**Soccer Player Transfer Fee Predictive Analysis**

Ryan Anderson Denison University April 12, 2019

Ryan Anderson is a senior undergraduate student at Denison University, studying Computer Science and Data Analytics with a concentration in Economics. As an avid soccer player and watcher, Ryan seeks to combine his academic research with his passion for soccer in this exploration of the economics and data behind the professional soccer player transfer market.

Contents

[Introduction 1](#_Toc8049981)

[Domain Review 2](#_Toc8049982)

[Methods 5](#_Toc8049983)

[Data Collection 7](#_Toc8049984)

[Data Cleaning 9](#_Toc8049985)

[Analysis 10](#_Toc8049986)

[Results 12](#_Toc8049987)

[Ordinary Least Squares Regression 13](#_Toc8049988)

[K-Nearest Neighbors 20](#_Toc8049989)

[Discussion 22](#_Toc8049990)

[Acknowledgements 25](#_Toc8049991)

[Appendix 26](#_Toc8049992)

[References 27](#_Toc8049993)

# Introduction

During the months of July and August, most club soccer teams are in their offseason, and do not have any game scheduled. So, one might believe that this is a boring time for a team’s fans. However, lasting from June 1st to August 31st, the transfer window is open for clubs to sign new players. Fans worldwide love to hear the news of star signings to their favorite club, and dread the news of a departure of a fan favorite. The transfer window is a great time for teams to look forward to the new season by signing players to strengthen their squads. In recent years, more data has become available for accessing soccer players’ abilities, giving teams a whole new perspective when evaluating a player that they wish to sign. Unlike many other professional sports, soccer players are often traded for sums of money rather than other players. A transfer fee is the sum of money paid between clubs when one club wishes to buy a player from the other club. Market value is a theoretical transfer fee that would be paid for a player, and the two terms are often used interchangeably.

The main goal of my research was to create a model that accurately predicts the transfer fee for professional soccer players. My model predicts a player’s transfer fee based off the following factors: age, the FIFA world ranking of the nation the player represents, goals, assists, total games played, the year the transfer occurred, and position. All the statistics were taken from the season prior to the player’s transfer because these statistics represent the most recent evaluation of a player’s ability and market value.

# Domain Review

In the past, teams would send scouts to watch matches and evaluate potential signings. Although in-person scouting is still in use, the availability of player data allows for much broader scouting, since clubs can use the data beforehand to narrow down the number of players they need to scout in person. Sending scouts out on scouting missions isn’t free, so a strong predictive model based on data will save clubs time and money. This literature review begins by discussing some potential key influencers of a player’s transfer fee by discussing research done by experts in the field.

My first hypothesis is that age will be one of the most influential variables for predicting a player’s transfer fee. Young players make exciting signings because they have room and time to reach their potentials. For example, in 2004, Manchester United, one of the historically best teams in the Premier League, the highest division of English soccer, signed 18-year-old Wayne Rooney for 26 million euros, quite a large fee for a teenager at the time (Transfer Records, 2019). He went on to become the club’s highest ever goal scorer with 253 goals. He also won 5 Premier League titles, a Union of European Football Associations (UEFA) Champions League title, one UEFA Europa League title, and countless individual awards during his thirteen seasons at the club (Wayne Rooney - Player Profile, 2019). Looking back on his career, his transfer fee was an absolute bargain. Although players in the early parts of their career have potential to become fantastic players, they may struggle adapting to a new team or even a new league.

Using another Manchester United player as an example, Memphis Depay was signed from PSV Eindhoven, one of the top teams in the Eredivisie, the top division of soccer in Netherlands. The 21-year-old was a fantastic player compared to his peers in Netherlands, scoring 28 goals in 39 games, and his fee was around 30 million euros (Transfer Records, 2019). However, the striker found the Premier League much more difficult, managing to tally only 4 goals in 36 games, before being sold for only 14 million euros (Transfer Records, 2019). As shown by these examples, there are high risks involved with signing young players, but sometimes that risk pays off. These players are more extreme cases of the risks and rewards involved with signing young players.

When it comes to scholarly research on determining market values of players, there are few studies that are both recent and extensive. Some researchers failed to create different models for each position (He, 2014), or only ran their model on only forwards (Mojewski, 2016). I will apply a detailed model to every position, while also having a unique model for forwards. It is important to have separate models since each position has a different role and different key statistics that show the player’s quality. One model was created that predicted a player’s transfer fee based on a player rating calculated himself based on stats prior to a player’s transfer (Sæbø & Hvattum, 2015). This model used nationality of players as one of its predictors, but in geographical groups (e.g. European Union, Africa). However, I think it may be more beneficial to look at the quality of the national team the player currently represents by using the player’s country’s FIFA world ranking.

