欧洲足球数据分析

聂宇舟 姚漪涵
韩子航 刘治列

模型改进与赌博策略分析

中国人民大学 RENMIN UNIVERSITY OF CHINA | 统计学院 SCHOOL OF STATISTICS

✓ 进一步改进

回顾. . . . . .

描述统计方面

- 挑选出各个位置具有代表性的球员六维模板
- 对球队排行做更加准确的分析，包括联赛内部和全体球队两种情况
- 针对博彩赔率的分析

统计建模方面

- 将球员的 22*6 个特征、球队的整体数据纳入考虑
- 使用深度学习方法和前述的机器学习方法进行对比
- 一个球员可能踢多个位置，在一场比赛中是否踢到了最适合的那个位置吗？
- 某个位置的球员，是某方面能力突出，还是各方面能力均衡更有助于球队取得胜利？

✓ 模型改进

- 对特征进行扩充
- 对模型本身进行改进

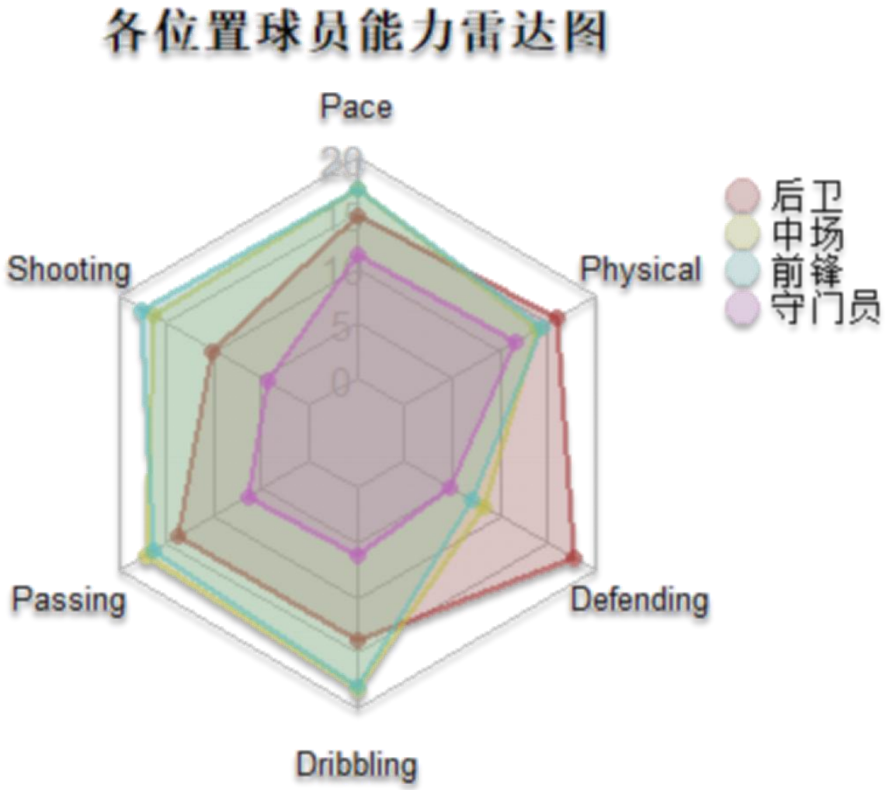
✓ 模型改进——扩充特征：球员

Pace	Shooting	Passing	Dribbling	Defending	Physical	GoalKeeper
Acceleration(0.45)	Att Position(0.05)	Vision(0.2)	Agility(0.1)	Interceptions(0.2)	Aggression(0.2)	gk_diving
Sprint Speed(0.55)	Finishing(0.45)	Crossing(0.2)	Balance(0.05)	Heading Acc(0.1)	Jumping(0.05)	gk_handing
	Long Shots(0.2)	Curve(0.05)	Reactions(0.05)	Marking(0.3)	Stamina(0.25)	gk_kicking
	Penalties(0.05)	FK Accuracy(0.05)	Ball Control(0.3)	Slide Tackle(0.1)	Strength(0.5)	gk_positioning
	Shot Power(0.2)	Long Pass(0.15)	Dribbling(0.5)	Stand Tackle(0.3)		gk_reflexes
	Volleys(0.06)	Short Pass(0.35)				

