Barclays Premier League Fantasy Football Data Analysis

Introduction

I've been really into BPL Fantasy Football this season and have been building a fantasy team since the start of the season last year.

I'm not very good at picking the right squad based on my instincts, so my basic goal is to use the data from the present and past weeks to determine the best picks for the upcoming week.

How Fantasy Football Works

I play Fantasy Football off the official BPL Fantasy site. For the uninitiated, Fantasy Football is an online game where users pick a squad of 15 real-life football players from a league. In this case, I'm playing the BPL, which is the English top division. Points are scored and collated depending on a player's actions in the actual game. The aim of the game is to amass the highest number of points each game week. The rules are here.

In general, players will register points for scoring goals, notching assists, clean sheets, penalties. They will lose points for getting yellow cards, red cards, missing penalties and conceding goals.

Friends tend to form mini-competitions to see who's better at picking the best squad week after week. At the end of the season, the winner of their league with the higest points gets the honour of having the best fantasy football instincts (although there's usually money involved in mini-leagues amongst friends). There's a budget allocated to each user so you can't always choose the best player because he might be too expensive. Players get transferred in and out of teams weekly, so there's an entire transfer market as well.

My Objective

The aim of my analysis is to predict a team of 15 players who will score the highest fantasy points within a budget of £100m in the upcoming gameweek.

Fantasy Football Data

Data is available from an API from the official site on players' basic information, team information, fixture information and performance stats. Each player's performance is aggregated every gameweek to give an overall view of the players' performance. There are nested data sets on each player's weekly performance history, as well as performance in past seasons.

The dataset contains a database of 550 players that are distributed amongst 20 teams in the BPL. There are a total of 63 data attributes for any single player. Each player has a unique player id, and data is updated once every *gameweek*. That is to say, a full round of fixtures played amongst 20 teams. Fixtures sometimes get postponed from one gameweek to another, so I'm only focusing on the first 26 gameweeks, with the full 10 games being played.

The dataset also contains the "fixture history" of a player, ie. the performance of each and every player in each and every game he has played so far. This is a t by 20 dataframe, where t is the gameweek number. I will be tapping on the fixture history of each individual player to construct features about their form. The description of the full attributes and feature set can be found in the Annex.

Preliminary Model Specification & Feature Selection

I conducted some preliminary modelling with a sample data set from Gameweek 26. The main purpose of my analysis was to uncover a base model specification and select a set of features for my model.

I came to the following main conclusions in my preliminary analysis.

1. **Form matters more than cumulative performance**

   - Based on regression that I ran on Gameweek 26 data, I inferred that features on cumulative performance are not as significant as "form" variables. These measure the “streakiness” of a players' performance.
   - However, defining a form variable is tricky and highly arbitrary. I'm going to have to restrict myself to certain specification of "form" for a start.

2. **Individual performance matters more than team performance**

   - I found that team performance variables, like team form, or whether a team is playing home or away is not as significant a feature as individual performance variable.

3. **The distribution of fantasy points is highly skewed**

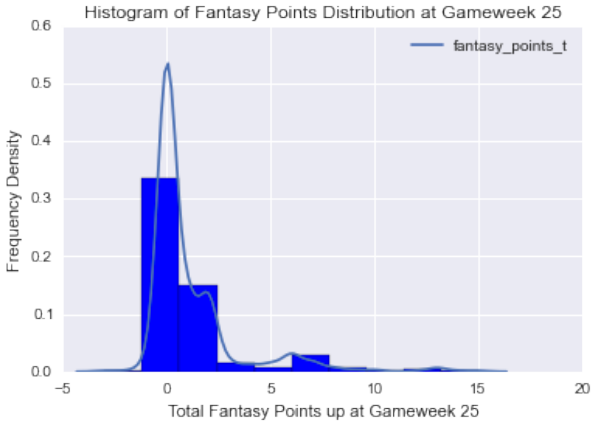
   - Figure 1 shows a distribution of the number of fantasy points per scored by each player in a typical gameweek. Approximately half of the players score below 2 points and about 30% of them score zero points.

