American University of Armenia

Capstone Project

Analyzing soccer's transfers and predicting footballers' transfer price

Narek Sahakyan
Tigran Avetisyan
Hayk Avetisyan
Ashot Khan-Aslanyan

Habet Madoyan

June 15, 2020

Abstract

The purpose of this project was to create a machine learning-based model which could predict the transfer fee of football players based on some range of data. Nowadays, concepts of AI are widely used in almost every sphere. Many football clubs use such techniques to maximize their profit. The paper demonstrates one way to develop such a technique. The first phase of the project included data collection for having a high precision model. Then various machine learning algorithms were applied such as linear regression, decision tree, deep neural networks, and others. The models were evaluated based on some specific measures. After model evaluations, respective conclusions and interpretations were made.

Keywords: Machine Learning, Deep Learning, Soccer, Sports Analytics, Regression, Networks Science, Transfers

Contents

1 Introduction

Nowadays, soccer teams buy and sell thousands of players during each transfer window, spending millions of dollars to buy them. The top players of the game are even worth a couple of hundred million dollars. With the increase in resources spent on football players during transfers, football teams aim to maximize the efficacy of each transfer. Some football teams use AI-based technologies to predict football players' transfer fees and market value. According to an article published on BBC, if a player transfers before their contract expires, the new club pays compensation to the old one. This is known as a transfer fee (Quick, 2017). While a player's market value is an estimate of the amount for which a team can sell the player's contract to another team (Herm, Callsen-Bracker, & Kreis, 2014). The market values attached to the players do not play a key role as in many cases a player is sold for a much higher price than his market value or much lower price. At the moment of his transfer from Barcelona to PSG Neymar's market value was 100 million euros, but PSG paid more than double the price to sign him. So, what are the main qualities of the player or other factors, that decide his transfer price? Also, when is the best time to sell the player? Which of the transfers paid off for the teams? And in general, are there any special connections among the teams or leagues of the transfer market? Those are some of the questions that we are going to provide answers to. We want to use simple statistical measures and ratios (such as goals, minutes per one goal and etc.) of the players' performance for predicting his transfer price. One of our main goals is to identify whether the easily interpretable statistical measures of the players are solid predictors for his transfer fee and how well can these variables predict the player's price. As the importance of a statistical measure varies depending on the position of the player on the field, we will implement the predictions for each position separately. In the end when we find the best prediction model we can investigate the underrated and overrated players based on their transfer fee.

1.1 Literature Review

Various studies and projects conducted by individuals or organizations for different purposes, which tried to investigate the features that may impact the transfer fee of the football players and conduct predictive models and methods.

First of all, there is a study conducted by CIES Football Observatory. They published a document called "Scientific assessment of football players' transfer value" in October 2018. The study tried to understand the predictive perspective of the transfer values of the footballers, mentioning that there is some predictable logic which is possible to model. The data that they have collected consists of over 2,400 transfers, involving top-5 league players between July 2011 and August 2018. The number of features is 36 including information about footballers' contract duration, year of transfer, book value, loan status, nationality and economic level of the releasing club. They used multiple linear regression which included only significant features, leading to the overall model to be very significant ($p - value < 0.00001$) with adjusted coefficient of determination evaluated as 0.86. The study concluded that the model is useful for defining the starting value, initial salary of the players, as well as understanding the value of the club in the future based on the price of the players (Polim, Ravenel, & Besson, 2018).

The next approach to understand the players' transfer value was done by Lukas Barbuscak with the study called "What Makes a Soccer Player Expensive? Analyzing the Transfer Activity of the Richest Soccer" in 2018. The data that the author used consists of quite a small amount of 49 players, including features such as number of google searches, number of years on contract left, rating of the players by the community, number of goals and assists and finally race of the players. The sample consists of the most valuable players on the market by the year of 2018. The models that he used are the two multiple linear regressions, which concluded that the most essential independent features to predict the transfer value are Contract Left and Race. First model included all the players, however the second one excluded Neymar, since he was the most valuable player at that time. The models are quite similar in terms of the adjusted coefficient of determination evaluated as 0.92 (Barbuscak, 2018).

Another study we examined was "Football player's performance and market value" published by Ricardo Cachuchino, Miao He and Arno Knobbe, published in 2015. The authors of the paper wanted to understand the relationship between the performance of the player and his market value. They also underlined the current problem of player's economic valuation and deviations in the market values and transfer fees. Their dataset included information about the player's performance, his ratings from WhoScored, market information from TransferMarkt and some other performance assessment metrics gained by juries. The scope of the project only included LaLiga players for the half of the 2014-2015 season. The

authors built their model of evaluating the market value using Lasso Regression. They emphasized the importance of choosing the right lambda in filtering the important features. Later they figured out the importance of evaluating the players based on their position and continued their work putting the main emphasis on evaluating the forwards. They came out with a simple linear model for evaluating the forward's performance based on his behavior on the field (fouls, yellow or red cards), shots from different areas, goals (underlining the goals from penalty area as a big plus), successful dribbles and other attributes relevant to the forwards position. Later, they studied the relationship of the performance and market value and came up with a conclusion that economical valuation of the player is dependent on his performance but his performance cannot be affected by economic factors (market value, transfer value). They also indicated that the undervalued or overvalued players can be found using the difference between the real and estimated market values of the player (M. He, Cachucho, & Knobbe, 2015).

We also found another similar study that tried to predict the market value of the footballers with linear models. This study included only EPL data about footballers in the season 2017-2018. It also included the ratings of the player from Fantasy Premier League and also the number of the player's Wikipedia page views. The overall formula of their model was based on the ability of the player and his performance. They regarded the number of page views as a proxy variable for ability. The model had R squared a little bit higher than 0.7. Overall, this model emphasized the importance of the proxy variables for accounting the player's ability (Maurya, 2018).

Another similar project was made by Yuan He. This project was the first case where the data was collected not only for one season, but for 5 seasons. They divided the data into multiple parts, one for the player's personal information such as his nationality year of birth, race, height, position and other similar attributes, and also, they had a dataset aligned for the player's performance metrics such as the number of games played, the number of goals, the number of clean sheets for the goalkeepers and other metrics. Also, this project included the national stats of the players. Yuan Hee's project used mainly two models, OLS (ordinary least squares), KNN (k nearest neighbors) and also Ridge, Principal Component Regressions. The authors used 10-fold cross validation techniques and used $RMSE$ as a main criterion for judging their models. The authors also added various performance ratios after the EDA, such as international caps to age and other interconnected attributes. According to the author the PCR model with a K value of 15 was the best among the chosen options. The authors advanced with that model in later developments. The authors started to do predictions and check the accuracy of the model using multiple approaches. The approaches chosen were.

- Taking overall mean as a sole prediction. $RMSE \approx 54.7$
- Taking the responsive model for each player. $RMSE \approx 27.27$

- Using value from last year. $RMSE \approx 25.64$
- Training original data matrix without cross validation. $RMSE \approx 21.27$

The authors concluded that the last year value of the player had the highest contribution to the response. Goals, Assists, International caps and other similar attributes also had high contribution, while the personal information of the player and the ratios added by the authors did not have high significance (Y. He, 2012).

There were also other similar studies done, but the results and scopes were in general similar to the mentioned works. Also, initially trying to find similar projects' implementation in other team sports, we did not find any similar projects in scope. The reason is that except soccer there are multiple team sports based in the USA, that have huge finances involved in the game. Those are Basketball, Baseball, American Football and others. However, in almost all of these sports, there are some limitations included on the team's budget from the according sport's federation of the country. A great example is the application of salary limits of the players in almost all the American team sport games. In addition to this in many sports the teams are getting their players from the junior leagues, and in baseball's case the worst team has the ability of picking the best junior player first. So, in general the "trade" practices in American sports and their occurring problem solutions will not be helpful for analyzing soccer's market.

Overall results

Year	Authors	Overall data scale	Methods applied	Final Results
2018	CIES Football Observatory	36 features, Top 5 leagues. Transfers from 2011 - 2018.	Multiple Linear Regression.	Predicting Economic level of the club: $R^2 \approx 0.54$ $RMSE \approx 0.138$ Transfer value: $R^2 \approx 0.86$ $RMSE \approx 0.203$
2018	Lukas, Barbuscak	49 transfers, 5 variables. The year 2017.	Multiple Linear Regression.	$R^2 \approx 0.92$
2015	R.Cachuchino ,A.Knobbe	Player's info, market value, performance data. Season 2014/2015.	Lasso Regression	N/A
2018	Shubham Maurya	Top 6 EPL, 2018. Personal info, market value, and google page views. FPL metrics.	Multiple Linear Regression	$R^2 \approx 0.7$
N/A	Yuan He	357 players, one season data. Personal information, performance metrics.	OLS, KNN, Ridge Regression, Principal Component Regression.	$RMSE \approx 43.29$

Table 1: Previous work summary

Most of the approaches used a small number of observations. Also, many projects used some external information about the players' performance such as FPL scores or WhoScored ratings. The most popular algorithm used in the mentioned projects was Multiple Linear Regression. Most of the projects used R^2 and $RMSE$ for estimating their model's predictive

power. The projects that provided the results had mostly pretty high R^2 (Table 1).

2 Methodology

2.1 Data Collection

Data collection was one of the most important phases of the project. As a model based on machine learning algorithms, a considerable amount of data was necessary to achieve high accuracy.

Data Scraping

After thorough research, we found all the necessary data on the Transfermarkt website [1]. Python modules were written in order to scrape data from the website. Overall, we obtained data about 18000 players from various leagues. Transfermarkt provided historical data of the player's performance since the beginning of his professional career and also his market value history starting from 2005. We also scrapped players' injury history and trophies from the Sofifa website [2], as this information was no structured in Transfermarkt. Later we connected the two sources by string distance algorithms in players' names, nationality, and clubs. The scrapping was done using python's Selenium and beautifulSoup packages.

Data Cleaning

The scrapped data contained a lot of messy information and gaps. We have cleaned up all the information and assigned the correct data types for each variable.Overall, there are around 4 categories.

- Datetimes(DOB of a player, date of a transfer, ...)
- Floats(points per game, minutes per goal, ...)
- Integers(height of a player, number of goals scored, ...)
- Categorical variables(strong foot of a player, continent of a player, ...)

[1] Transfermarkt website link
[2] Sofifa website link

2.1.1 Data Description

Main attributes

The main sources of information about the players are based on multiple categories each having according figures about the players' details.

- Player's physical,racial and playing attributes.
  - Height
  - Age
  - Strong foot
  - Continent
  - Position on the field
- Player's market value history
- Player's trophy history
  - Number of trophies won
  - Number of times player's team was a runner-up(second place)
- Player's performance attributes
  - Number of goals scored
  - Number of goals scored from penalties
  - Number of assists made
  - Number of yellow cards received
  - Number of second yellow cards received
  - Number of red cards received
  - Number of substitutions into the pitch
  - Number of substitutions off the pitch
  - Number of minutes played on the pitch
  - Number of elections into the team squad
  - Number of appearances in the starting eleven
  - Number of own goals scored

  - Number of clean sheets(only for Goalkeepers)
  - Number of goals conceded(only for Goalkeepers)
  - Points per game team earned when the player was playing

- Information about the transfer
  - Transfer's window type
  - Transfer's year type(Tournament year or not)

- Perfomance ratios
  - Winning percentage
    (Number of games won/Number of games played)
  - Playing percentage
    (Minutes played on the pitch/Minutes elected in the squad)
  - Minutes per goal
  - Minutes per assists
  - Minutes per yellow cards
  - Minutes per red cards
  - Minutes per conceded goals
  - Minutes per clean sheets
  - Number of clean sheets per conceded goals
    (Number of goals the goalkeeper has conceded per a game of not conceding any goals.)
  - Number of goals per penalty goal
    (Number of goals scored from penalty before scoring a game from the game)

Initially, we have also added a few more ratios underlying the frequencies of the players' substitutions both on and off the pitch, but the ratios contained a lot of infinity values due to often cases of 0 being the divisor. All of the performance metrics were used at the time of the transfer, so for example the age of the player was identified by substituting the date of birth of the player from the date of the transfers. In addition to the seasonal metrics of performance, the cumulative measures of the player were used at the time of the transfer. The operation used to accumulate the performance metrics were summation for all performance measures except points per game for which the running mean operation was used(Each two year's ppg were summed up and divided by two until the year of the transfer). The same operation was used for calculating the cumulative ratios.

Variable Abbreviations

Abbreviation	Description
g	Number of goals scored
a	Number of assists recorded
gc	Number of conceded goals(GK)
cs	Number of clean sheets(GK) (games with no conceded goals)
yc	Number of yellow cards
rc	Number of red cards
syc	Number of second yellow cards
son	Number of substitutions on to the pitch
soff	Number of substitutions off from the pitch
s	Number of selections into team squad
app	Number of appearances
mp	Minutes played on the pitch
ppg	Points per game the team earned when the player was on the field
mpg	Minutes per goals
pg	Number of goals scored from penalties
og	Number of own goals scored

Table 2: Abbreviations for seasonal perfomance variables.

The variables defined in Table 2, were available in Transfermarkt for all the players. The website contained historical seasonal performance data for each of the described performance measures for the players since the beginning of their career [3].

[3]The abbreviations for cumulative perfomance metrics were the same with **cum_** at the beginning

Abbreviation	Description
fmpct	Percentage of time spent on the field playing (mp)/(s * 96).
fwpct	Winning percentage of the team when the player was on the field (fmpct * ppg) / 3.
mpson	Per how many minutes the player was substituted on to the pitch (mp / son).
mpsoff	Per how many minutes the player was substituted off from the pitch (mp / soff).
apps	Per how many squad selections was the player playing on the pitch (s / app).
mpcs	Per how many minutes the goalkeeper had a game (90minutes), during which he didn't concede any goals (mp / cs).
mpgc	Per how many minutes does the goalkeeper allow a goal (mp / gc).
gcpcs	How many goals the goalkeeper has conceded per a game of not conceding any goals (gc / cs).
mpa	Per how many minutes does the player record an assist(pass after which a goal is scored) (mp / a).
mpyc	Per how many minutes does the player receive a yellow card (mp / yc).
mprc	Per how many minutes does the player receive a red card (mp / rc).
mprc	Per how many minutes does the player receive a red card (mp / rc).
ycprc	Per how many yellow cards the player has got a red card.(rc /yc)
gppg	How many penalty goals the player has scored per goal. (g / pg)

Table 3: Abbreviations for ratio variables.

The variables defined in Table 3 were all calculated using the variables in Table 2. The calculations required for calculating each ratio are described in Table 3. The minutes were chosen as a base metric as the number of games are strongly correlated with minutes

played, also the average length of a game is calculated using 96 minutes in order to take into account average delays in-game time due to extra-times and over-times [4].

Database structure

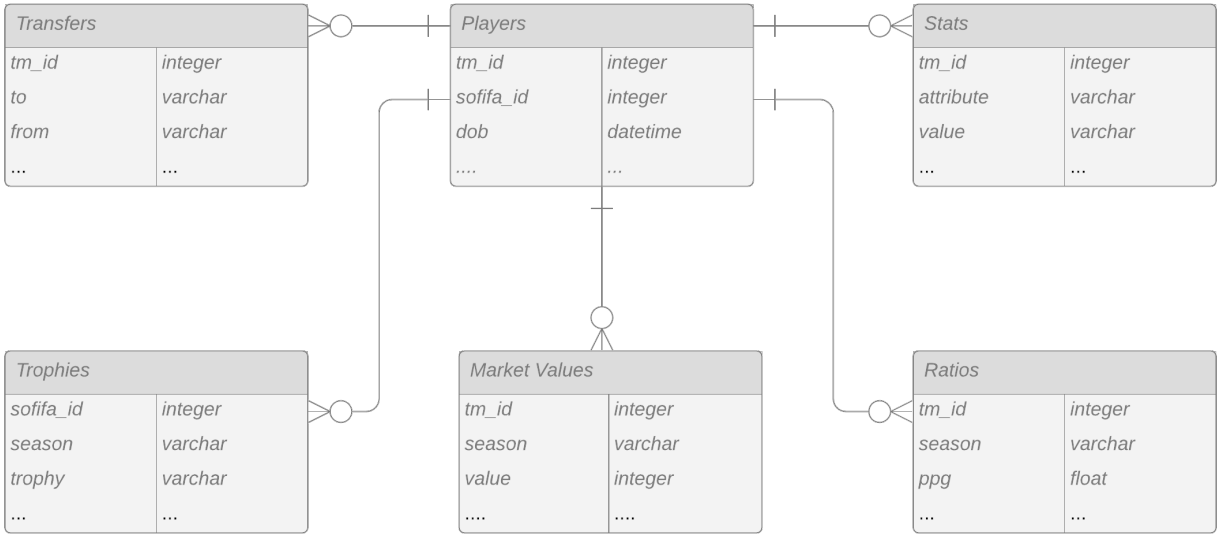

Figure 1: Entity Relationship Diagram

Figure 1 Shows the structure and relationships of the database. As no new records were added and all of the dataframes had a relationship only with Players' dataframe, we did not use any SQL framework. All of the entities were stored in a pandas dataframe and later the attributes that were relevant to the output variable were merged together into one dataframe for each position. The historical statistical data of the players was initially stored in a long format and later transformed into a wide format for analysis and modeling.

[4]The abbreviations for cumulative ratio metrics were the same with **cum_** at the beginning

2.2 Data Analysis

2.2.1 Exploratory Analysis

To identify the main characteristics of the players' that contribute or do not contribute to their transfer values and market values accordingly we have mostly used heatmaps, scatter plots, histograms, and boxplots each used with corresponding categorical and numerical variables. As there are 4 main positions for footballers, we have analyzed each position for finding the most relevant attributes specific to each position, that contribute to the transfer fee and market value of the players. However, while analyzing the performance ratios we have paid attention only to the ratios that make sense for the position in order to avoid missing values, as for example, the attackers receive yellow or red cards rarely thus they need more minutes for receiving a warning, which will lead to having a high number for the minutes per red or yellow card ratio for the player, thus having a correlation to the transfer price, and similar scenarios are present for each position. In order to avoid redundancy in data and missing values, only the ratios that are relevant to the position of the player were considered during the analysis. Also the early stages of the analysis showed that the player's injury history was not contributing to his transfer fee, so the information about the injury history was not later used.

Goalkeepers

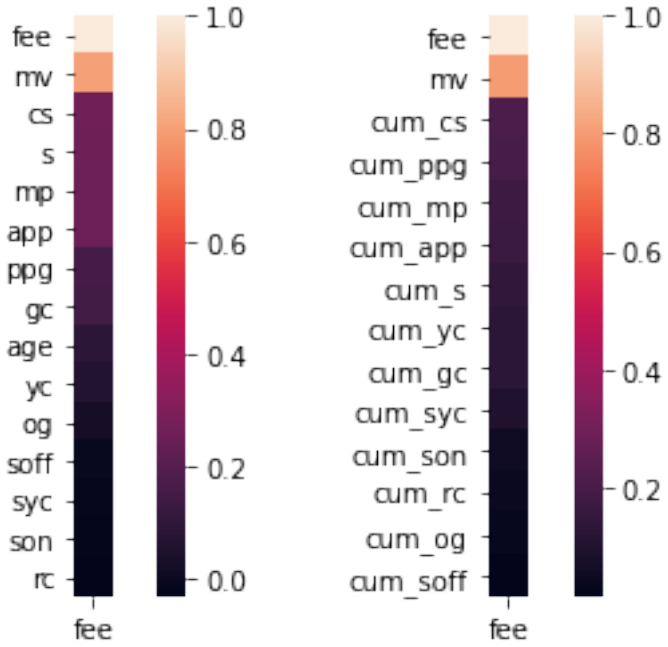

Figure 2: Price and statistics heatmap(Goalkeepers)

As we can see the market value of the player has the highest contribution to his transfer fee, and the statistical measures are not highly correlated with the transfer fee. The statistical measures of goalkeepers that have the highest correlation to their transfer price are the number of clean sheets the player made during the season of the transfers, the number of squad appearances, minutes played, the number of points per game the team earned when the player was on the field, and the number of goals conceded. When we take a look at the cumulative statistical measures at the time of the transfers we can see that most contributed ones to the transfer fee are the cumulative number of clean sheets, the cumulative value of points per game, minutes played, and appearances. However most of the statistical measures for the players are highly correlated to each other, as for example if the number of player's appearances on the field increase the number of minutes he played on the field increases correspondingly, and also we should take into account the ratios of some statistical measures as in general most of the measures are correlated to the number of appearances or minutes played for the players (Figure 2).

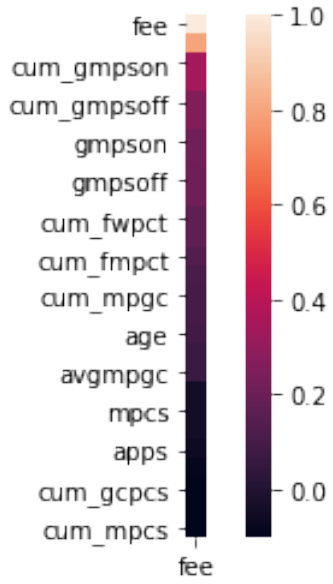

Figure 3: Price and ratios heatmap(Goalkeepers)

The statistical ratios that have the highest correlation to the player's transfer price are the cumulative and season based number of games per substitution on and off from the field, however, these metrics are not relevant to the goalkeepers as most of the time they are not often substituted off or on from the field. The ratios that are relevant to goalkeepers and have some contribution to the transfer price are the cumulative ratios of the field winning and playing percentage and minutes per conceded goal (Figure 3).