Looking at the FIFA ranking of the country a player represents, I believe that stronger nations will increase players’ transfer fees. Even if the player is a relatively poor player, I believe that representing a stronger national team creates a false sense of quality, which biases the transfer fee. One reasoning for this is the Homegrown Rule for the English Premier League, which states a team “must include eight Home Grown players out of a squad of 25 for that Premier League season” (Rapp, 2017). A Home Grown player is one who has spent at least three years with any English Football Association team before his 21st birthday (Rapp, 2017). This rule is implemented to limit the number of foreign players in the Premier League. The majority of these Home Grown players will be English, since players tend to grow up playing in their home country. This affects transfer fees because Premier League teams are willing to pay more for English players to fill this Home Grown quota. This is true for other leagues that have similar rules. Since countries with stronger leagues usually host stronger national teams, this is one reason why players from stronger nations draw larger transfer fees. The Chinese government is funding the CSL to spend large amounts of money to increase the competitiveness of the league (Panja, 2017). Transfer fees from clubs in this league are often extremely inflated. For example, 26-year-old Brazilian player Oscar, which is his full name, who was purchased by Shanghai SIPG from Chelsea for 60 million euros in 2016 (Transfer Records, 2019), the highest fee ever for a player by the Chinese Super League. Oscar was a decent player but struggled to get playing time for his old team, playing in only 9 games the season before his transfer (Transfer Records, 2019). To remove bias, I will have to standardize fees paid by clubs from the CSL or remove them entirely. I included a binary variable that was set to 1 if the player was transferred to a team in the CSL and 0 otherwise. However, I had a very small number of players who were transferred to the CSL, which meant including this variable had little effect on my final model.

I am confident that my research has brought knew knowledge to the domain of predicting a player’s market value. By analyzing past research, I aim to provide the most up-to-date and accurate model for determining a transfer fee for any player. I used appearances, goals scored, and assists provided to predict the player’s transfer fee. In addition, by using non-playing attributes such as a player’s age the strength of the soccer team for nation they represent to build my model, and if the player was transferred to the CSL, I will be able to capture the whole picture of the transfer, not just how good the player is. Predicting the market value of a player is a difficult but important task, as an accurate model can save clubs tens of millions of euros.

# Methods

I explored the idea that a player’s individual playing statistics are not the only thing that affects his transfer fee. Non-playing factors such as age and the player’s nationality may also have an effect. I believe age will have a negative effect on the transfer because as players get older, they have less time to play and room to improve. I can compare players with similar statistics to see if, on average, the younger player draws a larger fee. I also predict that if the player’s release clause was paid, it will increase their transfer fee. I believe that players from stronger nations, denoted by the country’s FIFA national ranking, will draw larger fees. Since the best team is ranked with a 1, this variable will have a negative effect on the player’s fee. For the individual statistics, I predict that for forwards, goals and assists will have strong positive relationships with the player’s fee. For midfielders and defenders, I believe there will still be a positive correlation for both, more for assists than goals, but not as strong as the relationship between the two for forwards.

For my project, I will be collecting data about professional soccer players and will attempt to predict their transfer fee from a variety of factors, including age, contract stipulations, FIFA ranking of the nation they represent, as well as individual statistics including goals, assist, games played, passing percentage, tackles, interceptions, clean sheets, saves, and goals conceded. My project does not require IRB approval. The data that I am working with is freely available to the public on the Transfermarkt website. After reading through the websites privacy policy and terms of use (Transfermarkt: Privacy Policy, 2018), I am certain that my use of the data is both ethical and legal. Although the data I am working with is human data, professional athletes are considered public figures, so there is no concern there. In addition, all player salary data and club financial data is all publicly available.

The methods I will be using for my prediction include linear regression and k-nearest neighbors. Linear regression is one of the most common statistical models used and is a basic but useful method. For these reasons, I see no concerns using a linear regression to predict players’ transfer fees. K-nearest neighbors returns a value that is aggregated from its k neighbors that are closest in distance (Hechenbichler and Schliep, 2004). This method allows me to find similar players from my training set and use their averaged transfer fees to predict the fee for the player I am testing.

In addition, there are no conflicts of interest involved with this project. Unfortunately, I have no personal connections to any of the professional soccer players in my dataset, nor do I have any relations to the data owners. I do not have any strong intent on finding a certain answer, which means my results will be unbiased. It is important to note that there are many factors not being measured in my dataset that can influence a player’s transfer fee. Soccer is known as the beautiful game for a reason and a player’s worth cannot be seen through his statistics. This inability to express a player’s true quality will most likely introduce bias into my models and decrease their accuracy.

## Data Collection

My final dataset is merged data from two sources: Transfermarkt.com and FIFA.com. Transfermarkt is the source for the transfer and player data, including player names, ages, positions, nationalities, transfer fees, the club they left, the club they joined, the season that the transfer occurred, goals, assists, and appearances (Transfer Records, 2019). Collecting data is often one of the most difficult parts of any research, but it is also the most important. There is rarely ever a nicely formatted csv that fits the exact specifications for your research. But there are websites that may have this information, and it is possible to extract that data through a technique called web scraping. I used the BeautifulSoup library for Python 3, which is an HTML parser that uses regular expressions to extract information from the website (Vargiu & Urru, 2013), and returns nicely formatted HTML that can be formatted into a dataframe, where I can begin to clean my data using pandas, a data cleaning package for python (McKinney, 2018).

