寻找未来足球之星

——基于 FIFA22 的足球运动员身价预测模型分析报告

小组成员：1.2020110301 郑强 2.2020110432 孙逸凡 3.2020111025 张凌玮

摘要：本报告通过 FIFA22 中球员的各指标数据来对足球运动员身价进行分析和预测，并以此为基础寻找各个位置在身价上最具有潜力的十位球员。我们选取了包括年龄、俱乐部、国籍、国际声誉、终结能力、射门力度、远射能力、罚点球能力、截断能力、逼抢能力、滑铲能力在内的各项能力指标来对球员的身价进行预测。因为能力之间存在着较大的相关性，因此我们采用了主成分分析的方式，将各项能力归纳为了进攻因子与防守因子。从我们的分析结果中可以发现，球员的国际声誉、防守能力以及所效力的俱乐部对于球员身价的影响最为显著。基于本报告的研究结果，得到如下规律：对于大部分不是顶尖的足球运动员来说，努力提高自己的防守能力，在场上积极参与拼抢和回防；树立良好的球员形象，提升自己的国际声誉是提升自己身价的不错选择。

一、研究目的

随着卡塔尔世界杯赛程的不断进行，在全世界范围内又一次引发了一轮新的足球潮。在 2022 年 12 月 19 日，梅西带领的阿根廷队在常规时间绝平的情况下通过点球大战战胜卫冕冠军法国队，时隔三十六年再次夺得世界杯冠军，梅西最终也如愿捧起大力神杯，卡塔尔世界杯也终于告一段落。回顾本届世界杯，在这一届比赛中既有如摩洛哥这样的黑马杀出，帮助非洲球队在世界杯历史上首次挺进四强，也有拥有超高全队身价的比利时、德国折戟小组赛……正如曾经的西德国家队主教练塞普·赫尔贝格曾经说过:“足球是圆的，什么结果都可能踢出来”，球场上各种各样的不确定性正是足球比赛的魅力所在。但是我们也同样也希望通过当今各种各样的数据分析手段，与收集得到的足球数据来探寻分析各种“偶然结果”背后的“必然原因”。正如此次世界杯就引入了 Var 裁判，足球芯片，智能图像分析系统等新技术并新增了保持控球权、受迫性失误等 11 项统计数据，可见通过技术与数据来赋能足球等体育运动具有越来越重要的意义。

本次报告我们小组成员以足球运动员身价为切入口，探究在球员身价背后潜藏的对身价产生影响的各种因素。球员身价综合反映了一个球员的技术能力，商业价值，发展潜力等核心，是一名球员个人实力以及在球队中战略价值的充分体现。同样的一个球队身价也可以很好地衡量其“战斗力”，因此每当一支球队或者一名球员可以正面对抗并战胜身价数倍于自己的对手时，其背后的因素除了运气，战术，球员磨合度等等奇妙的化学反应外，很容易让人怀疑一名球员的身价是不是被高估或者低估了；此外同时足球圈也一直流传着类似于“英格兰户口本”等说法，那么国籍是否真的会对足球运动员的身价产生影响呢？我们希望通过研究各类可能影响球员身价的因素来探索究竟哪些才是其价值背后的“Key Point”，并以此为基础寻找球场上那些充满潜力、球员价值还没有被充分挖掘的明日之星，同时也可以给出各个国家队实际“战斗力”的一个侧面反应——国家队身价，并在未来做进一步研究。

二、数据说明

本案例从 FIFA2022 游戏数据中选取了来自 168 个国家 16366 名球员的 14 个变量指标，

包括球员的姓名、身价、年龄、国际声誉及各项属性等数据。其中因变量为球员身价，自变量分为球员基本信息和球员属性两个维度，变量详细说明如表 1 所示：

变量类型		变量名称	取值范围	备注
因变量		身价		单位：欧元
自变量	球员基本信息	球员姓名	16366 个水平，例如：L. Messi	
		国籍	168 个水平，例如：Argentina	
		所属球队	856 个水平，例如：Paris Saint-Germain	
		场上位置	15 个水平，例如：RW	
		年龄	[16,45],取整数	
		国际声誉	20,40,60,80,100	数值越大表示国际影响力越大
	球员属性	Finishing (门前终结能力) Shot Power (射门力量) Long Shots (远射技术) Penalties (点球) Interceptions (拦截) Standing Tackle (抢断) Sliding Tackle (滑铲)	[0,100],取整数	数值越大表示该能力越强，反之越弱