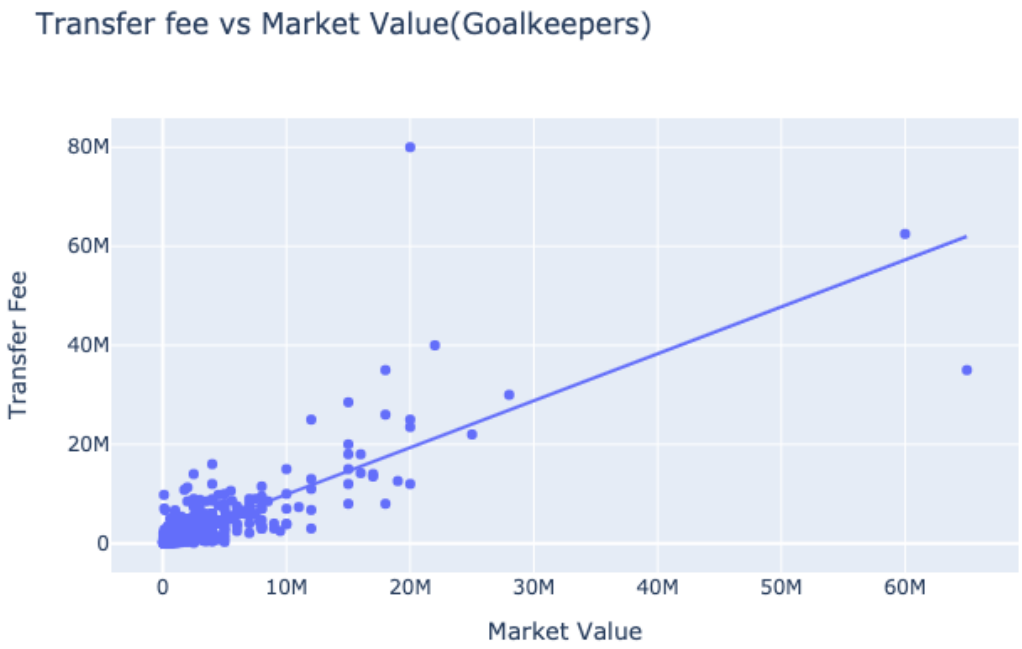

Figure 4: Transfer price vs market value(Goalkeepers)

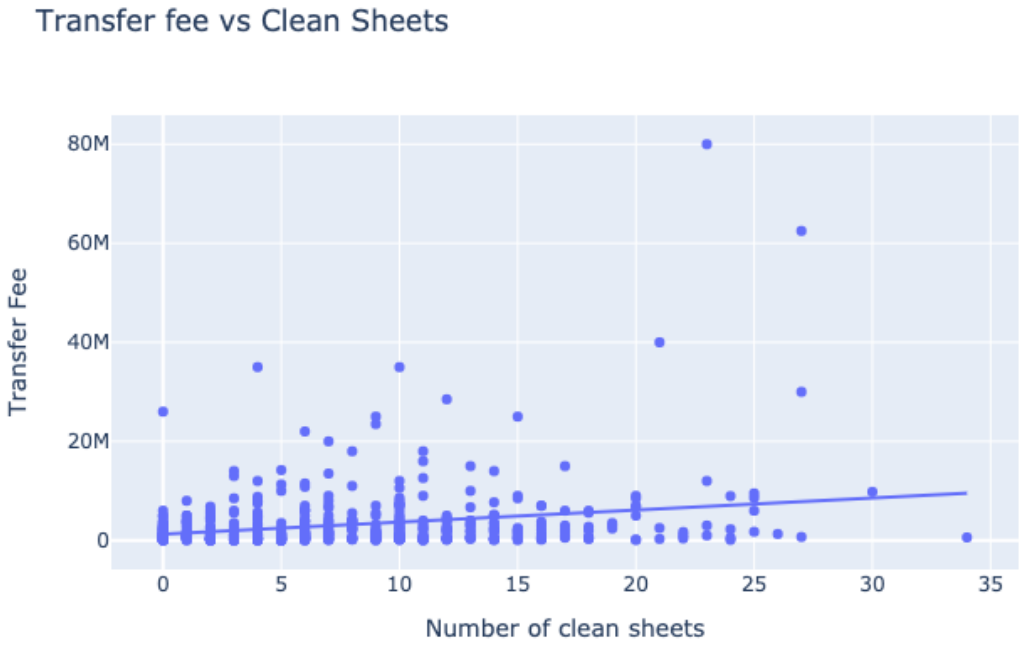

Figure 5: Transfer price vs clean sheets(Goalkeepers)

We can see that most of the time the transfer fee of the goalkeepers was lower than their estimated market values and there are only a few extreme cases when the difference between the transfer price and market values is significantly different, which happens mostly during transfers of young players or expensive players who have a few years left on the expiring contract (Figure 4). Also, the number of clean sheets tends to contribute to the transfer price, but the connection is not very strong as players as a goalkeeper may have conceded a lot of goals thus having a lower number of clean sheets, but also have many good save which can contribute to high transfer value (Figure 5).

Defenders

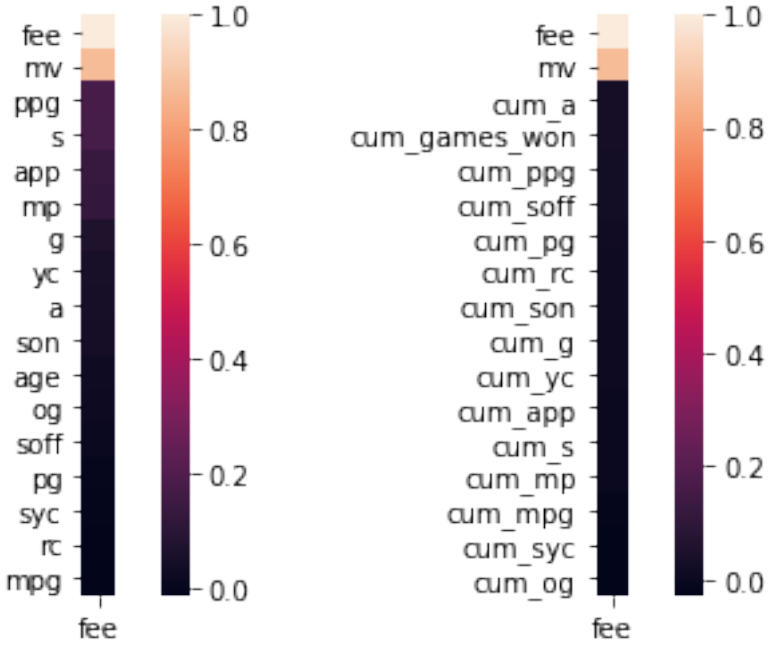

Figure 6: Price and statistics heatmap(Defenders)

The main attribute contributing to the transfer fee of defenders is also his market value, and the statistical measures have a very low correlation with the transfer fee of the player. The only season based performance metrics for the defenders that have a relatively higher contribution to their transfer fee in comparison with other metrics are points per game the team earned when the player was on the field, the number of appearances and the number of times the players was involved in the team's squad either as a starting eleven

player or substitute player. None of the cumulative statistical measures have even a relatively high contribution (Figure 6).

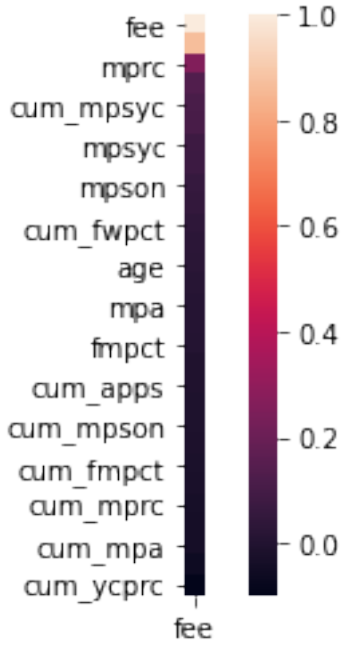

Figure 7: Price and ratios heatmap(Defenders)

For defenders, we can see that the highest correlation with the transfer fee is the minutes per red card statistic of the player, which shows on average how many minutes does the player does not a receive a red card after receiving one. As defender is a very aggressive position in soccer, it is excepted that defenders can receive many red or yellow cars, so a good indicator of a defender can be the interval of receiving red cards in minutes, and generally expensive players receive a red card more rarely. The ratio of yellow cards and red cards is a good indicator to find out, how many not so dangerous fouls does a player commit before committing a very dangerous foul, worth a red card. However, this ratio is not strongly correlated. Almost the same correlation value to the transfer fee has the winning percentage of the player when he is on the field. This is a little bit biased variable as the team's points are not dependent only on the defenders, they alongside the goalkeepers are responsible for the teams defending qualities and attacking attributes of the team is independent of the defenders' quality. (Figure 7)

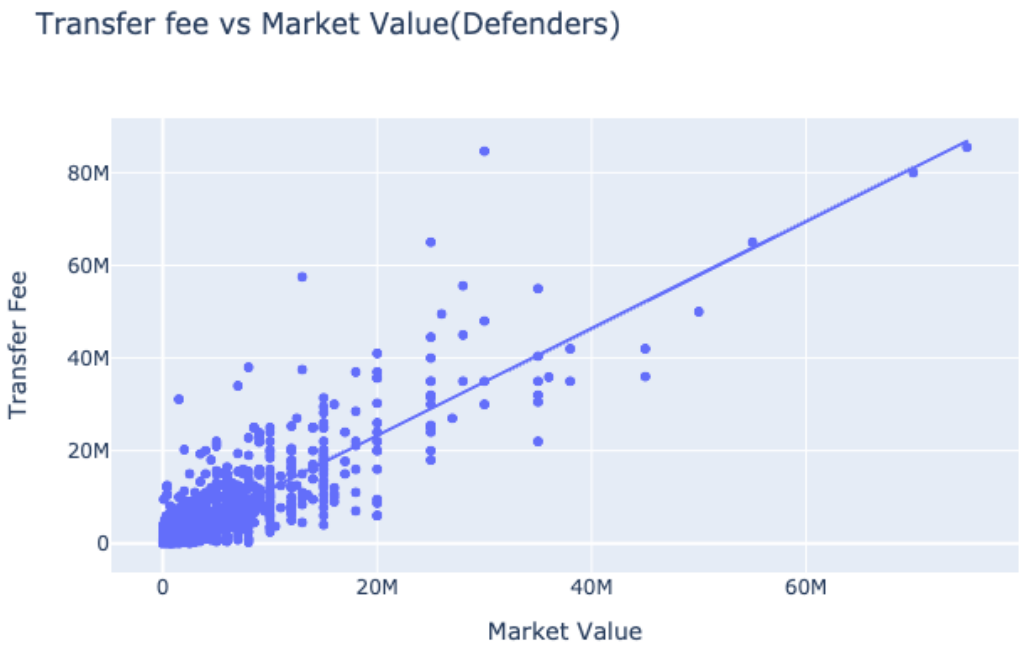

Figure 8: Transfer price vs market value(Defenders)

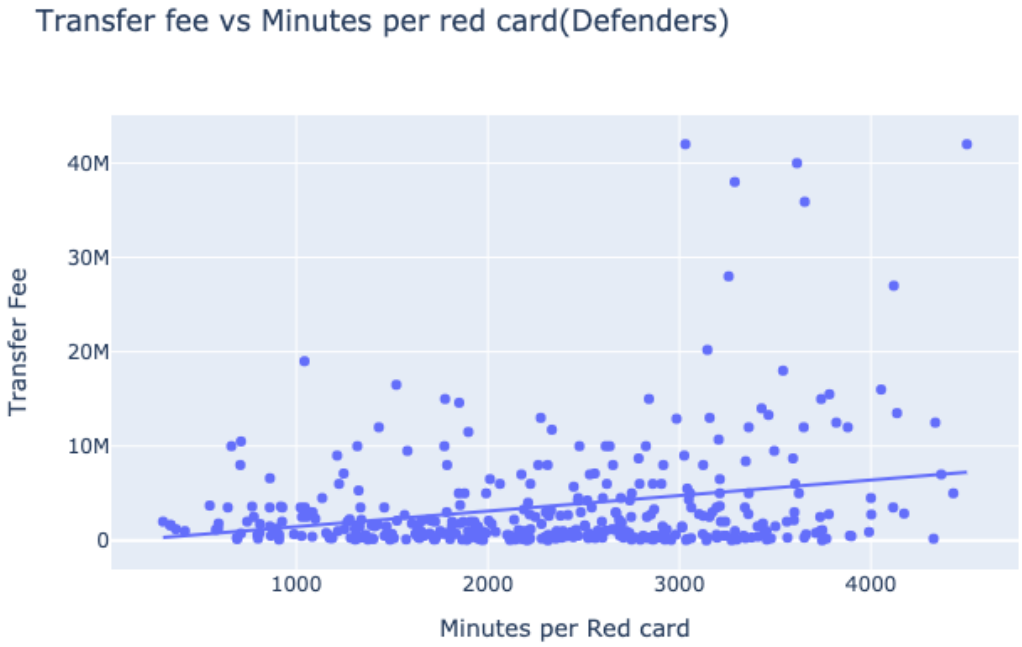

Figure 9: Transfer price vs minutes per red card(Defenders)

The number of defenders who were sold for a price higher or lower their estimated market value is pretty much the same (Figure 8).Also, we can see that the defenders who receive red cards not very often thus having a very high metric for minutes per red card are generally valued higher than those with a lower metric for minutes per red card who correspondingly receive more red cards on average (Figure 9).

Midfielders

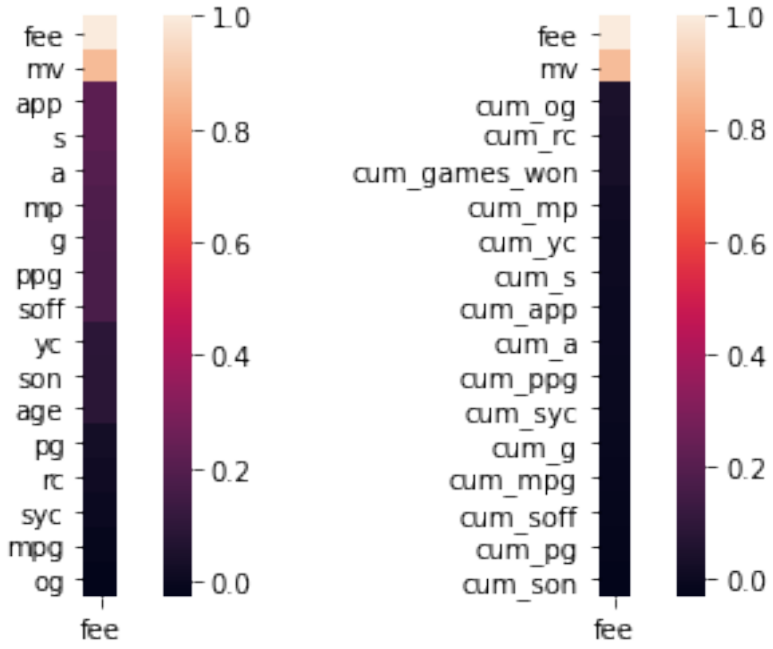

Figure 10: Price and statistics heatmap(Midfielders)

The seasonal statistical measures having the highest contribution to the transfer price are the number of appearances, squad selections, assists, minutes played, goals scored, and the average number of points gained by the team when the player was on the field. None of the cumulative statistical metrics have any contribution to the transfer fee of the midfielders (Figure 10).

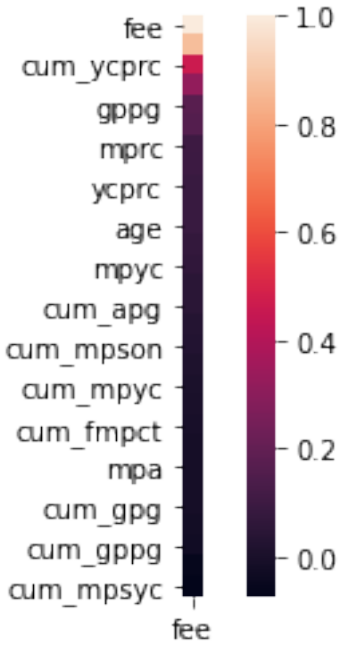

Figure 11: Price and ratios heatmap(Midfielders)

There are a few ratios of performance that have a contribution to the midfielders' transfer price, but the correlated attributes are not relevant to the position (Figure 11).

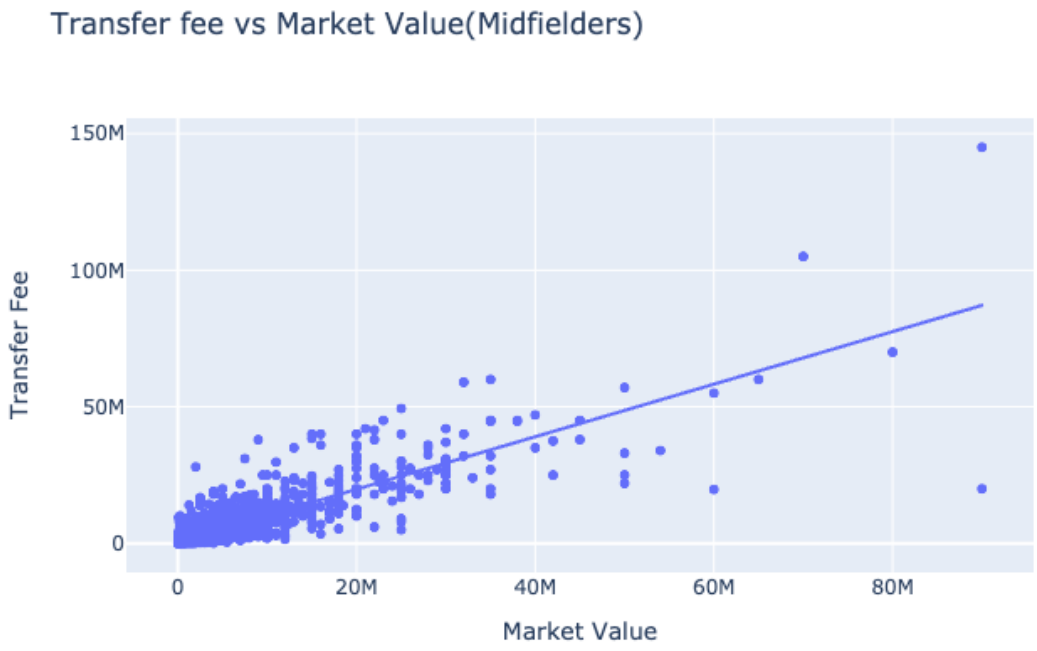

Figure 12: Transfer price vs market values(Midfielders)

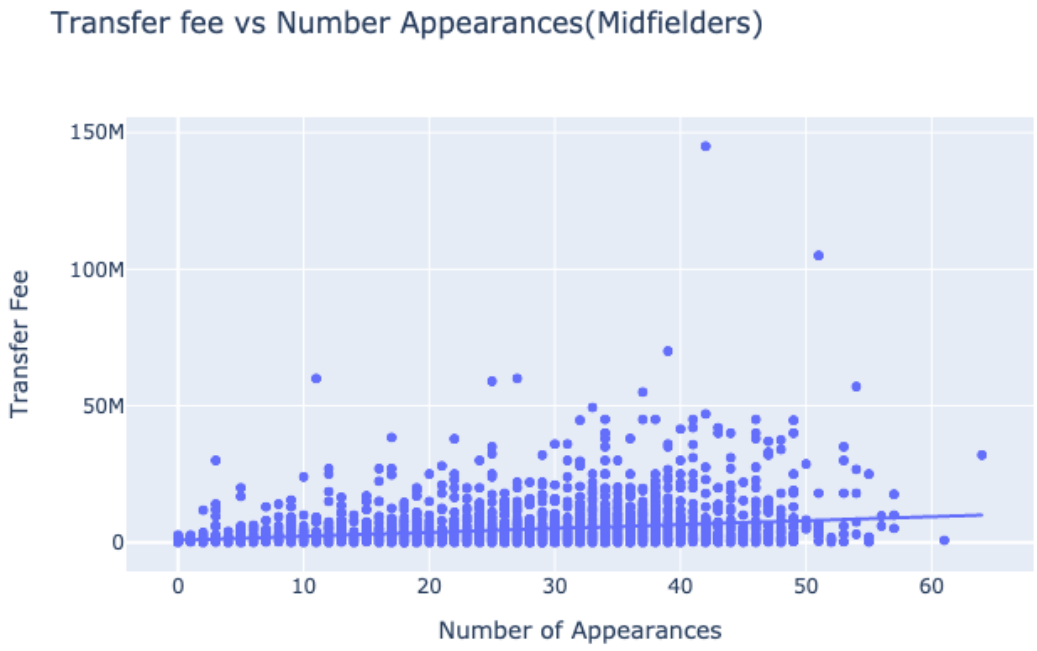

Figure 13: Transfer price vs number of appearances(Midfielders)

Most of the midfielders are sold for a price close to their expected market value, except some midfielders who were sold for a significantly higher or lower price than their market value at the time of the transfer (Figure 12).The correlation between the number of appearances and transfer fees of midfielders is not very high, but in general, the players with a higher number of seasonal appearances during the transfer's season are priced higher (Figure 13).