I used the BeautifulSoup package for Python 3.7.2. I took the 25 players with the highest transfer fees from each year since 2014, giving me 125 players to look at. I decided not to include any goalkeepers in my analysis, since they were such a small subgroup of the overall dataset. This was one of the limitations of my data collection, as these were the only players listed on the Transfermarkt website. This dataset is publicly available, and the site’s privacy and usage policies do not prohibit use of their data (Transfermarkt: Privacy Policy, 2018). Unfortunately, I was unable to get data permissions for more in-depth statistics, such as passing percentage or defensive statistics, which will hurt my final results. The data for the FIFA world rankings came from Kaggle and contain monthly rankings of every official FIFA national team (Fitzgerald, T. 2018). I took a subset of this dataset to only include the years 2014-2018, and only used statistics for the month of June, which is when the transfer window opens. Table 1 below shows the list of all variables in my dataset.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Nationality | String | The nationality of the player. Used to merge with the FIFA ranking dataset. |
| Year\* | Integer | The year the transfer occurred. |
| Player\* | Sting | The name of the player. |
| Age | Integer | The player’s age at the time of the transfer |
| Position | Categorical | The player’s position (12 levels: GK, RB, CB, LB, CDM, RM, CM, LM, CAM, RW, ST, LW) |
| Position\_Gen | Categorical | The player’s general position (4 levels: GK, DEF, MID, FWD) |
| Fee | Continuous | The transfer fee of the player |
| Left | String | The team the player left, or the selling club |
| Joined | String | The team the player joined, or the buying club |
| Appearances | Integer | Number of games the player played in the previous season in all competitions |
| Goals | Integer | Number of goals the player scored in the previous season in all competitions |
| Assists | Integer | Number of assists the player had the previous season in all competitions |
| CSL | Binary | Denotes whether the player was bought by a club in the Chinese Super League (1 for yes, 0 for no) |
| \*Unique Identifiers of one observation | | |

Table 1: Codebook for all variables in final dataset

The goal of my research was to create an entirely unbiased model. This meant using only objective variables when building my model. For example, the Transfermarkt website included an estimated market value for each player at the time of his transfer (Transfer Records, 2019). This value is a subjective estimate of how much a player is worth. I decided not to include this value in my model because it already includes some subjectivity, and I want my model to be completely objective. I also contemplated using a player’s rating from the popular FIFA video game series from Electronic Arts. Again, this is a subjective rating based on player performance, and would induce bias into my model.

## Data Cleaning

From the transfer data (Transfer Records 2019), I was able to get each player’s specific position, but I had to go through and manually mark each player’s general position. For example, center backs and outside backs were marked differently in the dataset, but I grouped both into the defender category. Each row in my dataset is uniquely identified by the player’s name and the season. I had to use both variables because there are a few players who have been transferred multiple times during the time period I am looking at.

Before I began my analysis on my data, I had to adjust for the trend of transfer fees over time. The idea of inflation rate of the transfer fees for soccer players comes from Dobson and Gerrard (1999), who predicted the transfer fees for soccer players in English professional leagues. Over the six seasons used in their study, they found an annual average transfer fee inflation rate of 19.4% (Dobson & Gerrard, 1999). This inflation rate is compared to the general inflation rate in the UK and the inflation rate of soccer-related goods, such as ticket prices and television revenue (Dobson & Gerrard, 1999). They found that although the inflation rate for player transfer fees exceeds the general inflation rate in the UK, it remained consistent with other football prices. In order to control for the average trend of the dependent variable over time, I included the year variable as a predictor to my linear model.

## Analysis

I used linear regression and k-nearest neighbors to make my predictions. Linear regression allowed me to see which variables are statistically significant and their effect on the player’s transfer fee. For example, I could see the effect that a player’s age would have on his transfer fee while holding all other variables constant. This means I could compare a 19-year-old and a 29-year-old who had similar statistics, and see the true effect that age has on a player’s transfer fee, since often players who are older are better players. Since my independent variables are a mix of continuous and categorical, linear regression is perfect since it allows multiple types of predictor variables.

Linear regression has five assumptions that the data must pass before it can be used: linear relationship between independent variables and dependent variables, multivariate normality, no multicollinearity, no auto-correlation, and homoscedasticity (Assumptions of Linear Regression, 2019). I believe that all my variables will have linear relationships with the player’s transfer fees. I will check for normality in my data with histograms. Multicollinearity occurs when two independent variables are linearly correlated with each other (Assumptions of Linear Regression, 2019). To test for this, I will look at a correlation matrix of all the independent variables to ensure that the correlation coefficients are small. Autocorrelation is when one observation is correlated to the other observations (Assumptions of Linear Regression, 2019). My observations are all independent, because one player’s transfer fee does not influence another player’s. Finally, homoscedasticity shows that the residuals are randomly distributed about the regression line (Assumptions of Linear Regression, 2019). For example, if residuals are higher for larger transfer fees, the data would be heteroscedastic. I plan to check for this using a scatter plot of the residuals and by running a Breusch Pagan test for heteroscedasticity.

K-nearest neighbors allows me to find similar players from my training set and use their averaged transfer fees to predict the fee for the player I am testing. My dependent variable, the player’s transfer fee, is a continuous variable, which makes both methods appropriate. One of the biggest advantages of k-nearest neighbors is that it doesn’t require any assumptions about the data distribution (Thirumuruganathan, 2010). I also ran tests in order to calculate an optimal value for k that maximizes the accuracy of the model.

There are three main ideas within my methods that I will have to tackle during this project: data collection, data cleaning, and data analysis. Since I was unable to find any datasets that contained all the variables necessary for my analysis, I had to collect my data from the web. This required research on the technique of web scraping, which efficiently allowed me to collect large amounts of data from my web sources (Transfer Records, 2019). Once my data was collected, I used pandas to clean the data. Once I have my data formatted fully, I can export it to a csv and upload it into R, where I merged the transfer and player statistics with the FIFA ranking data (Fitzgerald, 2018). I ran my analysis in R as well, which consisted of k-nearest-neighbor regression and linear regressions models. My research will provide a new perspective to the field of soccer player market evaluation by using the most cutting-edge data collection, cleaning, and analysis methods.

