Ukrainian Catholic University

FACULTY OF APPLIED SCIENCES COMPUTER SCIENCE BACHELOR PROGRAMME

Can a team picked by machine learning beat an avarage player's team in Fantasy Premier League?

Author:
Marko Mylyanyk

2020

Abstract—In the modern world, games become an important part of people’s lives. They help to spend time while waiting in line or riding to work, university, etc. Also, games that contain the possibility to play with or against other real players in them give a feeling of real-life competition. Fantasy Premier League is one of those games. It gives you a chance to pick a team of real-world football players and compete with other people. The purpose of the game is to pick a team, which by the end of the football season will get the biggest amount of points. I have been playing this game for the last three years and my attempts had not gone well, in spite of the fact that I have good knowledge of English football and analytical thinking. The goal of this paper was to describe my approach to create a program that picks a team that at the end of the season will be better than average, based on data from previous seasons. In this article, I give an overview of key points of my work, such as data exploration, extraction of main values and building a model that fits here the most.

Index Terms— Direct data manipulation, Games, Machine learning, Modeling and predictions

1 Introduction

AS I have already mentioned that I have played this game for three consecutive years and only once I have got results better than average. Having more than 5 million active users previous year, Fantasy Premier League is continuously growing, and attracts more and more attention. As a result, this year there are more than 7 million players. So, let’s see what actually FPL is.

1.1 Fantasy Premier League

English Premier League (EPL) is an English professional league for men’s association football clubs. It is contested by 20 clubs and operates on a system of promotion and relegation. Owing to its eminence in the sports arena as the most-watched sports league in the world, it generates high revenue and has, therefore, garnered broad attention from prominent industrialists and businessmen. Fantasy Premier League (FPL), a simulated version of EPL, is a chance for football enthusiasts to create their own squad of 15 players and battle across the entire world to atop the leaderboard. Organized by EA Sports, FPL aims at providing lucrative prizes and heaps of appreciation to the winners. The study is focused on helping a novice aficionado to obtain a head-start when contesting FPL and exhibit an approach better than the traditional naive ones.

1.2 Dream Team

As indicated in 1.1, the main objective of FPL is to maintain a team of 15 players whose performance will be the sole determiner of the leaderboard position. Ideally, the squad is formulated after proper analysis of footballers across all 20 participating clubs. Given a fixed budget of £100 Million, the team must consist of 2 Goalkeepers, 5 Defenders, 5 Midfielders and 3 Strikers. Additionally, it should generate the highest team score possible, which in turn, depends on the game-weeks in which at least one player has stepped on the field. Using data from the past seasons and few different forecasting techniques, I am going to make Machine Learning predict and pick a Dream Team, or at least close to it, instead of me.

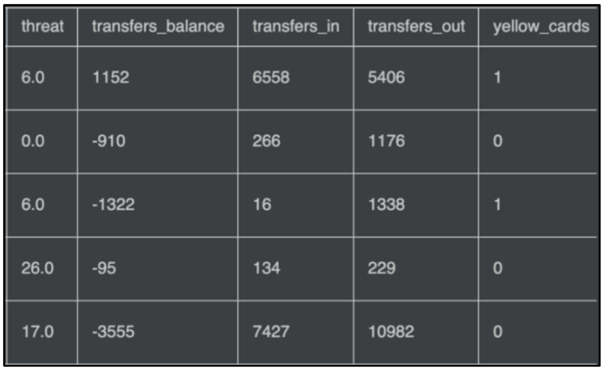
2 Observations

The first thing that I have noticed about most players of Fantasy Premier League was the player picking process problems. As all people have their own preferences and favorites in football, such as individual football players or teams, they often make a mistake - they try to pick their favorite, expensive and more popular players in a squad, instead of picking the cheaper ones with the same potential of point earning. To simplify, users make their decisions based on their favorite teams of English Premier League and most entertained players in real life. For example, at the end of the previous season Bournemouth FC player Callum Wilson got 168 points and Sergio Aguero from Manchester City got 201. Yes, 30 points of difference, but the price of Callum is $6.9 million, while the price of a famous Argentinian striker is a little higher than $12 million[1].