Attackers

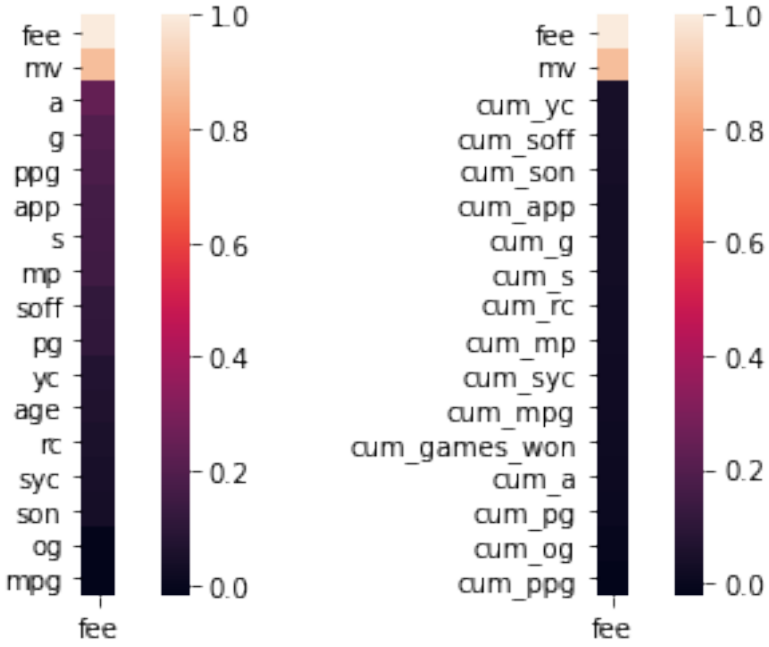

Figure 14: Price and statistics heatmap(Attackers)

The most contributed metrics to the transfer price of attackers are the number of assists, goals, the number of points earned by the team when the player was on the field, and also the number of squad selections and appearances. None of the cumulative metrics are connected to the players' transfer price (Figure 14).

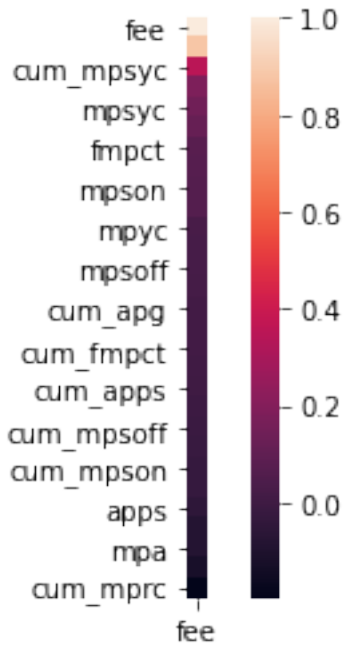

Figure 15: Price and ratios(Attackers)

The ratios are again not significantly contributing to the transfer price. The ratio that is related to the attacker's position and have a relative contribution to the price are player's field playing percentage(What proportion of time being elected in the squad of the team does the player spend on the field) (Figure 15).

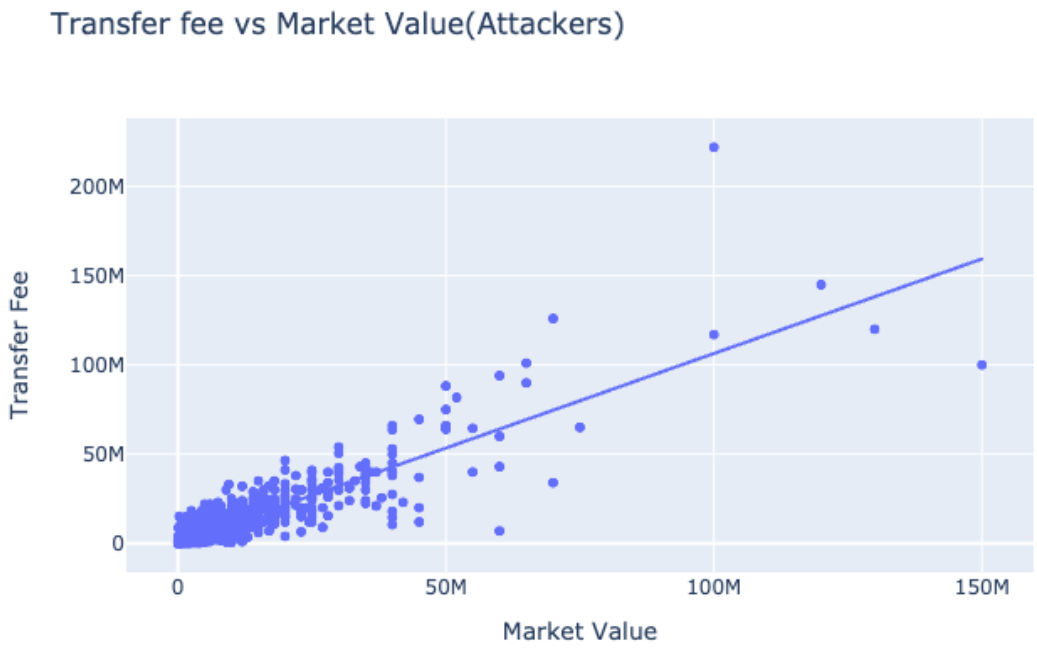

Figure 16: Transfer price vs market value(Attackers)

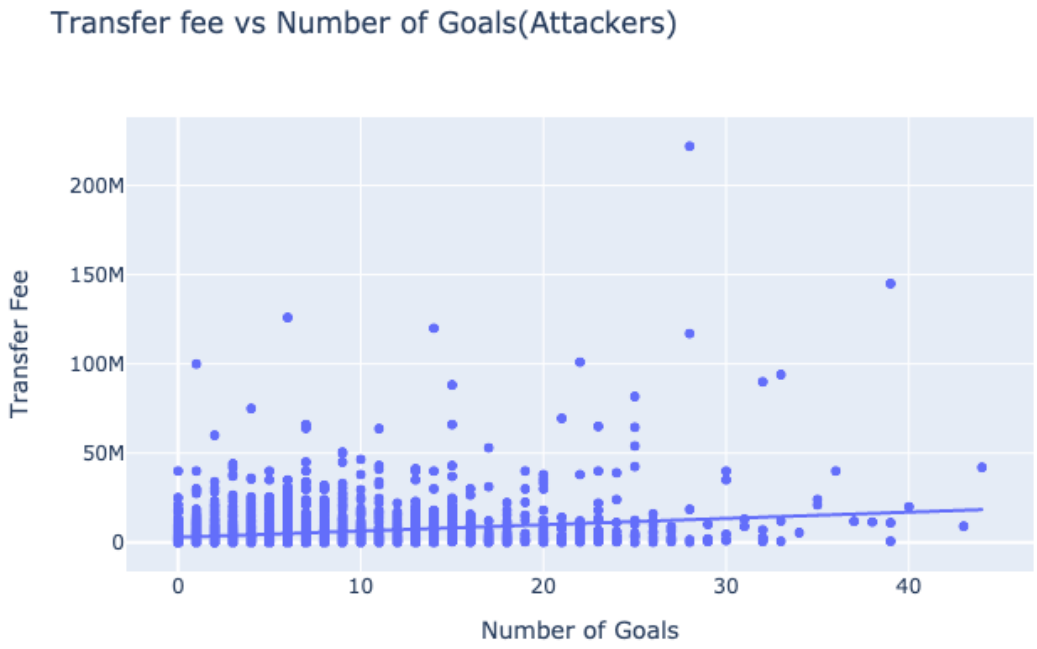

Figure 17: Transfer price vs number of goals(Attackers)

In general, the attacker's transfer price is more likely to be higher than his estimated market value rather than being lower, and only a few attackers were bought for a price significantly lower than their estimated market value (Figure 16). The number of goals the attacker scored during his transfer's season has a little correlation to his transfer price as in general the higher the number of goals the higher the price, but the connection is not strong as some players with a huge amount of goals were sold very cheap and some players with a few amount of goals were sold very expensive (Figure 17).

Market Value

As a result of analyzing each position's attributes to the transfer price, we found out that the market value of the player had the highest contribution to his transfer price in each position. Market value is a very close attribute to the transfer price, however, the market value of the player is not always a valid estimate for the player's price. Market value is also dependent on many attributes of the player. As we already analyzed the player's performance metrics contribution to their transfer price, and the contribution was also identical to the market value, we analyzed some other attributes of the players that had some correlation to their market value. We have analyzed the market value of the selected players during the 2019/2020 season to identify the currently important non-performance variables(Metrics that are independent or not highly correlated to the player's performance) that are contributing to the player's market value.

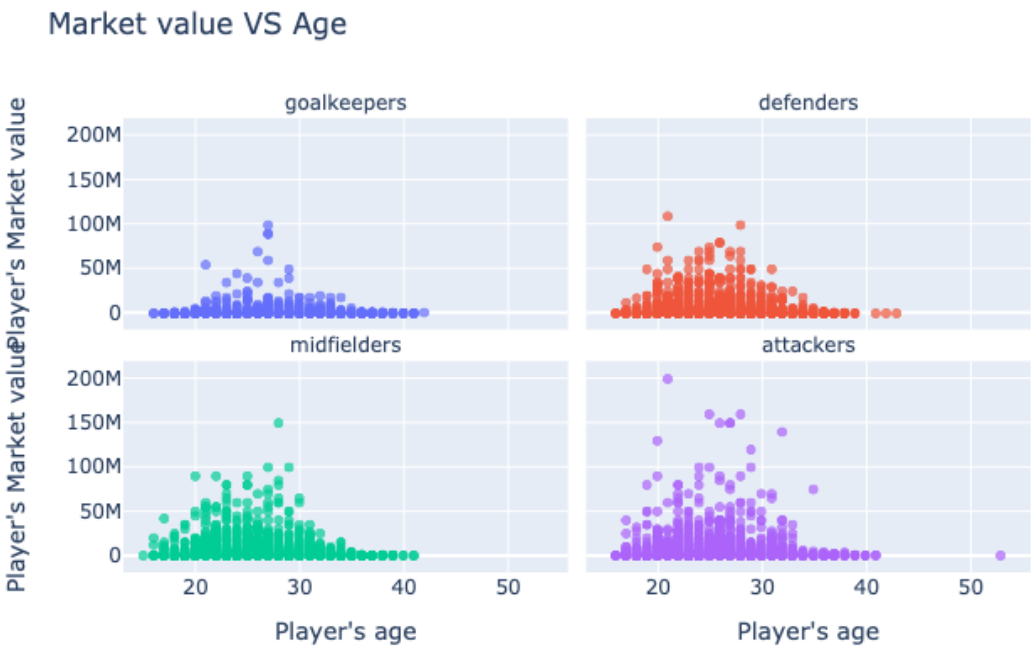

Figure 18: Market Value vs age

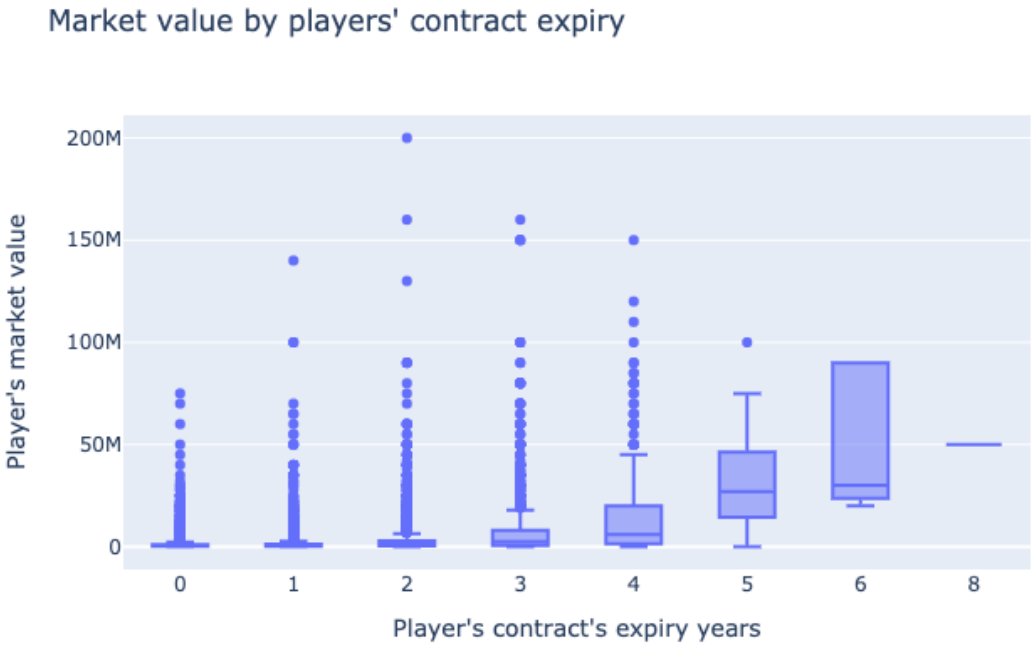

Figure 19: Market value by years on contract left

As we can see the correlation between the age and market value is not explicit. Starting from some moment the market value of the player decreases with the age increasing. Most of the expensive players are in the age period of 20-26 as in these years their market value increases alongside the age, as for this age period the players don't lose their physical attributes but gain more experience so their market value increases too. As we can see most of the players' market value is less than 100 million, and only a few players have higher market value. Most of these players are young superstars. For example, K.Mbappe who is the most expensive player in our dataset has market value of 200 million at the age of 20. According to his age, he has not reached his peak of performance yet but already has a very high market value. Most of the other expensive players are at their peak level of performance or very close to it, as they are in the 25-28 age period. Examples are Kevin De Bruyne or Antuan Griezmen. The market value of the players drastically drops after they turn 30, except some superstars and legends of the game such as Messi and Ronaldo who are aged 32 and 35 have market values of 140 million and 75 million accordingly. There are also other aged players who have pretty decent market value for their age such as Karim Benzema(35 million, 32 years old) or Angel Di Maria(40 million, 32 years old), but their market value does not overcome 40 million. Most of the expensive players are either attackers or midfielders. Only a few goalkeepers and defenders have significantly high market values. Also, in general, almost all the positions have the same age period where the players have their highest market values (Figure 18).

The players whose contract expires at the end of the season have the lowest market value, as at the end of the season if they do not expand their contract, other teams can sign them for free. A similar situation is for players who have one more year according to their contract. Those players are valued higher as their teams still have the chance of negotiating the player's contract or sell them by a relatively high price. The players who have 2 years remaining to their contract are usually those who can be sold by a very high price as if the team fails to agree on terms with the player, they can sell him on a very good deal. The players who have 3 or more years on their contract most of the time joined the team recently(1 or 2 years ago as based on age the teams mostly sign new players for averagely 5 years) or have recently extended their contract at the club (Figure 19).

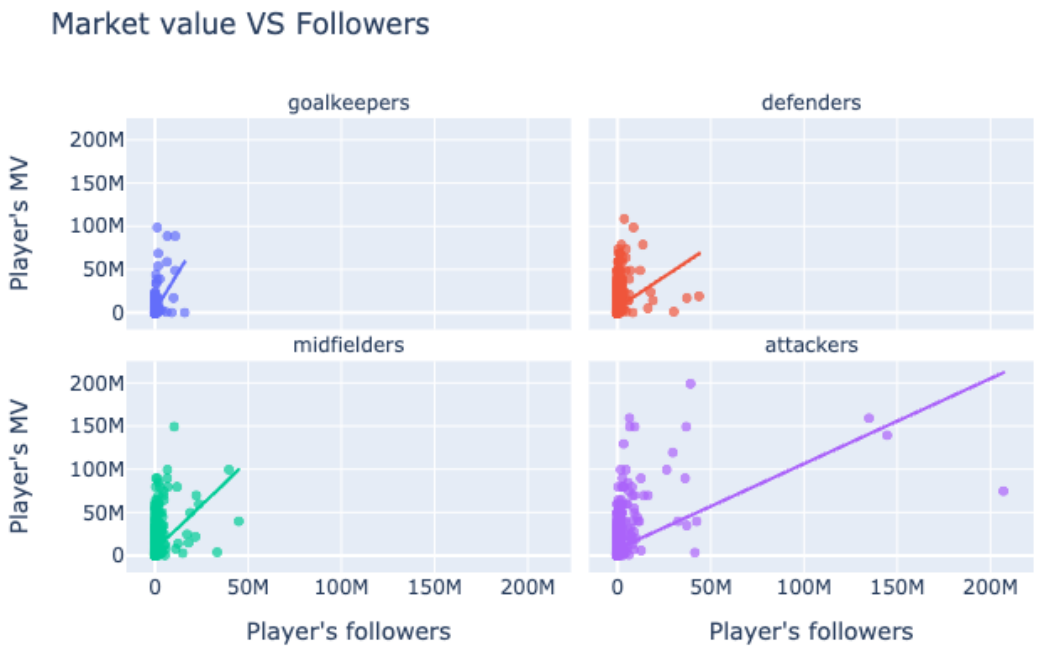

Figure 20: Market value vs instagram followers

Correlation	Goalkeepers	Defenders	Midfielders	Attackers
$Corr$(**Market Value, Followers**)	0.47	0.28	0.40	0.47

Table 4: Correlation table of followers and market value

Nowadays soccer is a very profitable business with lots of attractive contracts from huge companies, for both the players and the clubs. Taking that into account the famous soccer teams consider the player's popularity too when buying him, as if the player has good performance metrics and is also popular outside soccer then they can make huge profits from the contracts of the player. Taking this into account let's check the correlation between the number of player's Instagram followers and his market value. As we can see the correlation is positive meaning that in general the more popular(more followers Instagram) the player is the higher his market value. In this case, the age of the player is not a key factor as the number of the player's followers is mostly not dependent on his playing attributes. There only a few players who have a significantly high amount of followers. Most of these players

are from famous soccer clubs and have huge sponsorship contracts. Cristiano Ronaldo is actually the person with the highest number of followers on Instagram. We can see that the most popular position is attackers as players having an attacking position enjoy the highest popularity among social media users. The situation is very similar for defenders and midfielders as both positions players have more or less the same ratio of market value and followers. However, we can see that the goalkeepers are not that popular among social media users as for example one of the most expensive players of the dataset Yan Oblak who has a market value of 100 million euros has only about 1 million followers,(Table 4) whereas the attackers who have approximately the same market value enjoy double triple and even more number of followers (Figure 20).

Players' physical and racial attributes

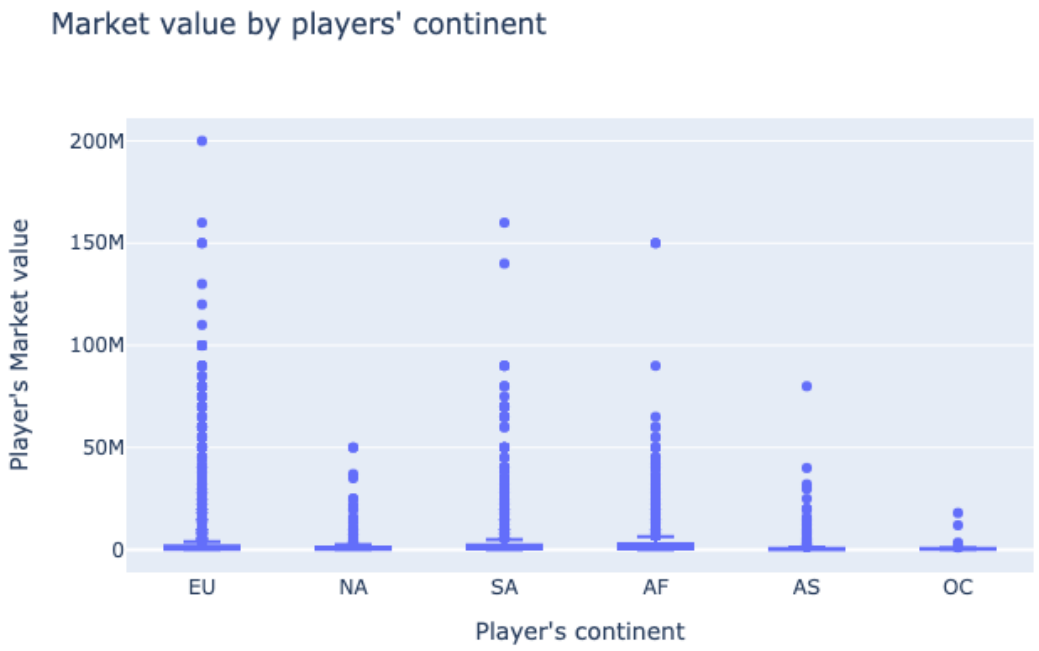

Figure 21: Market Value over continents

As we can see the players from European countries are valued the highest, as most of them play in European leagues. Players from South America and Africa are also highly

valued, as the South Americans are mostly very talented and have good technique and the African players have great physical conditions. The other continents are in general similar in terms of the market value of their players. Only a few Asian or North American players have high market values and the others are no different based on the continent (Figure 21).

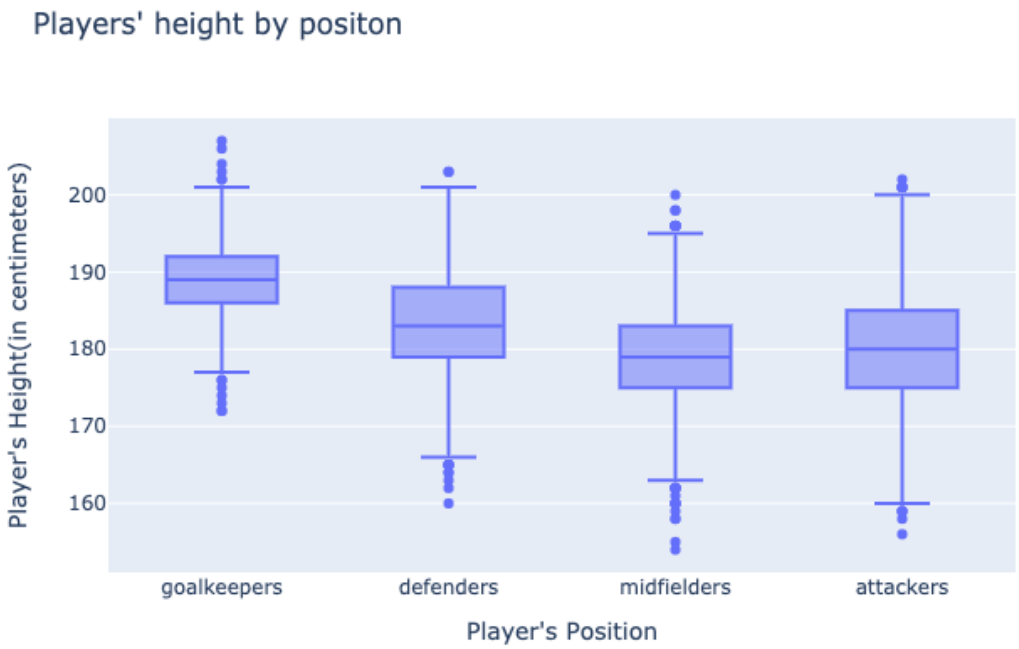

Figure 22: Player's height over positions

Another attribute that can be important for the player based on his position is height.As we can see high height is very important for the goalkeepers as the players with the highest height are goalkeepers and only a minority of the goalkeepers have a height of lower than 180 cm.Most of the defenders and attackers also have high heights but not as high as the goalkeepers and some of the players from these positions have heights lower or equal to 160 cm.According to the visualization, the height of the player is least important for midfielders (Figure 22).