表 1

关于数据可靠性与时效性说明：

本数据来自于 FIFA2022 官方数据，最后更新时间为 2022 年 8 月 18 日。FIFA 系列能力相关数据由 EA Sport 公司通过调查员与球探数据库进行打分并结合球员的客观数据来给出，此外还会捕捉球员的动态技术特点最终由公司相关编辑员形成球员的能力值；而身价、俱乐部等数据也是根据真实转会市场数据实时更新，数据通过 Python 爬虫获取并进行了整理。

由于本项目开始进行时 FIFA2023 游戏刚刚登录市场，其数据可能存在一定的不稳定性并且有较多数据难以获取，因此我们选用了 FIFA2022 的游戏最后一次补丁数据，数据在具备稳定性的同时同样还具备很强的时效性。模型的预测结果也可以结合当下的一些数据来验证，使得模型更具有现实意义。

关于数据筛选的说明：

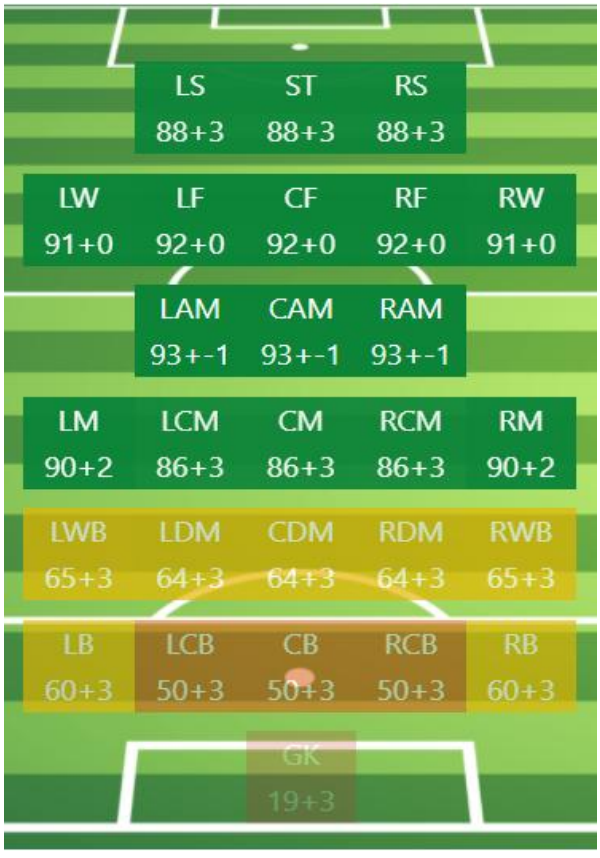
FIFA 官网上给出的球员能力值数据多达 30 多项，但是通过我们小组成员的初步分析发现能力指标间具有很强的共线性，并且通过简单回归之后其中很多数据的结果并不显著，考虑到一些技巧性比较高的技术能力实际意义不大，此外一些基本的球员脚下能力已经可以被其他能力值高低所反应，因此我们在攻防两端选出 7 项指标数据作为核心能力值模块来研究。

85 Crossing	96 Dribbling	90 Acceleration	86 Shot Power
91 Finishing	93 Curve	78 Sprint Speed	68 Jumping
70 Heading Accuracy	93 FK Accuracy	91 Agility	71 Stamina
91 Short Passing	91 Long Passing	93 Reactions	69 Strength
88 Volleys	94 Ball Control	95 Balance	93 Long Shots

44 Aggression	20 Defensive Awareness
40 Interceptions	35 Standing Tackle
93 Positioning	24 Sliding Tackle
95 Vision	
75 Penalties	
96 Composure	

（以球员 L.Messi 为例）

由于守门员与其他球员的能力值和身价判断标准差异较大，在本次研究中剔除了所有的守门员位置上的球员。此外当一名球员可以胜任多个细分位置时，我们选取了其在游戏中的推荐位置，虽然可能与现实俱乐部和国家队有一些出入，但这样划分一定程度上更能反应其真实的应用价值。我们还将所有位置分成了“前场”，“中场”，“后场”三大类，以便于研究和理解。如图 1 所示



