Fantasy Premier League

Stephen Perrine

2022-11-06

Introduction

The English Premier League (EPL) is the most popular soccer league in the world. During the most recently completed season (2021-22), the EPL generated approximately 6.5M euros in revenue. Spain's La Liga, the EPL's nearest competitor (among soccer leagues), generated approximately 3.7M in revenue over the same season.

Like many professional sports leagues across the world, the EPL runs an official 'Fantasy' league for its increasing fan base. In this fantasy league, fans build their own teams (composed of existing EPL players) and compete against one another in private groups. The chief challenge in managing such leagues is the **allocation of players**: how do you determine who gets whom?

Most fantasy leagues rely on a **draft**, whereby each team in a given fantasy group gets a fixed number of 'player picks'. Picks are conducted in rounds, such that each team team gets one pick per round. At the start of the first round, all players in the EPL are available for selection. However, after a player has been picked by a team in the given fantasy group, that player is no longer available to other teams.

The draft as an optimization problem

A fantasy team's success is, to a large extent, forged in these opening drafts: given the players available, a fantasy manager must make the best possible pick. In this sense, the best possible pick is the individual in the existing pool of players who will generate the most fantasy points in the upcoming season. Individuals generally rely on their intuition for this challenge– but to what extent can we use past performance to predict an individual player's future performance?

Intended Impact

This report will use data from the 2020-'21 season to predict player performance over the course of the 2021-'22 season. I will use the knowledge learnt from this exercise to optimize my player picks over the next fantasy premier league draft!

Methods/Analyisis

Getting the data

Data on individual player performance over the 2020-'21 and 2021-'22 EPL seasons were acquired through the github account of vaastaav, who was kind enough to provide cleaned and tidied data scraped from the Fantasy Premier League website. Fortunately, in this instance the partitioning of training data and test data is very simple.

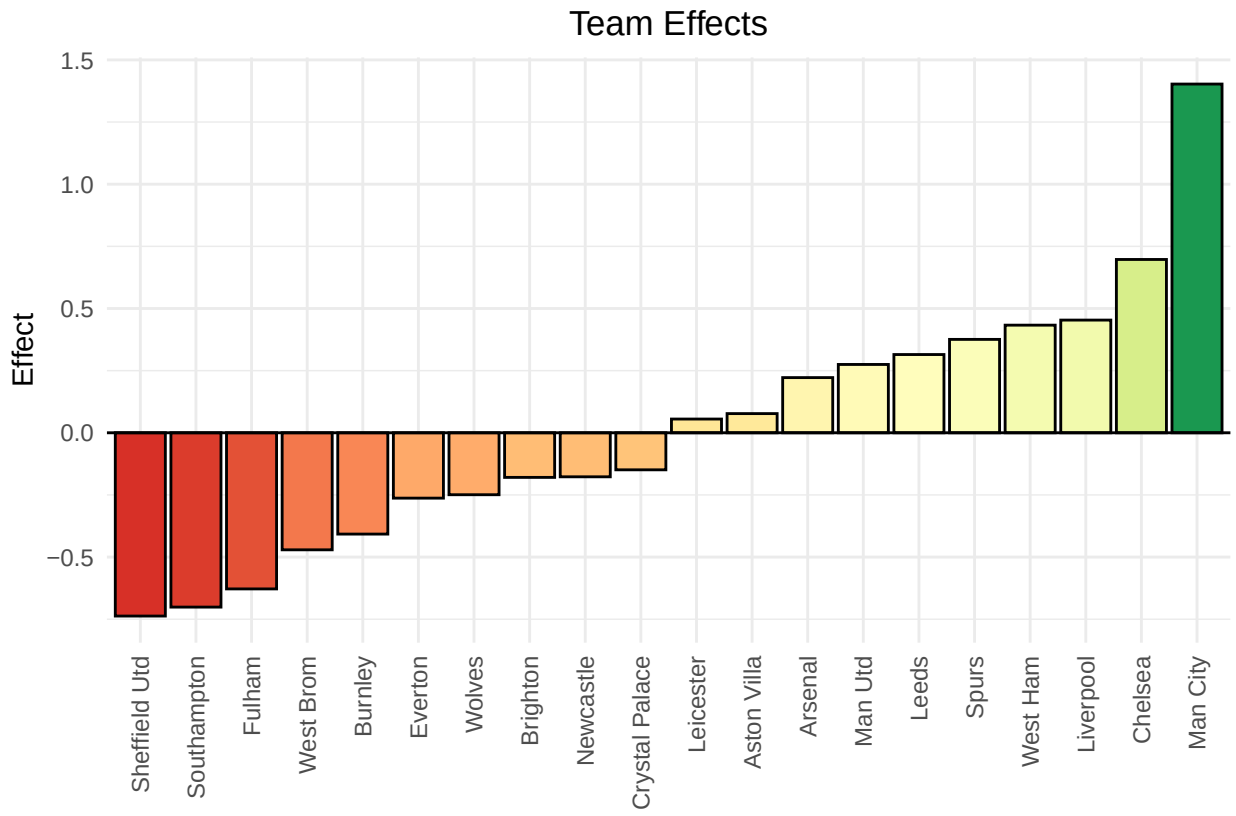
Recall our central research question: **To what extend can we use past performance to predict future player performance?** The training data on which we will build our model are therefore the data on player performance over the 2020-'21 season (season_tag == 1), and our test data are the data on player performance over the 2021-'22 season (season_tag == 2).

