A study of factors affecting the soccer Player Market Values in the Premier League



By:

Ziad M and Phil Nguyen

MATH 242 – Final Project

**I. Summary:**

The main goal of this project was to see what are the most important variables that contribute a significant effect on the market value of soccer players in the Premier League. We merged two datasets that combined statistics from the Fantasy Premier League game and player’s real-life performance from the 2016/2017 season. We decided to split the indicators we have into three main categories to see if there are significant predictors that we could point out. Those categories were split between, just Fantasy’s three main statistics (points, value and selection percentage), the individual player’s skills determined by FIFA pointing system (overall, pace, passing, shooting, defending, physicality, dribbling), lastly, the player’s information (plays for a big club, international reputation, age, height, weight, etc.). We have indicated that those sub categories were not strong enough to pin down what are the predictors that help us understand the player’s market value. As a result, we ran a stepwise model compiling all the predictor’s together and having the model output the most significant ones. We executed a model based on the results from the stepwise that we thought was best at predicting the player’s value. After coming up with the best model, we wanted to see if we could apply it to real life situations. The first application was to find overvalued players and how big is their overall contribution to their respective teams. The second application, was trying to find the optimal 15 players by using our observations based on the model and few further findings such as mean values and points for players in the Fantasy Premier League Game.

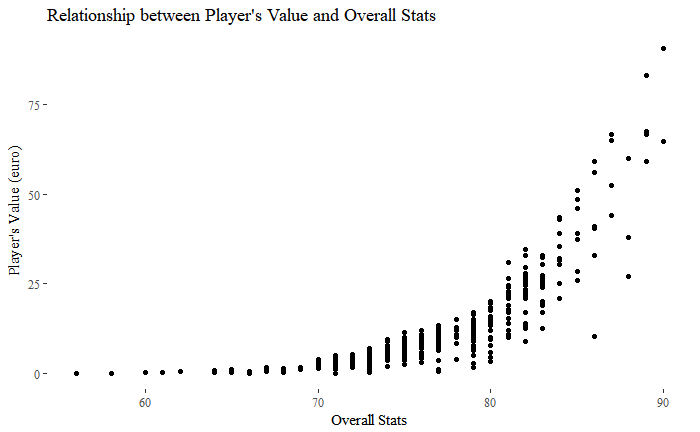
**II. Reason for Study:**

In recent years, we have seen some changes in the soccer world where teams are getting bought by wealthy owners and it has become an investment for multi-billionaire individuals. Therefore, the money element has changed not only how we look at the game, but also the function of player transfer market. Players transfer fees have skyrocketed and are still growing at an exponential rate. Clubs are now in a position where players are costing more than their actual value, and the price tags of the players are heavily inflated. For example, Neymar Jr, a Brazilian player, left Barcelona and joined PSG this summer for a record transfer fee of 222 million euros, two times the last record of 105 million euros, an overwhelming price tag for a 25-year-old player. We thought that the inflation in transfer fees in the soccer world is an interesting topic, and therefore we conducted this project to have a thorough analysis of how money has changed the perspective of the most famous sport in the world.