（以球员 L.Messi 为例）

图 1

三、数据探索与描述分析

1、描述性统计

变量中既有数值型的连续变量也有国籍，俱乐部等离散变量，因此我们先对国籍和俱乐部两个变量打分（打分规则见因变量描述分析），并对所有变量计算均值，标准差，中位数，最大值，最小值，变异系数等描述性统计量，得到的结果如图 2 所示。

	name_dat	mean	sd	media	max	min	cv
[1,]	"Age"	25.49428	4.874531	25	39	16	19.1201
[2,]	"Overall"	67.64167	6.349019	68	93	44	9.386253
[3,]	"ValueEUR"	3494149	8218515	1300000	194000000	1000	235.2079
[4,]	"Nationality_score"	62.83753	23.3719	60	100	40	37.19417
[5,]	"Club_score"	56.91412	15.23329	50	100	50	26.76539
[6,]	"IntReputation"	23.26962	9.420019	20	100	20	40.48205
[7,]	"Finishing"	52.46224	16.06342	56	95	10	30.61901
[8,]	"Shotpower"	61.98741	12.56858	64	95	15	20.27601
[9,]	"Longshot"	53.77298	15.42492	57	94	12	28.68526
[10,]	"Penalties"	53.65668	12.61453	54	96	13	23.50971
[11,]	"interceptions"	51.17055	19.05711	57	93	10	37.24234
[12,]	"StandingTackle"	52.73099	19.12394	60	93	10	36.26698
[13,]	"SlidingTackle"	50.16887	19.2333	57	95	10	38.33713

图 2

2、数据初步探索

为了更好地理解数据以及接下来步骤，我们对数据进行了一些初步探索，绘制出一些具有现实意义的图表来展示当今足坛的一些基本情况。

如图 3 和图 4 所示，当今足坛前场位置上身价最高的前三位是来自巴黎圣日尔曼的姆巴佩，来自曼城的哈兰德以及来自热刺的哈里·凯恩，同时这些球员的身价也是冠绝整个足坛，承担着球队锋线上进攻手的重要角色；而综合评分最高的则是我们熟知的梅西，C 罗和莱万，三人虽然年龄都比较高，但是仍然保持着不错的发挥，尤其是梅西更是在今年卡塔尔世界杯带领阿根廷队时隔 36 年问鼎世界杯，而上赛季 C 罗和莱万也在俱乐部有所表现。不过在数据的初步探索中我们发现在对球员的能力值进行评价时存在一定的“明星效应”，主要也是为了吸引更多粉丝来获取这些球员，这也印证了用 30 多个综合能力值变量来衡量球员价值会存在一定的误差，很大一部分原因是其中一些带有主观情绪的不客观评分导致的。