   - There are 550 players in the database but only a maximum of 11 per team, so 220 players can be starting in a round of games. Each team gets a maximum of 3 substitutes per game so there can only be a maximum of 260 point scorers out of 550 players in the database.

   - Players who start also typically score a minimum of 2 points, which is scored having played about 60 minutes of the game. Substitutions before 60 minutes are quite rare and usually due to injury.

   - I would therefore have to try to stay away from models that rely on normality assumptions.

Figure 1



I will specify my model as a regression problem. I will try to predict the number of fantasy points each player will score in the *t*-th gameweek, based on the features collected in the *t*-1-th gameweek. I will only consider features of lag 1 for a start. This means that I am assuming that a player's performance in gameweek *t-1* is relevant to his performance in gameweek *t*.

A simple linear model is thus specified as such:

$$Y_{t,i} = \beta^T . \boldsymbol{X}_{t-1,i} + \varepsilon_i$$

where $Y_{t,i}$ is the number of fantasy points scored by the *i*-th player in the *t*-th gameweek and $\boldsymbol{X}_{t-1,i}$ is the feature set associated with the *i*-th player recorded in the *t-1*-th gameweek. I will try to specify a model that will return the best predictions.

Bearing in mind that the overall objective is to choose 15 players that will make up my fantasy team, I will approach the problem as follows:

1. Subset the dataset into different types of players, ie. Goalkeepers, Defenders, Midfielders and Forwards.
2. Select a set of features based on each player type. This is because GKs score points for saves and clean sheets, while FWs do not, etc.
3. Model the data and narrow down to the best candidate model based on several model selection criteria.
4. Adjust model starting parameters like lag length until best candidate model is reached.
5. Predict the 5 best scoring Goalkeepers, as well as 10 Defenders, Midfielders and Forwards based on the model specified.
6. Pick 15 players (2 GKs, 5 DFs, 5 MFs and 3 FWs) from 35 subject to the budget cap of £100m.

Model Comparison and Selection [Work in Progress]

The current models under consideration are:

- Simple Linear Regression
- Multinomial Logistic Regression
- Random Forest Regressors

Given the skewed nature of the response variable, I will probably be more inclined towards the latter two models.

Optimisation Step

Next, we have to formulate the linear optimisation problem to obtain a set of 15 players that can be considered for next week's fantasy team. Our objective function is therefore the predicted scores for 15 players. I also have a set of constraints set by the game. I can only pick 2 GKs, 5 DFs, 5 MFs, and 3 FWs. I can also only pick 3 players from any single team. Lastly, I cannot exceed a budget of £100m to buy these players.The problem is as such:

$$\max_{x} \sum_{i,j,k} \widehat{p_{ijk}}.x_{ijk}\,, s.t.$$

$$\sum_{j} x_{1jk} = 2, \sum_{j} x_{2jk} = 5, \sum_{j} x_{3jk} = 5, \sum_{j} x_{4jk} = 3\,;$$

$$\sum_{j} x_{ijk} \leq 3 \;\text{ for } j = 1, \ldots, \text{N\_teams}$$

$$\sum_{i,j,k} C_{ijk}.x_{ijk} \leq 100;\; x_{ijk} = 0 \; or \; 1 \text{ for all } i, j, k.$$

where x_{ijk} denotes a binary variable of whether *k*-th player from playing in the *i*-th type of position (1 = GK, 2 = DF, 3 = MF, 4 = FW) and the *j*-th team from the teams that are feature in the 35 player set (N_teams). C_{ijk} denotes the cost of the *ijk*-th player, and $\widehat{p_{ijk}}$ is the predicted fantasy score of the *ijk*-th player.

Discussion of Results [Work in Progress]