- 修改球员六维属性
- 将全部六维属性作为代替总体评分作为特征

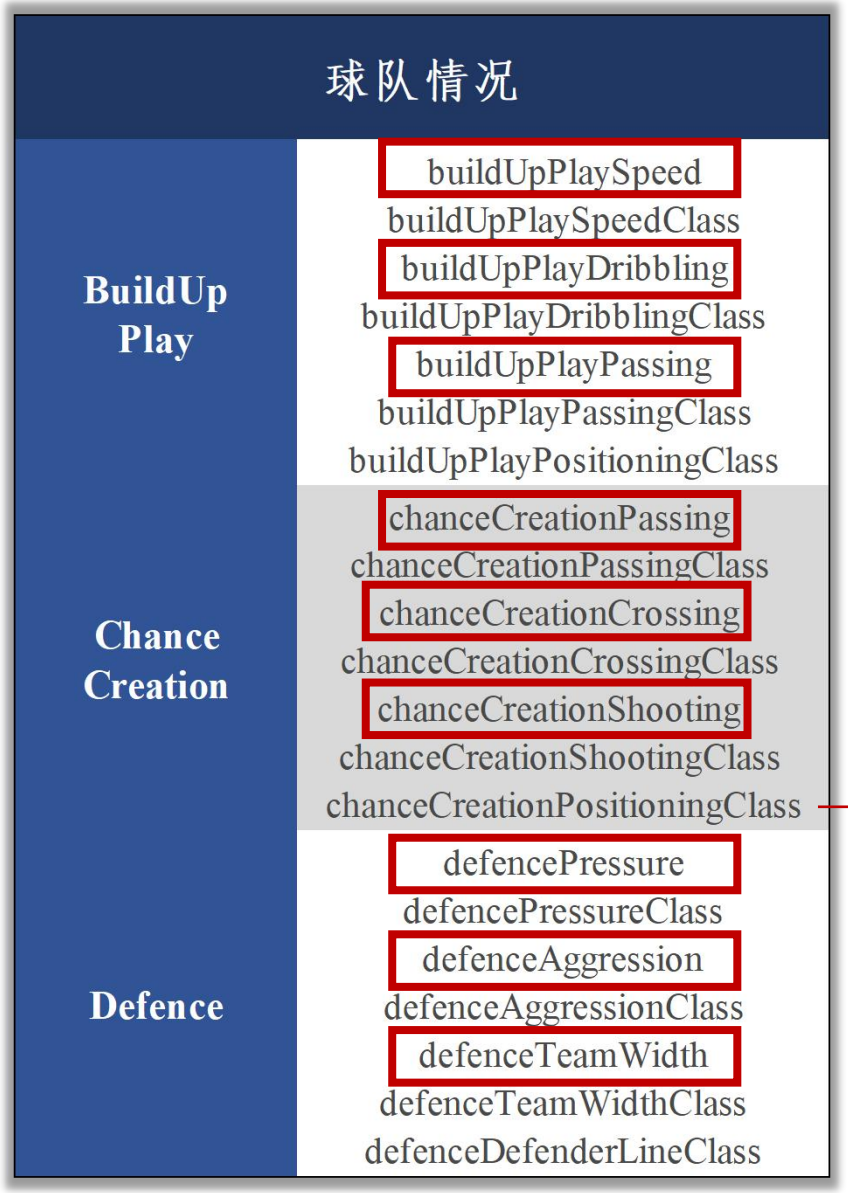
✓ 模型改进——扩充特征：球员

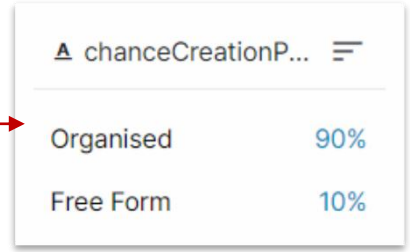
- 后卫球员的防守能力最佳
- 前锋球员射门和移动能力都最佳
- 中场球员传球能力最佳
- 后卫和中场在身体素质方面最佳
- **守门员相较于其他位置属性较低，故使用其位置自带五维特征作为变量**