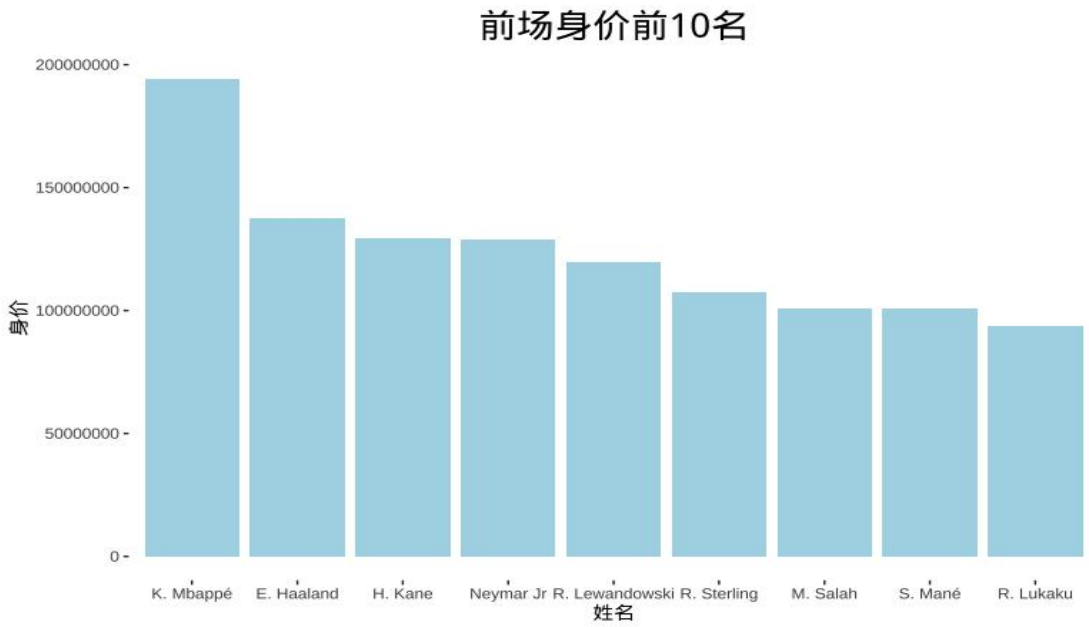


图 3

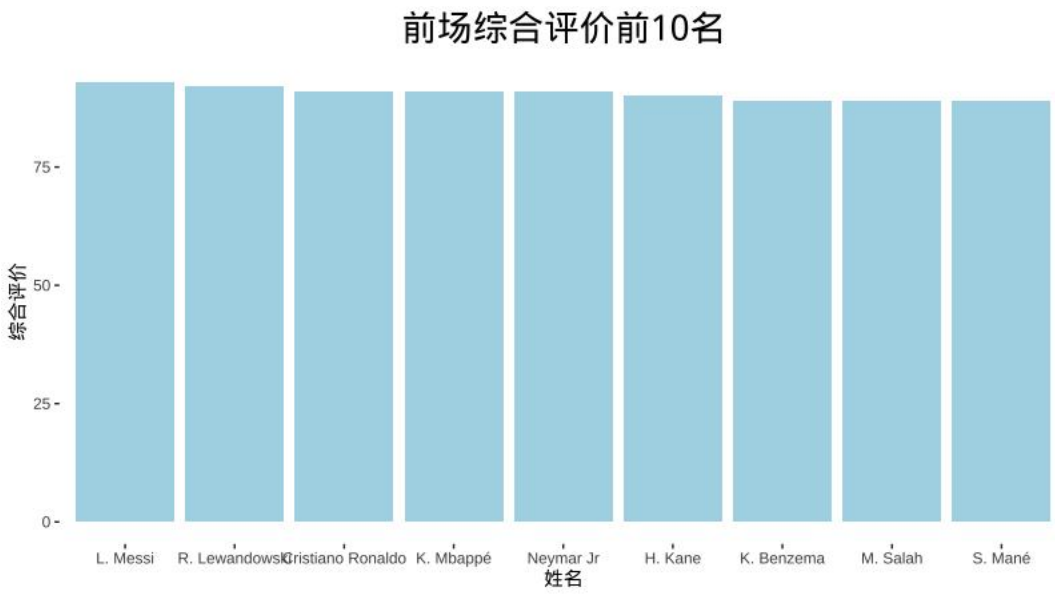


图 4

当今足坛中场位置上身价最高的前三位是来自曼城的德布劳内，来自巴萨的德荣以及来自曼联的桑乔，这些球员承担着球队的攻守转换中场发动机的任务；而综合评分最高的则是德布劳内，坎特和卡塞米罗，这些球员在联赛和杯赛中都有很好的表现。如图 5 和图 6 所示

