|  |
| --- |
|  |
| Examining how various Intangible Factors Influence Fantasy Players Points |
| DATA-231 Final Project |
|  |
| **Fungai Jani**  **11/20/2022** |

**Introduction**

Fantasy sports are a vastly growing addition to the sporting world. With over 9 million players the fantasy premier league game has become popular among soccer fans all over the globe. The game is based on analysis and prediction so what makes a good fantasy player? Being ahead of the curve and knowing what will lead a player to scoring a lot of points is the way to go. Influence is a new statistic that was added and is described as the degree to which a player has made an impact on a single match or through the season. This makes it an importance statistic but how closely related is it to points scored by the players.

From my early adolescence I have been participating in the Fantasy Premier League with family and friends. Taking the casual approach of just picking my favorite players always leaves me shorthanded when the season is concluding. Using analytics to figure out what makes players score more points would give anyone the upper hand in their mini leagues. Watching every match is not practical to decide which players to pick in your teams so picking variables that seem closely related to total points scored is the best way about going being the best fantasy Premier League player.

This project aims to explore the factors responsible for the total points scored by players in the English Premier League. I plan to run multiple linear regression models with total points scored as response variable and influence, BPS, minutes, goals scored and assists as explanatory variables, examining their relationship with total points scored. My primary research question for this project is, " Do players who have a high influence and play more minutes score more points in fantasy premier league?" I also plan to address questions like, "What are the best predictors for total points

**Methods**

The dataset was found from Kaggle and is titled FPL20\_21. The data was uploaded by Plavak Das after the conclusion of the 2020/2021 English Premier League. This dataset consists of the statistics of the English Premier League players in the fantasy game fantasy premier league. The data was extracted directly from the fantasy premier league site in the 2021-2022 season. Consisting of 19 variables and 714 observations.

This dataset is representative of the entire population of the English Premier League as it contains the statistic of every player in the league. The data is collected on all players regarding what they do or do not do in the matches, so it is random. The important variables in this study such as influence is defined as the influence of player on team. Influence Points are scored as various actions on the pitch are given a score between 1-10 and influence is calculated. BPS is Bonus Point System score of the player, bonus points are the additional points a player receives for an exceptional performance. Minutes is the total number of minutes the player played during the 2020/2021 season.

Since the dataset is a representation of the population of Premier League it includes players who did not play at all. These players would affect the model as the values are not missing but they are zeros so they would lower means. The players with zero minutes played were filtered out and removed before any models were ran. This took the observations down from 714 to 524 so removing 190 players.

Data analysis was done by running a multiple linear regression model and integrating various transformations and interaction terms to fit the best model with the available variables. I chose multiple linear regression instead of logistic regression as my dependent variable**,** net migration, is a numerical variable. I created a correlation matrix and various scatterplots between total points scored and all the other variables to learn more about the variables that I should include in my model and examine their relationships. All the models were also checked for having met all the inference criteria like linearity, normality, independence, etc., before making an inference.

**Results**

**Table 1:** Description of the most relevant numerical variables in the model

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Mean | Median | SD |
| Total\_Points | 60.015 | 50.5 | 49.008 |
| Influence | 308.069 | 253.1 | 262.526 |
| ICT\_index | 74.105 | 59.75 | 70.030 |
| Goals\_Scored | 1.882 | 1 | 3.357 |
| Assists | 1.727 | 1 | 2.619 |
| Minutes | 2333.75 | 1355 | 1039.827 |
| BPS | 261.601 | 227 | 209.825 |

A summary of the initial analysis of the numerical variables Total\_Points, influence minutes, goals scored, BPS and yellow cards can be found in Table 1.  The final model contains the variables seen in Table 1 as they are statistically significant as well as position.

***Table 2*:** Correlation matrix of all the relevant variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Total\_Points | Minutes | Goals\_Scored | BPS | Yellow Cards | Influence |
| Total\_Points | 1.000 | 0.8878 | 0.6832 | 0.9364 | 0.3899 | 0.9395 |
| Minutes | 0.8878 | 1.000 | 0.3957 | 0.9480 | 0.5962 | 0.9272 |
| Goals\_Scored | 0.6832 | 0.3957 | 1.000 | 0.4534 | 0.0971 | 0.5558 |
| BPS | 0.9364 | 0.655 | 0.4534 | 1.000 | 0.4944 | 0.968 |
| Yellow\_Cards | 0.3899 | 0.5962 | 0.0971 | 0.4944 | 1.000 | 0.4636 |
| Influence | 0.9395 | 0.9217 | 0.555 | 0.9681 | 0.4636 | 1.000 |

*Table 2* shows us some notable correlations between different variables in our model. Total\_Points correlates the strongest with Influence amongst all the other variables, having a correlation coefficient of 0.9395. As shown in the table, the most correlated predictors are influence and BPS, with a coefficient of determination of 0.9681. This strong positive correlation makes sense because more influential players are known to score higher in the bonus point system and receive additional points for their performance. This results in more total points also highlighting the strong correlation between influence and total points as well as BPS and Total\_Points. Total\_Points is moderately related to Yellow\_Cards with a correlation coefficient of 0.3899. This is logical as yellow cards decrease the total points score as they are a negative score on the game. We can see other interesting correlations between Goals\_Scored and Yellow\_Cards; they are weakly correlated with a correlation coefficient of 0.0971. These correlations are essential for summarizing a dataset as well as identifying and visualizing patterns in the data.