The meansure of player performance that we've chosen to predict is **points per game** (ppg), which is defined as [total fantasy points] / [matches played].

Insight 1: The 'Team' Effect

Followers of the EPL (or any sports league) will know that not all teams are created equal. Indeed, the EPL is more unequal than most professional sports leagues because there is no salary cap: teams are essentially free to spend whatever resources they have available. This creates a dynamic in which there are a few 'wealthy' teams who far outspend most competitors.

Knowing this, we should be able to find a 'team' effect in the data. To do so, we will subtract the player average (mu) from the team averages to determine how many additional points we should add to our player predictions given the team with whom a given player is associated. The following chart demonstrates the significance of the team effect:

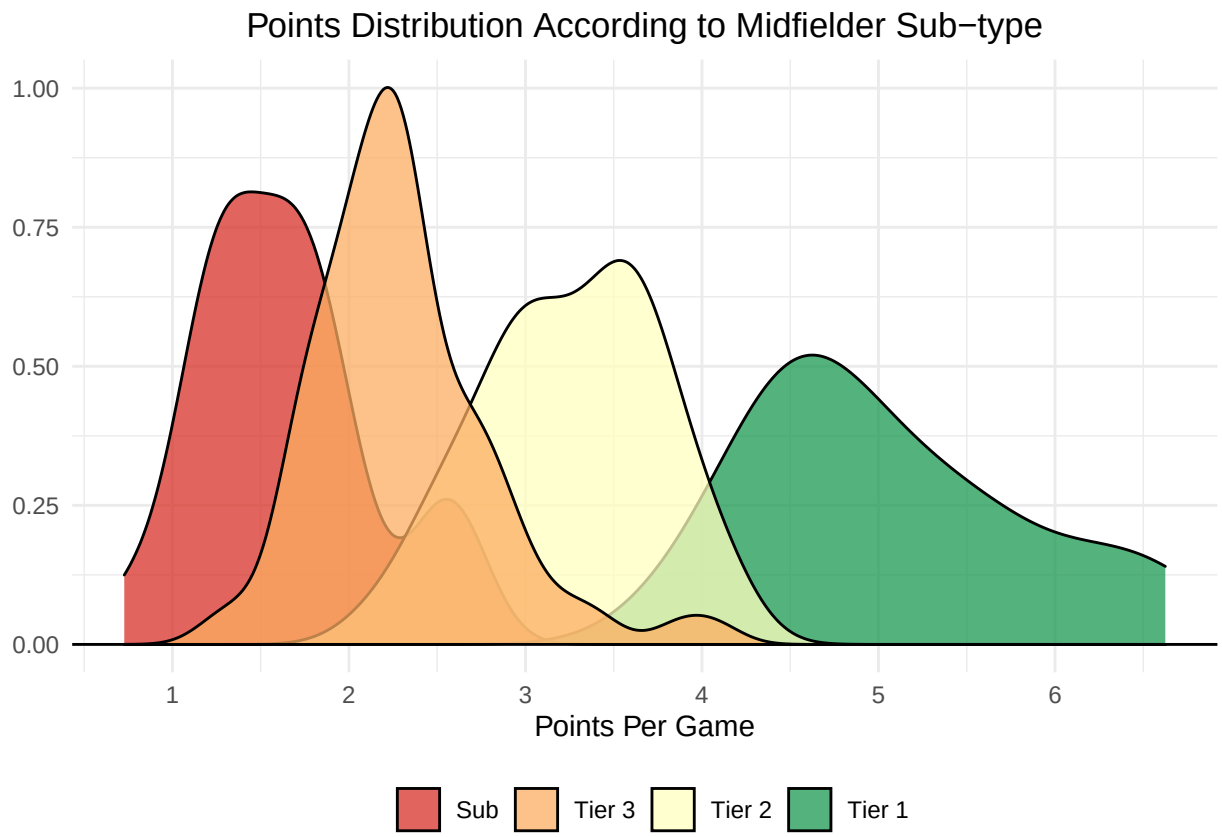


Insight 2: The 'Position' Effect

The Fantasy Premier League officially classifies players according to four positions: Forwards, Midfielders, Defenders, and Goalkeepers. However, followers of the game will know that there are sub-types within each of these broad classifications. For example, there are attacking midfielders whose role is to facilitate the offense, while there are defensive midfielders whose role is to break up the opponent's offense. The player stats for each of these sub-types will look drastically different.