✓ 模型改进——扩充特征：球队

- 将球队属性划分为三个部分，分别求均值得到球队属性
- 使用数值变量，不使用分类变量
- 将这3×2个变量和原先两只球队历史战绩变量拼接作为球队特征





✓ 模型改进——扩充更多模型

机器学习模型

- 使用更多集成模型：**ExtraTree、CatBoost、XGBoost**等等

深度学习模型：简单的三层神经网络

- 使用 **PyTorch** 和 **FastAI** 实现
- 均使用 **Adam** 优化器，运行直至收敛

其他操作

- 去掉赌博数据
- 在所有模型之前均使用 **PCA**（除开 **Logistic Regression** ）

✓ 模型改进——最终效果

- 增加特征的最终效果
  - **Accuracy：0.5439**
  - **Balanced accuracy: 0.4482**
- **深度学习模型取得了最好效果，但是速度非常慢**
- 综合来看，**ExtraTreeGini** 是**最具性价比**的模型
- 模型难以提升，不再做过多讨论

	model	test_accuracy	valid_accuracy	fit_time
0	NeuralNetTorch	0.5439	0.5303	14.6876
1	ExtraTreesGini	0.5305	0.5238	1.7374
2	LightGBMXT	0.5286	0.5349	3.8549
3	RandomForestGini	0.5284	0.5244	3.1672
4	CatBoost	0.5276	0.5329	14.2614
5	ExtraTreesEntr	0.5275	0.5143	4.4265
6	RandomForestEntr	0.5273	0.5251	1.8295
7	XGBoost	0.5270	0.5225	4.9992
8	WeightedEnsemble_L2	0.5257	0.5264	14.8083
9	LightGBM	0.5252	0.5395	7.2707
10	LogisticRegression	0.5244	0.5369	3.0577
11	LightGBMLarge	0.5223	0.5238	10.4874
12	NeuralNetFastAI	0.5200	0.5198	18.5978
13	KNeighborsDist	0.4615	0.4438	0.0460
14	KNeighborsUnif	0.4502	0.4248	0.0460

✓ 模型改进——模型解释

Logistic Regression 训练得到的部分参数

位置	变量	Defeat	Draw	Win
应当是后卫	home_player_2_pace	0.328	-0.156	-0.123
	home_player_2_shooting	0.084	-0.057	-0.007
	home_player_2_passing	-0.246	0.371	-0.035
	home_player_2_dribbling	-0.200	0.008	0.130
	home_player_2_defending	-0.292	-0.116	0.311
	home_player_2_physical	-0.558	0.279	0.224
应当是中场	home_player_6_pace	0.518	-0.667	0.074
	home_player_6_shooting	0.006	-0.266	0.026
	home_player_6_passing	-0.165	0.021	0.192
	home_player_6_dribbling	-0.234	0.300	-0.051
	home_player_6_defending	-0.518	0.592	-0.114
	home_player_6_physical	-0.029	0.157	-0.106
应当是前锋	home_player_11_pace	-0.345	0.114	0.181
	home_player_11_shooting	0.196	-0.839	0.445
	home_player_11_passing	-0.164	0.249	-0.100
	home_player_11_dribbling	-0.323	0.318	-0.030
	home_player_11_defending	0.239	-0.322	0.068
	home_player_11_physical	-0.100	0.105	0.020

- 防守队员：防守力提升将带来胜率巨大提升
- 中场球员：传球和移动能力提升将带来胜率巨大提升
- 前锋球员：射门能力提升将带来胜率巨大提升
- **各位置球员的某方面能力权重显著高于其他权重**

✓ 模型改进——模型解释

Logistic Regression 训练得到的部分参数

位置	变量	Defeat	Draw	Win
应当是后卫	home_player_2_pace	0.328	-0.156	-0.123
	home_player_2_shooting	0.084	-0.057	-0.007
	home_player_2_passing	-0.246	0.371	-0.035
	home_player_2_dribbling	-0.200	0.008	0.130
	home_player_2_defending	-0.292	-0.116	0.311
	home_player_2_physical	-0.558	0.279	0.224
应当是中场	home_player_6_pace	0.518	-0.667	0.074
	home_player_6_shooting	0.006	-0.266	0.026
	home_player_6_passing	-0.165	0.021	0.192
	home_player_6_dribbling	-0.234	0.300	-0.051
	home_player_6_defending	-0.518	0.592	-0.114
	home_player_6_physical	-0.029	0.157	-0.106
应当是前锋	home_player_11_pace	-0.345	0.114	0.181
	home_player_11_shooting	0.196	-0.839	0.445
	home_player_11_passing	-0.164	0.249	-0.100
	home_player_11_dribbling	-0.323	0.318	-0.030
	home_player_11_defending	0.239	-0.322	0.068
	home_player_11_physical	-0.100	0.105	0.020

