	-0.1083***	0.0326
Forward	0.4177**	0.1510
Log Market Value (Transfermarkt)	0.6462***	0.0226

Note. Table 2 shows the regression results where the log of the transfer fee is the dependent variable. For the p-values, * denotes p<0.05, **denotes p<0.01 and *** denotes p<0.001.

As we can see in Table 2 above, there are some interesting results. Age squared surprisingly does not have a significant relationship with transfer fee according to the multiple linear regression model. This is not as expected. According to the literature (Majewski (2017) and Müller, Simons, & Weinmann, (2017)), players who are very young are less likely to be worth a lot, due to inconsistency. Some young players grow and become excellent players, whereas others fail and slack behind. To mitigate this risk, football clubs are less likely to pay large sums in order to acquire the player. Hence, the transfer fee will often be lower. For old players, there is very little inconsistency in quality. However, older players are more prone to injury as the physical decline

increases with age. Furthermore, they will have less years left playing to their full potential regardless of injury. Therefore, clubs are also willing to pay less for players in this age category. The highest transfer fees are paid for players in their prime, who are both relatively consistent and at their peak physical condition for a long time. Yet these relationships are not present in the regression model. One explanation for this could be that age is of less importance when considering only the top 250 transfers per season in terms of transfer fees. This effect may be larger for the transfers in the middle and lower height of the transfer fee distribution, hence why this is not present with a dataset containing only high end transfer fees. Another interesting observation from Table 2, is the fact that the WhoScored rating is significant and positive. This means that if the WhoScored rating increases, the likelihood of a higher transfer fee for the player also increases. The WhoScored rating is based on the extensive and in-depth performance metrics. If a player's metrics are positive, they receive a higher WhoScored rating. This shows that in general if the statistics back up that a player is performing well, this coincides with a higher transfer fee. This indicates that statistics may indeed be a suitable route for clubs to find talented and valuable players. Clubs could exploit this by looking at players who according to their statistics should be very highly valued, yet play at a lower football level. They could then acquire these players before their valuation outgrows the financial level of the club.

Another interesting observation is that the coefficients of minutes played and goals scored are not significant. Although intuition would say that these two metrics are important in determining a player's value, the analysis shows this is not the case. In general, players who score a high volume of goals and play nearly every possible minute are very important and highly valued. This would also be the explanation as to why attacking players often move for larger transfer fees than defensive minded players. Nevertheless, the regression results show that these variable are not significant when it comes to the transfer fee. The main cause of this could also lie in the limited dataset used in the analysis. As only 1292 observations are used, there is room for noise. Furthermore, the fact that minutes played is not significant may have to do with the aforementioned multicollinearity issue. Since a multitude of the metrics are corrected to 'metrics per 90 minutes', the effect of minutes played may already be incorporated in these metrics. Another result that we see in Table 2, is that the metrics of creativity such as assists, total passes

per 90 and percentage of successful passes all have a significant and positive relationship with transfer fees. The defending statistics such as tackles per 90 as well as the negative creativity statistics such as timer a player is dispossessed per 90 all do have a significant relationship with transfer fees, but this relationship is negative. Both the observations on the creativity metrics and defensive metrics do align with the observations made in previous literature which also indicated the discrepancy between attacking and defensive metrics. Therefore, it does indeed seem that attacking and creative qualities are more beneficial for a player's transfer fee than defensive qualities, supporting the idea that offensive players are often more highly valued than defensive players. This is further supported by the fact that being a forward has a significant and positive relationship with the transfer fee. This aligns with the literature stating that attacking players are often more highly valued (Felipe et al. 2020). Lastly, it is interesting to note that a higher crowd-based market value assessment also coincides with a higher transfer fee. The coefficient of the market value variable is significant and strongly positive. Hence, the crowd-based market values by Transfermarkt may actually show a semi-accurate depiction of what the true market value of a player is. This also indicates that football clubs such as Olympique Lyon, Schalke 04 and FC Porto who use these crowd-based market values in their transfer policies (Keppel & Claessons, 2020) may be on the right path. Due to the close moderation on the website and the large number of crowd inputs, the market values on Transfermarkt end up being relatively realistic depictions of the actual market value and transfer fee. Hence, it would be wise to include this into the prediction model when considering the transfer fee of a player.