# Results

It is important to first cover all the assumptions before running a linear model. These assumptions include linear relationship between independent variables and dependent variables, multivariate normality, no multicollinearity, no auto-correlation, and homoscedasticity (Assumptions of Linear Regression, 2019). As seen in Figure 1 below, the transfer fees are not normally distributed for any position. See figure 6 in the appendix for a distribution of all players in the dataset. This could be adjusted by taking the log of the dependent variable, the player’s transfer fee, to see if this makes the data normally formatted. This is discussed in further detail in the methods section.

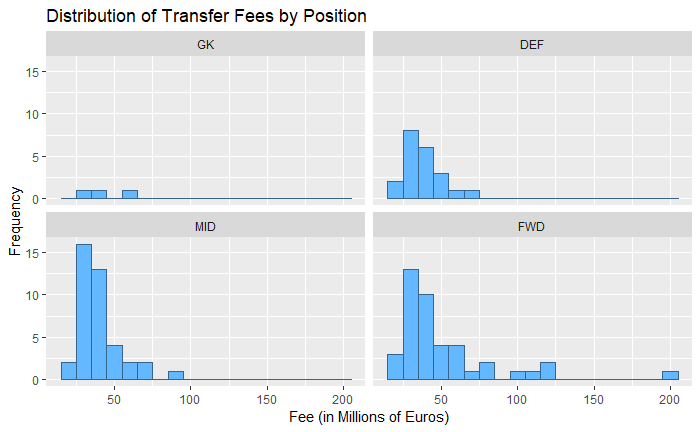


Figure 2: Distribution of transfer fees by position

Table 3 in the appendix shows a correlation matrix for all continuous independent variables: FIFA Ranking, appearances, goals, and assists. Because all the coefficients are below 0.3, there is no issue of multicollinearity. Multicollinearity occurs if two independent variables have a correlation coefficient above 0.8. Autocorrelation is when the observations are not independent of each other (Assumptions of Linear Regression, 2019). My observations are all independent, because one player’s transfer fee does not influence another player’s. This covers all of the assumptions needed to run a linear model. Testing for heteroscedasticity is done after running a model, and is discussed in the following subsection.

## Ordinary Least Squares Regression

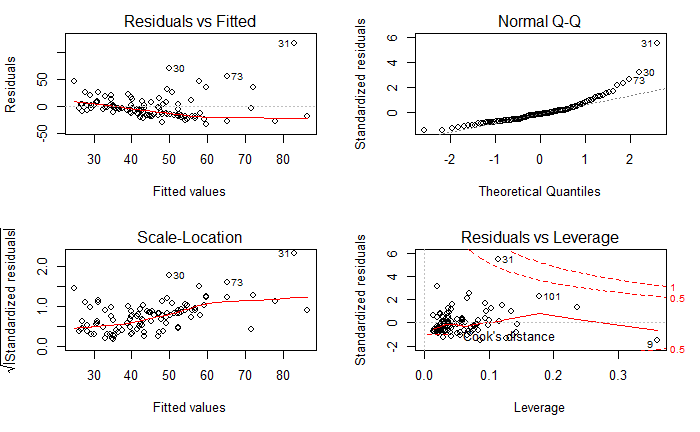
I ran an ordinary least squares regression to predict transfer fees for all outfield players, or non-goalies. Since no goalkeeper in my dataset has any assists or goals, it makes no sense to include them in the model. I used age, rank, appearances, goals, assists, a dummy variable for if the player was bought by a Chinese league team, year, and to dummy variables that represent if the player was a forward or midfielder, with defenders being the excluded group. The year variable, as mentioned earlier, is used to control for the trend of the data over time, which covers the issue of transfer fee inflation. The regression equation is shown below, labeled equation 1, with the standard error for each coefficient below it, with \* representing significance at the 10% level, \*\* representing significance at the 5% level, and \*\*\* representing significance at the 1% level:

*Table 2: Regression results predicting transfer fee (in millions of Euros)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | | Model 3\* | Model 4\* |
| Intercept | -11,070 | -11,020 | | -236.0 | -235.5 |
|  | (3102)\*\*\* | (3043)\*\*\* | | (49.95)\*\*\* | (48.85)\*\*\* |
| age | 0.6329 |  | | -0.018 |  |
|  | (0.762) |  | | (0.012) |  |
| rank | -0.2528 | -0.248 | | -0.004 | -0.0039 |
|  | (0.118)\*\* | (0.134)\*\* | | (0.0019)\*\* | (0018)\*\* |
| appearances | 0.069 |  | | 0.0034 |  |
|  | (0.319) |  | | (0.0051) |  |
| goals | 0.543 | 0.406 | | 0.0081 | 0.0062 |
|  | (0.330) | (0.227)\* | | (0.0053) | (0.0036)\* |
| assists | 1.846 | 1.709 | | 0.027 | 0.026 |
|  | (0.405)\*\*\* | (0.363)\*\*\* | | (0.0065)\*\*\* | (0.0058)\*\*\* |
| CSL | 4.416 |  | | -0.011 |  |
|  | (10.02) |  | | (0.061) |  |
| year | 5.515 | 5.481 | | 0.119 | 0.119 |
|  | (1.54)\*\*\* | (1.509)\*\*\* | | (0.028)\*\*\* | (0.024)\*\*\* |
| Position\_genFWD | -6.926 |  | | -0.109 |  |
|  | (8.368) |  | | (0.135) |  |
| Position\_genMID | -8.631 |  | | -0.120 |  |
|  | (6.472) |  | | (0.104) |  |
| N | 103 | 103 | | 103 | 103 |
| adj. R-sq | 0.2813 | 0.2954 | | 0.3146 | 0.3320 |
| F-statistic | 5.436 | 11.69 | | 6.201 | 13.67 |
| SSR | 21.39 | 21.18 | | 0.3444 | 0.34 |
| Standard errors in parentheses \* p<0.1, \*\* p<0.05, \*\*\* p<0.01 | | | \**In models 3 and 4, the dependent variable (transfer fee) is logged* | | |