So, I believed if I would be able to remove my own preferences and favoritism, then I could concentrate only on players' and teams' overall performance, resulting in correct team picking.

3 Data Description

Data used in this research paper was taken from official Fantasy Premier League website [1]. It is a widely detailed data which describes players performance for three previous seasons in FPL. It consists of 1500 entries or rows, 624 of which are unique. This means that there were 624 different players claimed in teams by different coaches. Each of that entry is described by 58 different parameters which was my main background for building predictions. As there were no possibility to get all information together, I had merged three different files into one, using players name as a key value, exploit package named Pandas[2]. This is a sample of described data.

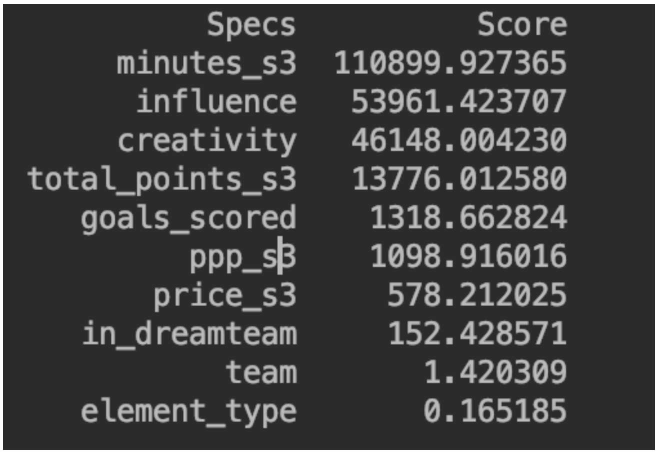


threat	transfers_balance	transfers_in	transfers_out	yellow_cards
6.0	1152	6558	5406	1
0.0	-910	266	1176	0
6.0	-1322	16	1338	1
26.0	-95	134	229	0
17.0	-3555	7427	10982	0

Fig. 1. Example of data received from Fantasy Premier League website.

4 Data Preparation

To make data appropriate to my needs, I had to found out which features are the most important. To do this I had used Univariate Selection, to determine the strongest relationship of all features with the output variable. As output variable I had decided to use "target" column, that have been created myself. It indicates players if player had more points than at least 75% of all players. If yes, then he has a value of 1, else – 0. I have selected ten, best-suited features using the chi-squared statistical test for non-negative values. The result of this operation is below.



```
            Specs          Score
       minutes_s3  110899.927365
        influence   53961.423707
       creativity   46148.004230
  total_points_s3   13776.012580
     goals_scored    1318.662824
           ppp_s3    1098.916016
         price_s3     578.212025
     in_dreamteam     152.428571
             team       1.420309
     element_type       0.165185
```

Fig. 2. Ten features on which depends player membership in to 75% of all players by total points gained.

So, I had understood that the most important parameter is minutes on a field spent by a player. And it makes sense, as the more you play, the bigger chance to commit an effective action, such as goal, assist or just to earn bonus points. Also, there are two more important features – "influence" and "creativity", so I had to count them when building a model. Features like "total_points" and "ppp" didn't count, as they already fully or partially consist of points, so there is straight correlation. Another feature "element type" means a position of a player on a field.

4.1 Data Cleaning

Dataset, what I had got after merging three others, consisted of a lot of weakly correlated to my aim information. And to perform predictions, I had to remove all redundant features. Mostly all of them were describing general and not important to points features, such as 'transfers_in", "transfers_out" which are showing how many users picked or dropped each player. Another example of dropped data, is "selected_by_percent" and "player_number" which didn't consist of any useful information to me. Even if, feature "selected_by_percent" sounds important, it only shows what percent of users have this player and as were mentioned, majority opinion is the first thing, we want to avoid.