We can therefore refine our position classifications by looking at the stats on minutes played, goals scored, and assists provided. We will use **k-means clustering** to determine the sub-type associations under each of our original four classifications, and then quantify the 'position effect' for each of these new sub-types.

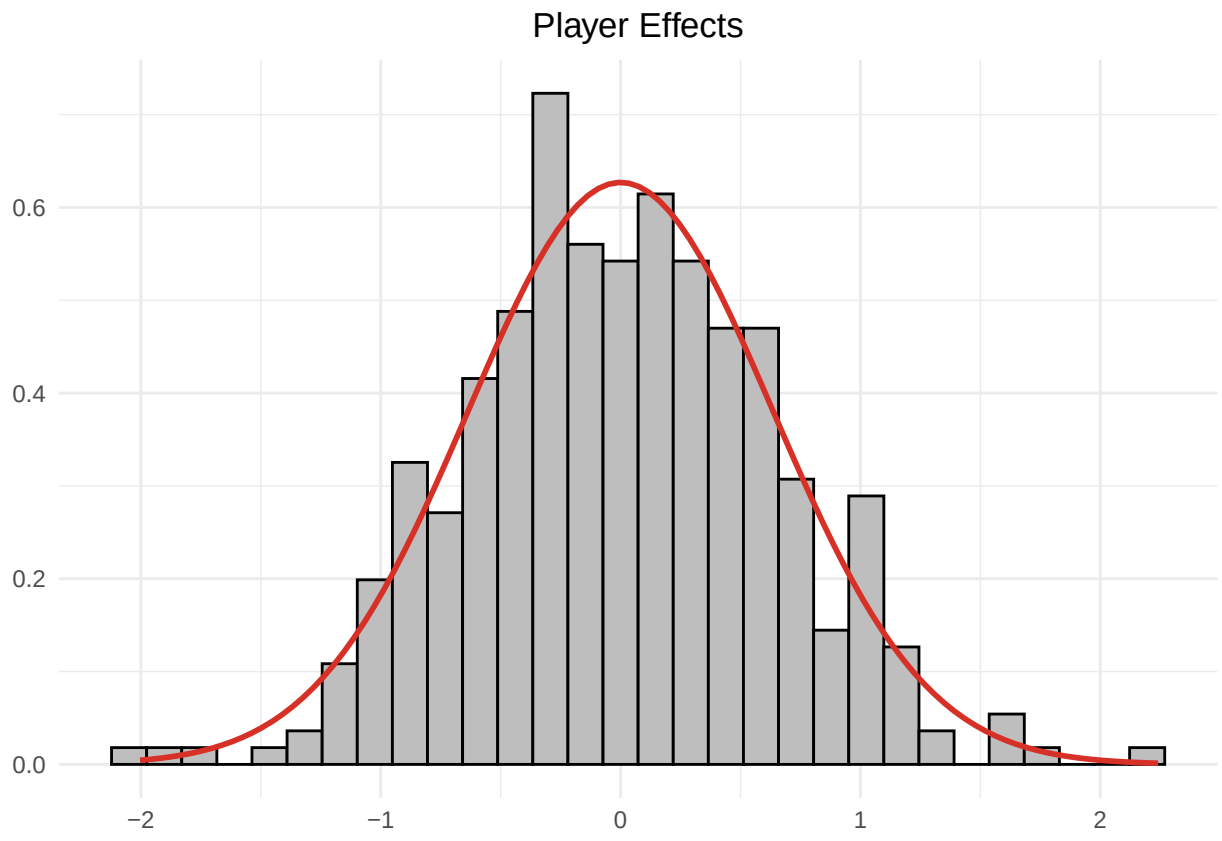
The following chart illustrates the significance of the sub-type effect as it relates to midfielders:



Insight 3: The 'Player Effect'

The residuals in player performance that remain after removing the team and distinct position effects can be attributed to an individual 'Player Effect'. This is the measure of the individual's contribution beyond their particular circumstances (the team to which they belong and the distinct position they play).

