Applied Statistics

Dr. Frazier

11/14/2018

Shivam K C

Individual Project: Predicting Total Points in Fantasy Premier League

Introduction

People feel good when their predictions turn out to be true. The feeling is even better when one wins a bet against his friends. This feeling is driving not only the purpose of my research, but also Fantasy Premier League (FPL)—the biggest fantasy football (soccer) game in the world with over five million players. FPL is an online game where people assemble a team of real-life soccer players and score points based on the actual performance of these players. The points are based on the playing time, the number of goals, assists, fouls, etc. in real life. For instance, a player gets four points (in FPL) for scoring a goal and two points for playing a match. The goal of FPL is to get more points than other participants. Since the points heavily depend on unpredictable factors like goals and assists, the points become harder to get. That is why the main goal of the research is to find a model that best predicts the total points a soccer player gets in FPL.

Either some FPL players have found the model and they are keeping quiet or there isn't much work done in this field. The reason why is that the articles that I researched are not directly related to FPL. An exception is *The wisdom of smaller, smarter crowds* by Goldstein et al., which recommends "…to make new teams based on popular players…" (488). Basically, I used other articles to help me guess and narrow down the variables that can predict the total points better. According to Matthews et al., the number of minutes a soccer player plays is crucial. Similarly, Lago-Peñas et al. found that in a soccer game "the variables that discriminate between winning, drawing and losing teams were the total shots, shots on goal, crosses, crosses against, ball possession and venue" (288). These research articles motivate my second research question: Does player's playing time, popularity, or influence in previous games best predict total points in FPL?

Materials and Methods

The data were collected of all the soccer players playing Barclays Premier League (BPL) for 2017-18 season by OptaSports, an international sports analytics company (*Statistics Explained*). It was made available in the official website of the Premier League. However, it could only be used through data scrapping. Vaastav Anand had cleaned the data and posted it posted on github. The target population for the analysis was the soccer players playing BPL for 2018-19 season. The sample was a good representative of the population because it consisted of most players playing the current season. However, three new teams had been promoted and all the teams had made some changes to their squads for the 2018-19 season.

The original dataset included 647 samples and 18 variables. I wanted my model to include predictors which were easier to predict. By that I mean predictors which are known to us before a game week. For instance, one can know beforehand whether a player will play the game or not. However, predictors like goals, assists, and cards are highly unpredictable in themselves. Thus, they were not used in the analysis. According to the dataset, only 79.4% of 647 players played at least a minute and the remaining 20.6 % did not play a single minute the whole season. Thus, there were many samples with zero Total Points. However, I decided to include all the samples (647) in the analysis because if I had selected a player for FPL and he ended up with zero points, I would have liked to predict that. Other notable information of the dataset is summarized in Table 1 (page 3).

Since the response variable (Total Points) and the predictors are numerical, running multiple linear regression models were considered. Correlations were calculated, and scatterplots were created to explore the association between each of the response variables with Total Points. Additionally, potential transformations of Total Points and Selected by Percent were carried out. Predictors were narrowed down to using summary regression. Simplicity and adjusted-R squared were used to narrow down to the best model. The t-tests of the coefficients were examined to determine the significance of each variable in the model. Also, F-statistic was used to determine the usefulness of a model. Confidence intervals and slopes were reported where appropriate for Total Points and the best predictors. Conditions of inference were also used to distinguish between models.

Harry Kane, a Tottenham Hotspur player, acted as an influential point when fitting the best model. When multiple linear regressions with and without the point were carried out, no significant change was observed (see in the Appendix for the comparison of the models with and without influential point). There was no improvement in the conditions of inference. So, I decided to include the influential point.

Table 1. The descriptive summaries of key variables. All variables are numerical except Dream Team which is binary.

	Definition	Units
Points	points a soccer player got in the next gw	points
BPS	utilises a range of statistics supplied by Opta that capture actions on the pitch, to create a performance score for every player	NA
Bonus	players with the top three BPS in a given match receive bonus points - three points to the highest-scoring player, two to the second best and one to the third	points
Threat	a value that examines a player's threat on goal	NA
Influence	evaluates the degree to which that player has made an impact on a single match or throughout the season	NA
Selected by Percent	percent of people that selected the player	%
Minutes	minutes a soccer player has played in the season	minutes
ICT Index	a single score for a soccer player for three key areas – Influence, Creativity and Threat	NA
Creativity	assesses player performance in terms of producing goalscoring opportunities for others	NA
Completed Passes	the number of passes a player completed in last GW	NA
Clean Sheets	clean sheet is an event when a player's team does not concede a goal. This variable is a number of such events	NA
Transfers Balance	it is the difference of number of FPL players that transferred in the player and number of FPL players that transferred out the player in a GW	NA

Dream Team	whether a player made it to the dream team in the next GW or not	Yes (1) or No (0)

Results

Table 2 shows that the standard deviations for most of the variables are greater than their means; so, these variables are not normally distributed.

