Building the most effective football team with limited amount of money

Henrik Sergoyan

12/15/2018

Introduction

Nowadays the money plays very big role in sports. Those club owners who have a great deal amount of money has no challenges to buy the best players for their teams. For example, when in the beginning of the season 2016-2017 Paris Saint Germain FC bought Neymar from Barcelona FC with record-breaking €222m. Moreover, today many clubs, such as clubs from Chine or USA, are buying players for not only winning matches but also for increasing their revenue by selling tickets or t-shirts of that players. Unfortunately these things greatly influence the competitive balance in leagues which causes the decrease of interest towards them. As an example may serve Juventus in Serie A, Bayern Munich in Bundesliga or Paris Saint Germain in France.
However, in football there remain clubs which aim to find players who have very good performance in their leagues and have comparably lower cost in transfer market. Therefore, my aim of this project to build some optimizer which will find the best player for the specified limited price.

The steps of the project

1. Data scraping

First of all, for building such an optimizer I need to have the most recent data of players: their performance and their value. Today, this kind of valuable data cannot be easily downloaded from the Internet and if we are speaking about the recent data it is almost impossible to find anywhere except on wepbages. So my first goal was scraping this kind of data from the most famous web pages.

1.1 Whoscored

The next question was finding the website which would give me all detailed statistics of player according I woul build my performance measure metric. I realized that the website *www.whoscored.com* after each game provides the rating of players who played in this game. As it is considered one of the most popular soccer web pages I relied on their analyzes. However, scraping data from it was not as easy it task as it may seem because this website is able to detect spider/bot user agents and block them.
So the standard ways of scraping could not solve this problem, and therefore, after some research I found the way to do it with the help of Python package called *Selenium*. This package has played significant role in my project since without of it I would not have the data and could not make my project.

Example:
The performance of players of La Liga Season 2018-2019 so far:
https://www.whoscored.com/Regions/206/Tournaments/4/Seasons/7466/Stages/16546/PlayerStatistics/Spain-La-Liga-2018-2019

1.2 Transfermarkt

The same problem I encountered with when I was trying to find the most recent data of transfer values of all players. Football fans will know that the most reliable data is providing *www.transfermarkt.com* which also blocked all my standard attempts of scraping. Fortunately, with Selenium I scraped the information of this website also.

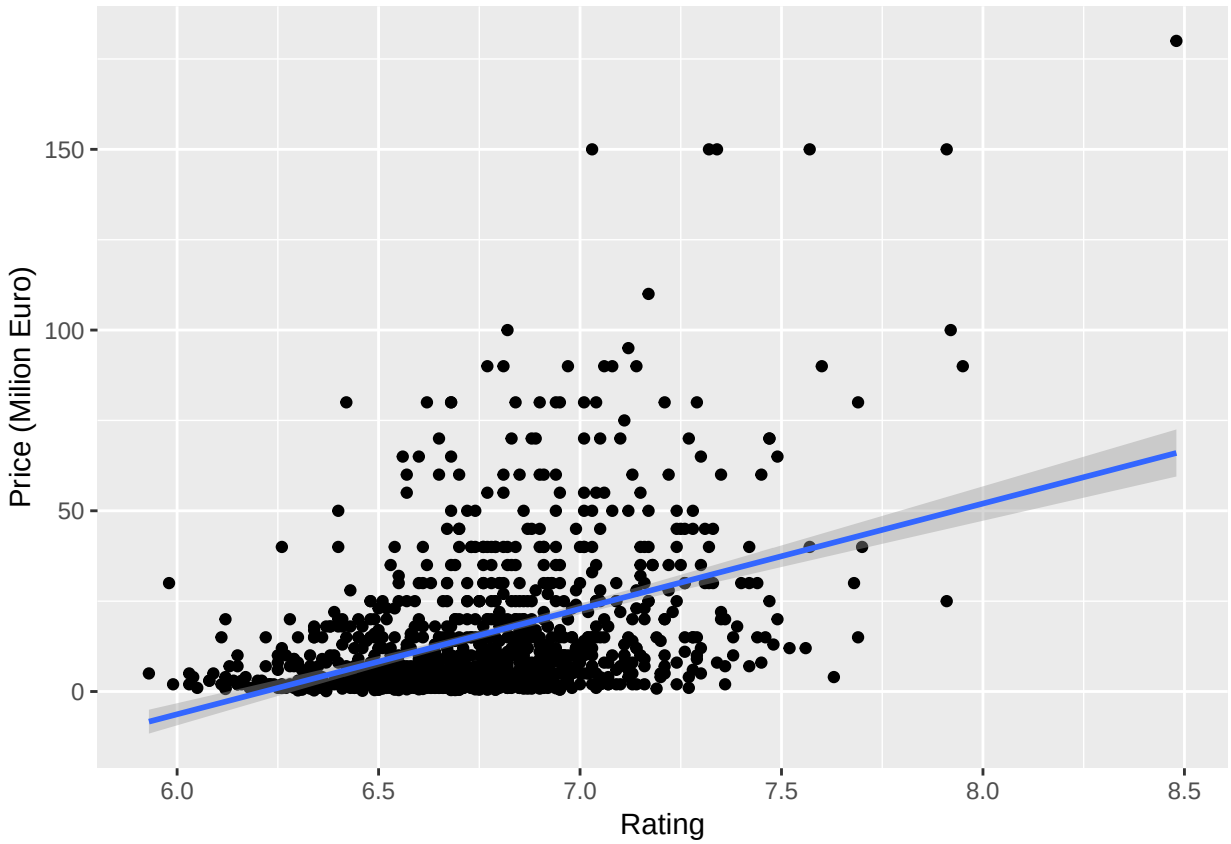
Example
Market values of players FC Barcelona:
https://www.transfermarkt.com/fc-barcelona/startseite/verein/131/saison_id/2018