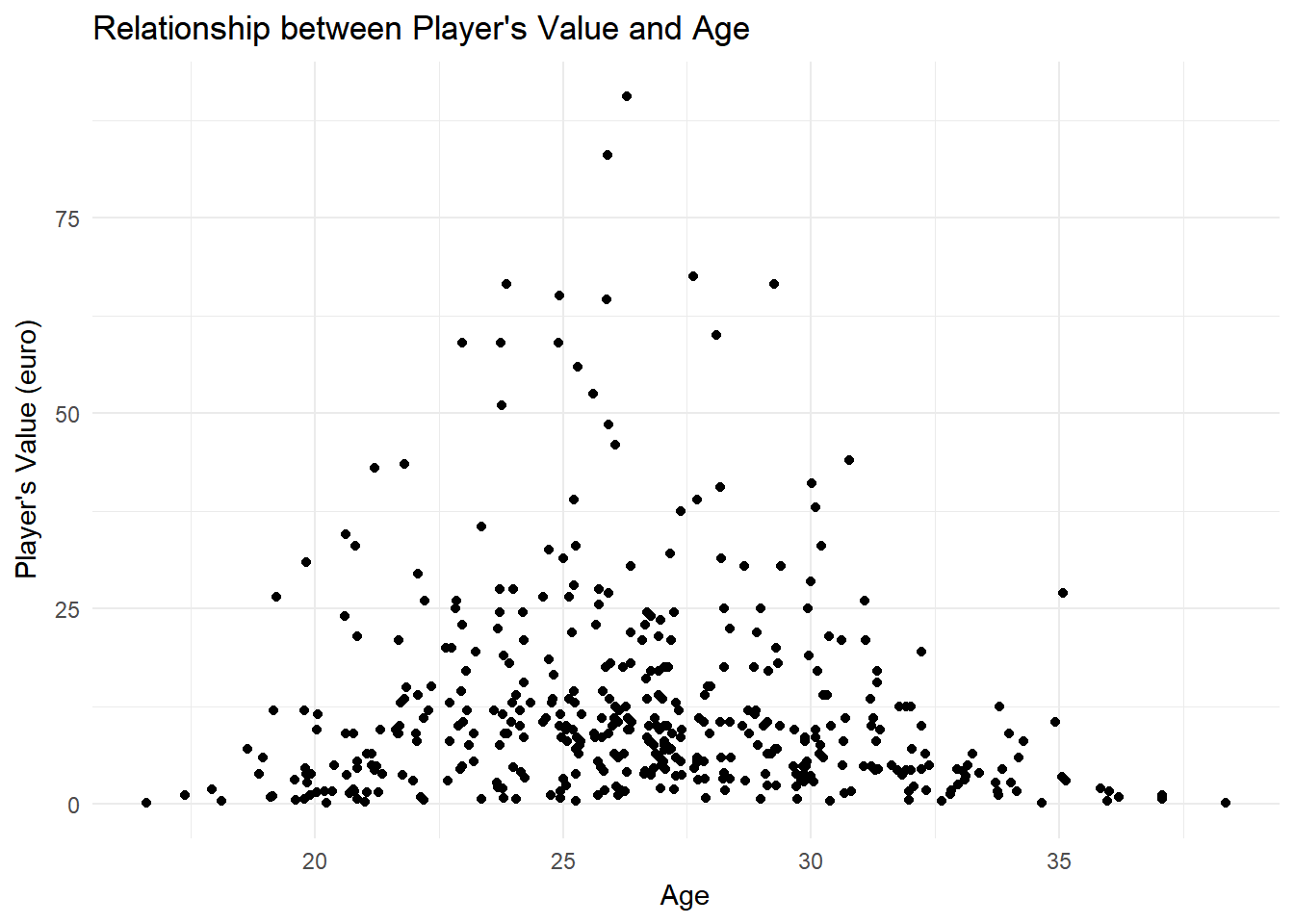
**III. Regression Analysis:**

Preliminaries

Our main method of setting up the regression model is the Stepwise Selection method. However, before reaching the point of conducting the method, we first must have a closer look into our interested independent variables that explain player’s value. First step, is to focus on our main interested predictor, the overall rating of players in the game FIFA 18. We believe that the overall variable would have great significance, due to the fact that it is a reliable indicator of how good a player is. Our initial assumption was that the player’s overall value has a positive correlation with the player’s market value, which is intuitive. To test our assumption, we made a scatterplot between the two variables:

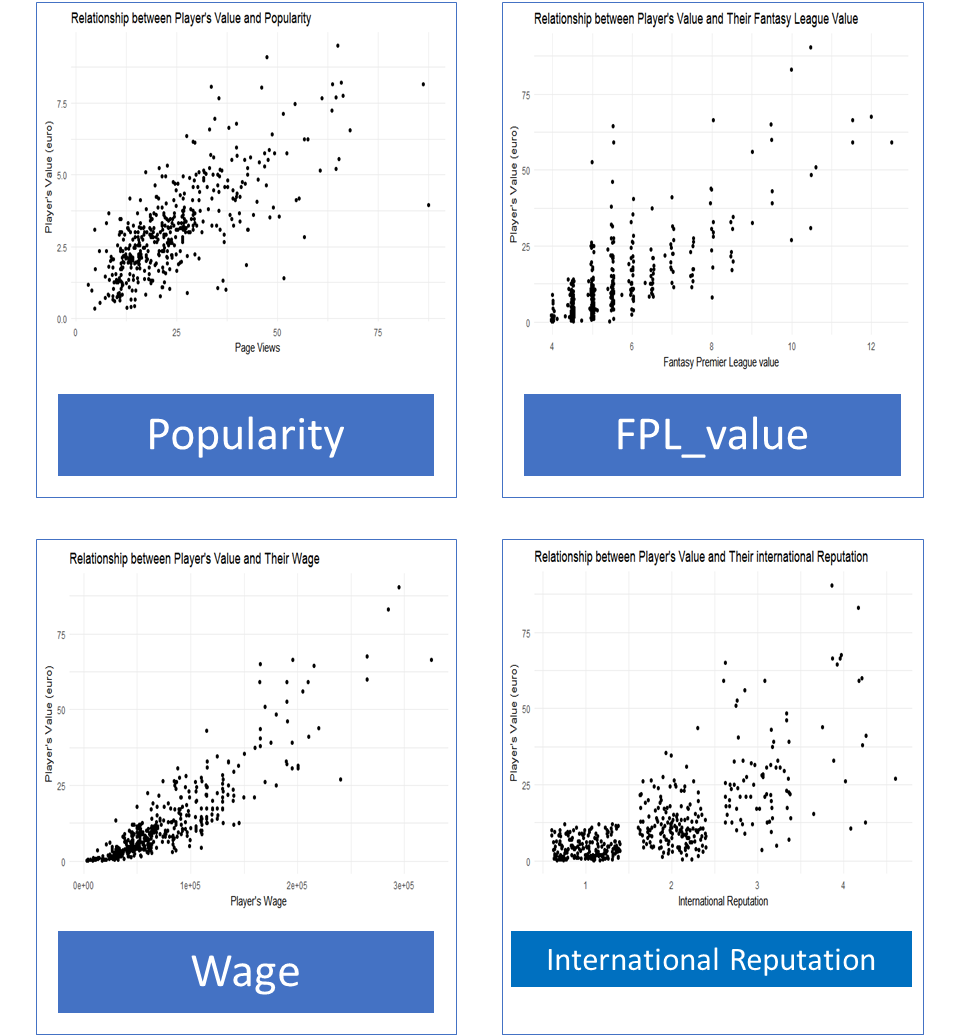
Looking at the plot, we notice that the relationship between player’s value and the overall stats seems to be exponential. Players who have overall stats from 60 to 75 seem to have similar market values, while players who have their overall above 75 have an exponential increase in their market values. To be able to fit the *overall* predictor in our final regression model, we apply a transformation by logging the overall predictor. 

The second interested predictor that we want to consider is *age*. We all know that physical strength and experience play an important part in determining a player’s value in every sport. Therefore, players who have the highest value would be ones who are astoundingly experienced but not too old. With that in consideration, we suspect that age has a quadratic relation with player’s value. To verify, we have the scatterplot between age and the player’s market value:



Just as what we suspected, the scatterplot has a bell-shape curve, with the highest values ranging from 23 to 28. Another way to control for confounding variables, besides adding the variable to the regression, is to directly adjust the variable that is being confounded. That being said, we will add in our regression model the adjusted version of *age* variable, *agesq*, which is the squared value of *age*. We then can fit age into our regression model without being concerned about the quadratic relationship.

With our two interested predictors being adjusted according to fit into our model, we want to consider some other predictors that we think would be significant in estimating player’s value. They are *page\_views* (number of views in wikipedia), *player’s wage*, *fpl\_value* (player’s fantasy premier league value), and *player’s international reputation*. We also expect positive correlation between each of these predictors with the player’s value. The scatterplots are below:



All four of our predictors, as expected, have a positive correlation with the player’s value. For the overall fitness of our model, we detect and exclude some the outliers for each of these predictors that we think will not explain the player’s value well. For example, Wayne Rooney’s market value is average, but due to his popularity, his *page\_views* is the highest of all.

The last modification we make before running the Stepwise selection is to transform player’s market value to sqrt, because *player’s value* is right-tail heavy, which could lead to severe heteroscedasticity. The modification of player’s value also leads to another change that is taking the sqrt of *page\_views*.