- 某个位置的球员，是某方面能力突出，还是各方面能力均衡更有助于球队取得胜利？

某方面能力突出

- 一个球员可能踢多个位置，在一场比赛中是否踢到了最适合的那个位置？

对于符合参数权重分布的球员是正确的

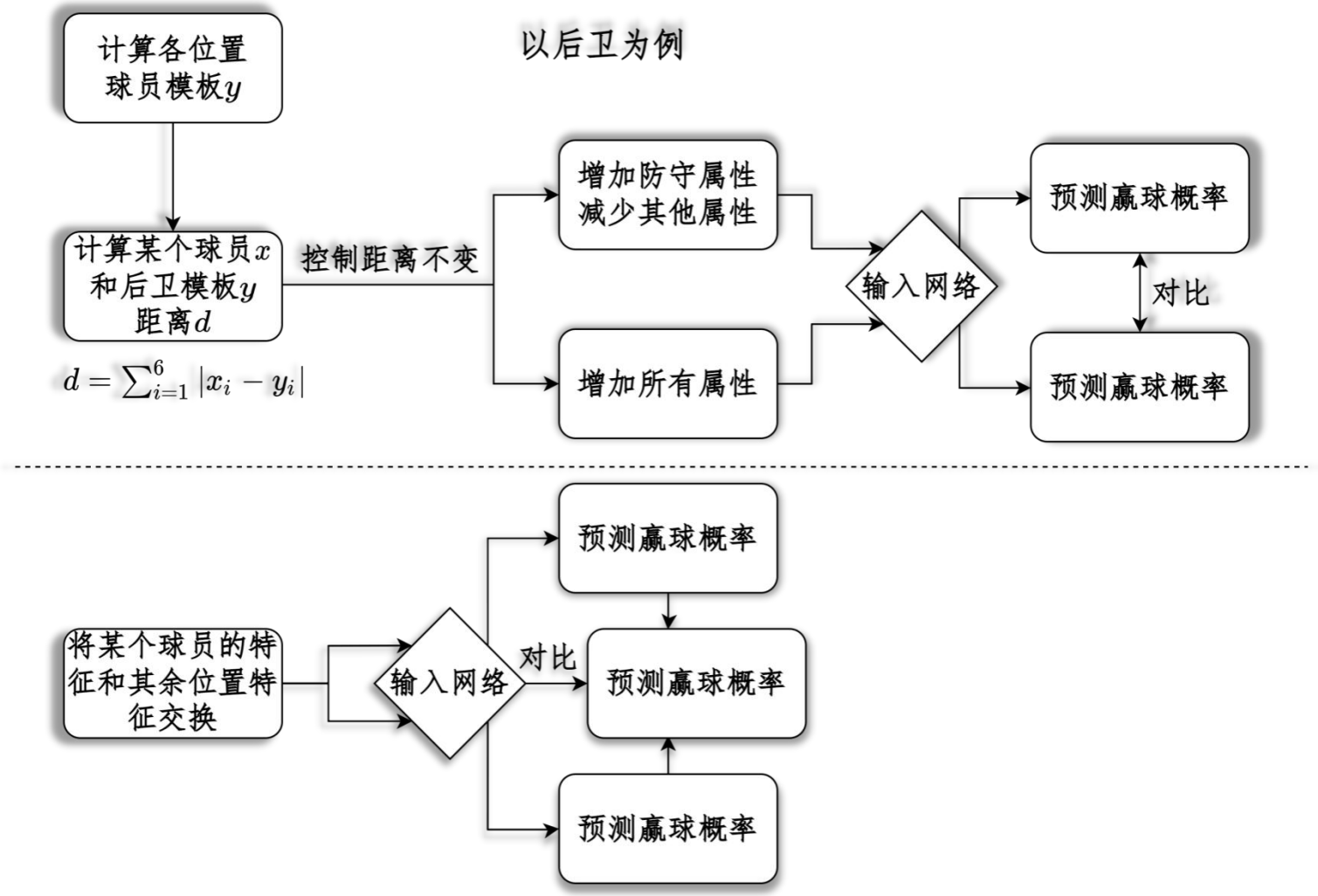
问题：

- Logistic Regression 预测效果并不是很好，甚至存在对胜率负增益的属性，参数不一定具有说服力
- 球员 1~11 使用坐标进行区分，不同比赛的前锋、中场、后卫、守门员四个位置人数和坐标不定