All independent variables were included in the multiple linear regression. As partially seen above, there are both variables which are significant and variables which are not significant. In the end, all independent variables were used in the subsequent prediction analysis. Although there are variables with insignificant coefficients, these were still procured for the final prediction analysis. The main reason for this is that exclusion of these variables led to significant information loss when it came to the predictions. Whilst some variables may not be significant on their own, they may play a part during the interaction with other variables and ultimately the predicted transfer fee. Furthermore, excluding them may cause some other variables to inadvertently acquire more weight than they should during the prediction process. Especially the boosting model can account

for these interactions through the use of regression trees. In order to paint the most complete picture and acquire the most accurate prediction model, all variables are thus kept in during the prediction phase of the analysis.

5.2 Gradient boosting

In addition to the multiple linear regression model, a gradient boosting model was created. For the gradient boosting model, the parameters used after cross-validation were as follows: the interaction depth is 9, the number of trees used is 100 and the shrinkage parameter is set at 0.1. The results on the relative importance of the 10 most important variables in the boosting model can be found in Figure 3 and Table 3 below:

Figure 3

Plot on relative importance of the top 10 variables in the boosting model

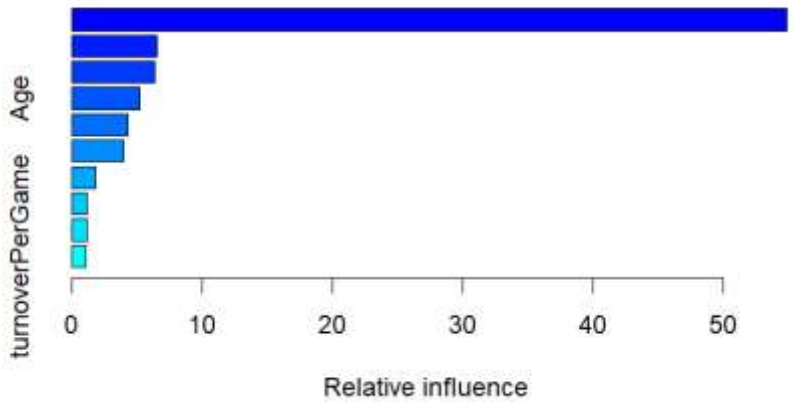


Table 3.

Relative importance of the top 10 variables in the boosting model

	Relative influence
Log Market Value (Transfermarkt)	54.344
League transferred from	6.272
League transferred to	6.062
Age	5.000
Tournament player performed in	4.247
Position	3.745
Successful dribbles per 90	1.836
Total turnovers per 90	1.302
Aerials won per 90	1.200
Total passes per 90	1.090

Note. Table 3 shows the importance of the top 10 most important variables in the gradient boosting model.

Table 3 shows the importance of the top 10 most important variables in the gradient boosting model. Here we can see that the initial market value estimation by Transfermarkt is the most important variable in the model. The market value from Transfermarkt is accountable for explaining around 54.3% of the variation in the transfer fee of the player. This is unsurprising, as the market value estimations from Transfermarkt are relatively accurate to reality (Peeters, 2018). Therefore, the boosting model will often make a split based on this pre-determined crowd-based estimation. For future use, this could pose a problem in case the market valuation estimates from Transfermarkt decrease in accuracy. However as mentioned, these values are closely moderated which should prevent the market values from Transfermarkt from being completely inaccurate.