The base model, shown as model 1 in table 2 above, was run with all variables. It gives an adjusted R-squared value of 0.2813, meaning that 28.1% of the variation in the transfer fees can be determined by the independent variables in this model. The coefficients for the constant, assists, and year of transfer are all significant at the 1% level, while FIFA rank is significant at the 5% level. Every rank higher (closer to 1) a player’s nation is in the FIFA ranking suggests an increase in his fee by around a quarter of a million euros. For each additional assist a player logs during the previous season, his transfer fee increases by 1.8 million euros. Finally, every additional year in the dataset increases the transfer fee by about 5.5 million euros. This shows that inflation occurs across the years of my dataset.

To test for heteroscedasticity, I ran a Breusch-Pagan test, which gives a BP test statistic and, depending on the degrees of freedom, gives a p-value determining if heteroscedasticity exists in the model. For my model including all outfield players, the test returns a p-value of 0.0041, which means that heteroscedasticity exists. In addition, I graphed the residuals vs the fitted values from my linear model to provide additional evidence that heteroscedasticity is present in this model, as shown in Figure 2. Since the red line does not line up with the dotted line showing consistent residuals, my data needs to be adjusted.



*Figure 3: Graph of residuals for predicted transfer values from a linear model using all outfield players*

To adjust for heteroscedasticity, I ran a generalized least squares regression. This regression estimates unknown variance in from the base linear model to normalize the residuals (Kariya & Kurata, 2004). After running the generalized least squares estimator, the r-squared value increase to 0.365, a slight improvement over the original. All of the variables had the same sign and significance levels, except number of goals scored is no longer significant at the 10% level. The model with the highest adjusted r-squared was Model 4, with the value being 0.332.

Since I was unable to find data that I could use that had detailed individual statistics such as passing or defensive statistics, my model cannot accurately judge the quality of midfielders and defenders. So, I decided to run a model using only forwards from my dataset. Forwards make up the largest percentage of my dataset by generalized position, with 42 of the 125 players listed as forwards. Because the main role for forwards is to score and assist goals, the variables in the dataset describe forwards more accurately. This model ran with better results, and is shown in table 3 below:

*Table 3: Regression results predicting transfer fee (in millions of Euros) for forwards*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | | Model 3\* | Model 4\* |
| Intercept | -17570 | -18070 | | 319.3 | 321.8 |
|  | (6563)\*\* | (6412)\*\*\* | | (98.45)\*\*\* | (94.62)\*\*\* |
| age | -1.254 |  | | -0.0097 |  |
|  | (1.585) |  | | (0.0238) |  |
| rank | -0.304 | -0.307 | | -0.0051 |  |
|  | (0.222) | (0.218) | | (0.0033) |  |
| Ln(rank) |  |  | |  | -0.137 |
|  |  |  | |  | (0.055)\*\* |
| appearances | -0.212 |  | | -0.00046 |  |
|  | (0.686) |  | | (0.0103) |  |
| goals | 1.033 | 0.706 | | 0.016 | 0.0096 |
|  | (0.595)\* | (0.434) | | (0.0089) | (0.0065) |
| assists | 3.304 | 3.026 | | 0.039 | 0.039 |
|  | (0.881)\*\*\* | (0.780)\*\*\* | | (0.013)\*\*\* | (0.011)\*\*\* |
| year | 8.735 | 8.969 | | 0.160 | 0.161 |
|  | (3.253)\*\* | (3.181)\*\*\* | | (0.049)\*\*\* | (0.0474)\*\*\* |
| N | 42 | 42 | | 42 | 42 |
| adj. R-sq | 0.339 | 0.3624 | | 0.3318 | 0.4186 |
| F-statistic | 4.504 | 6.826 | | 4.393 | 6.887 |
| SSR | 28.36 | 27.85 | | 0.4254 | 0.4147 |
| Standard errors in parentheses \* p<0.1, \*\* p<0.05, \*\*\* p<0.01 | | | \**In models 3 and 4, the dependent variable (transfer fee) is logged* | | |

This forward only model gave better results compared to all outfield players, with Model 4 having an adjusted R-squared of 0.3648. I decided to use to natural log of the rank variable because there are outliers in this variable, and logging this variable helps to stabilize and remove these outliers. In this model, assists, year, and the intercept are significant at the 1% level, the FIFA rank is significant at the 5% level, and all remaining variables are not significant at conditional levels. A Breusch-Pagan test on the forward-only model returned a p-value of 0.007, showing that heteroscedasticity does exist in this model at the 1% confidence level. Figure 3, shown below, shows the non-normal distribution of the residuals of the predicted transfer fees, providing further evidence that heteroscedasticity definitely exists in this model.