✓ 问题探究

- 某个位置的球员，是某方面能力突出，还是各方面能力均衡更有助于球队取得胜利？

- 一个球员可能踢多个位置，在一场比赛中是否踢到了最适合的那个位置？



利用神经网络进行测试

✓ 问题探究——球员位置分析：不同能力

- 针对门将、后卫、中卫、前锋，各随机挑出一场比赛中的某一球员，对其各维度得分按照上述方式进行调整
- 以后卫为例
  - 增加防守得分其他维度减少得分，相比于另一种方式，该球员所在球队的胜率有所上升
  - 后卫的防守能力相较于其他能力对比赛胜率的影响更大

后卫

决策	Match ID	Defeat	Draw	Win
增加防守	1	0.35463	0.24696	0.39841
均衡增加	1	0.35779	0.24988	0.39233
增加防守	4179	0.14066	0.23469	0.62465
均衡增加	4179	0.14050	0.23501	0.62448
增加防守	8801	0.33909	0.29733	0.36358
均衡增加	8801	0.33911	0.29743	0.36347
增加防守	20087	0.07688	0.06691	0.85620
均衡增加	20087	0.07712	0.06683	0.85605

✓ 问题探究——球员位置分析：最适合的位置

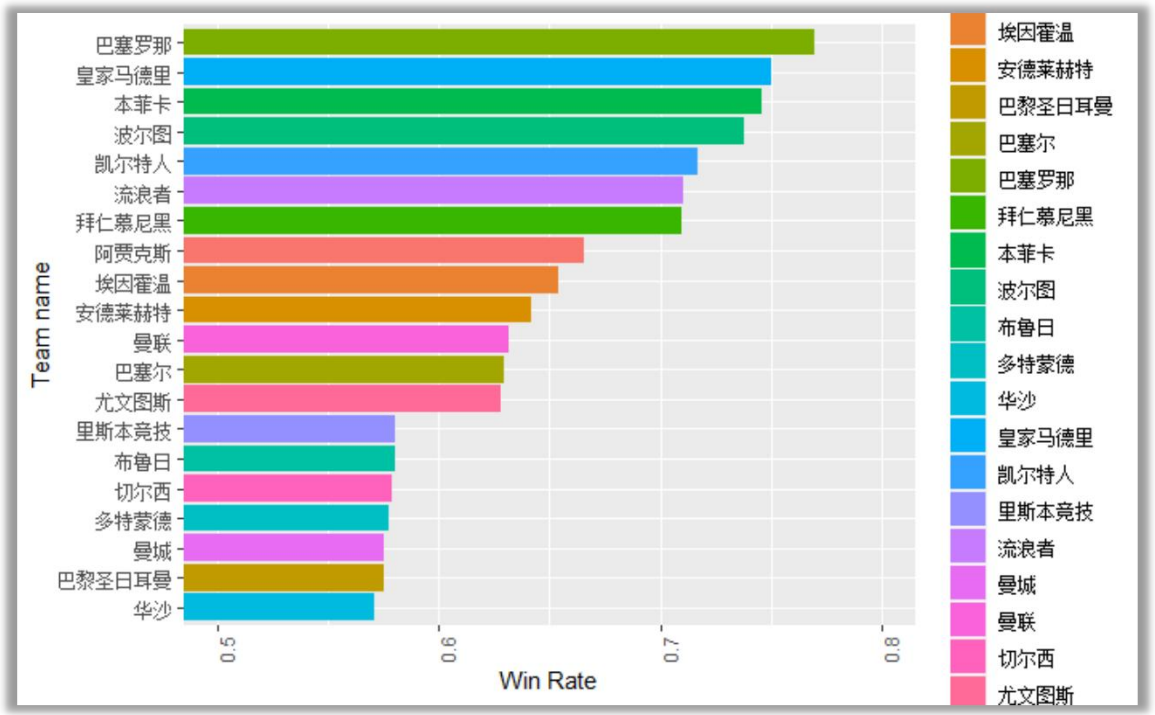
- 任选三支球队，**选取该球队中能力较强的球员**，对其球员位置进行调整
- 发现球员所在球队胜率**均有下降**
- 通过 Logistic Regression 参数得到的想法应当是正确的
- 而且现实中教练往往对球员较为熟悉，正常情况下做出的应当是最佳配置