Furthermore, we see that the league a player transferred from, the league a player transferred too and the league that the player performed in are relatively important as well. They account for around 6.3%, 6.1% and 4.2% respectively of the variation in the transfer fee. This suggests that the main predictor of quality and therefore transfer fee, is the general level of the buying and selling club, as well as the level the player has performed at. This aligns with the views projected by Müller, Simons, & Weinmann (2017), who also indicate the these factors are of importance when it comes to market value of a player. Although the variable importance does not indicate whether the relationship between these variables and the transfer fee is positive or negative, it is likely that generally speaking a higher league quality indicates higher individual player quality and a higher individual player quality indicates a higher transfer fee. Therefore, we see that it is important to acknowledge that both the buying and selling clubs need to take into account the quality level of the opposing party during transfer negotiations. Furthermore, the player's position also appears to be of relatively high importance when it comes to the transfer fee. Table 3 indicates that a player's position explains around 3.7% of the variation in transfer fees. From this, we cannot gather what positions are of importance and what positions have the highest transfer fee. However, literature shows that in general attacking players are significantly higher valued than defensive players. (Felipe et al., 2020) Hence, it is likely that this significant difference is reflected in the variable importance of the position of the players. Lastly, we see that age is relatively important as it accounts for around 5% of the variation in the transfer fee. This aligns with the findings of both Majewski (2017) and Müller, Simons, & Weinmann, (2017) which indicate that both older and younger players tend to be less valuable. Although the variable importance again does not indicate the sign of the relationship, a negative relationship is most likely.

The six most important variables in the gradient boosting model are all not performance metrics. However as seen in Table 3, there are also some performance metrics which are of great influence when it comes to transfer fees and are included in the top ten most important variables. We see that successful dribbles (1.8%), total turnovers (1.3%), aerial duels won (1.2%) and total passes per 90 (1.1%) are all relatively important in explaining the variation in transfer fees. Furthermore, we see that goals (0.3%) and assists (0.3%) are not in the top ten most important variables. In fact, these variables are positioned as the 30^{th} and 32^{nd} most important variables for explaining

variation in the transfer fee. They are both ranked lower than nearly all more in-depth performance statistics such as the four found in Table 3. Whilst literature has found that goals and assists are important explanations for difference in market value (Majewski, 2017 & Müller, O., Simons, A., & Weinmann, M., 2017), the boosting model indicates that taking the more in-depth match statistics are actually of more explanatory value. Goals and assists are created through the actions and decisions which accommodate the in-depth performance statistics. Hence, it seems that using the in-depth performance statistics to explain the variation in transfer fee allows for a more nuanced prediction. Therefore, it does seem that using the in-depth performance statistics in addition to the commonly used goals and assists statistics is actually beneficial for assessing market values.

5.3 Multiple linear regression and gradient boosting: Prediction comparison

After analyzing and interpreting the role of several variables, analysis was done to see if several models were effective in assessing and predicting transfer fees of players based on event metrics.

Table 4.

Transfer fee prediction model RMSLE

	Multiple Linear Regression	Gradient Boosting
Training set	0.4108	0.2369
Test Set	0.4110	0.4005

Note. Table 4 shows the RMSLE based on the prediction of the multiple linear regression and gradient boosting models on the training and test set.

In Table 4, the RMSLE results of the multiple linear regression and gradient boosting models on the training and test set are shown. The closer the RMSLE is to 0, the more accurate the model is in their prediction. As Table 4 shows, the gradient boosting model has a lower RMSLE when predicting on the training set than the simple multiple linear regression model. This means that the gradient boosting model performs better than the multiple linear regression model. This is most likely due to the fact that the gradient boosting model can capture the more intricate