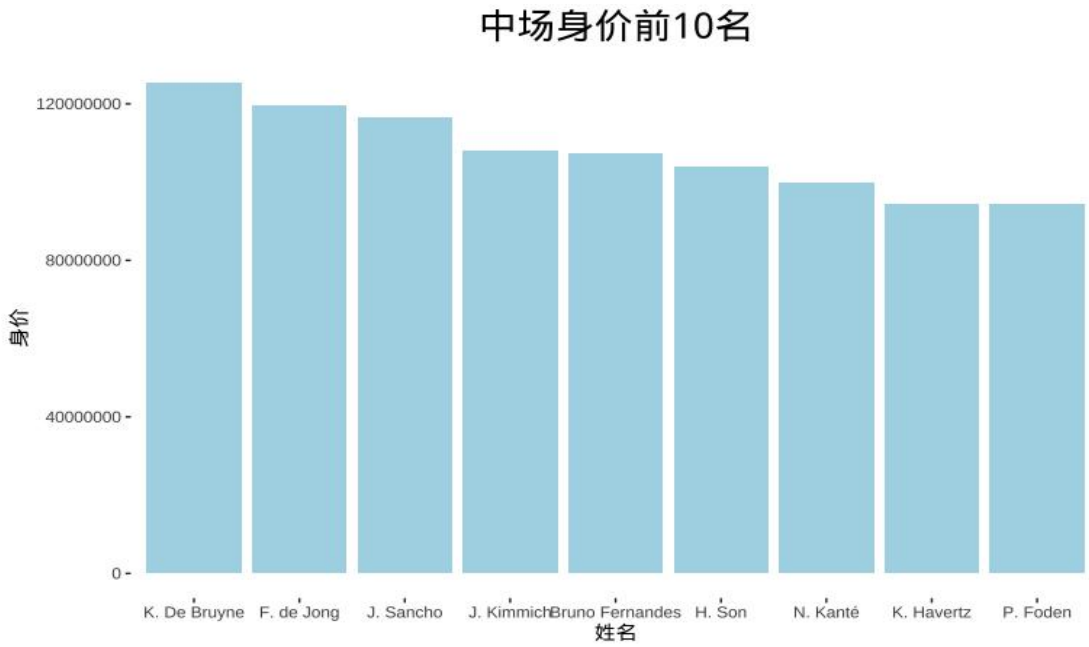


图 5

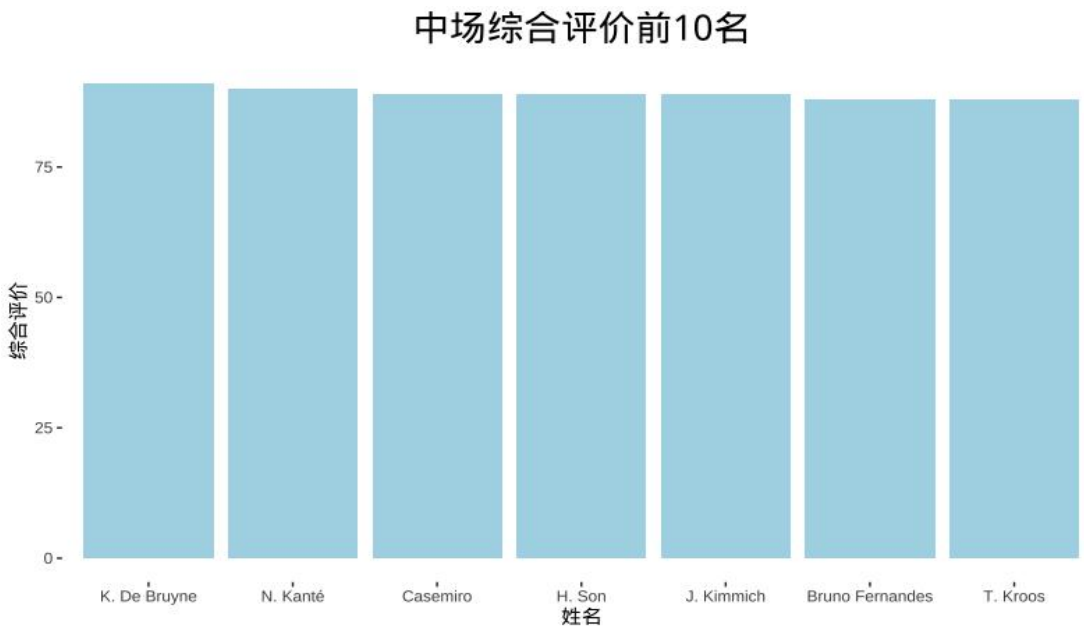


图 6

从图 7 和图 8 可以看出当今足坛后场位置上身价最高的前三位是来自利物浦的阿诺德，来自曼城的鲁本·迪亚斯以及来自曼联的马奎尔，这些球员承担着球队的防守和后场组织任务，同时后场中也不乏“带刀后卫”；而综合评分最高的则是范戴克，拉姆和拉莫斯。