位置	Match ID	Defeat	Draw	Win
前锋*	7	0.16627	0.24689	0.58684
中场	7	0.16624	0.24737	0.58639
后卫	7	0.18054	0.24134	0.57812
守门员	7	0.21149	0.23467	0.55384
前锋	3327	0.41523	0.30165	0.28313
中场*	3327	0.38765	0.31325	0.29911
后卫	3327	0.39776	0.31121	0.29103
守门员	3327	0.45607	0.30046	0.24347
前锋	19732	0.18826	0.23843	0.57331
中场	19732	0.18506	0.24209	0.57284
后卫*	19732	0.18275	0.23986	0.57738
守门员	19732	0.21187	0.23467	0.55346

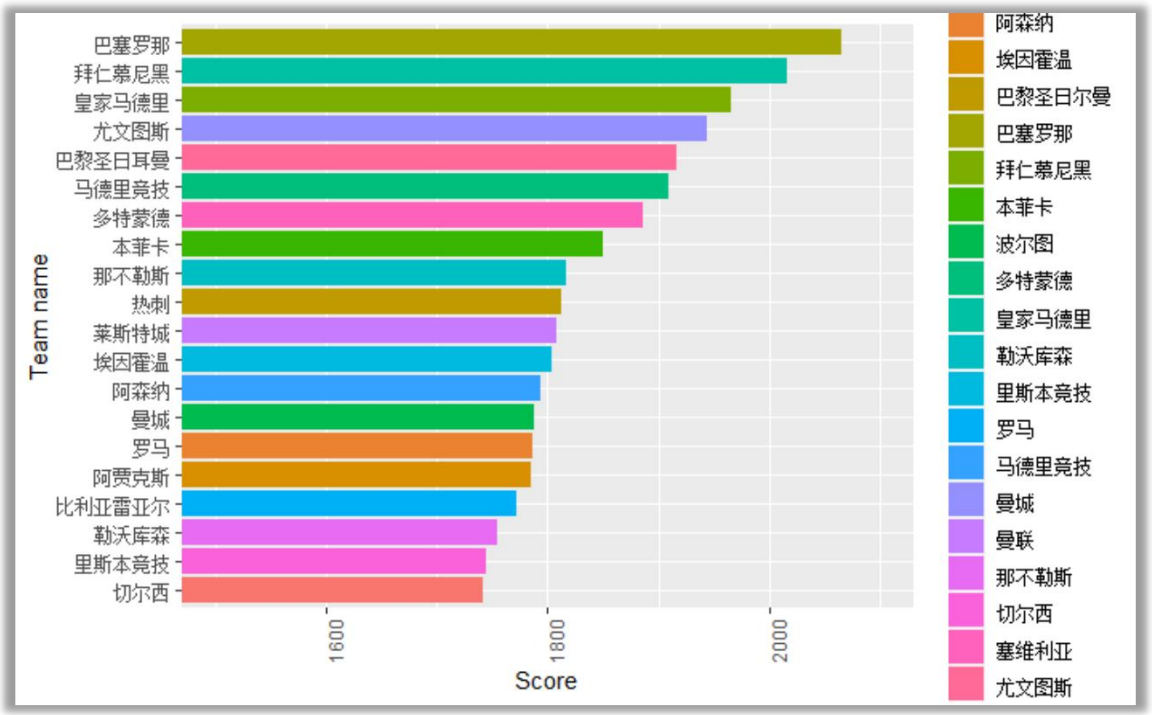
✓ 球队改进——胜率排行

- 之前的思路没有考虑到不同联赛的球队之间水平
- 将球队的本身属性、联赛规模、球队中的所有球员属性纳入考虑
- 以 2015-2016 赛季为例

原始胜率排序结果

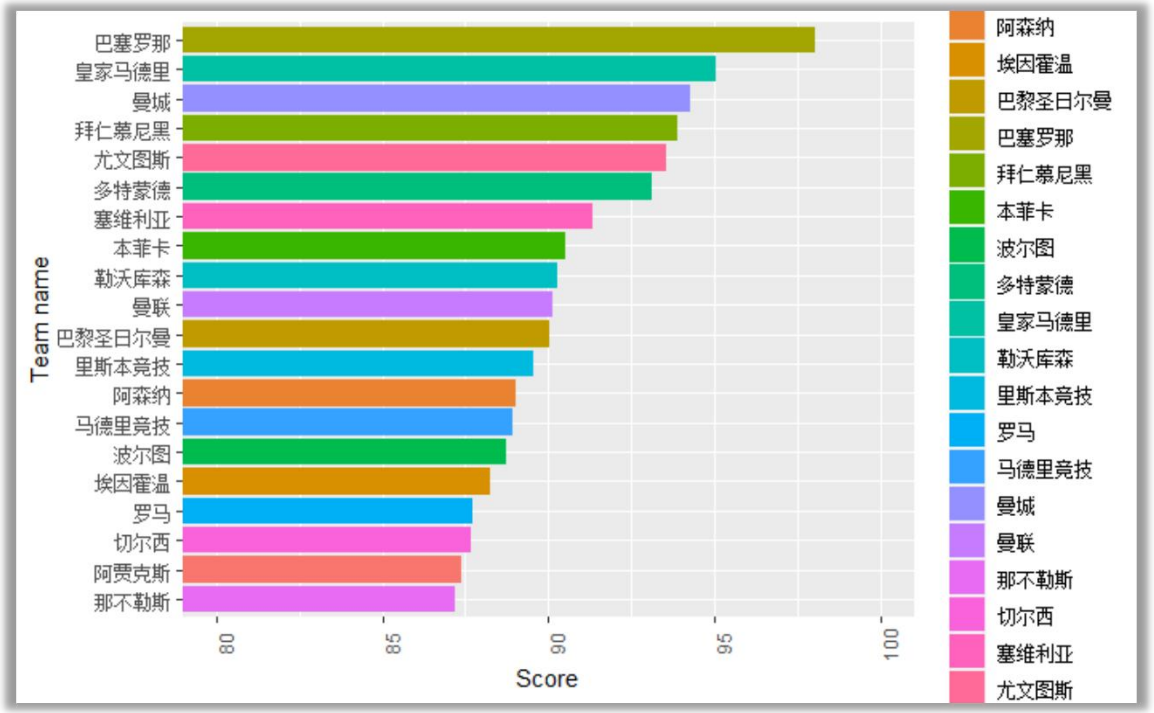


世界球队积分排序

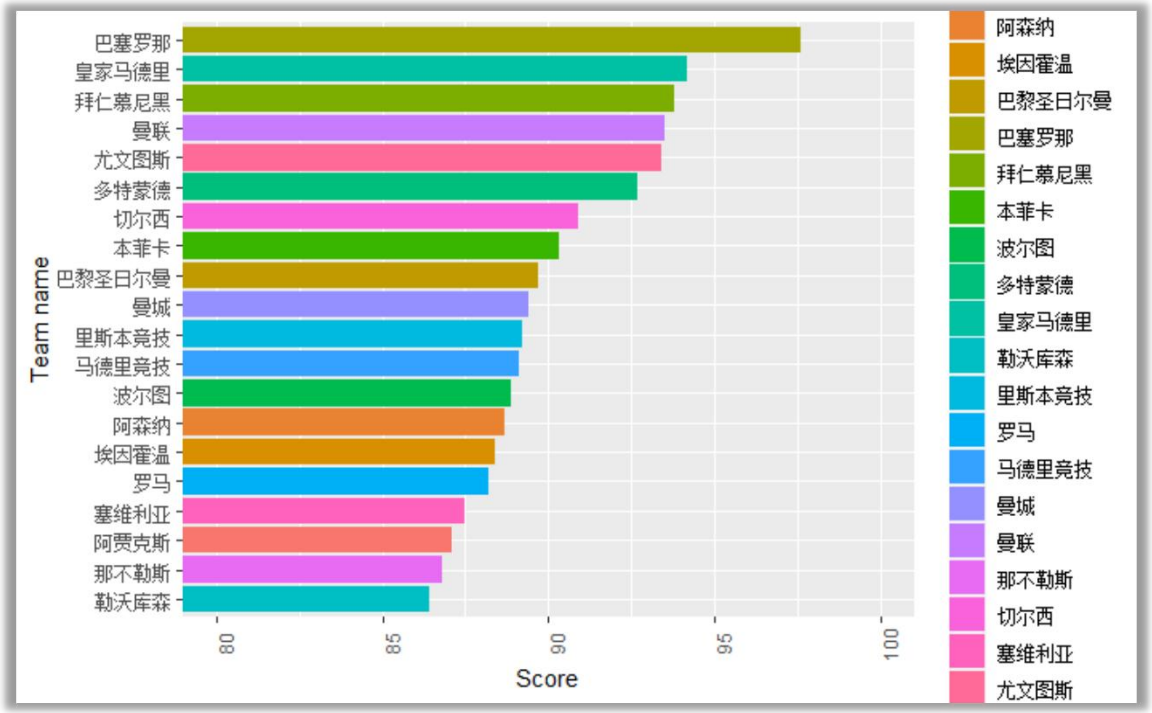