	Mean	Standard Deviation	Median	Interquartile Range
Total Points	48.6	49.8	36	80.5
Bonus Point System (BPS)	222.3	223.0	171	374
Threat	184.6	292.5	63	239
ICT Index	62.5	72.4	39.3	98.4
Minutes	1158.4	1089.5	987	2003.5
Selected by Percent	2.31	5.07	0.400	2.1
Influence	267.7	280.8	204.8	449
Bonus	3.74	5.54	1.00	6

Table 2. Spreads of key variables.

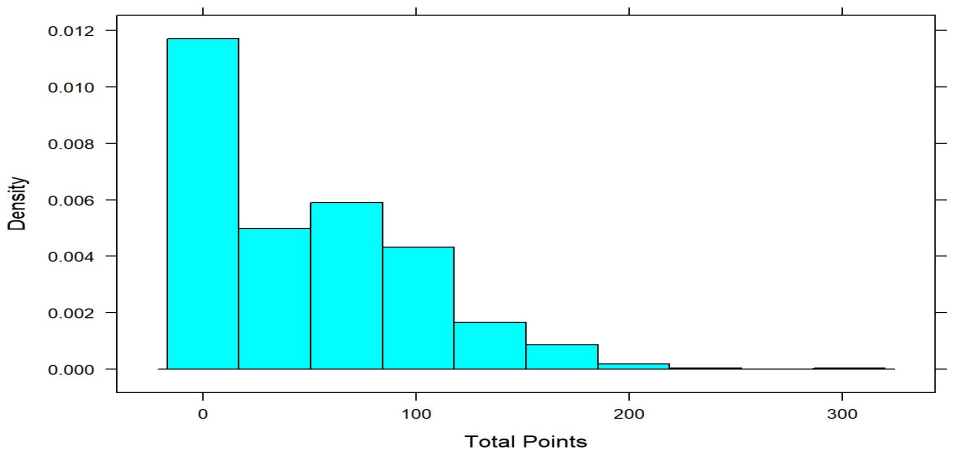

Figure 1. Right-skewed histogram of Total Points.

The histogram of the Total Points distribution is generated (Figure 1). The distribution is right-skewed. A few outliers are also observed at greater Total Points. Very high number of players with zero Total Points is observed because 20.6% of the sample did not play a single minute of soccer.

The results of this investigation are a model that uses the factors of BPS and Threat to indicate what a player's Total Points is expected to be in a given game week. The model is a linear regression on the Total Points variable of the following form:

$$Total\ Points = \beta 0 + BPS \cdot \beta 1 + Threat \cdot \beta 2$$

Table 3 shows the estimated parameter, 95% confidence interval, standard error, and the corresponding p-value for each factor. The coefficients for the two variables are found to be significant at the $p < 0.001$ level of significance. Therefore, BPS and Threat are significantly important predictors for Total Points in the model.

Terms	Estimated Parameter	Standard Error	p-value	95% Confidence Interval
Intercept, β0	0.369	0.515	0.474	(-0.642, 1.38)
BPS	0.172	0.00196	$<2x10^{-16}$	(0.168, 0.176)
Threat	0.0538	0.00150	$<2x10^{-16}$	(0.0509, 0.0567)

Table 3. The summary of the statistical results for the constructed model.

From Table 3, the model has the following form:

Total Points = 0.369 +0.172 BPS + 0.0538 Threat

From the model there are several expectations indicated by the coefficients associated with each variable:

- If BPS increases by one, Total Points will increase by 0.172 points holding other variables constant.
- If Threat increases by one, Total Points will increase by 0.0538 points holding other variables constant.

