EDA

han

2019/5/11

研究问题

研究问题的提出

足球是三大球之一，在全球有众多拥趸，赛事历来备受瞩目。因此，评估球员能力，发掘球员潜质，对于球队而言有着举足轻重的地位。其中，数据分析是能够客观有效的评估并预测球员表现的极其重要的一环。本次作业将会探索，是否能够通过本赛季球员的表现，来预测球员下一赛季球员的表现?

具体而言，研究问题是：通过本年度球员各项指标，能否预测下一年球员进球数?

数据说明

本次作业收集了两个年度（2012-2013以及2013-2014年度）的英超赛事中球员表现的各项指标，指标解释说明如下：

1.球员

2.年龄（单位：岁）

3.球队：所在球队名称

4.号码：球服号码

5.位置（分为前锋、中场、后卫）

6.出场（单位：次）

7.首发（即开场就上场，非替补）次数

8.出场时间：在场上总比赛时间（单位：分钟）

9.进球：进球数目

10.助攻：通过传球等方式帮助己方队友进球（单位：次）

11.传球：传球给己方队友（单位：次）

12.过人：通过一系列动作甩掉对手的（单位：次）

13.抢断：在规则允许范围内，夺取对方控制的足球（单位：次）

14.越位：越位是指该球员在接受己方进攻传球的瞬间a)在对方半场b)在球的前面c)与球门线之间不足两位对方球员，此指标为次数统计

15.犯规（单位：次）

16.红牌：被裁判红牌罚下次数

17.黄牌：被裁判黄牌警告次数

18.射门：球员射门次数

19.射正：射球在球门范围内的次数

20.射门成功率：进球数/射门数

21.头球进球：通过用头顶球进门次数

22.左脚进球：左脚射门进球次数

23.右脚进球：左脚射门进球次数

24.直接任意球进球：得到直接任意球机会的进球次数

25.点球：赢得点球机会时进球次数

26.赢得点球机会（单位：次数）

27.拦截：拦截对方进攻次数

28.解围：解围对方进攻次数

29.头球解围：对方球员进攻时通过头球方式解围次数

30.后场解围：在己方半场把球大力踢出，解围对方进攻次数

31.头球争顶成功（单位：次）

32.下一年进球（对应球员下一年进球数）

这些变量可以分成若干部分：

1.球员属性：包括球员、年龄、球队、号码、位置。

2.出场概况：包括出场、首发、出场时间、传球、抢断。

3.进攻数据：包括进球、下一年进球、助攻、过人、射门、射正、射门成功率、头球争顶成功。

4.犯规情况：包括越位、犯规、红牌、黄牌。

5.进球方式：包括头球进球、左脚进球、右脚进球、直接任意球进球、点球、赢得点球机会。

6.防守数据：包括拦截、解围、头球解围、后场解围。

描述统计

球员进球数~球员属性

位置

从图中明显看出，预测下一年球员进球数应该着重关注前锋的情况。而且前锋就是司职为球队取得进球的。
后面的讨论如不特别说明，都只讨论前锋的情况。

年龄

至于年龄，年龄和本赛季进球数有很弱的正相关，而与下一年的进球数有稍微强一点的负相关。

首先看一下“年龄~进球”和“年龄~下一年进球”的散点图，可以发现，数据点集中分布在中间年龄段，而且进球数比较多的球员也集中在25到30岁之间。用线性关系来描述进球、下一年进球与年龄的关系不是特别好，从散点图中可以猜测进球数与年龄有二次关系。

年龄与进球的关系可能可以反映在：

1.在一定年纪之前，随着年龄的增长，比赛经验会变得丰富，技术会变得娴熟，进球效率会提高。

2.过了一定年纪之后，年龄越大，运动机能下降，有的技术动作无法充分完成而导致进球效率下降。

球员进球数~出场概况

从出场概况的角度看，出场情况（出场次数、首发和出场时间）以及传球、抢断都对本赛季的进球有正相关的关系，这种正相关是很容易理解的：出场次数、时间越多，越有可能创造进球机会从而赢得进球；而本赛季的传球、抢断数据从某个角度体现了一个射手在场上的活跃程度，前锋在场上越活越，他在本赛季的进球数就可能越多。

然而这些变量对于下赛季的进球的正相关程度都不同程度地减少，甚至变成弱负相关。这是为什么呢？其中一个可能的理由是伤病，本赛季发光异彩的前锋球员如果劳累过度，下一赛季一旦出现伤病，球员数据例如进球数将会受到直接的影响；另外一个可能的原因是突然没有状态、或者由于能力出众受到对方球员重点盯防。