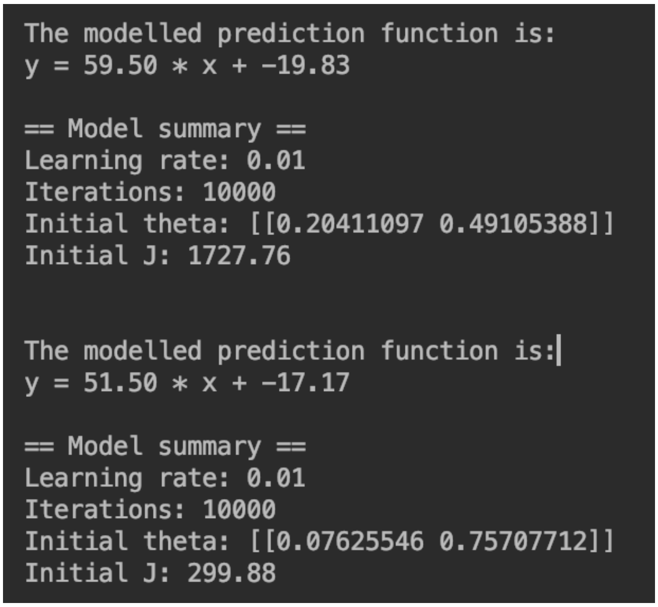
4.2 Missing Data

For each separate file, there were no missing values at all, but after merging, they appeared. It’s not a surprise, because obviously there are new players, that haven’t played one or all previous seasons, there are, players, that will not play in the next season by different reasons. It could be injury, transfer to another league, loss of coach respect and many other. Great news was that there are was only a few of entries that consisted of empty values. That’s why I have decided to deal with missing values, by substituting mean of previous or in some cases future seasons instead of Nan’s.

5 Modeling

5.1 Linear Regression

My first approach was to predict players future values, using Linear Regression. To perform this, I had written Linear model, that picks function that fits good, to parameters, I had put in it. In a simple word, I have written, an “next number prediction”, it generates the future feature value based on the same values, but in a previous season. In order to help minimize the errors associated with the prediction model and optimize it to better reality representation, I was using a Gradient Descent optimization algorithm. Based on data, that I had fed to the model, it had to develop simple function, that allowed to predict how would features evolve in the next season. It was performed to every player individually and to get good results, I had done 10,000 iterations for each player.



```
The modelled prediction function is:
y = 59.50 * x + -19.83

== Model summary ==
Learning rate: 0.01
Iterations: 10000
Initial theta: [[0.20411097 0.49105388]]
Initial J: 1727.76


The modelled prediction function is:
y = 51.50 * x + -17.17

== Model summary ==
Learning rate: 0.01
Iterations: 10000
Initial theta: [[0.07625546 0.75707712]]
Initial J: 299.88
```

Fig. 3. Example of functions, picked by a Linear Regression model, to predict total points in the next season.