Interestingly, these individual player effects appear to be normally distributed, with a mean of 0 and standard deviation of .64.



Results

Now that we have modeled three individual effects, we can make predictions on the test data to determine how far we've improved on our baseline RMSE of 1.285. Recall that, in the baseline model, we simply predicted the mean points per game total of 2.6 for all players.

Our new models involves the following:

prediction = mu + team effect + position effect + individual player effect

With the new model, we've produced a RMSE of 1.184. That is an 8% improvement on the baseline.

```
## [1] "RMSE with new three effect model = 1.18395243225139"
```

```
## [1] "RMSE when just guessing the average = 1.28491162695508"
```

Model Performance

Although our new model has improved on the baseline, the improvement itself is a bit underwhelming. Let's take a look at the most egregious errors to determine why.

Overprediction

The instances below illustrate two situations where our model **overpredicts**:

name	category	group	team	ppg	pred	error
Jesse Lingard	Midfielder	Tier 1	Man Utd	1.8125	6.466908	4.654408
Jack Grealish	Midfielder	Tier 1	Man City	3.0800	6.518050	3.438050
Conrad Egan-Riley	NA	NA	Man City	1.0000	4.005242	3.005242
James McAtee	NA	NA	Man City	1.0000	4.005242	3.005242
Kayky da Silva Chagas	NA	NA	Man City	1.0000	4.005242	3.005242
Liam Delap	NA	NA	Man City	1.0000	4.005242	3.005242
Cody Drameh	NA	NA	Leeds	0.0000	2.917097	2.917097
Pablo Mar	Defender	Tier 2	Arsenal	0.5000	3.400000	2.900000
Sead Kolasinac	NA	NA	Arsenal	0.0000	2.824333	2.824333
Bali Mumba	NA	NA	Norwich	0.0000	2.602417	2.602417
Christoph Zimmermann	NA	NA	Norwich	0.0000	2.602417	2.602417

The first situation occurs when a player moves from a lower performing team to a higher performing team. The errors for Jesse Lingard and Jack Grealish capture this situation. Both are players who had significant player effects based on their performance at West Ham and Aston Villa, respectively. However, having moved to a bigger club, more points are captured in their team effect, meaning the player effect would need to be downgraded accordingly.

The second situation where our model fails occurs when players only end up playing a handful of matches. This is either due to injury or lack of selection.

Underprediction

The instances below illustrate one important situation where our model **underpredicts**:

name	category	group	team	ppg	pred	error
Bali Mumba	NA	NA	Norwich	0.0000	2.602417	2.602417
Christoph Zimmermann	NA	NA	Norwich	0.0000	2.602417	2.602417
Sead Kolasinac	NA	NA	Arsenal	0.0000	2.824333	2.824333
Pablo Mar	Defender	Tier 2	Arsenal	0.5000	3.400000	2.900000
Cody Drameh	NA	NA	Leeds	0.0000	2.917097	2.917097
Conrad Egan-Riley	NA	NA	Man City	1.0000	4.005242	3.005242
James McAtee	NA	NA	Man City	1.0000	4.005242	3.005242
Kayky da Silva Chagas	NA	NA	Man City	1.0000	4.005242	3.005242
Liam Delap	NA	NA	Man City	1.0000	4.005242	3.005242
Jack Grealish	Midfielder	Tier 1	Man City	3.0800	6.518050	3.438050
Jesse Lingard	Midfielder	Tier 1	Man Utd	1.8125	6.466908	4.654408

The instances captured above generally lack a position effect. This occurs when 'new' players join the league, because they have no past data on which to classify them. This could also happen to players who missed lengthy periods of the past season due to injury (like Virgil Van Dyke)– these players were discounted from the classification algorithm because they lacked enough data.

Model Improvement

Thus, our predictions could be improved through the addition of human judgement:

- When players move from lower to higher performing teams, a human could manually downgrade the player effects accordingly.
- When new players join the league, a human could manually classify them according to position and estimate a player effect based on their performance in their previous league.
- When players return from lengthy injuries, a human could manually classify their position and estimate a player effect based on their last substantial season in the EPL.

Conclusion

This report has demonstrated that past player performance is indeed useful in predicting future performance. By building a model based entirely on one season's worth of data, we were able to achieve RMSE of 1.184 ppg on our predictions regarding player performance in the following season.

However, as illustrated above, there are specific situations where our model produces poor predictions. In these instances, our goal of optimizing player picks would be best achieved by combining our model predictions with human judgement.

Given that most fantasy managers rely exclusively on human judgement, the model should give us a competitive edge!