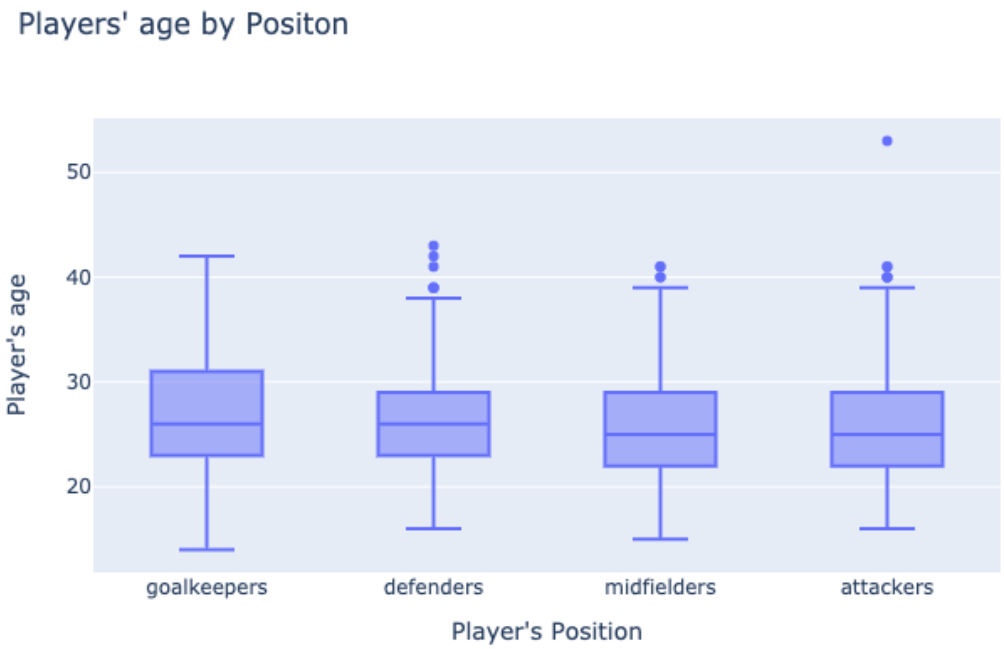

Figure 23: Player's age by positions

The positions where there are more players with generally older age are goalkeepers and defenders. These two positions in soccer require good experience in the game and some times the player's ability to understand the opponent, which is usually high among experienced players is valued even more than his physical attributes as the player can compensate those with his experience. Good examples are Italian goalkeepers and defenders who sometimes play up to 40 and do not lose their quality. The midfielders and attackers are usually younger as they do a lot of physical work on the field and old players mostly fail to keep their high level in these positions due to physical attributes that get worse by their age (Figure 23).

2.2.2 Time Series Analysis

In order to identify the overall patterns of the economic valuations of players in soccer, we have analyzed the transfer fees and market values of the player's over the years and over their age. We have also taken into account the effects of the categorical variables and other factors such as worldwide events taken place during the moment of the peaks of transfer prices and market values of the players. In order to summarize the overall prices of the players, we used different summarizers such as mean, median, maximum to identify the patterns for the players in each price category. The time period analyzed for transfers is around 20 years, starting from the late 1990s and early 2000s transfers. The market value of the players was available only starting from 2007, so the historical data of market value contains around 13 years and around 17,000 players.

Players' transfer fee over time

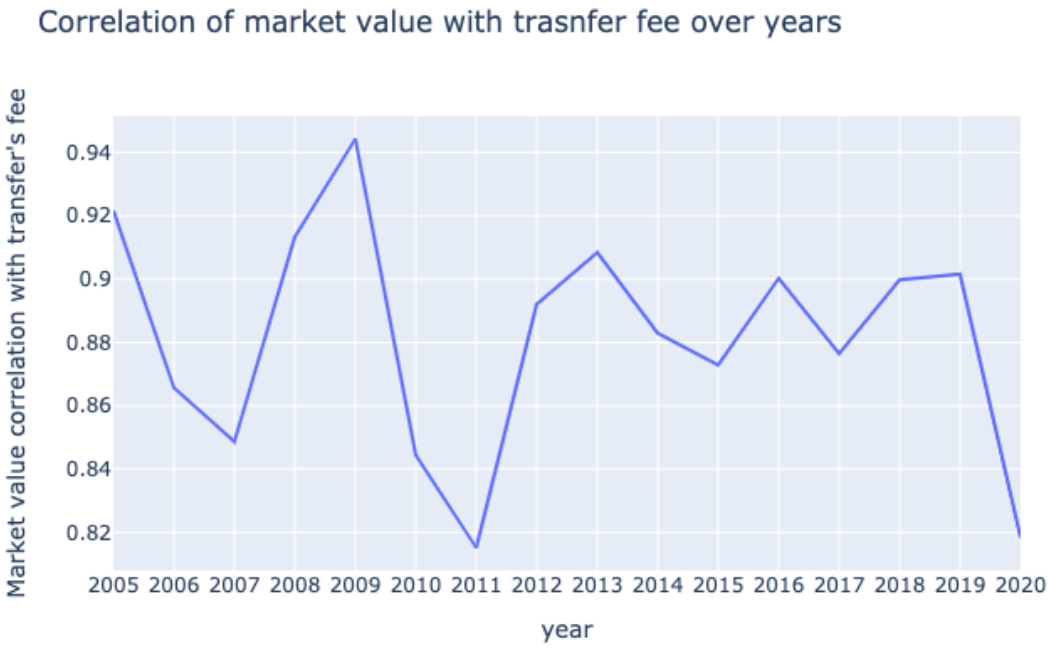

Figure 24: Transfer price and Market value correlation over years

As we can see the highest correlations between the transfer fees and market values were before 2010. Also, we can notice that after each world cup year(2006, 2010, ..., 2018), the correlation between the variables decreased as most of the time the players who played well in the tournament were sold for a price higher than their real value. And also we can see multiple cases, where the correlation decreased significantly the following year. Those are 2009, 2005, 2013, and 2016, 2019. Here again, We noticed a pattern. During the years 2009,2013, and 2016 transfer records were broken. So after a new expensive transfer, the fees for players became higher and thus less correlated to their market value (Figure 24).

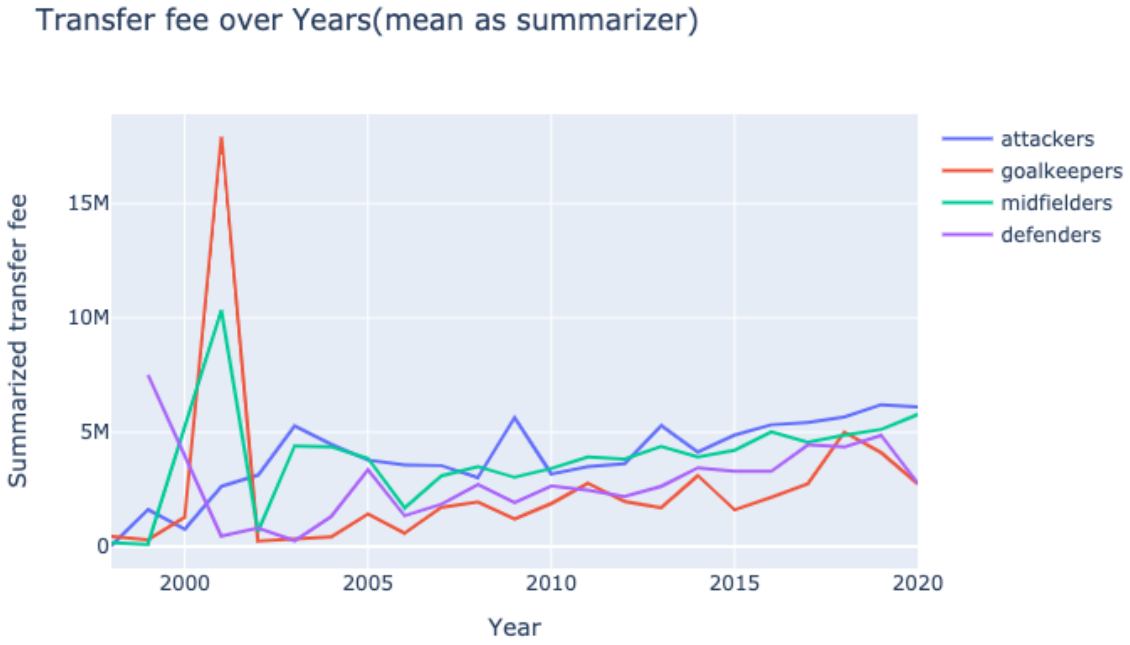

Figure 25: Transfer price over years(mean)

If we take mean as a summarizer for transfer fee over years, the highest point appears in the early years, as during those years transfers were not common things and in case of expensive transfers, the mean transfer fee became very high. We can see that the highest point was in 2001 for goalkeepers, as in that year the transfer record for goalkeepers was beaten when Buffon was bought for around 53m euros. Over time the high points seem to become less common as more expensive transfers happen (Figure 25).

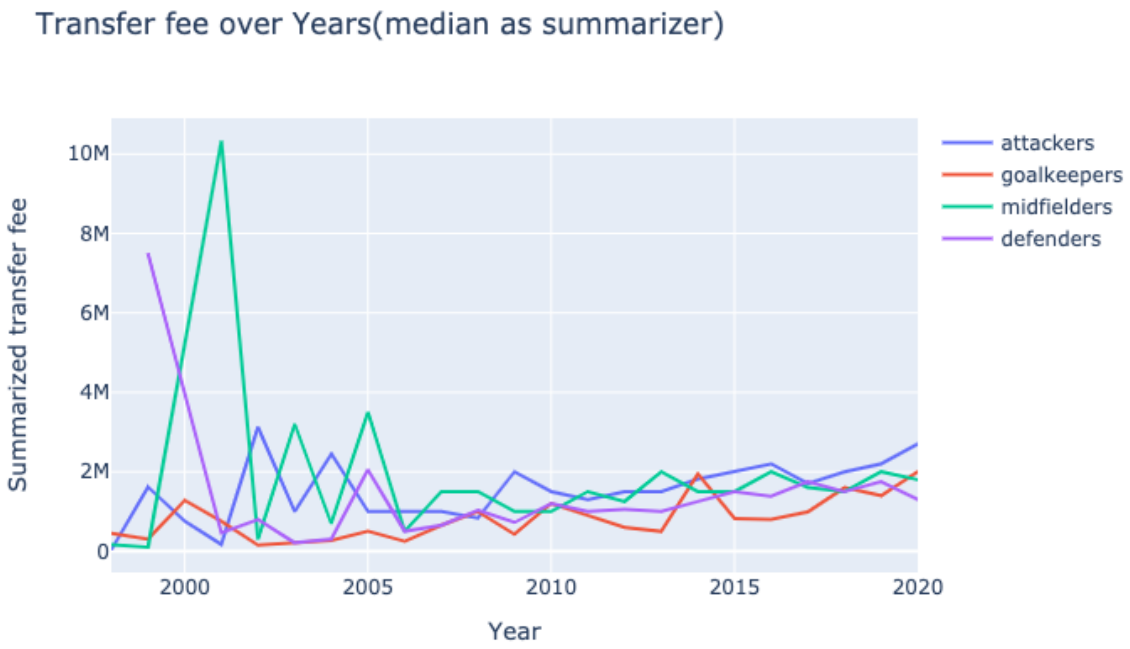

Figure 26: Transfer price over years(median)

After taking the median as a summarizer we can see the position for the highest point changed becoming midfielders and attackers in the 2000s. The goalkeepers do not have high value, taken median as a summarizer as the median is sensitive to extreme values and expensive goalkeeper transfers are a very rare phenomenon in soccer (Figure 26).

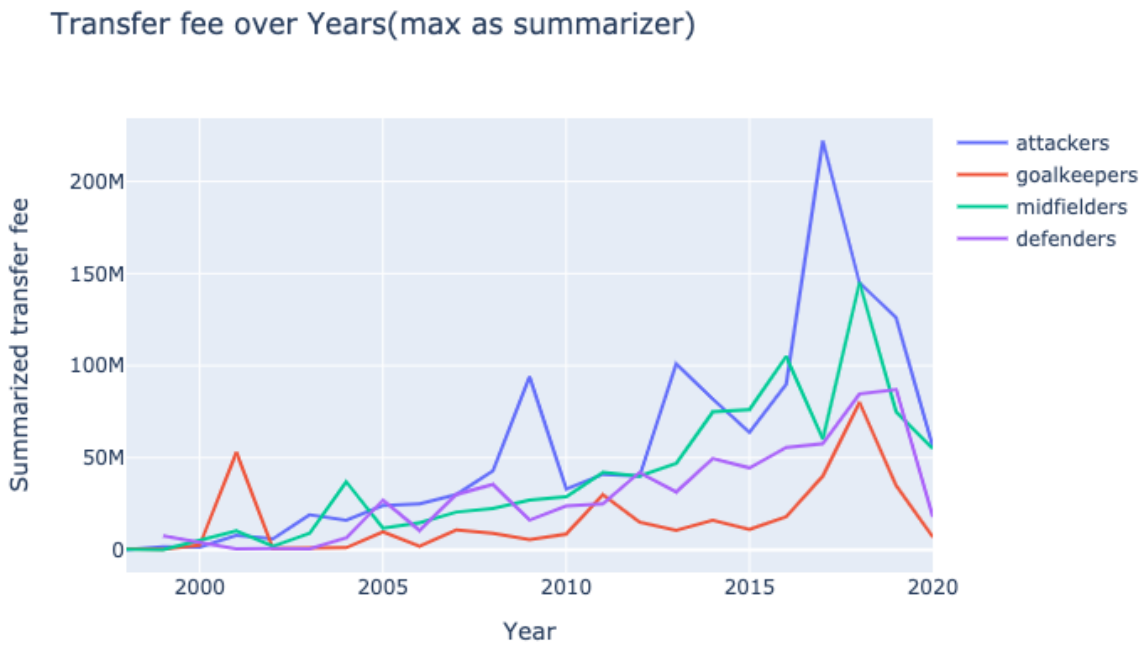

Figure 27: Transfer price over years(median)

Taking maximum as a summarizer we can see the transfer fee record-breaking years for each position. In this aspect, the goalkeepers are the most stable position as there were only 2 cases of goalkeeper transfers when the fee was record-breaking. Defenders are also mostly stable in this aspect with a few record-breaking transfers in this period. Midfielders also have only a few occurrences of record-breaking transfers and the periods between two record-breaking transfers were very long before 2014, after which the record was beaten twice in 4 years. Attackers have the highest transfer fee records and have the shortest periods among the record-breaking transfers. Also, the difference between two record transfers price was highest among attackers (Figure 27). That transfer took place in 2018 and the fee paid was two times expensive than the previous transfer record's fee. More details about Transfer Records [5].

[5]Link to historical transfer fee records

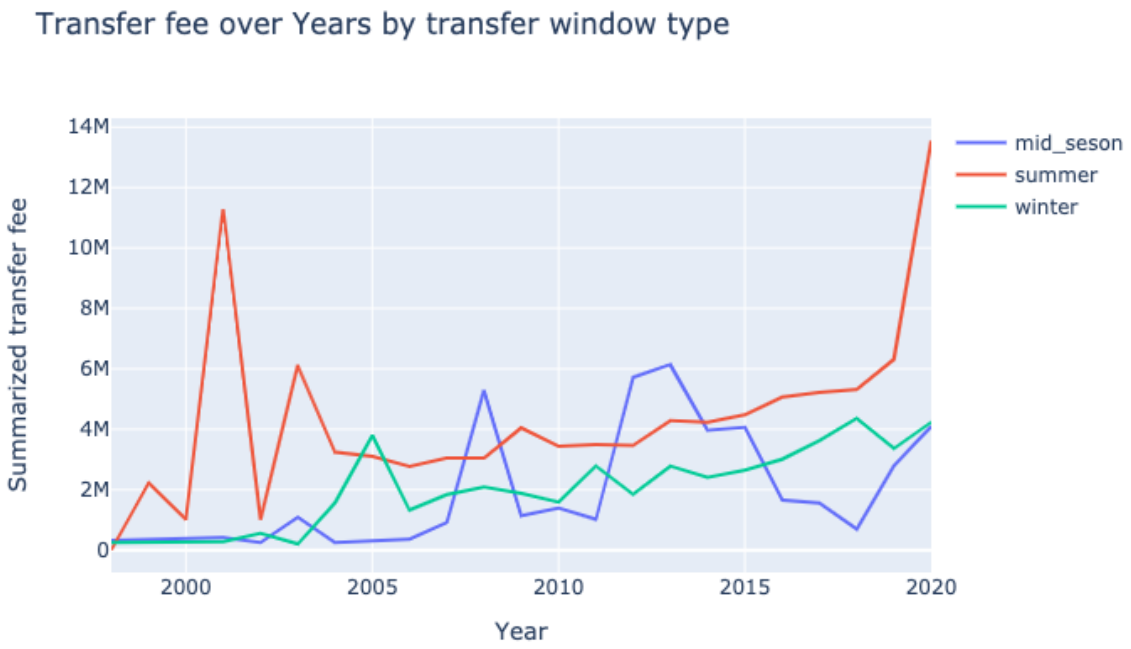

Figure 28: Transfer price over years by transfer window type(mean)

The most expensive transfers mostly happen during the summer transfer window, as the teams have more time to make the transfers. There are also some years when relatively expensive transfers happened during the middle of the season. But most of the time transfers in the middle of the season are rare events (Figure 28).

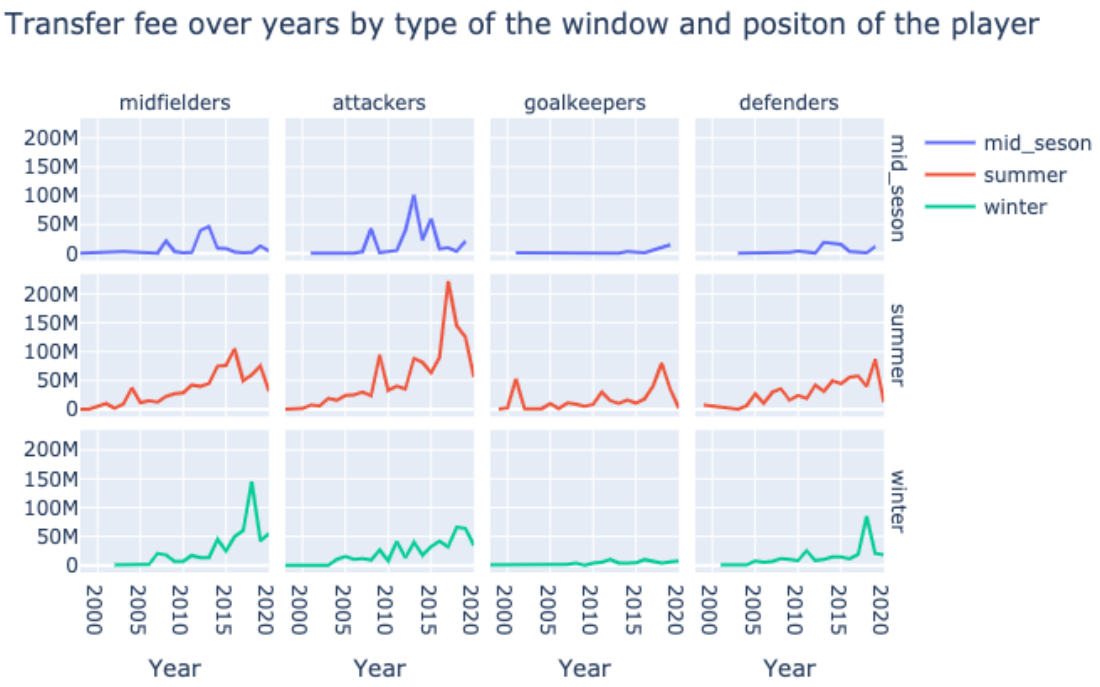

Figure 29: Transfer price over years by transfer window type and position(max)

Taking max as a summarizer we can see that almost all the records were beaten during summer transfers' window, except for midfielders, for which the transfer record for that position was beaten during a winter transfer window. Also, one of the records for attackers was beaten in 2013's midseason as Gareth Bale's transfer happened on September 2 (Figure 29).

