Predict To Succeed: Optimal Sequential Fantasy Football Squad Formation Using Machine Learning Tools

Arseniy Stolyarov, Gleb Vasiliev
New Economic School
National Research University Higher School of Economics
Toros Research

18 November 2017

1 Introduction

Fantasy football is a popular game in which participants assemble a roster of real-world athletes and gain points for their successful performance in football matches. Fantasy sports appeared as a unique phenomenon in the USA in 1980s [2]. The rapid evolution of the Internet turned fantasy sports into an easily accessible entertainment for millions of people worldwide. Fantasy games are usually timed to coincide with major championships and events, either single (i.e. the FIFA World Cup) or season-long (e.g. the English Premier League). The most popular fantasy sport subject in Europe is association football. Fantasy tournaments are often run either by media (e.g. `www.sports.ru/fantasy`) or by leagues' organizers (the Fantasy Premier League [1], the Official Bundesliga Fantasy).

In this paper, a model of optimal sequential decision making in the Fantasy Premier League (FPL) is presented. There were more than 4.5 million participants in the 2016/2017 FPL season. This fact makes the FPL the most competitive fantasy league in the world. A high level of competitive intensity requires making the best decisions over time in order to succeed in the FPL. This task is particularly demanding as there is a spanless set of possible actions, moreover, the actions should be made under uncertainty. Thus, there is a need for a model that will be powerful enough to perform well under these circumstances. The general idea is to maximize the number of points scored in a given round by solving an integer programming problem such that an objective function sums number of points predicted by a machine learning algorithm and constraints describe corresponding game rules. To predict expected points, we apply XGBoost, one of the most effective modern data science methods. We then build an automated manager that utilizes this approach and achieves rather decent results in team formation.

2 Related Work

The growth of fantasy sports unsurprisingly motivates researchers to tackle the problem of playing optimally in the fantasy tournament.

Lutz [6] studies American football fantasy leagues. He attempts to predict fantasy NFL scores (expected points) of quarterbacks in a given match using historic match-level data. To this end, he uses two popular machine learning algorithms, support vector machines and neural networks.

Belien et al. [2] build a general linear optimization model for fantasy sports maximizing the total sum of points earned. Authors make an overview of popular fantasy games and introduce a set of constraints suitable for many of them. This model is the best possible model if one knows beforehand the exact number of points scored by each player in each gameweek. The study is illustrated by managing a fantasy cycling team squad.

Bonomo et al. [3] state an optimization problem similar to the one in [2]. In contrast to [2], this study focuses on maximizing the ex ante objective function that is explicitly set by the authors.

Matthews et al. [7] concentrate on the optimal performance in the Fantasy Premier League (FPL). They consider the problem of managing a fantasy team as a Bayesian reinforcement learning one and build a competitive automated FPL manager.

Goldstein et al. [5] investigate the wisdom of crowds phenomenon by analysing the FPL data and find that smart small crowds can be identified. Furthermore, a FPL manager can improve his strategy by copying the decisions of these smart players. The empirical research is supported by a theoretical consensus captain model.

3 Methods

Suppose that we know in advance the number of points scored by each player in each gameweek. In order to select the best possible team for the upcoming FPL gameweek, we have to solve the zero-one LP

$$\sum_{i \in I} \left(\mathsf{pts}_\mathsf{i}^\mathsf{gw} \left(x_i + k_i \right) + \left(\mathsf{pts}_\mathsf{i}^\mathsf{gw} - 4 \right) m_i + \varepsilon \left(\mathsf{pts}_\mathsf{i}^\mathsf{gw} \left(y_i + l_i \right) \right) \right) \to \max_{x_i, y_i, k_i, l_i, m_i} \tag{1}$$

subject to the constraints on the maximum total number of players in the roster, the size of the starting squad, the maximum total number of players representing the same Premier League team, the total number of players having the same position, and the maximum total cost of the whole team imposed by the FPL rules, where $\mathsf{pts}_\mathsf{i}^\mathsf{gw}$ denotes the number of points scored by the player i in the gameweek gw, x_i^{gw} equals to 1 iff the player i was in the squad in the previous gameweek and is a starter in the current gameweek, y_i^{gw} equals to 1 iff the player was in the squad in the previous gameweek and is a substitute in the current one, k_i^{gw} equals to 1 iff the player i came on a free transfer in the current gameweek and is a starter, l_i^{gw} – iff the player i came on a free transfer in the current gameweek and is a substitute, m_i^{gw} – iff the player i came on a paid transfer in the current gameweek and is a starter, and $\varepsilon > 0$ is arbitrarily small.

However, in a real situation we don't know beforehand the number of points that a player will score in the upcoming gameweek. Nevertheless, we can build a predictive model, make predictions for all players and treat $\mathsf{pts}_\mathsf{i}^\mathsf{gw}$ as *expected* number of points scored by the player $i \in I$ in the gameweek gw.

Our approach to solution is to use an algorithm that does not depend on multiple hypotheses, does not assume a certain type of dependence (e.g. linear), performs well in similar situations and has a built-in feature selection procedure. We apply tree boosting method called XGBoost [4] to the dataset from the official FPL website [1] that includes data for each unique combination of a player and a gameweek. It turns out that the most important features are the player's current form, the total number of incoming transfers, the player's current price, and points scored by the player in the previous gameweek.

Combining the two steps, the predictive model and the linear program, we are able to reach the top overall ranks comparable to the result of [7] for the manager considering only the upcoming gameweek. This result is achieved in the environment of more complicated game rules and increased competitiveness while using a simpler model.

References

[1] The official Fantasy Premier League website. `https://fantasy.premierleague.com`. Online; accessed: 2017-05-25.

[2] Jeroen Belien, Dries Goossens, and Daam Van Reeth. *Optimization modeling for analyzing fantasy sport games*, volume 4. KULeuven, Faculty of Business and Economics, June 2013.

[3] F Bonomo, G Durán, and J Marenco. Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game. *International Transactions in Operational Research*, 21(3):399–414, May 2014.

[4] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, March 2016.

[5] Daniel G Goldstein, Randolph Preston McAfee, and Siddharth Suri. The wisdom of smaller, smarter crowds. In *the fifteenth ACM conference*, pages 471–488, New York, New York, USA, 2014. ACM Press.

[6] Roman Lutz. Fantasy Football Prediction. *arXiv.org*, May 2015.

[7] Tim Matthews, Sarvapali Ramchurn, and Georgios Chalkiadakis. Competing with humans at fantasy football: team formation in large partially-observable domains. July 2012.