relationships between the independent variables better, resulting in more accurate predictions. This is because the boosting model uses regression trees, which not only assesses the effect of one independent variable on the outcome variable transfer fee, but also takes into account relationships between variables due to the multiple paths within a regression tree. Hence, it therefore seems that a boosting model is more effective in predicting the transfer fee of a player more accurately than the multiple linear regression model. However, a RMSLE value of 0,2369 is decent, but still not an extremely accurate model. Although we cannot directly derive any straightforward interpretation from this number, this does mean that the relative error between the predictions and the actual transfer fees are still relatively high. It means that using gradient boosting we can predict the transfer fees of players using performance data reasonably, but sufficiently well to directly take as the targeted transfer fee. Nevertheless, it can prove to be a useful guideline for clubs to have a better idea when setting the initial transfer fee. On the other hand, one could argue whether this predicted value is any better than simply taking the crowd-based market valuation by Transfermarkt.

Furthermore, we see that the difference between the multiple linear regression model and gradient boosting model is much smaller when it comes to the test set. The gradient boosting model still performs better than the multiple linear regression model. However, this is only a very limited difference. Therefore, this difference is almost negligible, especially given that the gradient boosting model is much more advanced than the simple multiple linear regression model. In addition to this, the drop-off in accuracy between the training and test set is much smaller for the multiple linear regression model than for the gradient boosting model. This suggests that the latter model maybe suffers from overfitting. A possible solution to overcome the overfitting problems in the future, would be to get a data set of a more considerable size. Variables such as ‘league from’ and ‘league to’ possess important information regarding the level the player performs at and thus the level their transfer fee should be with respect to their abilities. Furthermore, we have seen in Figure 3 and Table 3 from section 5.2 that these variables are of very high importance when explaining the variation in transfer fees. However, in this dataset there are multiple leagues who have between 5 and 10 observations. Perhaps if more observations are added per league, the true role of each league regarding transfer fees can be found. In addition,

the more accurate picture for these variables can help overcome the overfitting problems for the general model. Another possible solution would be to increase the cross-validation window of the shrinkage parameter. This parameter is the strongest factor in combating overfitting problems. Perhaps increasing the number of shrinkage parameters tested during cross validation can combat the majority of overfitting problems. This was also tested. Although this did decrease the discrepancy between the training and test set, the overall results were worse. We cannot say with certainty that solving these overfitting problems is enough to significantly improve the gradient boosting model's performance, though. Hence, it begs the question whether the gradient boosting model genuinely provides an increase in accuracy when assessing football player transfer fees using extensive performance metrics when compared to a simple multiple linear regression model.

5.4 Performing boosting on subsets by position

In order to attempt to increase the accuracy, the dataset was split in four groups: Attackers (513 observations), Midfielders (682 observations), Defenders (363 observations) and Goalkeepers (55 observations). A gradient boosting model is performed on the first three groups. The Goalkeepers group was not used, as this group lacks too severely in terms of observations to create a solid boosting model. The results can be found below in Table 5. For the attacking group, the interaction depth was 5, the number of trees used was 100 and the shrinkage parameter was 0.1 For the midfielder group, the interaction depth was 4, the number of trees used was 100 and the shrinkage parameter was 0.1 Lastly for the defender group, the interaction depth was 5, the number of trees used was 100 and the shrinkage parameter was 0.1.

Table 5.

Transfer fee gradient boosting prediction model RMSLE based on three subgroups

	Attackers	Midfielders	Defenders
Training set	0.1848	0.2415	0.1689
Test Set	0.3928	0.3920	0.4489

Note. Table 5 shows the RMSLE based on the prediction by gradient boosting models on the training and test set of the three subgroups.

