MACHINE LEARNING FINAL PROJECT:

EVALUATING MARKET VALUE FOR PREMIER LEAGUE PLAYERS



Group 25

Deeptish Mukherjee

Jordan Deng

Miaoqi Yang

Yvette Peng

Contents

Executive Summary………………………………………………………………………………….3

Background: The Greatest Show in the World….....……………………………………..4

The Inefficiencies and the Low-hanging Fruit in the Current Market…………..4

Scope of the Study…………………………………………………………………………………….5

Analysis……………….…………………………………………………………………………………..6

Recommendations and Business Insights…………………………………………………11

Conclusion………………………………………………………………………………………………12

Appendix…………………………………………………………………………………………………13

**Executive Summary**

The English Premier League is one of the most widely watched sporting events in the globe. A large part of the business of managing the teams competing in this league has to do with player transfers between clubs. Each of the last few Premier League seasons saw the clubs spending $1B on transfers. Despite the high-stakes nature of these transactions and the impact on a club’s treasury, the transfer market is rife with inefficiencies, with teams routinely overpaying or undercharging for players.

In this report we took stock of all of the factors that drive player value for each position on a football pitch, and attempted to quantify the major factors that drive market value of Premier League players, that ultimately are a significant determinant of transfer fees. Through our supervised learning methods, we tried to disentangle the effects of their experience, abilities and roles in their teams on their market value, and to make actionable recommendations for football managers operating in the transfer market. We also created predictive models using linear regression and random forests to predict the market values of these players.

Through our analysis we found how the effects of playing for a top club or being a domestic player can seem misleading when aggregated. We also narrowed down specific performance metrics that were deemed most significant contributors to player value by our model, such as goals, assists and passes for forwards and midfielders, which were obvious, but some that were counterintuitive such as recoveries for defenders and take-ons for midfielders. Finally, we used these findings to provide recommendations to management personnel about being cautious about overvaluing English forwards, and undervaluing English defenders and midfielders, to stay clear of certain heuristic traps and to strategically factor experience and age while operating in the transfer market. Simply choosing carefully between domestic players and English players in the right position can save a team **$20M** over a season while buying 3 outfield players, without needing to compromise for quality.

**1. Background: The Greatest Show in The World**

The English Premier League is the most watched competition in association football, with a [74% increase in viewership](https://worldsoccertalk.com/2019/08/21/premier-league-viewership-grew-74-last-decade-us-tv/) in the US alone in the last decade. It is broadcast in 212 countries to 643 million homes and a potential TV audience of 4.7 billion people. 20 teams compete each season to compete for the top prize while trying to bolster their teams to the best of their capabilities by paying huge transfer fees to buy players. This results in teams paying amounts as high as $110M for a single player with a median of $12M being paid for players in the Premier League. In 2018 and 2019, Premier League clubs spent a combined amount of $1.9B, which is 200% times higher than the amount spent in the first 2 years of the decade. In the year 2018-2019, European

**2. The Inefficiencies and the Low-hanging Fruit in the Current Market**

While it is not uncommon for teams to invest a lot of money to find players repaying the favor in the form of on-pitch performance, an overwhelmingly large number of players join a team and fail to perform at a level commensurate to the price that was paid for them. In this backdrop, there is a lot of value in trying to understand the factors that contribute to a player’s market value, and exploring how any such findings would help football decision-makers come transfer season.

Stakeholders in the soccer business acknowledges the rapid inflation in the transfer market over the last decade, and only over half of all teams competing in Europe can claim to be profitable. A large part of the problem are player transfer fees and wages, and for a number of smaller teams in the Premier League struggling to compete in the same league with financial behemoths like the 2 Manchester clubs (United and City) are trying to leverage analytics to find competitive advantages by using analytics. One of the areas of application are to identify over-valued players in the market and stay away, while identifying undervalued players that may be picked up for a low price. This is the area we wish to focus on for this report.

**3. Scope of the Study**

Most transfer fees are also influenced by a large number of “*non-footballing” factors outside of market value as well, such as the number of years spent left in their current contracts, negotiating power of player agents, free agent status, family relocation, cultural rehabilitation and fit within the team, to name a few*. However, for this report we will work with the assumption that transfer fees closely follow the market value, and will consider that predicted market value is an indicator of the transfer fees that should be paid for the player.

The set of players used for this study was all the players who had played more than 5 games in the Premier League and had a market value listed on the [TransferMarkt](https://www.transfermarkt.us/) website, our source for the market value data. This furnished us with a dataset of **432** players across the 20 teams, with as many as 50 different factors to evaluate them across.