***Table 3:*** Model Coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coefficients: | | | | |  |
|  | Estimate | Std. Error | t value | P value | 95% CI |
| (Intercept) | 0.83736 | 0.3635 | 2.292 | 0.0222 | (0.1201, 1.5546) |
| Influence | -0.02497 | 0.0056 | -4.461 | ~0 | (-0.3596, -0.0140) |
| Minutes | 0.0118 | 0.0010 | 11.343 | ~0 | (0.0097, 0.0138) |
| Goals Scored | 4.1937 | 0.1438 | 29.172 | ~0 | (3.911, 4.4759) |
| Assists | 2.0135 | 0.1654 | 12.174 | ~0 | (1.6888, 2.3383) |
| BPS | 0.1612 | 0.0073 | 22.107 | ~0 | (0.1469, 0.1755) |
| Yellow Cards | -1.8178 | 0.1840 | -10.134 | ~0 | (-2.1700, -1.4657) |

*Table 3* depicts the results of the multiple logistic model that was developed, which reveals the results of the overall efficiency of the significant variables. The final model contains the best variables shown above with some interaction terms. It contains interaction term between Influence and Position. This interaction was modeled because these variables are not independent of each other, and thus one can enhance the other. The Position of the players affects their influence. The final model shown in *Table 3* produced an R2 value of 0.9804 indicating that 98.04% of the variation in the model can be explained by these final selected variables. For every 1 point scored in Fantasy Premier League, the influence decreases by -0.055 holding all the other variables constant. holding all the other variables constant. For every 6.415 goals scored the total points earned by a player increases by 1, holding all other variables constant. Position is a categorical variable and interpretation is the difference between the response group and the variable group, holding all the other variables constant. The reference group for position is defense.

The 95% confidence interval for all the significant variables are shown in *Table 3*. For influence, we are 95% confident that a one unit increase in influence will decrease the total points scored between -0.0676 points and -0.04308 units, holding all other variables constant. The 95% confidence intervals for the goals scored and assists, which is consistent with our findings that they only provide strong evidence for the model based on the p-value that we saw earlier as well knowing that players that score and assist more earn more points. For position, we are 95% confident that the difference between defense and forward is a log odds ratio of 1.792 units and 2.231 units.

It is crucial to ensure that the final model satisfies all the assumptions of a multiple linear regression model. The first assumption is constant variance, and it is checked by the residual’s vs. fitted values plot of the final model, and it shows reasonably constant variance. Normality of the residuals is satisfied by the standardized residuals vs. theoretical values, residuals lie reasonably close to the theoretical line. Therefore, we can say that the normality condition is met. Another assumption of a multiple linear regression model is the independence of each outcome for total points. We cannot say that the variables are completely independent of each other. For instance, the more minutes a player plays the more goals they will score. Forwards and midfielders are more likely to score than defenders and goalkeepers. In this case, we cannot say that it is entirely independent but not to the level that could be concerning or problematic. Another assumption is the influential points, and we do not have any influential points in our data. The random assumption is met as the data is collected as players play in every game. Finally, linearity conditions were met in the model. The table in the appendix showing the relationship between Influence and Position shows why interaction terms were used.

**Discussion**

The final model with the variables Influence, minutes, goals scored, assists, BPS and yellow cards is ultimately the most significant predictor of total points. Initially, the primary aim of this research paper was to find out if influence positively impacted a player’s total points. However, as I was moving forward with my research, I realized that I was asking the wrong question and I could have improved my question. The primary findings of this research were that influence has a negative effect on total points in this model albeit a very weak negative effect according to the dataset that we are working with. An interesting finding of this paper was that BPS is, in fact, the best predictor of total points, and this is interesting to me because before influence would also positively affect the bonus point system of a player thus being a positive predictor for total points. From my final model, influence had a was a significant predictor along with the interaction term. I also wanted to explore other relationships like the relationship between club and total points as well as teams and the other predictor variables. I was unable to explore all these relationships in-depth because I got rid of some of the players that had played zero minutes, so some teams had very small sample sizes for players in comparison to others.

One implication from these results is that the position is closely related to all other variables. The challenge for Fantasy Premier League managers is to rank players based on positions and choose from there. Defenders may do well in comparison to other defenders but compared to midfielders they may not be doing as well. Future researchers should explore this in-depth with advanced statistical tools like the interaction plot.

Limitations of this study include that the data is just from one season of the Premier League. There are a lot of different variables that could predict total points better, like team and price of the player but price was not included in the dataset. Future researchers in this field could include this variable in their model for a better result. Another limitation can be that the sample size of the data is small with a limited number of variables. It could lead to a higher error of margin and question the reliability of the findings. The removal of players with zero minutes could be due to certain players being injured the entire and not playing so picking data from a range of seasons would be better in constructing a model.

**Bibliography**

Micahel Potts 2022, Fantasy Premier League Tips 2022/23: Midfielders To Sign, Radio Times Magazine, <https://www.radiotimes.com/tv/sport/football/premier-league-2022-23-midfielders/>

Woodhard Robert, *How to win your fantasy football league, by our data scientists* [*https://www.frontier-economics.com/uk/en/news-and-articles/articles/article-i7337-how-to-win-your-fantasy-football-league/#*](https://www.frontier-economics.com/uk/en/news-and-articles/articles/article-i7337-how-to-win-your-fantasy-football-league/)