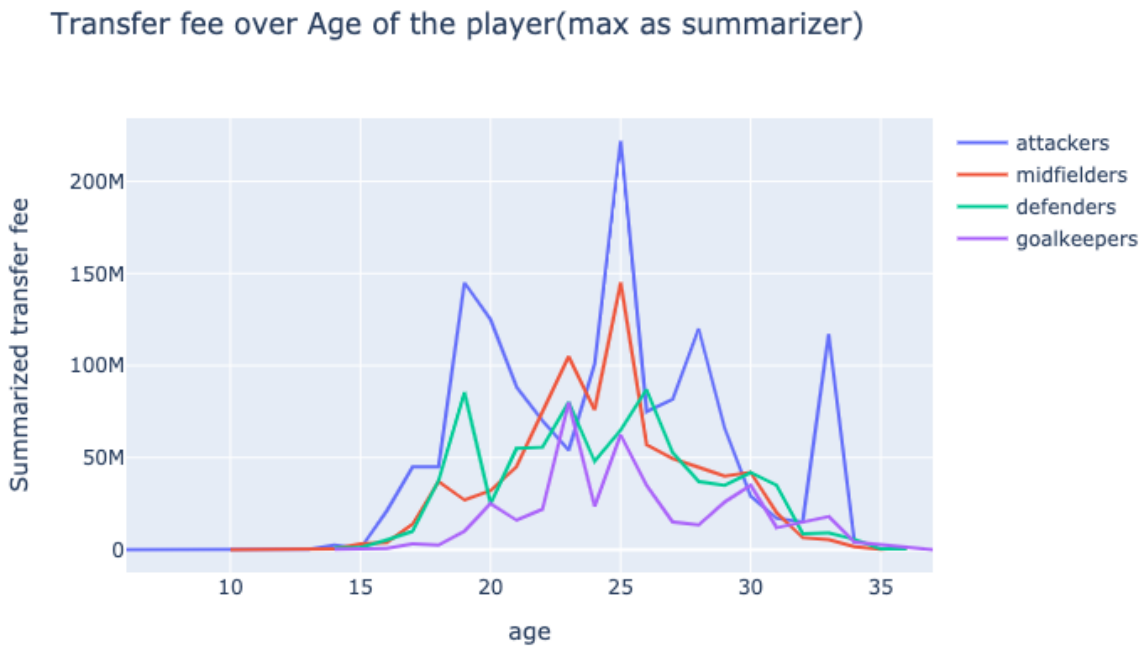

Figure 30: Transfer price over age(max)

As we can see most of the expensive transfers take place with players aged from 20 to 25 for all positions. Players belonging to this age category are the ones with huge potential, that increases over time till they reach their peak age, which is for most of the player's around 29-32 years old based on the position on the field. We can also see a huge increase for 33 years old attackers, but this peak happened because of Cristiano Ronaldo's transfer to Juventus for around 100m when he was aged 33. This is an exception rather than a general behavior as a player at his age are valued low, and top transfers of players in this age category is a very rare event (Figure 30). More details on the player's based on age can be found on transfermarkt's most valuable players page [6].

[6]Link to the most valuable players

Players' Market value over time

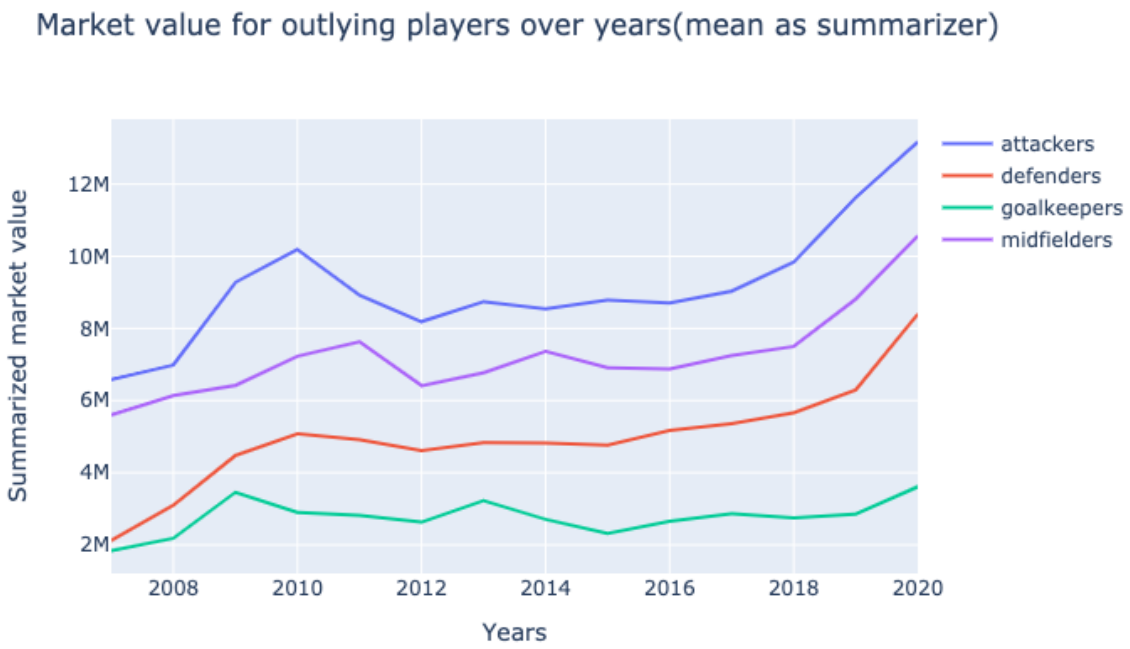

Figure 31: Market value over years(mean)

We have firstly analyzed the players who have a very high market value, which most of the time means that they play in the top championships and most probably for good national teams. Almost all the positions started to have a rise in market value starting from 2007. Goalkeepers and defenders did not lose track and increased in value most of the time up to 2020. However, there is a noticeable pattern. As we can see in general almost every two years the player's market value rises compared to the previous year. The increase happens due to International tournaments. Most of the major international tournaments take place every 4 years, so every 2 years, there is either a world cup or a continental cup. During the tournaments, most of the player's who play good earn high market value increases, and we can see that in 2014 when the world cup in Brazil took place, the midfielders had an increase in their market value compared to 2013 and then a decrease in 2015. The same pattern is repeated for attackers in 2010. However, after 2016 all the players started to have an increase in their market value in general, with only once having a decrease(for goalkeepers

in 2018). The goalkeepers were mostly stable in terms of rising their market value, having a notable drop in 2015(again after world cup).The reason for the drop in the market value of goalkeepers is that the world cups that took place during the analyzed years were the ones with the highest number of goals scored on average in our dataset, so in general, the goalkeepers allowed many goals during those tournaments and as a result had decrease in their market value after the tournaments (Figure 31).The average goals statistics over the world cups [7].

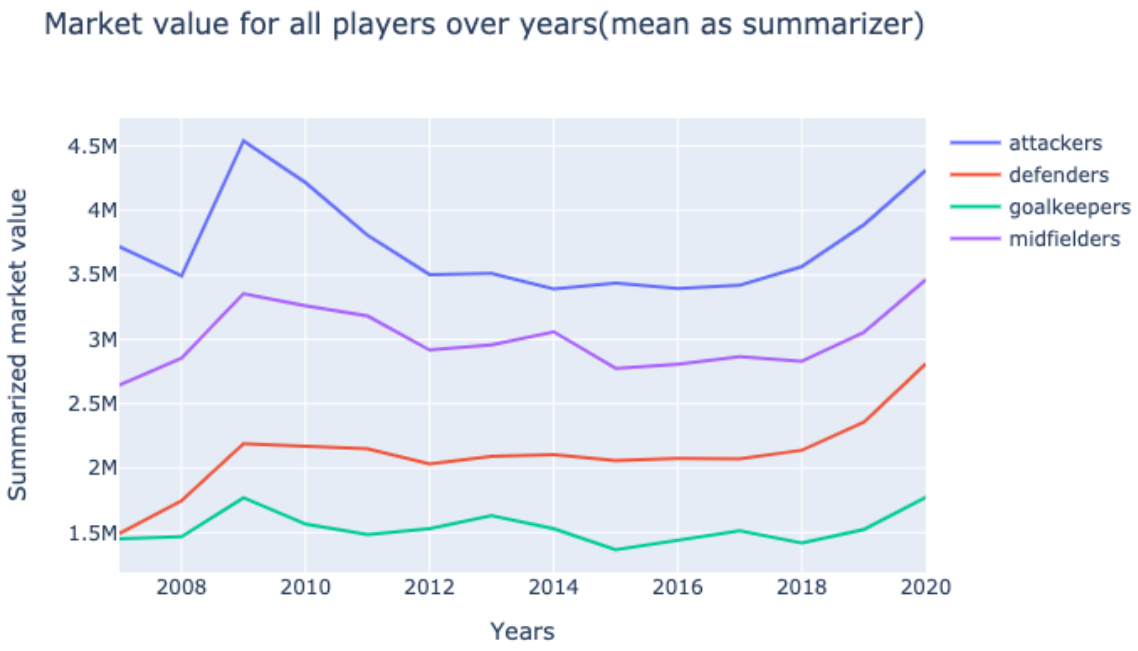

Figure 32: Market value over years(mean)

As we can see if we include all the players from the dataset, the international cup effect is not very often anymore, as most of the players are playing for teams from not popular leagues and weak countries which play in international or continental tournaments very rarely. In general players from all the positions had their peak value in 2009, (during that year a record-breaking transfer of Cristiano Ronaldo took place), after which the values

[7]Link to an infographic about the world cups

seemed to be stable and started to rise dramatically from 2018 when Neymar's record-breaking transfer took place (Figure 32).

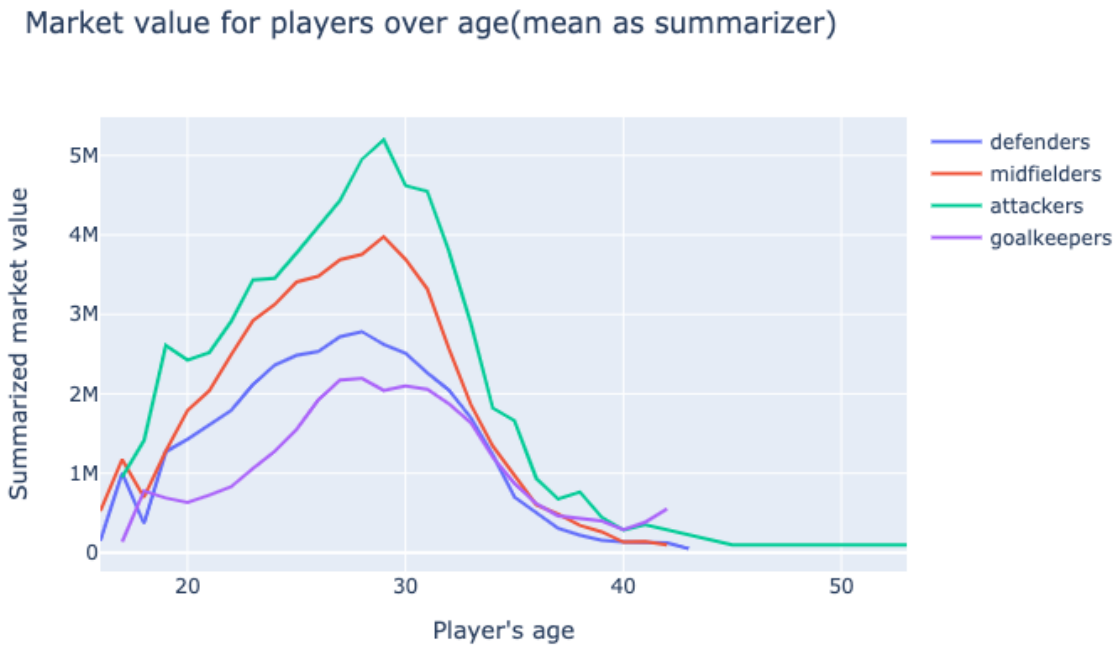

Figure 33: Market value over age(mean)

As we can see the players have the lowest market values at the beginning and at the end of their career and most of them reach their peak value at the age of 27-31(based on the player's position. The goalkeepers are able to have the least amount of decrease in market value as they get older compared to the other positions. Also, we can see that the market values of the players started to increase dramatically from the late teens and early twenties of the player up to his peak (Figure 33).

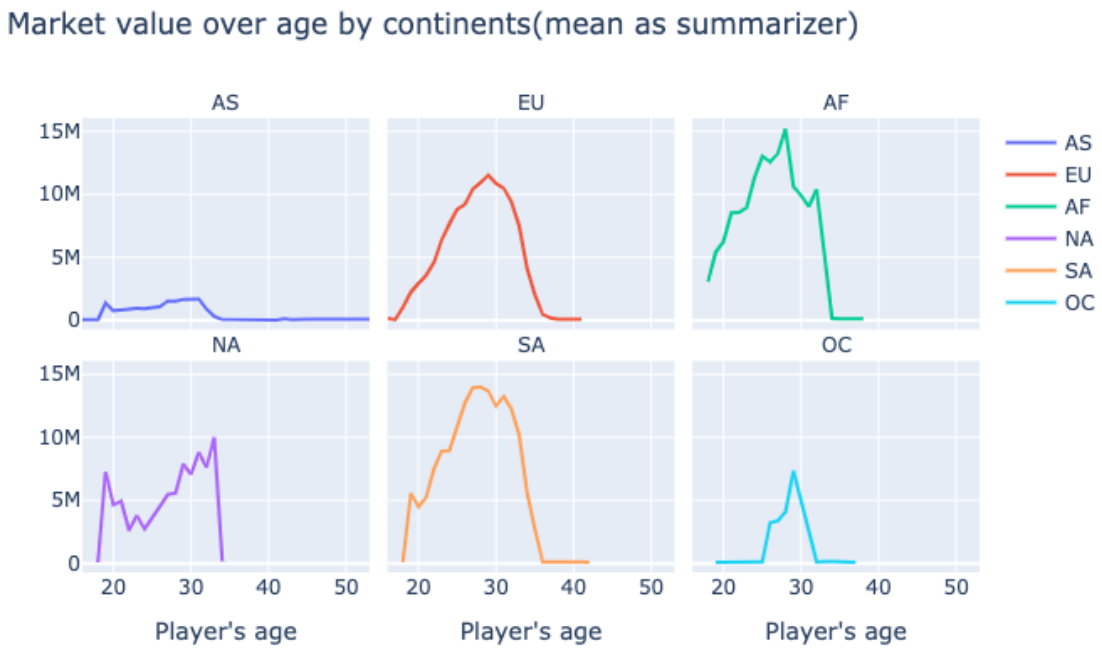

Figure 34: Market value over age by continents(mean)

If we take mean as a summarizer for market values, we can see that the African players have the highest market value in comparison to players from other continents. Asian players have the lowest values as the majority of them play in their domestic championships. The players from North America and Oceania do not have a high market either, as they also mostly play in their domestic leagues. North American players are the youngest to retire in comparison with other leagues, whereas Asians are the oldest to retire, our dataset contains a Japanese player who is 53 years old. The market value's behavior is very similar for South American and European players, except the fact that young players from South America are valued more expensive than the European players of the same age (Figure 34).

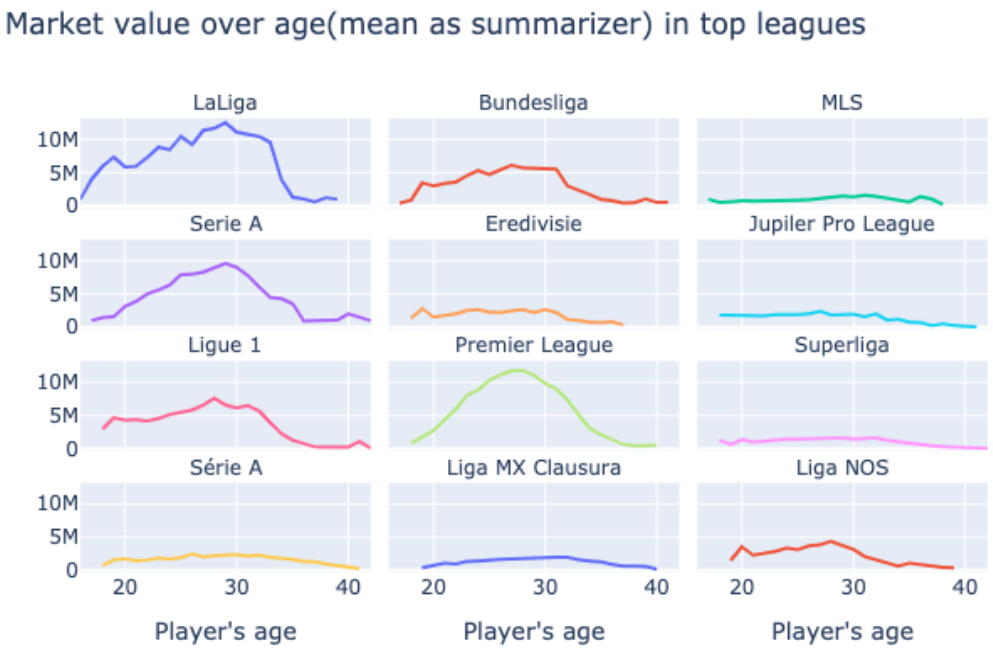

Figure 35: Market value over age in top leagues(mean)

If we take a look at the market value change over player's age in the most popular leagues, we can see that players in League 1(French league) have the highest market value at the youngest age(mean taken as summarizer). However, players older than 18 do not have a higher market value than their agemates from other leagues. The young players also have a good starting point in Liga NOS and Serie A.(Portuguese and Brazilian). These leagues are famous to the whole world for exporting the youngsters to top European leagues. Old players have no value in any league. We can see that players have a significant increase in their market value only in the top 5 soccer leagues. The top 5 league player's market value has almost the same path in terms of increasing and decreasing along with age, and the most expensive players are in LaLiga and Premier League. The situation in other leagues does not change almost at all as the mean values of the players reach only about 2-3 million in the best cases (Figure 35).

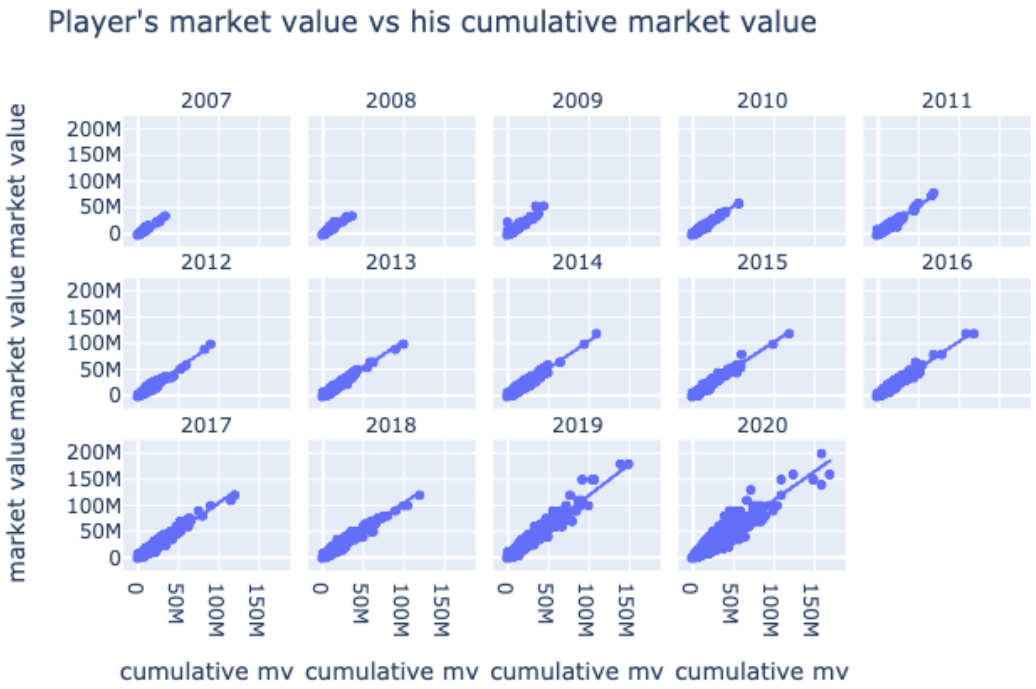

Figure 36: Market value vs cumulative market value over years)

As we can see before the middle 2010s the correlation between the player's seasonal market value and the cumulative mean of his market value at the time of the transfer was very high, meaning that seasonal based increases or decreases in the players' market value were rare. However, we can see that after around 2011, the seasonal market values of the players started to have a little deviation from their cumulative market value meaning that over time more seasonal based overvalued or undervalued players occurred (Figure 36).

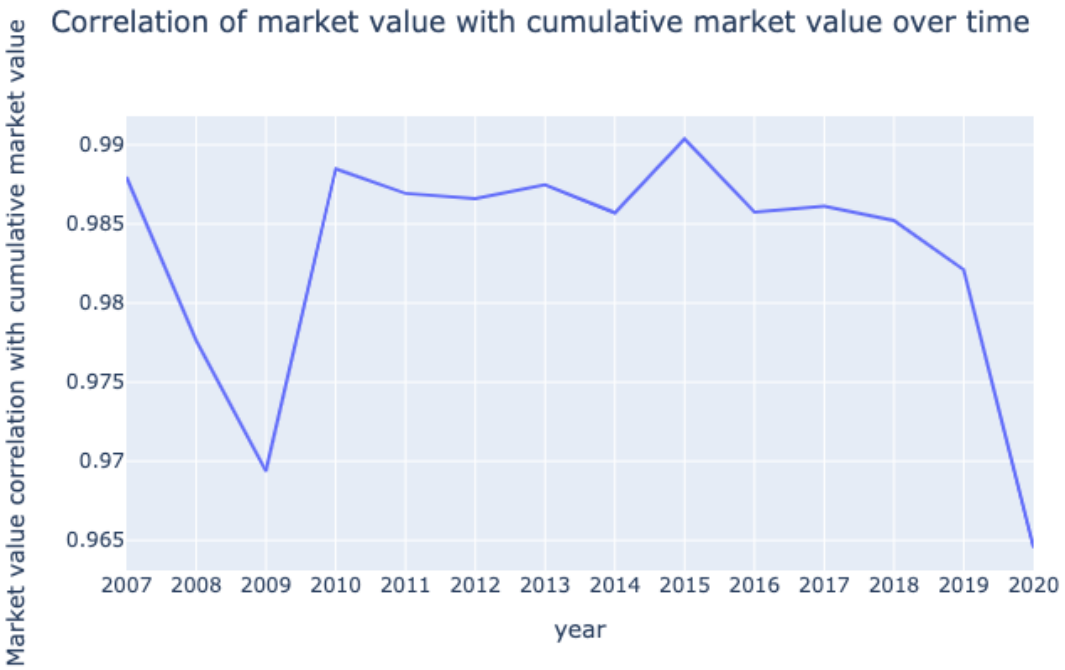

Figure 37: Market value correlation with cumulative market value over years

As we can see the correlation between the variables is solid and there were only a few peaks , and we can see that the current trend is decreasing in the correlation between the variables as currently more and more one season wonderers appear, who get attached high market value to them, having no significant market value history. The decrease in the rate of correlation emerged in 2017, nevertheless the decrease rate is not significant yet. (Figure 37).