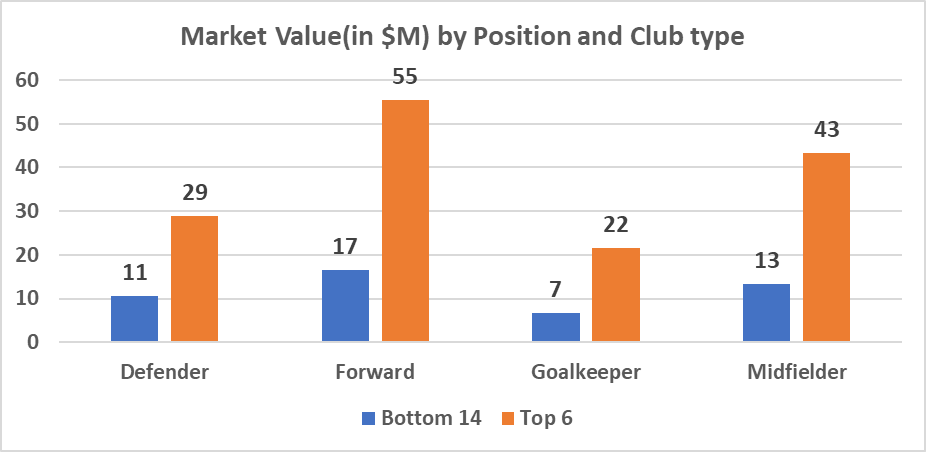
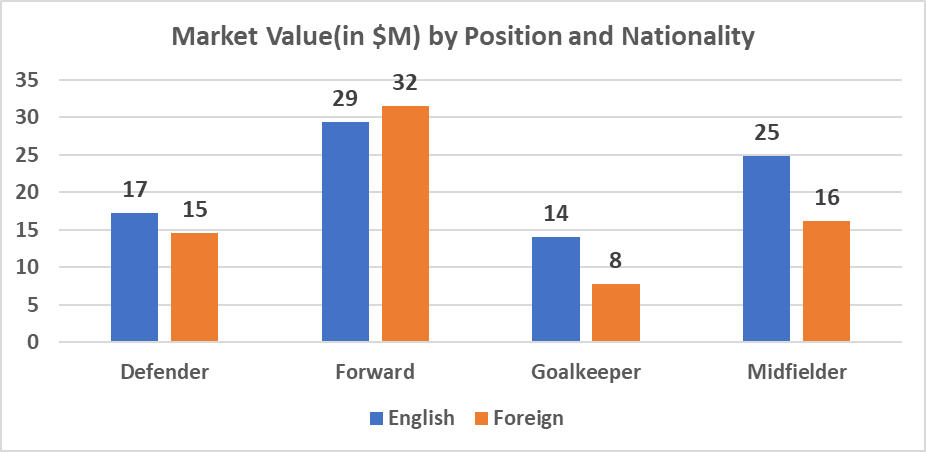
The nature of football is such that the traits that determine the market value of a player vary greatly across positions. This led us to break the datasets down by positions (Keepers, Defenders, Midfielders and Forwards) so that we could evaluate them on the basis of different traits that made footballing sense. This provided us with **42 Keepers, 154 defenders, 152 midfielders and 84 forwards** to work with.

Within defenders and midfielders, there are differences in traits that generally distinguish full-backs and center-halves (for defenders) and central midfielders and wingers (for midfielders), but this tagging was unavailable in the data we sourced.