✓ 球队改进——综合排行

将全体特征纳入考虑



去掉球队属性特征



球队属性是 **2008-2016** 的综合结果，

在 **15-16** 赛季使用不够合理

✓ 球队改进——最终结果

- 使用 **spearman 秩相关系数**算法，得到不同排行榜的差异
- 事实证明，**不使用球队数据的排序**结果最为准确

世界球队积分排名	原排序法	综合排序（去除球队数据）	综合排序
巴塞罗那	巴塞罗那	巴塞罗那	巴塞罗那
拜仁慕尼黑	皇家马德里	皇家马德里	皇家马德里
皇家马德里	本菲卡	拜仁慕尼黑	拜仁慕尼黑
尤文图斯	波尔图	曼联	曼城
巴黎圣日耳曼	凯尔特人	尤文图斯	尤文图斯
马德里竞技	流浪者	多特蒙德	多特蒙德
多特蒙德	拜仁慕尼黑	切尔西	塞维利亚
本菲卡	阿贾克斯	本菲卡	本菲卡
那不勒斯	埃因霍温	巴黎圣日尔曼	勒沃库森
热刺	安德莱赫特	曼城	巴黎圣日尔曼
莱斯特城	曼联	里斯本竞技	曼联
埃因霍温	巴塞尔	马德里竞技	里斯本竞技
阿森纳	尤文图斯	波尔图	马德里竞技
曼城	里斯本竞技	阿森纳	波尔图
罗马	布鲁日	埃因霍温	阿森纳
阿贾克斯	切尔西	罗马	埃因霍温
比利亚雷亚尔	多特蒙德	塞维利亚	罗马
勒沃库森	巴黎圣日耳曼	阿贾克斯	那不勒斯
里斯本竞技	曼城	那不勒斯	切尔西
切尔西	华沙	勒沃库森	阿贾克斯
和积分排行榜对比	0.517	0.665	0.618

✓ 赌博策略分析

博彩公司	含完整赔率的比赛数