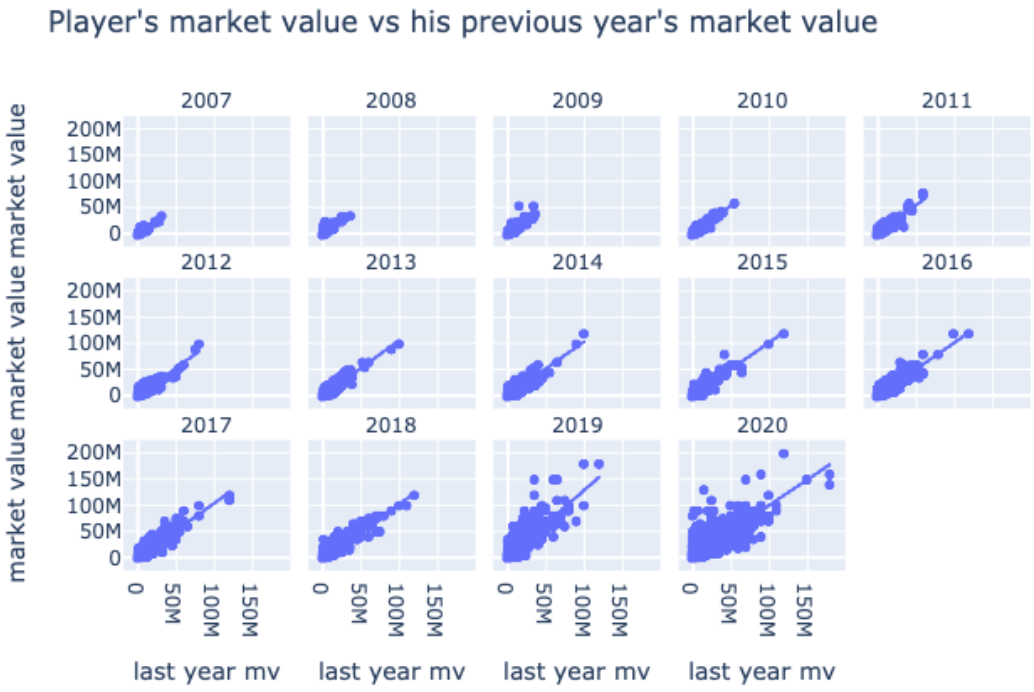

Figure 38: Market value vs previous year's market value over years

As we can see the players' previous year market value is strongly correlated to their current market value. The correlation seems to be higher in the early years and started to get lower after the early 2010s late 2000s, similar to the case of market value and cumulative market value correlation(Figure 38).

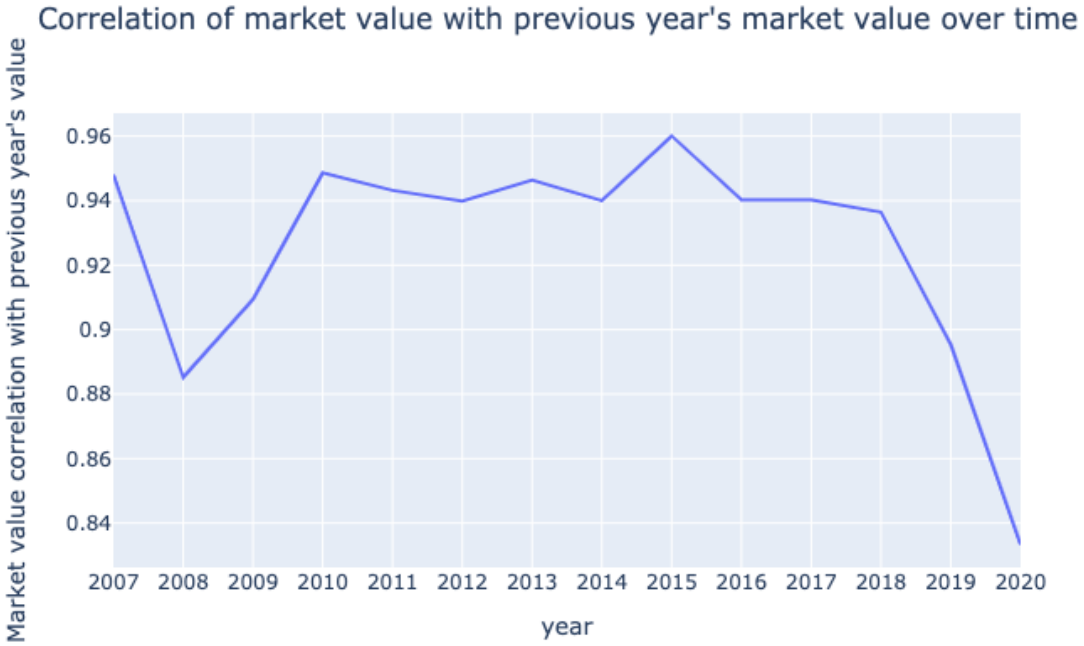

Figure 39: Market value correlation with previous year market value over years

As we can see the correlation between a player's market value and his previous year's market value was the highest in the early years and started to drop drastically starting from 2018. This behavior is explained by the fact that in the early years it was easier to value the players and mostly their valuation was done using trusted approaches build over time, but over time more factors influence the market value of the player so the correlation with previous years value gets lower and lower. Again we can see breakouts in some international competition years and after Neymar's transfer (Figure 39).

2.2.3 Network Analysis

In order to identify the main ruling clubs and countries of the transfer market, we have analyzed the transfers' network by visualizing and interpreting different network statistics. We have used each transfer of the dataset as an edge, taking the team from which the player was bought and the team the player was sold to as nodes. The same approach was used for analyzing the loan network and winter transfers. As, in soccer transfer between two same clubs can happen many times, so as a result the network's type is directed multigraph. We have also used the Node2Vec algorithm for finding the similar clubs in the network.

Node2Vec is a feature learning algorithm for networks that transforms the network's information into vectorized form using random walks and a combination of tree-based BFS and DFS (Grover & Leskovec, 2016). In order to visualize the similarity scores of the nodes we have used PCA dimensionality reduction algorithm (Qu, Ostrouchov, Samatova, & Geist, 2002).

The framework used for building the networks was networkx and as the networks contained a lot of nodes, and interactive visualizations tool bokeh was used. In most of the visualizations, the color of the node represents the content of the club, the size of the node represents the node's degree(number of incoming and outcoming edges), the color of the edge represents the transfer window type, the width of the edges represents the fee of the transfer and the opacity of the edges represents the age of the players involved in the transfer.

Soccer's Transfers Network by teams

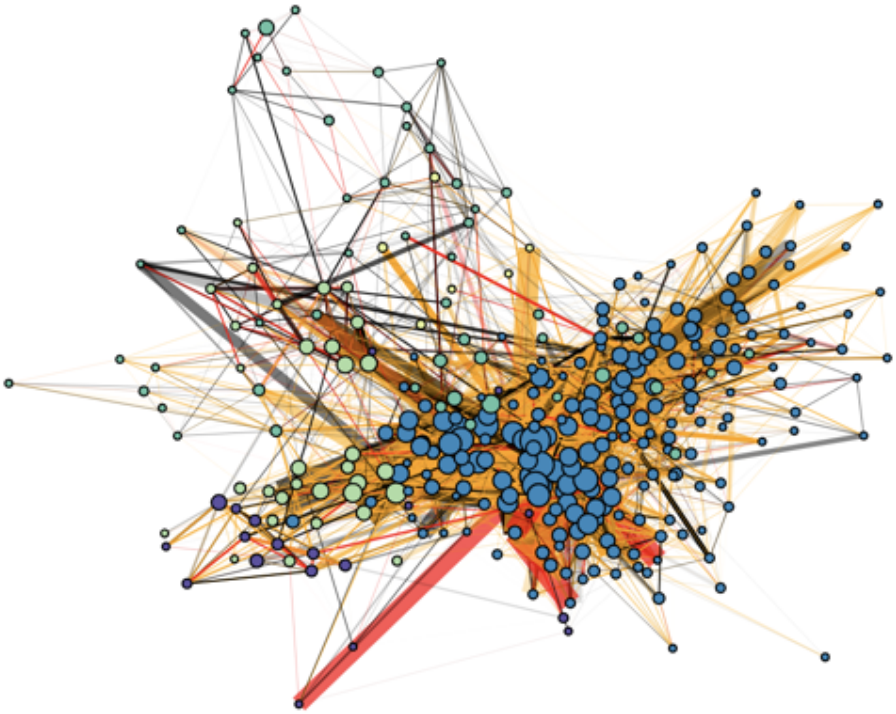

Figure 40: Transfers network using teams as nodes

As the networks contain a lot of nodes, we cannot get much information from the visualization. However, we can see that the European teams(node color blue) dominate in the market, with the majority of them connected with each other, also the majority of the transfers happen during the summer transfer window(edge color orange), except for most of the Asian teams, which make most of their deals during the winter transfer window(edge color black) as the primary pre-season transfer window for most of the Asian leagues takes place during January-February. We can also see that the young players(low opacity of the edge) tend to be transferred to big teams from small teams for high transfer fees(high width of the edge) (Figure 40).

Soccer's Loan Network by teams

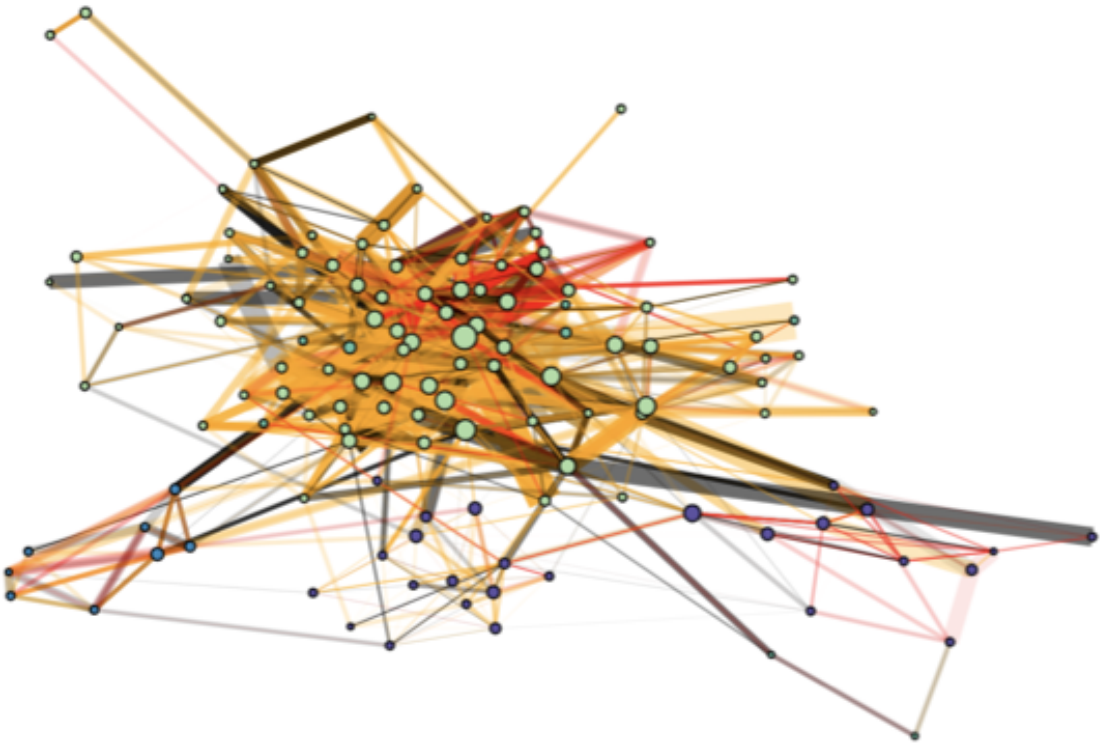

Figure 41: Loans network using teams as nodes

The main players of the market are again the European teams(node color green) and most of the loans again take place during the summer transfer window. We can see that there are only a few teams with many connections(big node size) and most of the other teams have similar levels of connections. We can see that most of the players loaned are young(low opacity) as most of the time big teams cannot provide enough gaming practice for their youngsters so they send them to loans to teams where they can gain enough gaming practice and develop their skills (Figure 41).

Network Statistics

Statistic	Transfers	Loans
Network's density	0.08	0.12
Network's reciprocity	0.64	0.95
Network's assortavity based on continent	0.627	0.76
Network's assortavity based on league_class	0.379	0.417
Network's assortavity based on country	0.563	0.593
Network's assortavity based on degrees	0.270	0.15

Table 5: Networks Statistics on transfers and loans networks

Let's first analyze the transfers network. As we can see the network's density is very low, which is logical as we have many teams, and not all of them have connections between each other. However, the reciprocity of the network is relatively high, as most of the teams that make deals with each other have transferred in opposite directions too. The main attributes for the assortativity of the teams are their continent and country, as it is easier for players to move to another team that is in the same continent where they play, and even more when it happens in the same country. The metric is around 0.5, as most of the talented players from other continents and non-EU countries tend to move to European soccer clubs, as there they have higher chances of succeeding. League's class has the lowest effect on the assortativity as most of the time players from leagues with lower-ranking tend to move to higher-ranked leagues. The degree of the node also has a relatively low connection to the assortativity of the nodes, as teams with a low number of connections not always are connected to teams with a lot of connections (Table 5).

If we investigate the networks statistics for the loans network we can see almost the same metrics as for transfers network, except almost maximal value for reciprocity, which is logical as in most of the cases player who is loaned to another club comes back to his club, and only in some cases the club that loaned the player buys him.

Community Detection

Figure 42: Communities in the transfers network

Girvan Newman's community detection algorithm was used to identify the communities of the transfer network (Girvan & Newman, 2002). The algorithm's generator was iterated 5 times and as a result, it found 7 communities. In the figure above the node color attribute is the community of the team. We can see that there is a very big community with most of the teams involved in it. If we compare this visualization of the network with the one using the content of the team as the node color, we can see that many teams from South America migrated into a community with most of the European clubs. We can clearly see the young talent suppliers of the European teams. (Small nodes in green with wide edges and low opacity). With the help of the graph's interactivity on IPython, we were able to find out that the Asian clubs have two communities, which contain the Japanese and South Korean teams accordingly. The other 4 communities detected by the algorithm were very small and not visible on the visualization (Figure 42).

Node2Vec

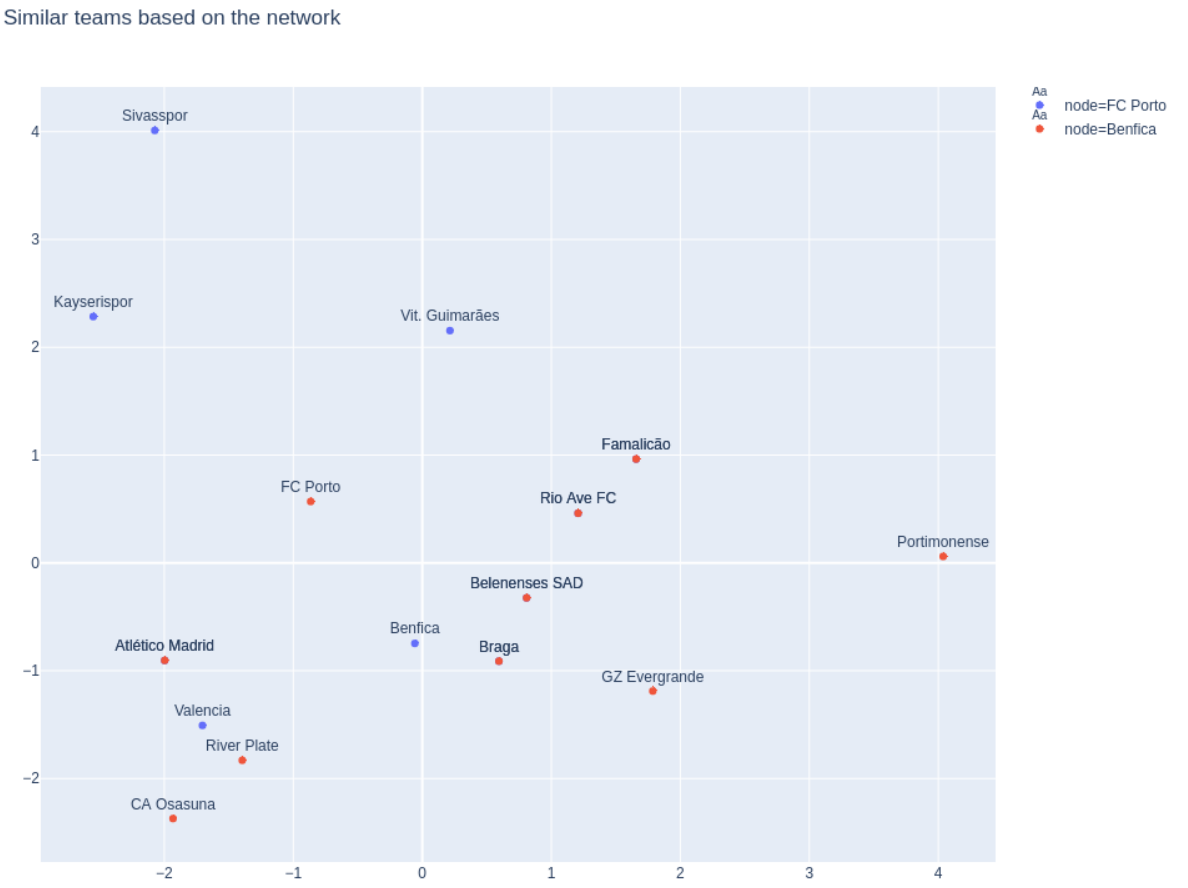

Figure 43: Similar teams according to Node2Vec

Node2Vec identified the team's similarity mostly based on the country of the teams, frequency of transfers between the clubs, and similarity in terms of making transfers. We can see that one of the most profitable making teams of the network, Benfica is similar to many Portugues teams and also some famous teams as the club is famous for providing the top leagues with high-level youngsters. Another famous youngster provider Porto has almost the same situation as Benfica. Porto is very similar to some Portuguese teams and also some teams from outside Portugal (Figure 43).

2.2.4 Insights

EDA insights

Through exploratory data analysis we have identified that the most significant variable to the reponse was the player's market value. We have also identified that the categorical variables also make impact on the player's market value which in its turn impacts the potential transfer fee. Also we have found out that the perfomance metrics and ratios were not strongly connected to the output variable.

Time Series Analysis insights

As a result of analyzing the historical data of transfer fees and market values of the player over the years and their age, we have found out the periods of the players' peak valuation and also identified that external events, such as world or continental tournaments taking place during the transfers' years make an impact on the fee.

Network Analysis insights

The analysis of the transfers and loan networks gave general information about the main active participants of the transfer market and showed the main communities of the network. The implementation of Node2Vec showed an overall criterion for the similarity amongst the teams and the results were pretty similar to the community detection algorithm's findings, as most of the similar teams were part of the same community.

2.3 Data Preparation

2.3.1 Data Transformation

After analyzing the data in various aspects, we have started the process of preparing our data for the modeling stage. In order to gain most of our data without losing too many observations and information, we have made some changes to our dataset. The transformations were made on both the output variable and on some input variables.