In Table 5 we can see similar problems arising with the gradient boosting models. When split up into the three subgroups, the gradient boosting models become very accurate for the training sets. The RMSLE of the Defenders subgroup even goes as low as 0.1689, indicating a relatively accurate prediction. However, the overfitting problems remain throughout all groups when going from the training to the test set. With these smaller datasets, the difference between training and test set becomes even larger than with the general gradient boosting model. These overfitting problems could have multiple causes. One cause could be that the parameters found during the 10-fold cross validation are too focused on the training set and not accurate for test sets. However, in principle the existence of the shrinkage parameter should counteract these issues. More likely is the fact that the overfitting problems are again due to very limited data for certain variables. For some variables such as the 'league to' and 'league from' have very limited occurrences. Therefore, a transfer from the Dutch Eredivisie in the training set may differ a lot from another transfer from the same division in the test set. As there are very limited observations in the dataset on this (especially for the three subgroups), this can cause a multitude of overfitting problems. This idea is additionally supported as absolute difference in relative error is the highest for the smallest subgroup (Defenders) and lowest for the largest subgroup (Midfielders). Deleting the 'league to' and 'league from' variables can be an option. The analysis was also performed without these variables. Although this did cause the training and test set to be closer together in terms of RMSLE, the overall performance was lackluster. Therefore, the information loss when excluding these variables is a larger problem than the overfitting problems caused by a lack of observations.

This is also supported by the importance of the 'league to' and 'league from' variables in the general boosting model.

5.5 General conclusions

These results are not strong enough to strongly conclude whether football player transfer fees can be accurate assessed using extensive performance data in combination with a stronger machine learning model. The limitations with the dataset harm the models to a significant degree. Nevertheless, it would be wise for clubs to consider using extensive performance data during transfer negotiations and scouting. We have seen with the WhoScored rating that solid underlying performance metrics can be indicative of a highly valuable player. Clubs could use this knowledge to find hidden talent based on data, such as the WhoScored performance metrics. This evidence is also supported by the fact that the in-depth performance data was more important when explaining the variance in transfer fees than basic metrics such as goals, assists and minutes played, according to the gradient boosting model. This shows that using the in-depth performance statistics to explain the variation in transfer fee allows for a more nuanced prediction. Therefore, it does seem that using the in-depth performance statistics in addition to the commonly used goals and assists statistics is actually beneficial for assessing market values. Furthermore, the results from the gradient boosting method are not bewildering, but they are reasonable. Especially when splitting the observations based on position (Attacker, Midfielder, Defender), the gradient boosting method does show sign of promise. If clubs are to incorporate this method, additional research needs to be done on what factors are of importance when it comes to transfer fees, both performance and non-performance related. Furthermore, tests with larger datasets should be run. It is interesting to see whether an increased dataset in combination with the manual split based on position can gather sufficient results when applying both extensive performance data and an advanced machine learning model. Nevertheless, clubs should in principle only use the machine learning model as guideline and starting point when it comes to transfer negotiations instead of adopting the predicted transfer fee one on one. Performance metrics should mostly be used as guidance when regarding player quality and transfer fee

assessment, rather than directly applying them when making administrative choices at executive level.

6. Conclusions & Limitations

To sum up, multiple conclusions can be derived from this research. First of all, there do seem to be indications that using extensive performance metrics to asses a player's market value can be useful. When the dataset is large enough, the gradient boosting model can be used to get a more accurate picture of the estimated market and transfer value. If football clubs were to incorporate this, this may be able to prevent overpayment for a player during transfer negotiations. However, further research with a larger database must performed in order to derive strong conclusions on the success of a more data focused strategy. As for now, the limited number of observations when it comes to the league where the player performed in, the league where the player transferred from and the league where the player transferred to make it difficult to gather strong evidence from this research. Hence, it remains that the role of extensive performance metrics should be a guiding one rather than a direct decisive role when it comes to transfer fees and negotiations. Second of all, more research needs to be done on what factors play a part in the accumulation of the player's transfer value. This not only concerns performance metrics, but also exterior factors such as popularity and transfer history. Although this research has found that in general performance metrics regarding attacking and creativity are beneficial to market and transfer value and defense metrics have a negative relationship with market and transfer values, these relationships need to be explored further. To aid in this process, it may be a consideration for football clubs willing to use data to create different models based on position. This research has established that there is evidence to suggest that using a different model for each of the different positions on the field may improve the accuracy of the model. Understanding what metrics make a player in each position more valuable, may be precious information when deciding on players to scout or transfer. However, this also needs to be assessed on a larger scale in future research as the dataset in this paper has proven to be insufficiently large in order to derive very strong conclusions. Given this research, the main conclusion is that clubs should mainly use the extensive performance metrics and advanced machine learning model as guideline and starting point when