2. Data preparation

The next task was merging those two datasets into one final dataframe. Fortunately, the incosistency between player names in Transfermarkt and Whoscored was not much and this task did not take long time. The next you may find the head of final dataframe

```
## # A tibble: 6 x 8
##   Players      Clubs        Mins   Age Values    Rating Position League
##   <chr>        <chr>       <dbl> <dbl> <chr>      <dbl> <chr>    <chr>
## 1 Lionel Messi Barcelona     871    31 180,00 M~   8.48 AM(CR),~ La Liga
## 2 Raheem Ster~ Manchester~  1026    23 90,00 Mi~   7.95 M(CLR),~ Premier ~
## 3 Cristiano R~ Juventus     1250    33 100,00 M~   7.92 M(L),FW  Seria A
## 4 Eden Hazard  Chelsea       963    27 150,00 M~   7.91 M(CLR),~ Premier ~
## 5 Thorgan Haz~ Borussia M~  1073    25 25,00 Mi~   7.91 AM(CLR)~ Bundesli~
## 6 Suso         AC Milan     1259    25 40,00 Mi~   7.7  AM(CLR)~ Seria A
```

For my own interest I plotted the scatterplot of player ratings and their values to see the correlation between these two variables.



From the graph we can see there are a lot of undervalued players (Below of blue line) which can be part of your team.

3. Optimization problem

The next task was to build an optmizer to maximize the overall rating of the team. So the optimization problem can be formulated as follows:

maximize sum of the ratings of a players of a club
subject to sum of the values of a players less than equal to your budget
number of players = 11

In addition to this, there is also one constraint which is number of players in each position which depends on the formation of your team (4-4-2, 4-3-3, e.t.c).
If you are familar with the most famous optimization problems in Linear Programming you may admit that this is the same as Knapsack problem (0/1) with item limit. We know that Knapsack problem can be easily solved with dynamic programming but when we are adding item limit constraint to it we would need to implement multidimensional dynamic programming.
Fortunately, R has a package called *lpSolve* the functions of have all necessary optimization skills to solve the problem. All the codes you may find in the file *final_shiny.R*.

4. Shiny Application and conclusion

As the beautiful conclusion of my project I decided to build one small shiny application for helping all interested people to see the effectiveness of my project. My shiny application consists of two parts. The first part is a simple search engine which helps to find the players and their ratings and values. The next part is the squad builder. In this part you are specifying your budget and number of players in each position. After clicking "Submit" button the corresponding squad of players will appear. You may choose your preferred number of players in each position, for example 0 GK and 0 defender, 2 midfield players and 1 forward. The link to the application:
https://henosergoyan.shinyapps.io/squad_builder/