Output variable transformation

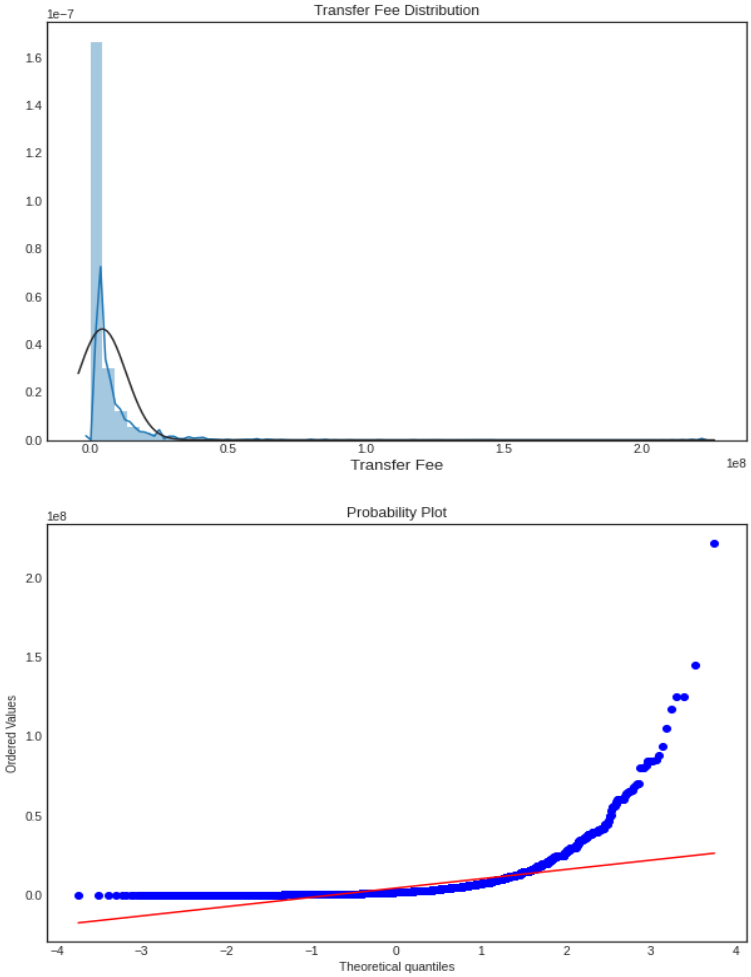

Figure 44: Histogram and probabilty plot of transfer fee(original scale)

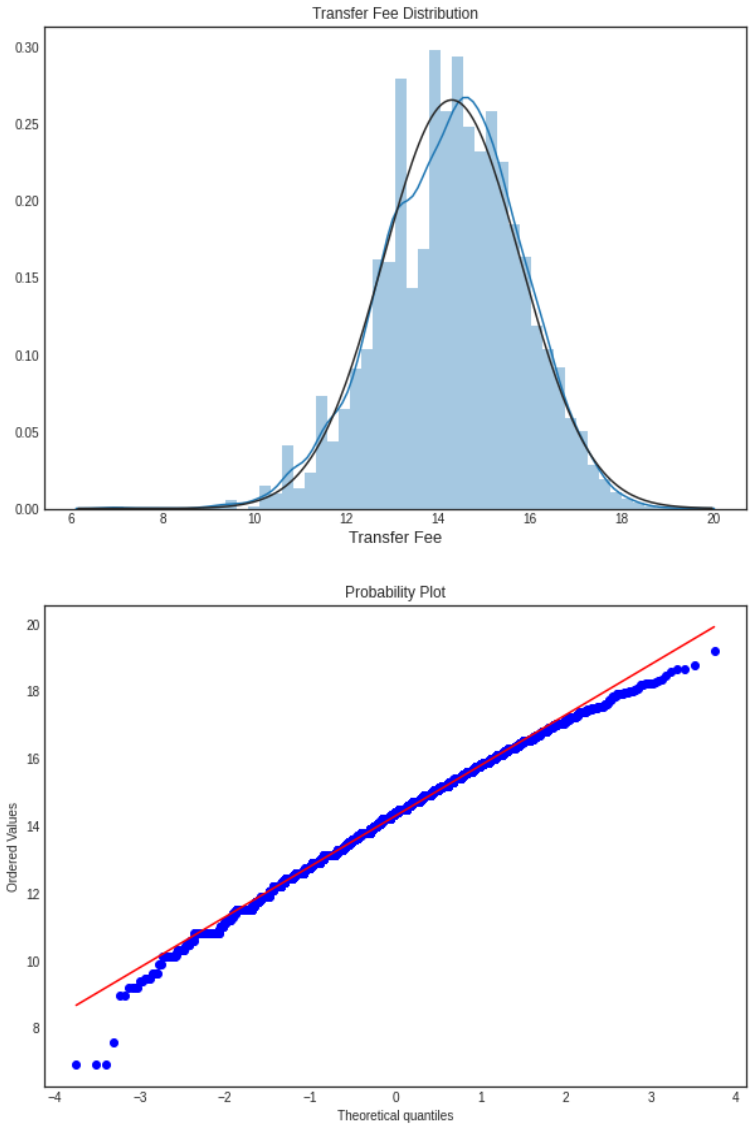

Figure 45: Histogram and probabilty plot of transfer fee(logarithmic scale)

As the output variable was highly skewed and we were losing a lot of observations when removing the outliers, we have decided to scale the output variable to logarithmic scale in order to normalize the output's distribution (Figure 44). We used $\log(1+x)$ to scale the output variable from millions scale to log scale, the same transformation was also applied to players' seasonal and cumulative market values at the time of the transfer as those variables were also from the same scale as the output variable (Figure 45). After applying

the transformation we have removed the observations for which the output variable was an outlier using the z-score approach.

One hot encoding of categorical variables

As the dataset contained some categorical variables that had contribution to the output variable we have encoded those using one hot encoder as some of them contained more than two levels and we did not want to use one variables for defining each as in that case we would have to assign levels of importance for each group for each categorical variable, we rather wanted the model to identify the importance of each group for each categorical variable. The categorical variables used were

- The continent of the player(6 groups)
- The strong foot of the player (Right,Left,Both or unknown)
- The window of the transfer(Summer, Winter, Mid-Season)
- The type of the year of the transfer(Tournament year or non-tournament year)
- The tactical position of the player on the field. (This variable was not used for goalkeepers, and had from to 3-5 levels depending on the position.)

2.3.2 Missing Value Imputation

After selecting the most significant variables based on the insights from the data analysis stage, we have imputed the missing values for each position. Most of the missing values were in ratios as the ratios were calculating using divisions of different statistical measures and sometimes the divisor was 0 leading to na values. Before applying an imputer algorithm we have identified the variables which we were going to drop. The criterion for keeping a variable with missing values was that it should have no more than around 30$ missing values. After identifying the variables missing data of which can be imputed according to our criterion, we have proceeded with the imputation algorithm.

Missing values imputation with KNNImputer

We have used the KNNImputer algorithm on the selected input features (García-Laencina, Sancho-Gómez, Figueiras-Vidal, & Verleysen, 2008). KNNImputer fills the missing values in the dataset by taking the mean of the k nearest neighbors of the observation, where k is

the parameter for how many neighbors to look for. The neighbors are identified based on the similarity of the non-missing features of the observation with other observations.

2.3.3 Feature Selection

There were a few stages of feature selection. We have defined a general pipeline that was used for all the datasets and there were two approaches to choosing the features inputted to the models. Firstly we have identified the most significant variables to the output variable using the insights gained during the analysis stage. As most of the features were strongly correlated with each other we later removed correlated input variables by keeping only one of them. Later we imputed the missing values. After these stages, the observations where the input variables were outliers using z-score were removed. After accomplishing all of these stages we have used 2 approaches for selecting the final input variables that were used for training the models and predicting. The two approaches of feature selection for the models were applied for each position and the one with the best results was kept.

Feature selection using *p*-value

We have used an iterative approach of keeping all the variables that have a significance for the output variable lying in a confidence interval of higher or equal than 95%, thus picking the variables that have p-value lower than equal to 0.05. The algorithm iterates over all the input variables each time removing the variable with the highest p-value and keeps the variables that meet the threshold criteria for the p-value calculated by fitting the input variables to the output variables using Ordinary Least Squares method. While using this approach we have given only the numerical features of the dataset as an input, without calculating the significance of the categorical variables in order to not face miscalculations during the p-value calculation process, as the columns with encoded categorical features take values of either 1 or 0.

Feature selection using RFR's feature importance

The second approach was using Random Forest Regressor's feature importance parameter. Random Forest Regressor shuffles each column while predicting and if as a result the predictive power decreases the variables' importance increases. We have inserted the dataset including the categorical variables and have chosen n columns sorted by their importance, where n was the number of desired input variables to use for the models

2.4 Modeling

Data preparation and exploratory descriptive analysis helped us to gain some crucial insights about the data. At this moment, we will implement commonly used machine learning and deep learning algorithms to predict footballers' transfer price. Since the characteristics of the football players may differ for each position, we decided to implement same set of models on each position separately (goalkeepers, defenders, midfielders, attackers). Since the transfer price is a continuous variable, we can confidently state that we need to solve a regression problem.

To start the analysis, firstly we should define the success metric for this particular problem, which will eventually help us to compare different models and analyse their performance. Understanding interpretation and differences between several regression metrics, we decided to use root mean squared error to evaluate different models. $RMSE$ stands for the square root of the average of squared differences between the actual and predicted values.

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

Since $RMSE$ squares the difference between actual and predicted values, it may eventually give huge weight to large errors, which will help us to choose model that predicts the output as precise as possible. After training the models, we have predicted the output using the test dataset and calculated the $RMSE$ for the outputs converted back to the original scale in millions in order to compare the models' prediction powers.

2.4.1 Machine Learning

Before starting to implement machine learning models, first we need to split our data into training and testing sets. Each dataframe will use 80 percent of observations to perform training and 20 percent of observations to validate models. At the end, we will also implement 5-fold cross validation in order to check how models will work on different splits of the data. Now let us see the set of machine learning models that were used to solve the problem.

1. Multiple Linear Regression.
    - No transformations were made on the training set.
2. Polynomial Regression.

- Training set was transformed by including different interactions of variables with the degree of 2.

3. Elastic-Net Regression.
   - The algorithm is used to perform regularization of features.
   - The penalty parameter alpha was tuned with cross validation from 500 different values.
4. Decision Tree Regression.
   - The parameters like maximum depth of the tree, minimum number of samples required to split internal node and maximum number of features were tuned using cross validation.
5. Random Forest Regression.
   - The number of trees and maximum depth parameters were tuned using cross validation.
6. Voting Regression.
   - Voting regression combines all models by taking the mean of all predicted values as a final output.

Above mentioned 6 models were implemented on all positions. To summarize the performance of the models, we calculated RMSE using 5-fold cross validation and took the mean of all folds to finalize the performance of the models. At the end, we can state that the model that outputs the minimum cross validated score can be considered as a best model for the following problem.

2.4.2 Deep Learning

After applying various machine learning algorithms using the structured representation of the data, we applied deep neural networks regression in order to identify whether the predictions can get any better when using unstructured data and different neural networks. We again applied the neural networks to each position using the same metric for success as in machine learning models. The only significant difference between the applications of the models on each position is that we used a simple neural network for the goalkeepers, as the number of observations was the fewest in this position and the experiments with more

flexible neural network structures did not improve the accuracy over the simple network. The structure of the network for the other positions was more flexible in comparison to the network used for goalkeepers and contained 4 layers. As the training process of neural networks is very costly, we did not implement cross-validation for the models, and in order to estimate the generalization of the models, we have used validation datasets. As the goalkeepers contained the least amount of observations we have used the test dataset in order to evaluate the performance of the model on unseen data during the training process. As for the other positions the number of observations was significantly higher we have used a 20 percent split of the training data as a validation data. In order to prevent overfitting, we have implemented early stopping callback on the training process, which will stop training if the loss of the model on the validation data does not have significant improvements over $n/10$ epochs, where n is the number of epochs used in the training and will restore the weights from the epoch with the best validation loss. We have also added L_1 regularizations on the kernel weights, activity function outputs, and on the bias.

$$L_1(x, y) = \sum_{i=1}^{n} (y_i - h_\theta(x_i)^2 + \lambda \sum_{i=1}^{n} |\theta_i| \tag{2}$$

The neural network built for goalkeepers also contains a dropout before the first hidden layer. The optimization algorithm used for both networks was Adam and the loss function was mean squared error.(Kingma & Ba, 2014) The activation function used for all the layers in both networks was relu. The networks for each position were trained on 2000 epochs.

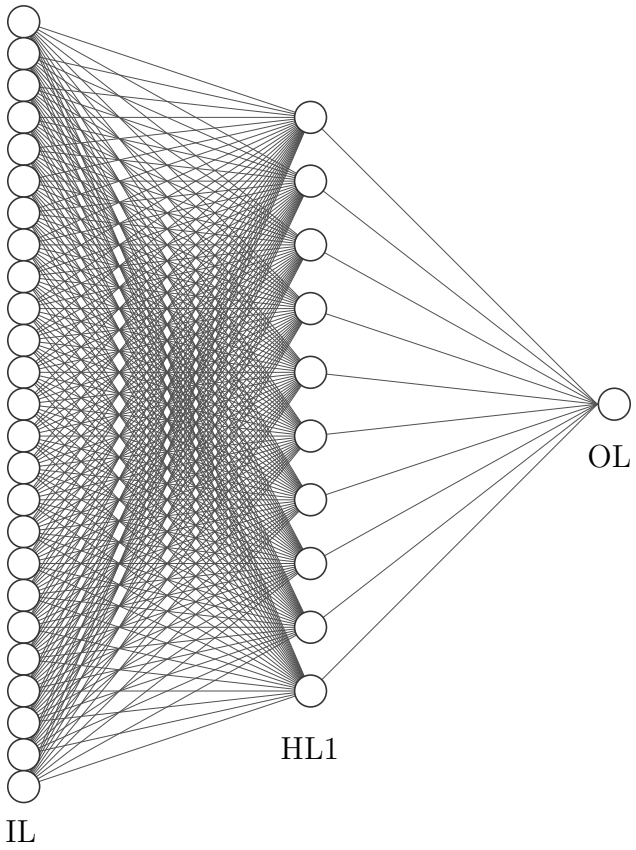

Figure 46: Neural Networks structure for goalkeepers

The neural network used for goalkeepers, contains 1 input layer, 1 hidden layer, and an output layer. The input layer contains 25 neurons, the hidden layer contains 10 neurons and the output layer contains 1 neuron (Figure 46)

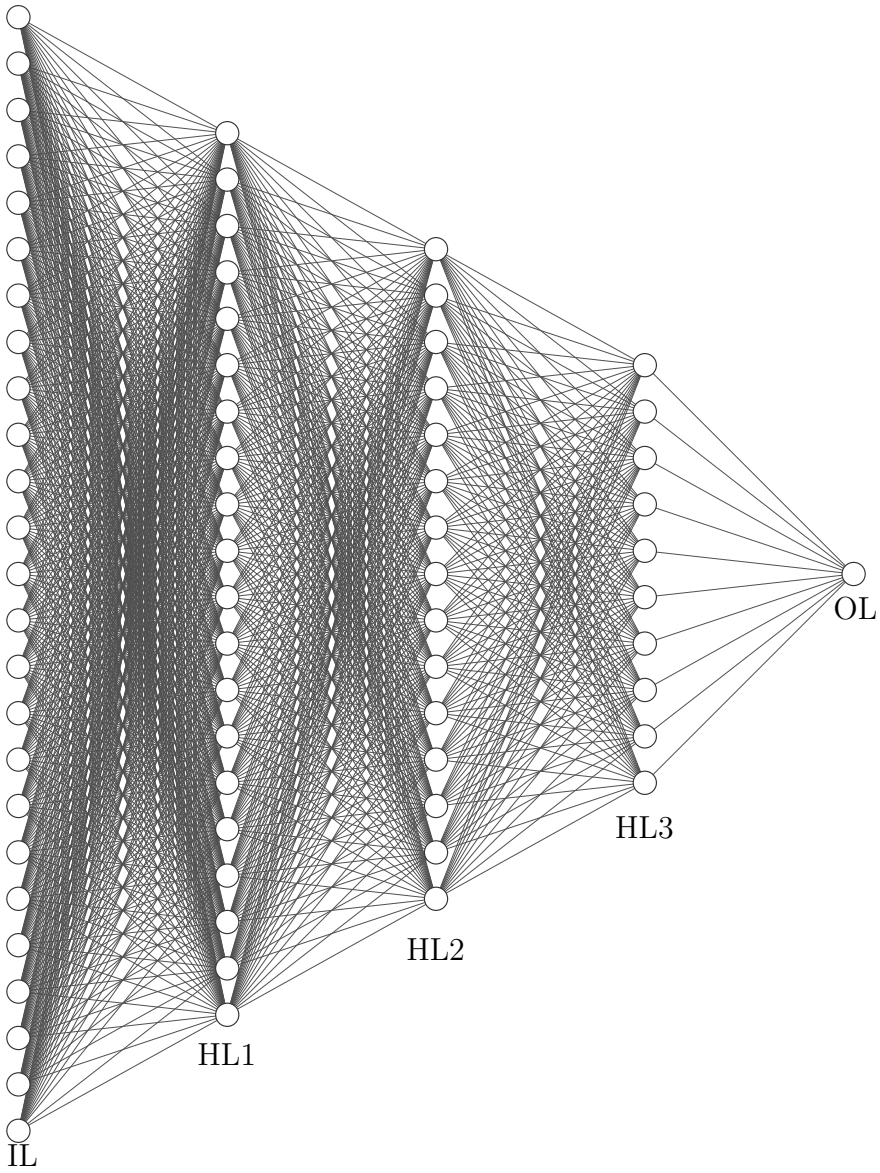

Figure 47: Neural Networks structure for other positions

The structure of the network used for other positions was more flexible containing one input layer with 100 neurons, 3 hidden layers with 50,25 and 10 neurons accordingly, and an output layer containing 1 neuron (Figure 47).

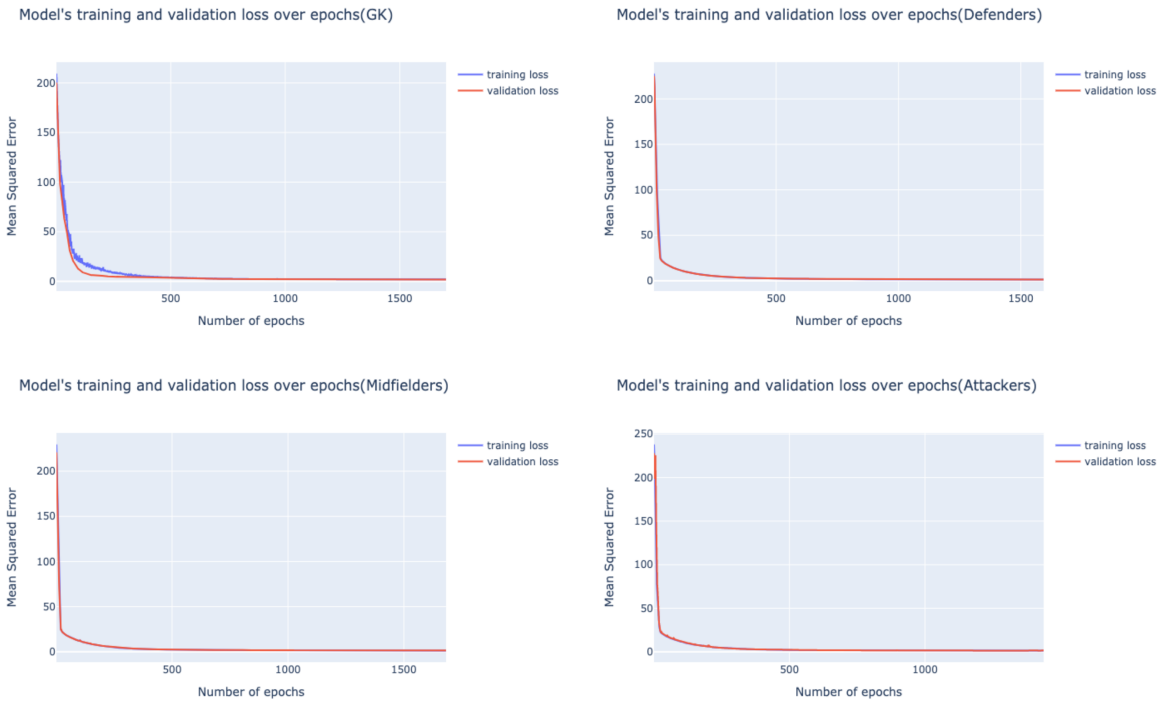

Figure 48: Models' loss for all positions

As we can see all of the models have generalization power and the validation loss decreases alongside training loss. Also, we can see that Early Stopping prevented the training at some point before the full number of iterations which was 2000 for each position, after seeing no considerable improvements on the model's loss on validation data (Figure 48)

2.5 Findings

2.5.1 Predictive Power and Interpretations

Machine Learning Results

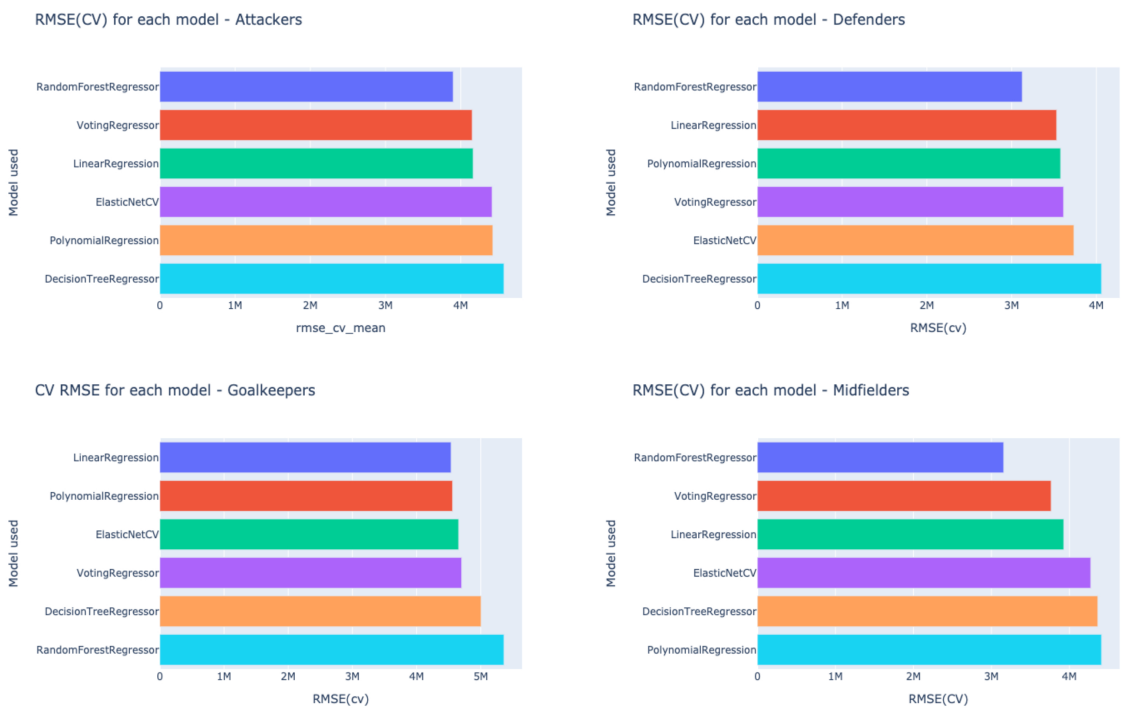

Figure 49: Cross validated RMSE for each position and model

After performing machine learning models we can see that the best performing models are Linear Regression on goalkeepers and Random Forest on other positions.

Due to the small number of observations in the goalkeepers' dataset, we can clearly see that the results are more biased compared to the other positions. Here we can also see that the best performing model Random Forest works best for defenders and midfielders (Figure 49).