it comes to transfer negotiations instead of adopting the predicted transfer fee one on one. Extensive performance metrics should mostly be used as guidance when regarding player quality and transfer fee assessment, rather than directly applying them when making administrative choices at executive level.

As mentioned, the main limitation of this research is the limited data. Variables such as the league the player performed in, the league the player transferred from and the league the player transferred too are important to get an accurate prediction. Therefore, future research should focus on creating a more sizeable dataset, including multiple transfers from and to the same leagues. Then the true effect of each league can be assessed much better than in the current research. By accounting for this effect better, the true role of extensive performance metrics can also be researched further. In addition to this, there are other variables to consider in future research such as measures of popularity and historic transfer fees. These variables can aid in understanding why some paid transfer fees are significantly higher than the in-depth performance metrics would suggest. Furthermore, there should be additional experiments with other machine learning methods. This research has focused mainly on multiple linear regression and gradient boosting methods. However, perhaps other machine learning methods prove to be more effective for this type of prediction and truly unearth the potential value of adding extensive performance metrics. Lastly, a deeper dive into crowd-based estimations should be made. It seems that crowd-based estimation sources such as Transfermarkt are often relatively accurate, despite this being purely subjective. It would be interesting to see to what extent a crowd-based estimation method could provide an alternative for clubs who either do not want to or simply cannot afford to invest in the data department for transfers. Hence, a more thorough study of the inner workings of crowd-based estimations would also be interesting to explore.

7. References

Asif, R., Haque, S. I., Zaheer, M. T. & Hassan, M. A. (2016). Football (Soccer) Analytics: A Case Study on the Availability and Limitations of Data for Football Analytics Research. *International Journal of Computer Science and Information Security, 14*(11), 516-518. https://www.academia.edu/30927325/Football_Soccer_Analytics_A_Case_Study_on_the_Availability_and_Limitations_of_Data_for_Football_Analytics_Research

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis*. Research Synthesis Methods, 1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/bf00058655

Dejan. (2020, May 27). Soccer (football) dataset from whoscored. Kaggle. https://www.kaggle.com/datasets/dejoski/soccer-football-dataset-from-whoscored

Felipe, J. L., Fernandez-Luna, A., Burillo, P., de la Riva, L. E., Sanchez-Sanchez, J., & Garcia-Unanue, J. (2020). Money talks: Team variables and player positions that most influence the market value of professional male footballers in Europe. *Sustainability*, *12*(9), 3709. https://doi.org/10.3390/su12093709

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation, 121*(2), 256–285. https://doi.org/10.1006/inco.1995.1136

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5). https://doi.org/10.1214/aos/1013203451

Hall, P. (2023, January 6). *FIFA to introduce cap on fees in widespread agent rule changes.* Reuters. Retrieved April 24, 2023, from https://www.reuters.com/lifestyle/sports/fifa-introduce-cap-fees-widespread-agent-rule-changes-2023-01-06/

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

Keppel, P., Claessens, T. (2020, December 18). How the volunteers of data website Transfermarkt became influential players at European top football clubs. Follow the Money - Platform for investigative journalism. https://www.ftm.eu/articles/transfermarkt-volunteers-european-football#:~:text=Marseille%20are%20not%20the%20only,data%20in%20a%202016%20report.