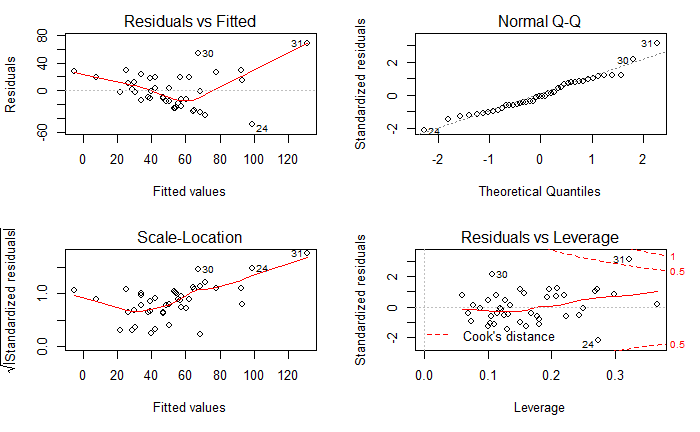


Figure 4: Fitted transfer fee vs residuals for forwards

However, running a generalized least squares regression gave an adjusted r-squared value of 0.3574, which unfortunately is not an improvement over the original model. Therefore, I will refer to the ordinary least squares model for my analysis.

Based on model 4 above, every higher FIFA rank leads to a player’s average transfer fee increasing by 13 percentage points. Each additional goal scored in the previous season increases a forward’s average transfer fee by 1 percentage points. Each additional assist a forward has in the previous season leads to a 4 percentage point increase in that player’s average transfer fee. Finally, each additional year in the dataset increases a forward’s average transfer fee by 14.7 percentage points.

Over the course of the current (2018/19) season, one of the biggest transfer rumors is the transfer of Chelsea’s star forward, Eden Hazard, to Real Madrid (Simpson, 2018). This season, the 28-year-old Belgian player has scored 17 goals and has 12 assists in 41 appearances across all competitions (Transfer Records, 2019). Currently, Belgium sit as the top ranked nation in FIFA’s rankings (FIFA.com, 2019). So, if this transfer were to occur in 2019, by plugging in the above stats in to the model for forwards, Hazard’s predicted transfer fee would be 94.38 million euros. Considering Chelsea’s asking price is 100 million euros (Mokbel, 2019), this is a very accurate prediction of Eden Hazard’s market value. However, there are many unaccounted for variables in my model, so this prediction should be taken with caution.

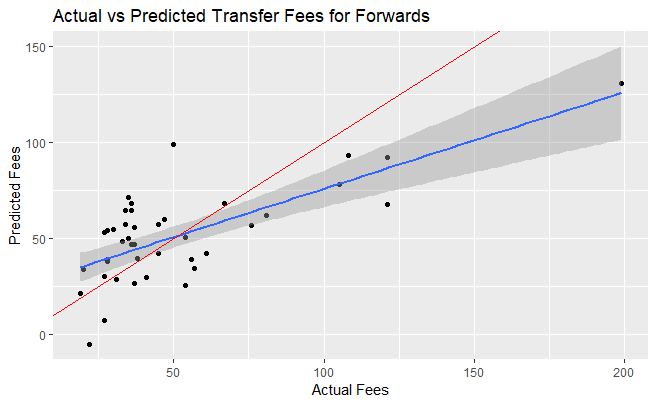


Figure 5: This figure graphs the actual vs. predicted transfer fees for my linear model that included only forwards. The blue line represents the best fit line from my model, while the red line represents y=x.

Figure 4, shown above, compares the actual transfer fee to the transfer fee predicted from my linear model for forwards. The blue line represents the regression line for these values, while the red line represents the line y = x. If my model was perfect, the predictions would fall right along this red line. However, my model’s line of best fit’s slope is not as steep, showing that my model underestimates fees for larger actual fees, and overestimates small fees. This could be due to the bias in my data, which I will discuss in more detail in the next section.

## K-Nearest Neighbors

The next step of my analysis is to perform a k-nearest neighbors regression to predict transfer fees. I used the same dependent variables here that I used in my OLS regression: age, FIFA rank, appearances, goals, assists, and year. Since k-nearest neighbors using a constant k to find nearby data points, I had to run several models to find an optimal k. Figure 5 below graphs the chosen value of k verses the mean squared error for that iteration. The minimum mean squared error occurs around a k value of 12 or 13. I decided to move forward in my analysis using 12 for my k-nearest neighbors models.

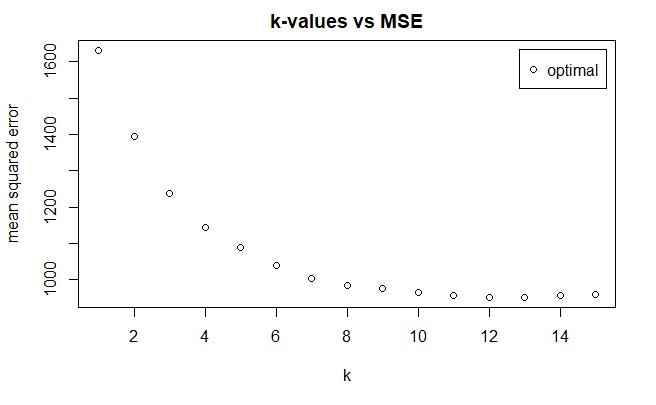


Figure 6: Finding an optimal k-value by minimizing the MSE

Once I had an optimal value for k, I began the k-nearest neighbors regression. With an optimal k value of 12, this technique takes a data point from my test set and find the 12 other data points that are closest in Euclidean distance, which is a sum of the differences between every variable (Thirumuruganathan, 2010). Then, to calculate the predicted values, I trained a model on 80% of the data, with the last 20% of the dataset to be used to test the model and return summary statistics. I ran the k-nearest neighbors regression on my test set, which produced figure 6 below, which shows the actual transfer fees versus their predicted values for my test set:

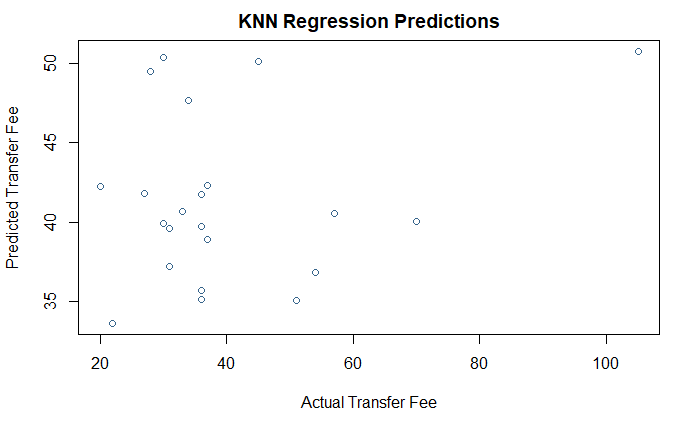


Figure 7: Actual vs predicted transfer fees for knn regression

The k-nearest neighbors regression method in R does not return the r-squared value. So, I had to calculate it by hand using the predicted and actual values from my test set using equation 1 below. In this case, the x values were the actual transfer fees while the y values were the predicted fees from my k-nearest neighbors regression.

*Equation 1: Calculating R^2 manually*

This returned an r-squared value of 0.238, meaning that 23.8% of the variation in a player’s transfer fee can be explained by the variables in the regression model. Compared to the generalized least squares regression that included the same dataset and players, which had an r-squared value of 0.365, this model is not an improvement. I decided not to run a k-nearest neighbors regression on only forwards, because the subset of players that are forwards is only 42, so taking 20% of this would only be about 8 players to run my regression on, which is too small of a sample size to get reliable results.

# Discussion

The goal of my research was to create a model that accurately predicts the transfer fee for professional soccer players based off the following factors: age, the FIFA world ranking of the nation the player represents, goals, assists, total games played, and the year of the transfer. I ran two separate linear models: one for all non-goalkeepers and one for only forwards. I also ran a generalized least squares regression for each model to correct for heteroscedasticity in the model. This yielded an improvement in the r-squared of the first model, but not the model that included only forwards.

Although my best model, with an r-squared of 0.504, was not a perfect predictor of players’ transfer fees, it still confirmed many of my predictions for how each of my dependent variables would impact transfer fees. My model showed that younger players drew larger transfer fees, holding all other variables constant. This makes sense since young players have many years left to play and often have room to improve before they reach their prime. My hypothesis that players representing higher ranked nations in the FIFA national rankings would draw larger fees was also confirmed to be true. Also, as predicted, scoring more goals and providing more assists increases players’ average transfer fees significantly.

My research also revealed what was lacking in my model. Some potential sources of bias in my model include limited sample size, not enough variables, and data issues. As with most research, more data is always better. Because of the limitations of web scraping and the Transfermarkt website, I was only able to collect data on 25 players from each year, which were the 25 most expensive. My dataset consists of data from the years 2014-2018. I decided to use this time frame because 2014 was where transfer fees started seeing a sharp incline in their amounts. For example, the most expensive transfer in the 2012/13 season defender Thiago Silva for 37.8 million euros (Transfer Records, 2019). The drastic change in transfer fees since then led me to believe that including years further in the past would bias the data even more. If I were able to collect more than 25 players for each year, I would be able to accurately predict the fees of less valuable players. In addition, it is important to note that the players in my dataset are all players who have been transferred to a new club in the past 6 seasons. The subset of players who have been transferred is not a representative sample of all professional soccer players. For example, Lionel Messi has been with FC Barcelona in Spain since 2001, when he was just 13 years old (Transfer Records, 2019). Loyal players like Messi are not represented in my dataset, because they have no transfer fee to analyze.

One of the biggest downfalls of my research was the lack of enough variables, especially regarding individual statistics. Originally, I had planned to use very in-depth statistics such as passing percentage, key passes, interceptions, and tackles to accurately represent a player’s ability. With my only individual player statistics being appearances, goals, and assists, my model was inaccurate, especially for midfielders and defenders. Unfortunately, I was unable to find a resource with this level of detailed statistics that I had permission to use.

In addition, there are many other non-playing characteristics that can influence a player’s transfer fee. I could not find data on the contract stipulations of players at the time of their transfer. I wanted to use the number of years remaining on a player’s contract when he was transferred as a dependent variable in my model. I believe that, if a player’s contract was expiring soon, his club would be willing to accept a lower fee rather than let him go for free once his contract is up. Another contract stipulation that could have influenced transfer fees is whether a release clause was paid. A release clause is a fee that is set in a player’s contract, that if a club offers this amount to the player’s current club, the current club must accept, and the player will be transferred if the player himself agrees to the move. Release clauses speed up the timeline of a transfer but can lead to inflated prices. KPMG Football Benchmark, a division of financial advising company KPMG focused on soccer, discusses how the idea of a release clause can increase a player’s transfer fee (2018). Another factor that could have an effect would be how long a player has been with their current club. As I mentioned earlier, Messi has been with his current club for 18 years. Because of his loyalty and extraordinary talent, it would take a massive fee for Barcelona to be willing to sell him.