**4. Analysis**

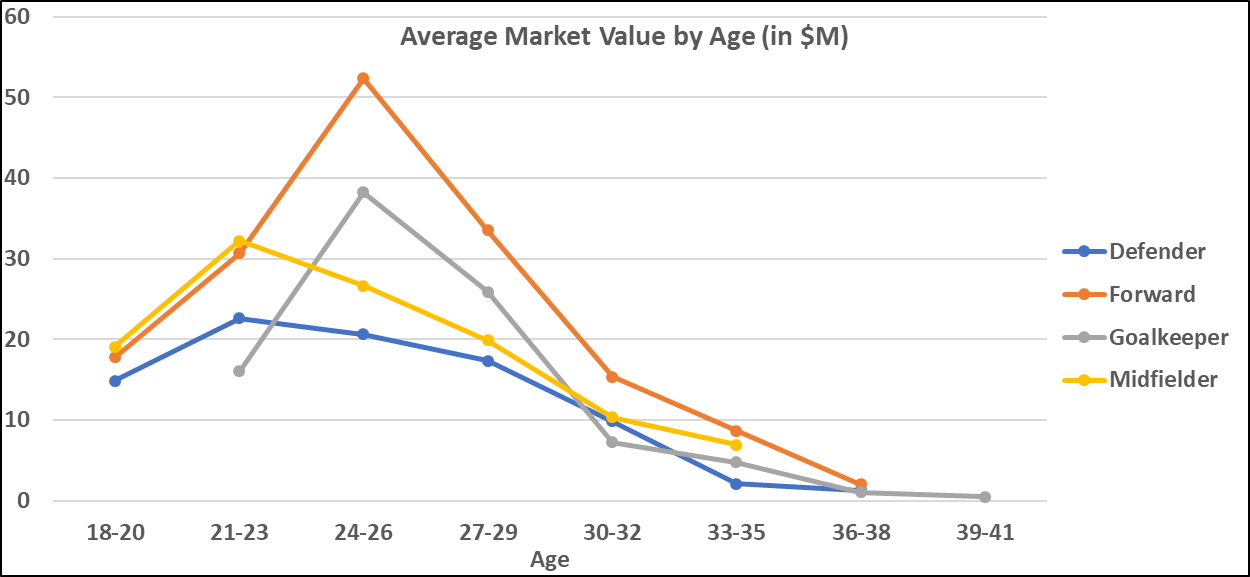
**4.1 Getting the Data Ready**

**4.1.1 Handling Dimensionality**

An important matter that needed to be dealt was the high-dimensionality of the data. Given that we had decided to treat this as **4 separate analyses** based on positions, **50** factors were far too many to run a reliable learning technique on for any 1 of the groups. We used our understanding of the game to eliminate features that we assessed to be trivial determinants of market values for a specific position. For example, number of goals scored per games is not a relevant field for a keeper’s market value to be dependent on, while the number of goals conceded per game is not a factor which determines a forward’s market value. There were also 2 categorical fields that needed to be treated, namely the Nationality (**54 levels**) and the Club (**20 levels**). These were invaluable fields, as **top 6 clubs** have an average market values more than **200% higher** than the **bottom 14** clubs, while the domestic English players (who make up **66%** of all players) across the board cost **23%** less than foreign national players on an average. To handle the high-dimensionality of these categorical variables we created 2 flags: **the top club flag** which indicated whether the club was one of the top 6 clubs, and a **flag to indicate whether the player was English or of a different nationality**.

**4.1.2 Handling Survivor Bias**

An illustration of how significant the position-level breakdown of our analysis was is provided in the chart below, where we show the variation of player value with age, broken down by position.



The next step for us after narrowing down the dimensions of interest for each position, we had a take care of **survivor bias**, which ensures that players with higher number of games played would clock higher statistics and would confound our understanding of which metrics actually affect market value. So, **we normalized the metrics to per-game metrics** to ensure that we were dealing with intensive quantities.

**4.2 Defenders**

To disentangle the effect of various metrics in an interpretable way, we chose to create a multiple regression model. Using a step-wise elimination process, we identified the key metrics that drove defender market value, as described below.

**4.2.1 Age and Appearances:**

For defenders the peak market value for players keeps going up till the age of **23** by about **$2.9M** on average for every incremental year and then it keeps going down by **$2.6M** on average for every additional year, all other things being equal. Every additional appearance adds **$50K** to a player’s market value on an average.

**4.2.2 Club and Nationality:**

All other things equal, the market value of an English defender is **$4.3M** less than that of a foreign player, contrary to it being **$2M** higher. *Interestingly, all things equal, a top club defender is not evaluated any higher than a defender in any other club*.

**4.2.3 Performance Metrics:**

All things equal a defender valuation increases by **$480K** for every incremental *pass per game*, **$320K** for every *clean sheet kept every 100 games* and by a staggering **$2.3M** with every additional *recovery per game*.

**4.3 Midfielders**

**4.3.1 Age and Appearances:**

For midfielders the peak market value for players keeps going up till the age of **23** by about **$7M** on average for every incremental year and then it keeps going down by **$2.9M** on average for every additional year, all other things being equal. *Additional appearances are not a significant driver of a midfielder’s value*.

**4.3.2 Club and Nationality:**

All other things equal, the market value of an English midfielder is **$4M** less than that of a foreign player, a difference far less than the apparent **$9M**. All things equal, a top club midfielder is evaluated **$14M** higher than a midfielder in any other club, not **$30M**.

**4.3.3 Performance Metrics:**

All things being equal, a midfielder’s valuation increases by **$380K** for every incremental *pass per game*, **$358K** for *every assist per 100 games*, by **$260K** for every additional *game won in a 100 games played*, by **$500K** with every *goal scored every 100 games* and by a massive **$2.2M** for every *additional take-on in a game*.

**4.4 Forwards**

**4.4.1 Age and Appearances:**

For forwards the peak market value for players keeps going up till the age of **25** by about **$9.4M** on average for every incremental year and then it keeps going down by **$6.4M** on average for every additional year, all other things being equal. Additional appearances are not a significant driver of a forward’s value.

**4.4.2 Club and Nationality:**

All other things being equal, the market value of an English forward is **$10.6M** more than that of a foreign player, which is very different from the aggregated **$3M** difference. All things equal, a top club forward is evaluated **$17M** higher than a forward in any other club, which is far less than the apparent **$39M** difference at an aggregate level.