Figure 50: Actual vs predicted values of the test set

Correlation	Goalkeepers	Defenders	Midfielders	Attackers
$Corr(y_{test}, y_{predicted})$	0.75	0.84	0.80	0.79

Table 6: Correlation table for actual vs predicted price (Machine Learning)

We can see that the predicted values and actual values of the test set have a strong correlation, which indicates the strong predictive power of the models (Table 6). The correlation coefficients between predicted and actual values were bigger than 0.7 (Figure 50).

Deep Learning Results

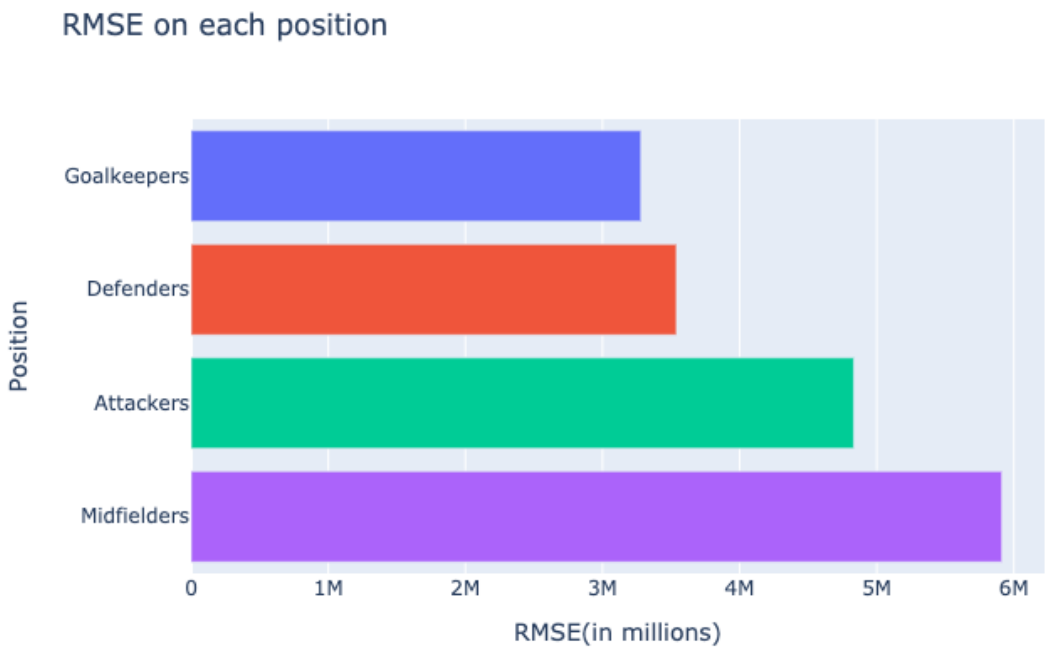

Figure 51: RMSE on each position(Neural Networks)

As we can see only the predictions for the goalkeepers have a noticeably smaller score of $RMSE$ in comparison to the best model from Machine Learning algorithms in each position. The $RMSE$ score for defenders is also very similar to the results of the best Machine Learning model for that position. The $RMSE$ scores for attackers is a little higher than the same score obtained using RFR. However, the $RMSE$ score for midfielders using deep neural networks regression is significantly higher compared to RFR for midfielders (Figure 51).

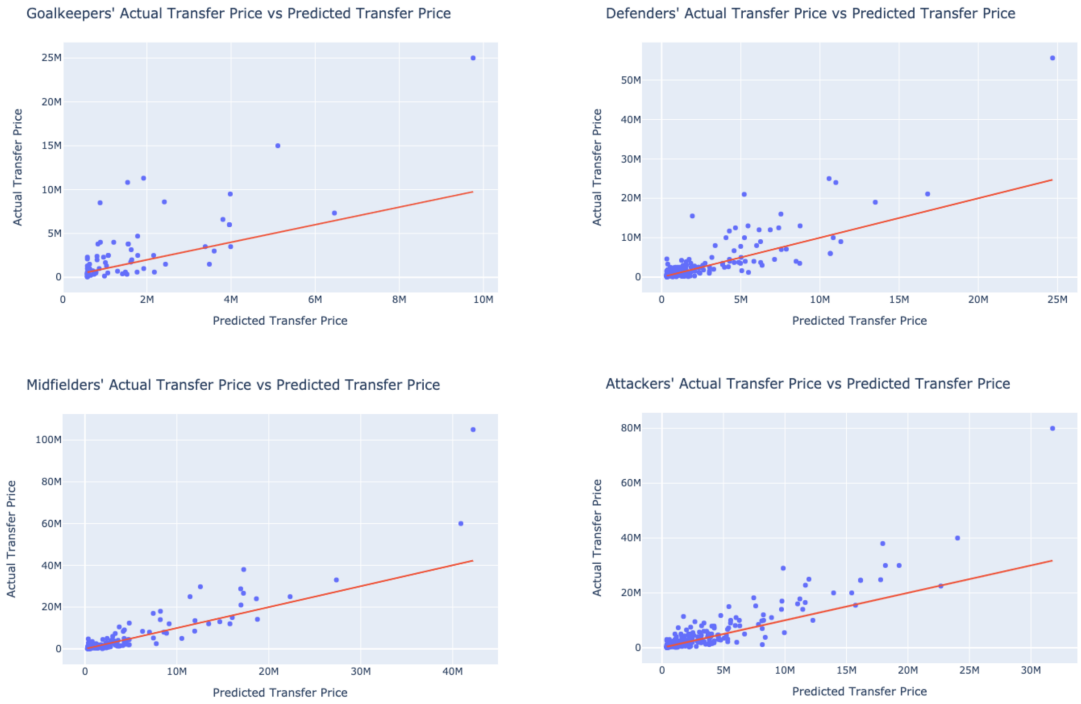

Figure 52: Actual vs predicted transfer price(Neural Networks)

Correlation	Goalkeepers	Defenders	Midfielders	Attackers
$Corr(y_{test}, y_{predicted})$	0.77	0.87	0.91	0.91

Table 7: Correlation table for actual vs predicted price (Deep Learning)

As we can see the predictions made using the neural networks regression have a higher correlation with the actual output, and the variance amongst the predictions is lower in comparison with the predictions made by the best machine learning algorithm for each position (Table 7). However, we can also see that the model generally tends to undervalue the players, and most of the time the predicted price for the players using neural networks regression was lower than their actual price (Figure 52).

The correlation score of predicted and actual variable increased for all the positions. The improvements in the correlation score were significant only for midfielders and attackers in comparison with the correlation score obtained by the best machine learning approaches used for each position.

2.5.2 Comparisons and best model selection

Before comparing and interpreting the results of the Machine Learning and Deep Learning approaches, let's first identify the main differences in the way we implemented them.

One of the key differences between the Deep Learning and Machine Learning approaches was the pipeline used for preparing the data for inputting to the models. The common parts of the pipeline amongst the approaches were the process of removing multi-colinear variables, removing the outliers observations amongst the input features using z-score. In contrast to Machine Learning algorithms, Deep Neural Networks did not use categorical data as the results seemed unchangeable in the case of both using and not using those, whereas the categorical variables made a huge impact on Machine Learning models. Another key difference between the pipelines of data preparation amongst the approaches was the process of final feature selection for the model.The 5 most important features according to RFR's feature importance criteria were selected and later transformed into $z - score$ for all positions in neural networks regression approaches. The approach of the final stage of the data preparation pipeline,which was the process of selecting the features for the training and test datasets was different amongst the positions in Machine Learning approaches.

- Goalkeepers (p-value based significance) + ppg
- Defenders (p-value based significance)
- Midfielders (10 most important variables according to RFR)
- Attackers (10 most important variables according to RFR)

The described approaches were a result of various experiments and the best results in terms of lower $RMSE$ score are represented. For goalkeepers the attribute of points per game was also added to the 95% significant variables as the initial p-value score for this attribute was very close to the threshold, and the model had noticable improvements after adding this variable. The input variables were not transformed in the training and testing datasets.

Interpretations and best model selection

We can see that in general, the Machine Learning algorithms have given better results in terms of lower $RMSE$. The only position for which the Deep Neural Networks regression usage leads to lower $RMSE$ was goalkeepers. The scores were also similar for Attackers and Defenders. So, overall the Deep Neural Networks regression approach worked considerably well for the positions for which the relationship was very close to linear. If we check the cross-validated $RMSE$ scores in Machine Learning Results we can see that for the mentioned positions the obtained $RMSE$(CV) score in Multiple Linear Regression algorithm was close to the best model's score and for goalkeepers, the Multiple Linear Regression usage leads to the lowest $RMSE$(CV) score. Taking those facts into account we can conclude that for this problem, the deep neural networks do not work particularly well, and had closer predictions to Machine Learning algorithms only for positions for which the relationships between the input variables and output variable were close to linear.

According to our main criteria for selecting the best models, the Random Forest Regressor algorithm works best for all the positions except Goalkeepers. For goalkeepers the lowest $RMSE$ score was obtained using deep neural networks, so according to our criteria it is the best model for this position, but if take into account that the differences amongst the $RMSE$ scores obtained by the Multiple Linear Regression approach and Deep Neural Networks approach were around 1million, but the cost of training the neural networks is much much higher than the cost of training the Multiple Linear Regression model, we choose the Multiple Linear Regression as the best model for goalkeepers, as it is much more simple and the predictive power is not very significantly lower compared to Deep Neural Networks Regression.

Feature Importance

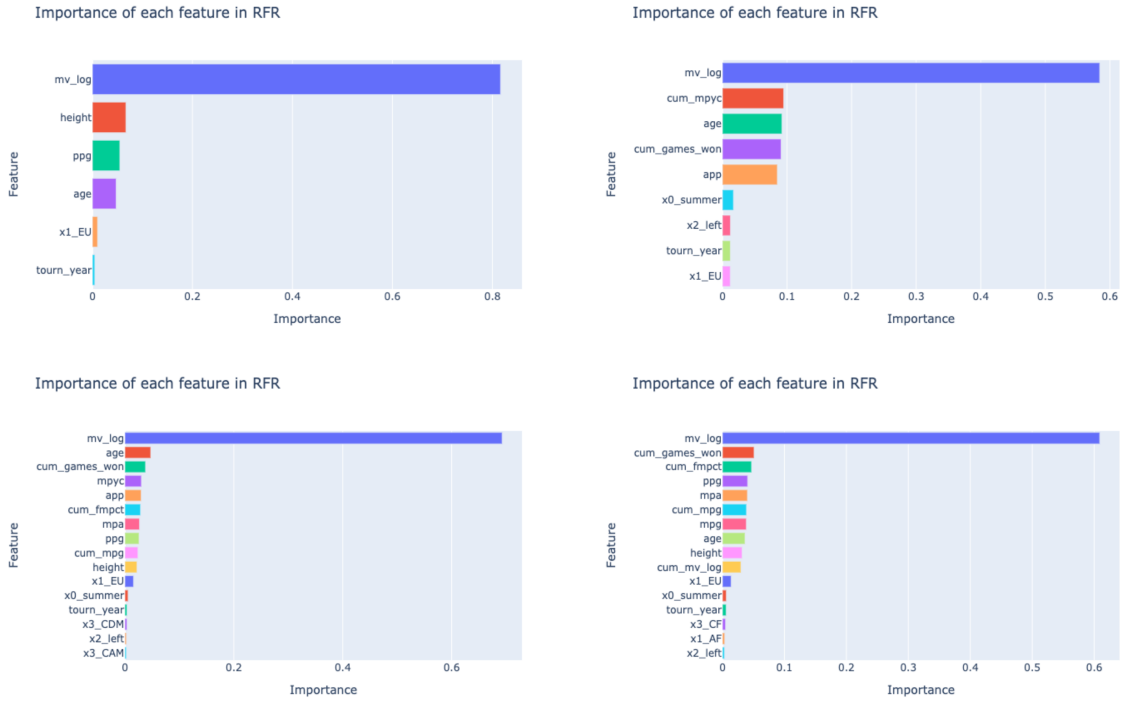

Figure 53: Most important features according to RFR

We have used Random Forest Regressor's feature importance for each position in order to identify which of the features have the highest contribution to the fee according to the trained model. As expected after the data analysis stage, the market value of the player is the most important feature in all positions. We can also see that the points per game has relative importance for each position, as this is mostly a team attribute and not specific to any position, but rather its a combination of each positions' players effort on the pitch. The importance of the other features is considerably lower. Let's interpret the most significant variables except market value for each position according to RFR.

For goalkeepers, the most important attribute after the market value is their height, as taller goalkeepers generally have better goalkeeping skills. The age also seems relatively significant. From the encoded categorical variables, only two have importance, and the importance is very low. The variables are goalkeeper's race(it is important whether the player is European or not), and the transfer's year type indicating whether a tournament took place during the transfer's year. (Figure 53, upper left figure)

For defenders age, the cumulative number of won games and minutes per yellow card and the number of appearances during the transfer's year are the most important. The strong foot of the player(left-footed defenders only) and European nationality are the important categorical variables, alongside with the transfer specific variables, from which the important ones for defender are whether or not the transfer's year was a tournament year and whether the transfer took place during the summer transfer window.As described in the analysis the defenders usually gain many yellow cars, and the expensive ones gain yellow cards in a lower frequency, thus having high numbers for mpyc.(Figure 53, upper right figure)

For midfielders the age, the cumulative number of won games, minutes per yellow card during the transfer season, appearances, cumulative playing percentage, and minutes per assists are the most important, alongside height and cumulative minutes per goal. The relatively important positions for the price of midfielders are central defensive midfielder and attacking midfielder and the other categorical variables' importance is similar to defenders. We can see that for midfielders both attacking and defending attributes have a contribution to their price, as midfielders are generally the most balanced players. (Figure 53, lower left figure)

For attackers, the important performance attributes are pretty similar to midfielders, except defensive performance metrics not being important for attackers. We can also see that for attackers most of the attributes are important on the cumulative basis, including the player's cumulative mean of his market value before the transfer. So attackers, in general, are priced based on their whole career performance rather than season based performance. (Figure 53, lower right figure)

As for goalkeepers the model with the lowest cross validate $RMSE$ score was Linear Regression let's also interpret the coefficients for each feature.

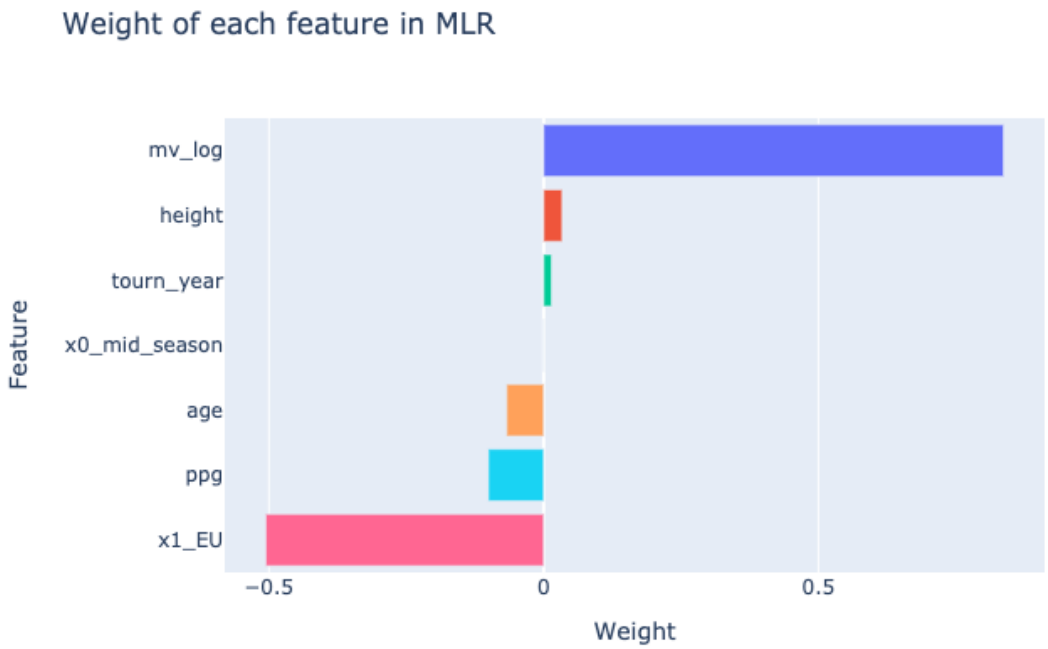

Figure 54: Weights of each feature according to MLR(Goalkeepers)

As we can see only the height of the player, his market value's log, and the type of the transfer's year have positive weights, and the other variables have negative weights. The mid-season transfer window type has a very unsignificant negative coefficient, as in general transfers made during the Midseason are rare and especially for goalkeepers, for which transfers, in general, happen rarely. We can also see that age, points per game and the Europan nationality of the goalkeepers have negative coefficients. Age's relationships to the fee were expected to be negative as in general old players are valued lower. We can also see that the European nationality of the goalkeeper has a negative coefficient, which is explained by the fact that most of the European players play in EU leagues, and most of the EU teams changed their goalkeepers very rarely, and for many teams, the goalkeepers stay in the starting eleven for even around 10 years, and later as they get older they are sold to other teams for low values, most of the European goalkeepers are also bought when they are young, thus, in general, having a lower price. The negative weight for the points per game variable indicates that if the team has ppg rates, thus a high winning percentage, the goalkeepers probably are not the main players responsible for that (Figure 54).

Overall interpretation

To conclude we can see that with the usage of the only performance-based variables, the model mostly undervalues the players rather than overvalues them, meaning that according to the models we have used the transfer fee of the player is not only dependent on his performance-based statistics. We can see that the transfer specific categorical variables such as tournament year and transfer window were important for each position. So, our findings in the analysis can be confirmed as in years during which a national tournament takes place the prices of the transfers seem to get higher. Another common measure amongst all the positions was the race of the player indicating the importance of only the European players, which again shows that the players from other continents are generally undervalued and for European players their nationality is an advantage. We can also see that the relationship between the performance metrics and position gets more complex over positions, being very close to linear for goalkeepers and defenders and getting more complex for midfielders and attackers. A crucial insight from the analysis about the popularity can be related to this, as the midfielders and attackers gain the most popular amongst social media, and taking into account the current tendencies of soccer, their popularity can be a key factor for their price. We did not include the player's popularity rate as the data of Instagram followers was missing for most of the players, and in order to be more precise about the player's popularity his contracts with brands and his desirability in the advertisement market should be taken into account, but the data for those attributes was not available.

3 Summary

3.1 Interactive Dashboard Application

As a visual representation of collected and analyzed data, we used plotly.dash to create a simple dashboard. This dashboard contains table representations of dataframes we have used, as well as the scatter plots and line charts of market value and transfer fees of the players in relation to their age, current year, last year's, and cumulative market values as well as positions, continents and major leagues. The dashboard uses Dash Core Components (DCC) and Plotly Express (PX) for converting the data frames to a more user-friendly appearance. The dashboard is deployed using Heroku servers.

3.2 Recommendations

Although we saw deep analysis of the football players' data, there is always a room of improvement that can be done to improve projects results and further the studies. Here you can find some important points that may increase the value of the project in the future.

- **Adding more features**.

  There exist wide variety of data that can be added to the main dataset, which may eventually increase predictive power and quality of interpretations. This step is heavily based on strong domain knowledge, as well as applying creativity in the feature engineering process.

- **Adding more observations**.

  As you can see at the beginning of the paper, we exclusively used publicly available data to make the following analysis. There are numerous private and public data sources, using which will greatly impact the analysis by adding more observations.

- **More hyperparameter tuning**.

  Tuning the large set of hyperparameters may sometimes be limited by time and computational resources. Thus, adding computational power will also help to tune more hyperparameters, which may eventually lead to better performance of machine learning and deep learning models.

3.3 Conclusion

With this project, we aimed to fully analyze the data by implementing in-depth exploratory descriptive analysis, which also included some crucial insights from network and time series analysis. With the help of a large amount of information obtained about the data, we tried to predict football players' transfer fee, unlike other projects which were more concentrated in predicting football players market value. Based on the domain knowledge it was decided to apply algorithms separately based on positions (goalkeepers, defenders, midfielders, attackers). As a result, different models were selected for each position based on cross-validation obtained $RMSE$ score. The insights gained from the analysis stage and the predictive power of the models used for each position can be used in order to identify the undervalued and overvalued players based on their performance and to identify the best conditions and time for selling the players to gain the maximal profit.

References

Barbuscak, L. (2018). What makes a soccer player expensive? analyzing the transfer activity of the richest soccer. *Augsburg Honors Review*, *11*(1), 5.

Football transfers, rumours, market values, news and statistics. (n.d.). Retrieved from `https://www.transfermarkt.com/`

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2008). K-nearest neighbours based on mutual information for incomplete data classification. In *Esann* (pp. 37–42).

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 855–864).

He, M., Cachucho, R., & Knobbe, A. J. (2015). Football player's performance and market value. In *Mlsa@ pkdd/ecml* (pp. 87–95).

He, Y. (2012). Predicting market value of soccer players using linear modeling techniques. *University of Berkeley (working paper)*.

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.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.*

Kuper, S. (2014). *Soccernomics: Why england loses, why spain, germany, and brazil win, and why the us, japan, australia and even iraq are destined to become the kings of the world's most popular sport.* Nation Books.

Leone, S. (2019). *FIFA 20 complete player dataset.* Retrieved 2020-05-03, from `https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset`

Maurya, S. (2018, Jun). *Can linear models predict a footballer's value?* Retrieved from `https://towardsdatascience.com/can-linear-models-predict-a-footballers-value-33d772211e5d`

Polim, R., Ravenel, L., & Besson, R. (2018, Oct). *Scientific assessment of football players' transfer value.* Retrieved from `https://football-observatory.com/IMG/pdf/note01en.pdf`

Qu, Y., Ostrouchov, G., Samatova, N., & Geist, A. (2002, 04). Principal component analysis for dimension reduction in massive distributed data sets..

Quick, M. (2017, Aug). *How does a football transfer work?* Retrieved from `https://www.bbc.com/worklife/article/20170829-how-does-a-football-transfer-work`