This method of predicting player transfer fees is extremely important for clubs during the transfer window. A strong transfer window can turn a team from a mid-table finisher into a title contender. In today’s data-driven world, using hard, objective statistics to evaluate players has become a reality. Clubs now have the ability to quickly analyze which players they are interested in signing before having to send scouts to observe players in person. Using models that evaluate players’ abilities will save clubs time and money when searching for a new player during the transfer window. And although there is no way to numerically represent every quality about a player, these predictive modeling techniques are still a very important method for signing new players, and the introduction of more data into the world of soccer will help similar models’ popularity in the future.

# Acknowledgements

I would like to express my gratitude to Dr. Sarah Supp for her continued support and valuable feedback throughout the semester. Her advice paved the way to make this project possible.

I am also very grateful for all of my fellow Data Analytics 401 students whose feedback, participation, and support helped me get through the most difficult times of my research, and I am excited to see what the all accomplish after graduation.

# Appendix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Rank | Age | Appearances | Goals | Assists |
| Rank | 1.0 |  |  |  |  |
| Age | 0.099 | 1.0 |  |  |  |
| Appearances | 0.187 | 0.133 | 1.0 |  |  |
| Goals | 0.189 | 0.253 | 0.249 | 1.0 |  |
| Assists | 0.191 | 0.075 | 0.273 | 0.289 | 1.0 |

Table 3: Correlation matrix of continuous variables to test for multicollinearity

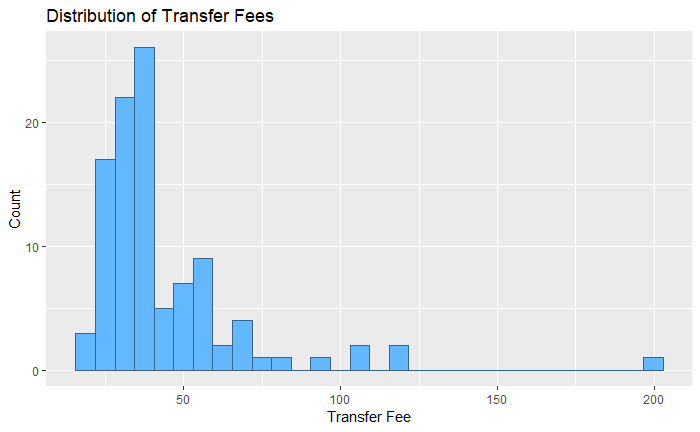


Figure 8: Histogram showing distribution of transfer fees.

# References

Assumptions of Linear Regression. (2019). Retrieved from https://www.statisticssolutions.com/assumptions-of-linear-regression/

Dobson, S., & Gerrard, B. (1999). The Determination of Player Transfer Fees in English Professional Soccer. *Journal of Sport Management*, 13(4), 259-279. doi:10.1123/jsm.13.4.259.

Fitzgerald, T. (2018, June 08). FIFA Soccer Rankings. Retrieved from https://www.kaggle.com/tadhgfitzgerald/fifa-international-soccer-mens-ranking-1993now

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), 788-797. doi:10.1093/bib/bbt026.

Hechenbichler, K., & Schliep, K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. *Institut Fur Statistik*, 386(399). Retrieved from https://epub.ub.uni muenchen.de/1769/1/paper\_399.pdf.

Kariya, T., & Kurata, H. (2004). Generalized least squares. Hoboken, NJ: Wiley.

KPMG. (2018, September 18). Player Valuation: Putting data to work on transfer market analysis. Retrieved from https://www.footballbenchmark.com/library/player\_valuation\_putting\_data\_to\_wor \_on\_transfer\_market\_analysis.

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). doi:https://doi.org/10.7206/jmba.ce.2450-7814.177.

McKinney, W. (2018)*. Python for Data Analysis: Data Wrangling with pandas, NumPy, and IPython*. Sebastopol, CA: OReilly Media.

Mokbel, S. (2019, April 08). Real Madrid won't meet Chelsea's £100m asking price for Eden Hazard but will ask him to force dream. Retrieved from https://www.dailymail.co.uk/sport/football/article-6898489/Real-Madrid-wont-meet Chelseas-100m-asking-price-Eden-Hazard-ask-force-dream.html

Python (Version 3.6.2). (2017). Retrieved from https://www.python.org/downloads/release/python-362/

R (Version 3.4.1). (2017). Retrieved from https://cran.r-project.org/bin/windows/base/old/3.4.1/

Rapp, T. (2017, October 03). Premier League Homegrown Rules: Explaining EPL Player Quotas. Retrieved from https://bleacherreport.com/articles/2137974-premier-league-home-

grown-rules-explaining-epl-player-quotas

Simpson, C. (2018, June 18). Eden Hazard: Real Madrid 'Know What They Have to Do If They Want Me'. Retrieved from https://bleacherreport.com/articles/2781619-eden-hazard-real madrid-know-what-they-have-to-do-if-they-want-me

Thirumuruganathan, S. (2010, May 18). A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. Retrieved from https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction to-k-nearest-neighbor-knn-algorithm/

Transfer Records. (2019). Retrieved from https://www.transfermarkt.co.uk/transfers/transferrekorde/statistik?saison\_id=2017&lan \_id=0&ausrichtung=&spielerposition\_id=&altersklasse=&leihe=&w\_s=+=1

Transfermarkt: Privacy Policy. (2018, May 25). Retrieved from https://www.transfermarkt.co.uk/intern/datenschutz

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising*. Artificial Intelligence Research*, 2(1). doi:10.5430/air.v2n1p44.

Wayne Rooney - Player Profile 2019. (2019). Retrieved from <https://www.transfermarkt.us/wayne-rooney/profil/spieler/3332>