B365	19691
WH	19678
BW	19675
VC	19669
LB	19662
IW	19635
SJ	14627
BS	11856
GB	11854
PS	10446

- 共有十家博彩公司
- 一场比赛包含他们对胜负平预测的三个赔率，然而存在缺失值

✓ 赌博策略分析

简单利用赔率信息进行博彩预测（胜负平）

所有比赛统一投注10元

- 安全策略

  每场比赛都购买**最低赔率**的一方

- 冒险策略

  每场比赛都购买**最高赔率**的一方

- 随机策略

  每场比赛都**随机购买**胜负平中的一种

✓ 赌博策略分析——最终结果

只使用 **England France Germany Italy Spain Netherlands Portugal** 联赛进行分析

策略方法	成本（元）	正确预测比例	盈利（元）	盈利比例
随机选择	169540	33.35%	-12291.4	-7.25%
最安全的策略	169540	53.44%	-6208.2	-3.66%
最冒险的策略	169540	20.84%	-15059.8	-8.88%
我们的模型	152160	59.29%	26439.7	17.38%

- 我们的模型在大型联赛上取得了59%的准确率！
- 没有考虑比赛顺序关系，无法应用，后续将继续修改，或许可以在22年的联赛赌球:）

谢谢大家！

代码供参考：https://github.com/rucnyz/soccer