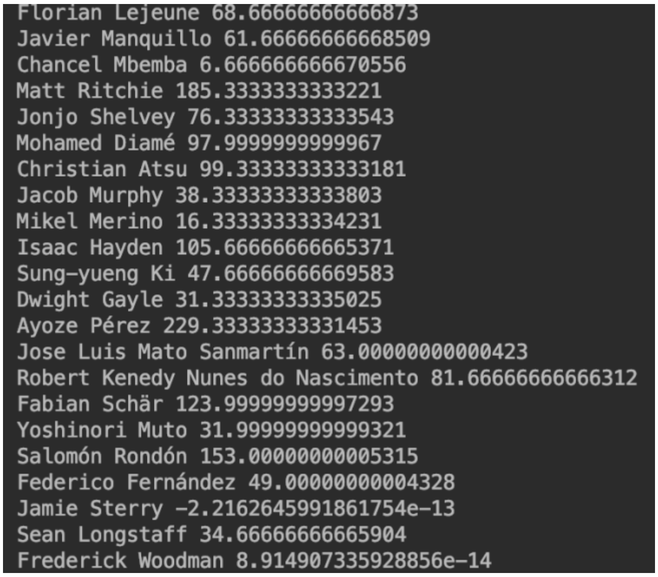
5.2 Binary Classification

To perform this type of prediction, I used as training data, data from two first available seasons. And as target column, I used column I had created before – “target”, which shows if player in top 25% of all players by total points. I did so, because I decided, that players from top 25% had the biggest potential to enter next year Dream Team. As model was trained, I used model on whole dataset, to get predictions on the following season. After that, I get all players that had potential to be in among the best in the next season.

6 Results

Result of a model, was new columns with predicted prices and points, based on which my team was picked.

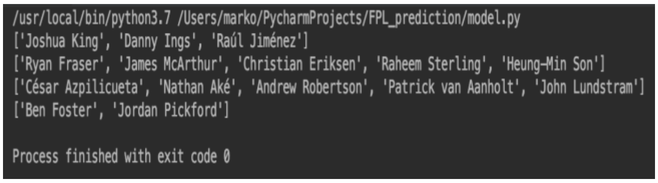
[1] All prices are presented in scope of a game equivalence, and do not match players actual market value.
[2] Pandas is a package of programming language Python.



```
Florian Lejeune 68.66666666666873
Javier Manquillo 61.66666666668509
Chancel Mbemba 6.666666666670556
Matt Ritchie 185.3333333333221
Jonjo Shelvey 76.33333333333543
Mohamed Diamé 97.9999999999967
Christian Atsu 99.33333333333181
Jacob Murphy 38.33333333333803
Mikel Merino 16.33333333334231
Isaac Hayden 105.66666666665371
Sung-yueng Ki 47.66666666669583
Dwight Gayle 31.33333333335025
Ayoze Pérez 229.33333333331453
Jose Luis Mato Sanmartín 63.00000000000423
Robert Kenedy Nunes do Nascimento 81.6666666666312
Fabian Schär 123.99999999997293
Yoshinori Muto 31.99999999999321
Salomón Rondón 153.00000000005315
Federico Fernández 49.00000000004328
Jamie Sterry -2.2162645991861754e-13
Sean Longstaff 34.66666666665904
Frederick Woodman 8.914907335928856e-14
```

Fig. 4. Sample of predicted total points by my model for the next season.

The value of received squad is $98.9 million, which is even slightly less that given budget of $100 million. To fairly estimate obtained results, there is a need to wait a few months, as Fantasy Premier League is depending on real time actions, which lasts for approximately 9 month.



```
/usr/local/bin/python3.7 /Users/marko/PycharmProjects/FPL_prediction/model.py
['Joshua King', 'Danny Ings', 'Raúl Jiménez']
['Ryan Fraser', 'James McArthur', 'Christian Eriksen', 'Raheem Sterling', 'Heung-Min Son']
['César Azpilicueta', 'Nathan Aké', 'Andrew Robertson', 'Patrick van Aanholt', 'John Lundstram']
['Ben Foster', 'Jordan Pickford']

Process finished with exit code 0
```

Fig. 5. Result of a team picking process. There are listed names of players, which have a great potential to get as a team in top 10% next season.

7 Conclusion

In this paper, I have shown how we could use Data Processing and Machine Learning techniques, to play an online game against real people. The problem of my model and approach in general is in dataset size. It is pretty small, to make an accurate predictions. In spite of this, I believe that my team picked by my program is good enough to complete the original target.

8 References

1. Official Website of FPL:
   https://fantasy.premierleague.com/
2. Data:
   https://fantasy.premierleague.com/drf/bootstrap-static
3. Source code:
   https://github.com/mylyanykmarko/FantasyPremierLeague
4. Gradient Descent:
   https://www.coursera.org/lecture/machine-learning/gradient-descent-8SpIM