Kucharčíková, A. (2011). Human Capital-Definitions and Approaches. *Human Resources Management & Ergonomics*, *5*, 60-70. https://frcatel.fri.uniza.sk/hrme/files/2011/2011_2_05.pdf

Kumari, K. & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences, 4*, 33-36. https://doi.org/10.4103/jpcs.jpcs_8_18

Majewski, S. (2016). Identification of factors determining market value of the most valuable football players. *Journal of Management and Business Administration. Central Europe*, *24*(3), 91–104. https://doi.org/10.7206/jmba.ce.2450-7814.177

Monteiro, R. K., Prates, R. C., & Frota, L. M. (2022). The determinants of player transfers in Brazil: The role of expectations in the football market. *Applied Economics*, *55*(26), 2964–2977. https://doi.org/10.1080/00036846.2022.2107989

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

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt Valuations and international soccer results. *International Journal of Forecasting, 34*(1), 17–29. https://doi.org/10.1016/j.ijforecast.2017.08.002

Refaeilzadeh, P., Tang, & L., Liu, H. (2018). Cross-validation. Encyclopedia of Database Systems, 677–684. https://doi.org/10.1007/978-1-4614-8265-9_565

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227. https://doi.org/10.1007/bf00116037

Seltman, H.J. (2014). *Mixed Models*. Retrieved from https://www.stat.cmu.edu/~hseltman/309/Book/chapter15.pdf

Slehkyi. (2019, November 8). *Football transfers 2000-2018.* Kaggle. https://www.kaggle.com/code/slehkyi/football-transfers-2000-2018/input

Wald, A. (1947). A note on regression analysis. *The Annals of Mathematical Statistics, 18*(4), 586–589. https://doi.org/10.1214/aoms/1177730350

8. Appendix A

Table 6.

Variable index

Variable Name	Description
tournamentName	Competition the player played in. The Whoscroed data was collected in this competition.
height	Player height in centimeters.
weight	Player weight in kilograms.
positionText	Position of each player. Consists of 4 levels, namely Goalkeeper, Defender, Midfielder, Attacker.
apps	Number of match appearances.
subOn	Number of times player was substituted on from the bench.
minsPlayed	Number of minutes played.
rating	Average Whoscored rating per match
yellowCard_x	Number of yellow cards (whole season)
redCard_x	Number of red cards (whole season)
aerialWonPerGame	Number of aerial duels won per 90 minutes
manOfTheMatch	Number of man of the match awards (whole season)
tacklePerGame	Number of tackles per 90 minutes
interceptionPerGame	Number of interceptions per 90 minutes
foulsPerGame	Number of fouls committed per 90 minutes
offsideWonPerGame	Number of offsides won per 90 minutes
clearancePerGame	Number of clearances per 90 minutes
wasDribbledPerGame	Number of times dribbled past by opposition player per 90 minutes

outfielderBlockPerGame	Number of blocks by outfield players per 90 minutes
goalOwn_x	Number of own goals (whole season)
goal_y	Number of goals (whole season)
assistTotal_y	Number of assists (whole season)
shotsPerGame_y	Number of shots per 90 minutes
dribblesWonPerGame	Number of dribbles won per 90 minutes
foulGivenPerGame	Number of fouls won per 90 minutes
offsideGivenPerGame	Number of offsides given per 90 minutes
dispossessedPerGame	Number of times player was dispossessed per 90 minutes
turnoverPerGame	Number of turnovers per 90 minutes
keyPassPerGame_y	Number of key passes per 90 minutes
totalPassesPerGame	Number of passes per 90 minutes
accurateCrossesPerGame	Number of accurate crosses per 90 minutes
accurateLongPassPerGame	Number of accurate long passes per 90 minutes
accurateThroughBallPerGame	Number of through balls per 90 minutes
passSuccess_y	Average percentage of successful passes per game
Age	Player age (at time of the transfer)
League_from	League player transferred from. Consists of 41 levels.
League_to	League player transferred to. Consists of 30 levels.
Market_value	Market value of the player according to transfermarkt.de (at time of the transfer)
Transfer_fee	Reported transfer fee paid.