**4.4.3 Performance Metrics:**

All things equal a forward’s valuation increases by **$840K** for every *incremental pass per game*, **$1M** for *every assist per 100 games* and by **$990K** with *every incremental goal scored per 100 games*.

**4.5 Keepers**

**4.5.1 Age and Appearances:**

For keepers the peak market value for players keeps going up till the age of **26** by about **$5.8M** on average for every incremental year and then it keeps going down by **$2.2M** on average for every additional year, all other things being equal. Additional appearances are not a significant driver of a keeper’s value.

**4.5.2 Club and Nationality:**

All other things equal, the market value of *an English keeper or a top club keeper is not significantly different from a foreign keeper or any other club*, which is far less than the apparent **$6M** difference between English and foreign keepers or the **$15M** difference between a top club keeper and a bottom club keeper.

**4.5.3 Performance Metrics:**

All other things equal, a forward’s valuation increases **$1M** for every *win per 100 games* and very counter-intuitively goes down by **$400K** with *every clean sheet every 100 games*.

**4.2.2 Predicting Market Value:**

To predict the market value of all players in each position we used 2 competing models: linear regression and random forests. The linear regression model was trained on randomly selected training sets, from **2000** different samples. The random forests were made from **500** trees over **50** different samples. **65:35** split between train and test was carried out for both models.

*The linear regression outperformed the random forest model for defenders and midfielders, but the random forest model outperformed the linear regression model for forwards*. The 42 data points for the keepers did not allow us to create fruitful predictive models using either of the techniques, and the OOS performance for both models was extremely poor.



**5. Recommendations and Business Insights**

1. All other factors held constant, it makes far more sense for mangers to target English midfielders (*as they cost* ***$4M*** *less than foreign*), forwards that are foreign (*as they cost* ***$11M*** *less than English*) and English defenders (*as they cost* ***$4.3M*** *less*).
2. All other things equal, experience in terms of number of games only increases that value of defenders. *Team management personnel should be wary of extracting as much value as they can of young and experienced defenders that they sell to other clubs, as each additional game adds a price premium of* ***$50K***.
3. *There is a tendency to overvalue top 6 club midfielders by* ***$14M*** *and forwards by* ***$17M***. This implies that clubs can save money by *targeting players of similar caliber in these positions without paying an unnecessary price premium*. In contrary, defenders from top 6 clubs do not cost significantly more than their lower table counterparts, if all things are the same, which means that the difference in $18M in price is due to the difference in quality. The market for defenders by this evidence is the most efficient.
4. Age plays a massive part player valuation, and for cash-strapped team managers should try to *sell their players as close to their peak years to derive the maximum value possible*. *The difference of selling a promising forward at 26 instead of 25 could be as high as* ***$6.4M***.
5. Player traits such as successful take-ons for midfielders and ball recoveries for defenders drive massive value, while often seemingly important traits such as accurate crossing for midfielders and blocks and interceptions for defenders do not appear significant. This may be because recoveries and take-ons have a certain visibility element which grabs a certain amount of attention. Caution must always be exercised so that such traits are not over-valued over direct contributors such as assists and completed passes.
6. Forwards gain more value with every incremental assist as they do with every incremental goal, and more than midfielders do with every assist. This implies that managers with forwards that assist regularly should look to capitalize of them consciously.
7. **Conclusion**

There is a lot value to be mined from the data available surrounding soccer. The scope of this report was narrow, but the study could be extended to all of the top European leagues to get a more comprehensive flavor. Additional to the ones already mention, there are a few more limitations in this study:

* The number of appearances of the players was limited to the Premier League, even if they had played in other leagues
* There was a dearth of useful metrics on goalkeepers which hindered their analysis
* There are problems regarding treating all goals and assists the same, as some events occur at key and crucial times, and are thus more vital to a team’s success. An example would be the difference between a goal scored to win a game 3-2 in the last minute of a game vs the 6th goal of a 6-0 win.
* Players who play with better players are more likely to have better stats, a problem that is only partially tackled by the top club flag.

A more holistic study that can account for these shortcomings would be able to provide even more value than we hope to have delivered through this report.

**Appendix**

**Definitions:**

* Assist: The last pass in a move that leads to a goal.
* Recovery: The act of dispossessing an opponent player and emerging with possession
* Interception: The act of a player from an opposition team intercepting a pass by moving into the line of the intended pass
* Block: The act of stopping a shot intended for the goal
* Cross: The act of putting the ball into a potentially goal-scoring area from the flanks of a football field
* Successful Take-on: The act of beating a player one on one by speed on trickery
* Clean Sheet: The act of not conceding a goal in a game