The model meets only some conditions of inference. Its linearity is good. However, constant variance and normality are problematic. The data fulfills representativeness. However, each data is not independent of another in that if a player scores pass another player, the scorer gets positive points, but the defender gets negative points. To see if constant variance and normality of the model would improve, natural log of Total Points was plotted against BPS and Threat. However, linearity, constant variance, normality and adjusted R-squared got worse and thus was rejected. The final model itself was found to have an adjusted R-squared of 0.966, meaning that it was found to 'explain' approximately 96.6% of the variation of Total Points within the data set. Likewise, Selected by Percent showed log relationship with Total Points; so, the former was logged and included in the summary regression. However, the variable did not appear in the simple models of summary regression and thus was rejected. From Table 3, we can see that both of our predictors are statistically significant as their p-values are approximately equal to zero, however the intercept seems to be statistically insignificant as its p-value of 0.474 is greater than zero. Similarly, we can see that our model is useful because the F-statistic of 9030 is significantly greater than 1. Additionally, we can make the following statements from our model:

- We are 95% confident that if BPS increases by one, the average Total Points will increase between 0.168 and 0.176 points holding other variables constant.
- We are 95% confident that if Threat increases by one, the average Total Points will increase between 0.0509 and 0.0567 points holding other variables constant.

Discussion

According to the results, BPS along with Threat best predicts Total Points. This answers both my research questions. The results contradict with Goldstein et al. (488) and Matthews et al because they respectively suggested Selected by Percent or Minutes to better predict Total Points; however, that is not the case. The results do somewhat agree with Lago-Peñas et al. in that attacking variables determine the course of a game (288): Threat is an attacking variable in the model that best predicts Total Points. So, the results recommend assembling more attacking players in your FPL team than defensive players to score more points. Similarly, the presence of BPS in the model tells us to assemble players that regularly contribute to their team.

The strengths of my analysis are:

- Simplicity of the model: My model of Total Points against BPS and Threat is simple in that it has only two predictors and the model is easy to interpret.
- Linearity: The linearity of my model is really good with data points lying closer to the regression line (see in the Appendix for the linearity of the model).
- Representativeness: The target population for the analysis was the soccer players playing BPL for 2018-19 season. Even though three new teams had been promoted and all the teams had made some changes to their squads for the 2018-19 season, the sample was still a good representative of the population in that it consisted of most players playing the current season.
- High adjusted R-squared: The final model was found to have an adjusted R-squared of 0.966, meaning that it was found to 'explain' approximately 96.6% of the variation of Total Points within the data set.

The weaknesses of my analysis are:

- Normality: The normality of the final model is problematic as most data points do not lie on the line in the normal Q-Q plot. (see in the Appendix for the normality of the model).
- Constant variance: The constant variance is also problematic as the data points flare out when going towards right of the residual plot (see in the Appendix for the constant variance of the model).
- Independence: The independence of my dataset is problematic. If a player scores pass another player, the scorer gets positive points in FPL, but the defender gets negative points. So, each data is not independent of another.
- Results unreliable: Since most conditions of inference are not met, the p-values reported might not be accurate. Thus, the results are not reliable.

As you can see, my model is not reliable due to the weaknesses of my analysis. Therefore, to improve my analysis and build on my research, I suggest the following:

- Use other dataset of the past or present season and calculate MSPEs of different fitted models (for instance, models in the summary regression). Then judge models based on the lowest MSPE which considers the simplicity and adjusted-r squared of the

models. Then check the conditions of inference so that the model is reliable unlike mine.

- Consider difficulty of games in the future analysis. An attacking player playing against a weaker team might perform better than an attacking player playing against a strong team.
- Consider Minutes in the future analysis. Whenever I play FPL, I feel players who get more game time (Minutes) do really well. Minutes impact all the three variables present in the model: more minutes, more Total Points and BPS, and more chance for a player to be threatening in attack. Therefore, maybe in the future research, Minutes can be somehow incorporated in the model.

References:

Anand, Vaastav. “Fantasy-Premier-League/data/2017-18/cleaned_players.csv”, github.com/vaastav/Fantasy-Premier-League/blob/master/data/2017-18/cleaned_players.csv. Accessed 3 Nov. 2018.

Lago-Peñas, Carlos et al. “Game-Related Statistics That Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League.” *Journal of Sports Science & Medicine* 9.2, 2010,: 288–293.

Matthews, Tim, et al. "Competing with humans at fantasy football: team formation in large partially-observable domains." *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012, dl.acm.org/citation.cfm?id=2900925. Accessed 20 Nov. 2018.

Statistics Explained, Barclays Premier League, www.premierleague.com/stats/clarification. Accessed 20 Nov. 2018.