**Stepwise Selection:**

After having all the predictors in careful consideration, we now conduct the stepwise selection. In the process, we also add in other dummy variables like *new\_signing, region, preferred foot, big club*, etc. These variables indicate the player’s background details, as we believe some the variables would be significant in predicting the value of a player. After running the stepwise selection, the resulting final regression model is below:

## Call:

## lm(formula = sqrt(eur\_value) ~ sqrt(eur\_wage) + log(overall) +

## age + agesq + fpl\_value + international\_reputation + as.factor(position\_cat) +

## big\_club + sho, data = EPL\_Data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1.06079 -0.25135 -0.02319 0.19492 1.50974

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -4.846e+01 2.702e+00 -17.930 < 2e-16

## sqrt(eur\_wage) 4.966e-03 7.435e-04 6.679 7.60e-11

## log(overall) 1.126e+01 6.479e-01 17.374 < 2e-16

## age 1.196e-01 5.730e-02 2.088 0.03742

## agesq -4.481e-03 1.045e-03 -4.287 2.24e-05

## fpl\_value 2.790e-01 2.586e-02 10.788 < 2e-16

## international\_reputation 2.232e-01 4.318e-02 5.168 3.65e-07

## as.factor(position\_cat)Midfielders 1.827e-01 6.121e-02 2.985 0.00300

## as.factor(position\_cat)Defenders -6.122e-02 7.792e-02 -0.786 0.43244

## as.factor(position\_cat)GoalKeepers 4.378e-01 8.401e-02 5.212 2.93e-07

## big\_club -1.159e-01 6.099e-02 -1.900 0.05808

## sho -6.874e-03 2.472e-03 -2.780 0.00567

## Multiple R-squared: 0.9445, Adjusted R-squared: 0.943

## F-statistic: 654.1 on 11 and 423 DF, p-value: < 2.2e-16

Observing the summary table above, our final regression model for predicting player’s value is a very good model, with a really high adjusted R-squared value of 0.943, indicating an accurate predicting model. Moreover, with the p-value of the overall significance test being small (<2.2e-16), we have a statistically significant and fitted model. Of all the predictors, only Defenders dummy variable and *big\_club* are statistically insignificant, while all other variables are significant at all level.

**IV. Applications and Implications of the Model**

With the prediction model of the player’s value we built in the previous part, we conducted two separate applications:

**Identification of overvalued players**

For this application, our initial hypothesis was that overvalued players affect the team’s performance. To test that hypothesis, we want to focus particularly on the overvalued players in the “top-tier” teams, teams that are considered as the “big fish” in the league based on their overall strength of the squad and historical records of winning titles. These teams are Manchester United, Manchester City, Chelsea, Arsenal, and Liverpool, which can be filtered by the variable *big\_club* in our dataset.

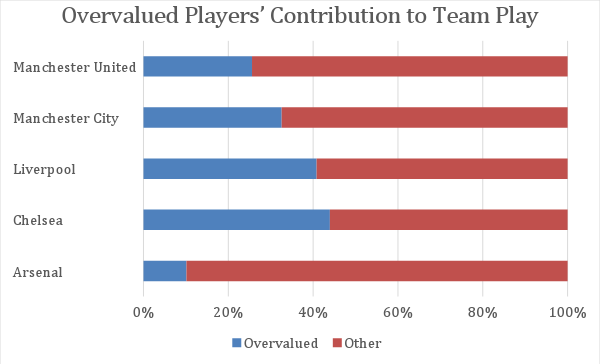
The method of defining overvalued players is to find the residual of each player’s value. The summary of the residuals is below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1st Quantile | Median | 3rd Quantile | Max |
| -14.16 | -1.98 | -0.36 | 2.27 | 17.57 |

With the summary table of the residuals, our interest is its 3rd quantile. We define in our project that any player that has value’s residual larger than the 3rd is considered as overvalued. With that criterion in consideration, we have compounded a summary of table of the number of overvalued players in the top 5 teams in Premier League:



We notice that Chelsea has the highest number of overvalued player in the team, while Arsenal has the lowest. An interesting finding is that Chelsea is also the champion in last season. This leads us to believe that the number of overvalued players has a positive correlation to the standing of the big team in the league. In other words, we hypothesize that overvalued players, even though impact negatively to the team’s financial status, can be vital in the team’s performance. To have a more detailed look into the contribution of overvalued players in each of the top teams, we have a summary stacked bar plot below:



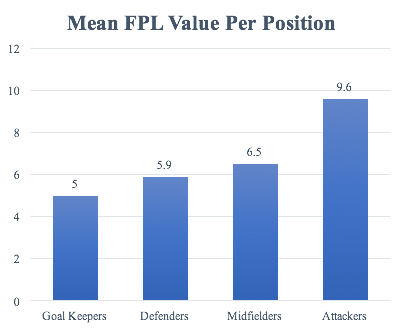
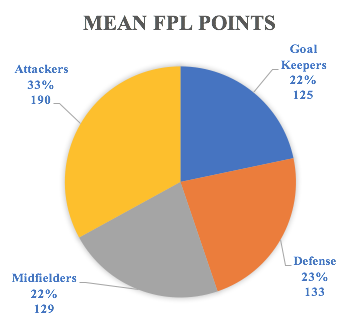
As we have found earlier, Chelsea has the highest number of overvalued player, and is also the champion of the last season. It can be explained the high percentage of contribution from overvalued player from this team, over 40%. Manchester United, even though having a sufficient number of overvalued players, 6, they only contributed fairly to the overall team play, 23%. The team’s overvalued players underperformed last season. That is one of the reason why the team ended up in a bad position, 5, lowest in the top teams. Overall, the main finding of this application is that we can see how important it is for top teams to spend an excessive amount of money on overvalued players, as they are the keys to winning the league. Through the application, we got to apply our model to identify the overvalued players, as well as understand the role of money in determining the success of a team.

**Identifying the optimal 15 players for a Fantasy Team**

After doing our analysis and identifying the key contributors to a players’ price tags we now have a clearer perspective on how to pick a Fantasy Premier League with a caution to their game value and the constraints of the game itself. Those limitations are as follows:

* Three players at most from the same team.
* Budget of the user is 100 at the start of the game.
* Only allowed 2 goalkeepers, 4 defenders, 5 midfielders and 3 attackers.

Our further analysis was finding the top 15 players in the game per position based on their overall points in the season. Then we wanted to find the mean value of points and mean price tag of each player per position of those top 15 players. This will allow us to use the regression model as a reference when understanding the price tag on those players and that we should pick players who have points greater than the average of the top players in the league for their respective positions.



Those two graphs represent our mean findings for both the average points and average price tag for players, so we do not get a player who is too far beyond the average value and not too below the average points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Team** | **Points** | **Value** | **Position** |
| Tom Heaton | Burnley | 149 | 5 | GK |
| Artur Boruc | Bournemouth | 120 | 4.5 | GK |
| Charlie Daniels | Bournemouth | 134 | 5 | LB |
| Gareth McAuley | West Brom | 131 | 5 | CB |
| Nathaniel Clyne | Liverpool | 129 | 5.5 | RB |
| Ashley Williams | Everton | 127 | 5.5 | CB |
| Dele Ali | Tottenham | 225 | 9.5 | CM |
| Geroginio Wijnaldum | Liverpool | 149 | 7.0 | CM |
| James Milner | Liverpool | 139 | 6.5 | CM |
| Etienne Capoue | Watford | 131 | 5.5 | DM |
| Cesc Fabregas | Chelsea | 121 | 7.0 | CM |
| Alexis Sanchez | Arsenal | 264 | 12 | LW |
| Eden Hazard | Chelsea | 224 | 10.5 | LW |
| Romelu Lukaku | Manchester United | 224 | 11.5 | CF |

As we can see most of our players have points that are above the average except for few who are not that much lower than the average keeping in mind that the average includes only the best players in the league and not the whole player list. In terms of budget management, we have few players that are a bit higher than the average value and only one over valued player which is Alexis Sanchez. The table represents the best players you can have on your team that would collectively get you the most points throughout the season, because it works out that the best 5 midfielders are the ones in that table who are inexpensive which allowed us to have more funds for other positions that are a bit more expensive than the average value per position in the top 15 player list. This could be a strategy used by many users who play Fantasy Premier League which will help them with their decision making when picking out their team for the season at hand.