因而，如果要预测前锋下一年的进球情况，就不太适合用有关球员本赛季出场情况的变量了。

球员进球数~进攻数据

从进攻数据的角度看：

1.助攻、过人、射门和射正与本赛季进球数有较强的相关性，其中从图4的散点图中可以看出,射门、射正与进球数呈现出非常明显的线性关系，然而图4中助攻、过人与进球数的散点图中就感觉可能会有强影响点（特别是“过人~进球”散点图中的右上方——当时还在利物浦的苏亚雷斯），用pearson相关系数可能会有偏差。而图5中的相关系数是kendall相关系数，助攻、过人与进球的相关性变成弱正相关了。另外，射门成功率、头球争顶成功这两个变量与本赛季进球数也有一定的弱相关性。

2.助攻、过人、射门和射正与下一年的进球数也有不小的相关性，其中射门，射正与下一年进球数的相关系数相比于本年的进球数明显下降了，但正相关性还是比较强；助攻、过人与下一年进球数的相关系数非常稳定，相比于本年的进球数的相关系数差别很小，这两个变量是稳定的衡量球员下一年进球潜力的变量。

3.本年进球数与下一年的进球数从图4的散点图中能够看出是有比较强的正相关的，这也符合常理。pearson相关系数达到0.62，而kendall相关系数也有0.3，正相关还是挺明显的。

4.虽然助攻、过人、射门和射正与下一年进球数的相关性，特别是kendall相关系数才0.3、0.4左右，但是助攻、过人、射门和射正不失为衡量球员下一年进球潜力的变量，因为这四个变量都是进攻进球的常见且重要的步骤，体现了一个射手的基本素质。而射门成功率以及头球争顶成功次数与进球数的相关性较弱且随着时间的推移相关性下降，不是衡量球员下一年进球潜力的很好的变量。

图5是助攻、过人、射门、射正、进球与下一年进球的kendall相关系数矩阵图，也体现“助攻、过人、射门、射正、进球”这五个变量与“下一年进球”变量的明显的正相关关系。

球员进球数~犯规情况

从犯规情况看，犯规、红黄牌与进球数量的相关性弱，因为球员是否进球与球员犯规情况几乎不搭边。

而越位与进球数量的相关性略强。可能的原因是，越位通常发生在进攻时，而积极进攻的前锋更有可能在无意中触碰“越位”犯规，与此同时积极进攻的前锋可能在更多的前插中创造出更多进攻机会，从而可能会带来更多的进球。

球员进球数~进球方式

从进球方式看，各种进球方式与本赛季进球数以及下一赛季进球数都有较强的相关性，特别是左脚进球和右脚进球与进球数的相关性较高，因为这两种进球方式是进球的主要方式，左脚或右脚射门是一个球员的基本素养；而获得点球机会越多，说明这名球员进攻经常能够深入禁区，使得对手不得不在禁区里犯规，同样可以反映球员的进球能力可能是比较高的。

因而，左脚进球、右脚进球和获得点球机会这三个变量可能是预测球员下一年进球潜力的很好的变量。

球员进球数~防守数据

最后几个变量都是考量防守队员（中后卫）的常见指标，而作为前锋通常不会经常去拦截解围。这四个变量与进球数量的关系是负相关，可能是因为如果前锋不得不着力参与防守时球队通常在比赛时就陷入困境了，这时球员想进球就比较艰难了。

总结与后续解决方案

综合上述球员能力六个角度的数据的分析，可以初步筛选出可能能够用来预测球员下一年进球能力的变量，例如年龄、助攻、过人、射门、射正、进球、左脚进球、右脚进球和获得点球机会等。

从parallel coordinate plot中可以看到，除了年龄之外的每一个维度，从上往下看总体而言颜色从深到浅，意味着在除年龄之外的维度上数值越高，下一年进球数就倾向于越高；而年龄这个维度从上往下看总体而言颜色从浅到深再到浅，这可以表明下一年进球数与年龄可能呈现二次关系。

总而言之，经过描述统计，我们对数据有了初步而全面的了解，接下来如果要定量预测下一年球员进球数，可以运用建立回归模型的方法进行。

Machine Learning Method

```
library(reticulate)
use_python("/anaconda3/bin/python3.6")
```

Random Forest

SVM

Neural Network

Conclusion

Well, SVM method is better than Random Forest, which makes sense, for that random forest is better for larger samples. As for this task, we have only less than 200 samples, where SVM is more suitable.(Neural Network is to be implemented.)