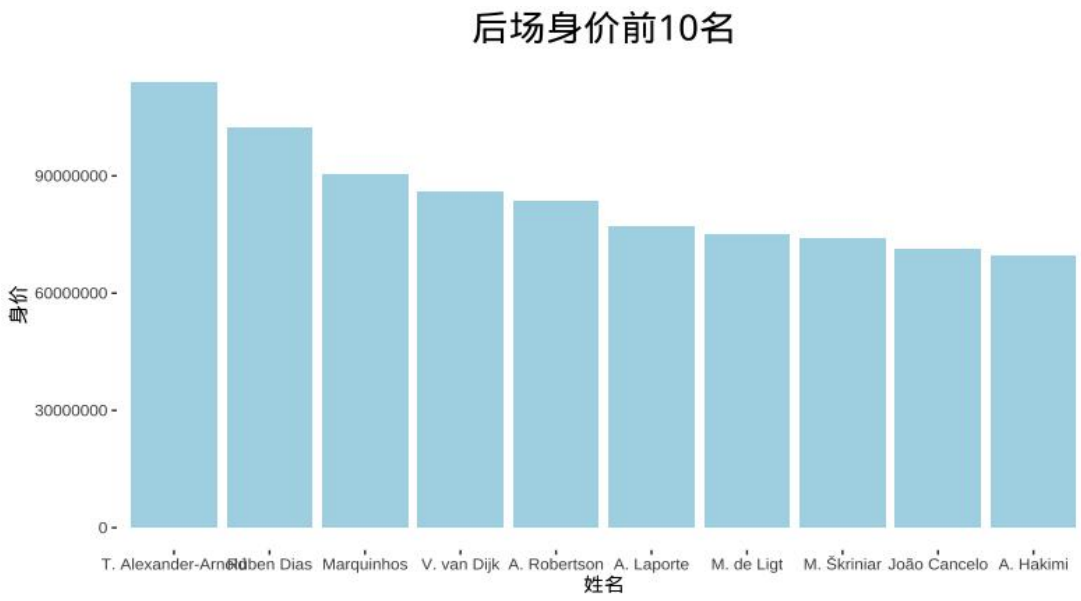


图 7

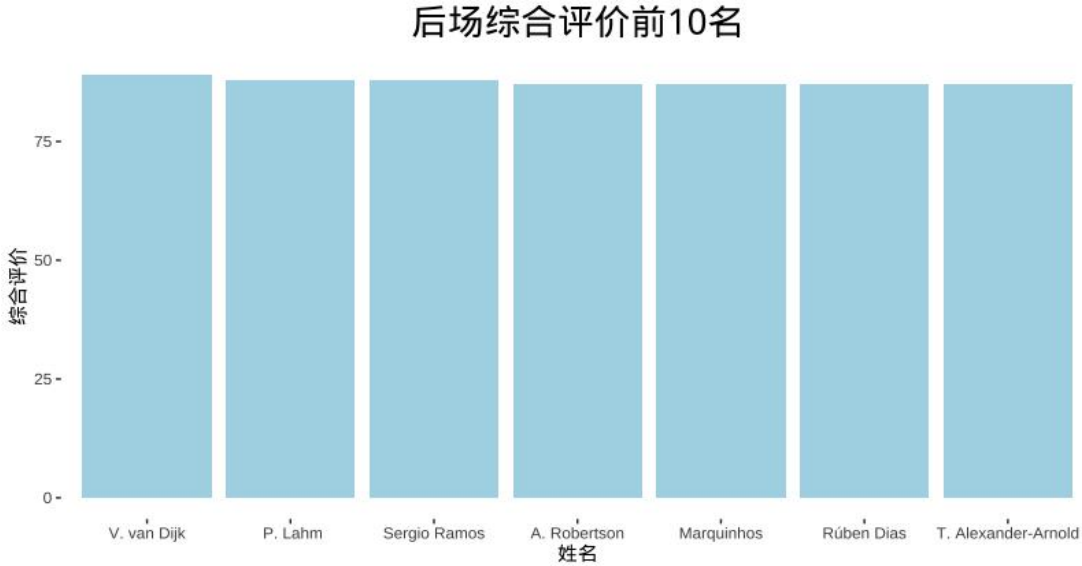


图 8

综上可以看到在超级巨星中前场球员的身价往往更高，在游戏中的能力也更突出，这一点主要也是因为人们往往更喜欢看到球员更激情的进攻而不是摆大巴式防守；此外也可以分析出 FIFA 在球员能力评分时考虑往往会给一些传奇巨星更高的综合评分，从而也造成忽视了现实世界中正处于当打之年的年轻球员的结果。这从侧面应证了对于球员能力值变量进行筛选的必要性，另一方面也说明现实中的身价比游戏里的评分更能反应球员的当前状态，通过对身价进行回归并预测是具有现实意义的。

接下来我们着眼于球员的数量，球员数量分布如图 9 所示。从整体上来看各个位置中，前场球员最少，中场球员最多，主要是游戏中将很多可以胜任边前锋的球员位置设定在了前腰等中场位置。前中后场的细分位置上前场以前锋，中场以前腰，后场以后腰人数最多，边路球员的数量较少，也可以看出如今可以胜任边路位置的球员的珍贵性。

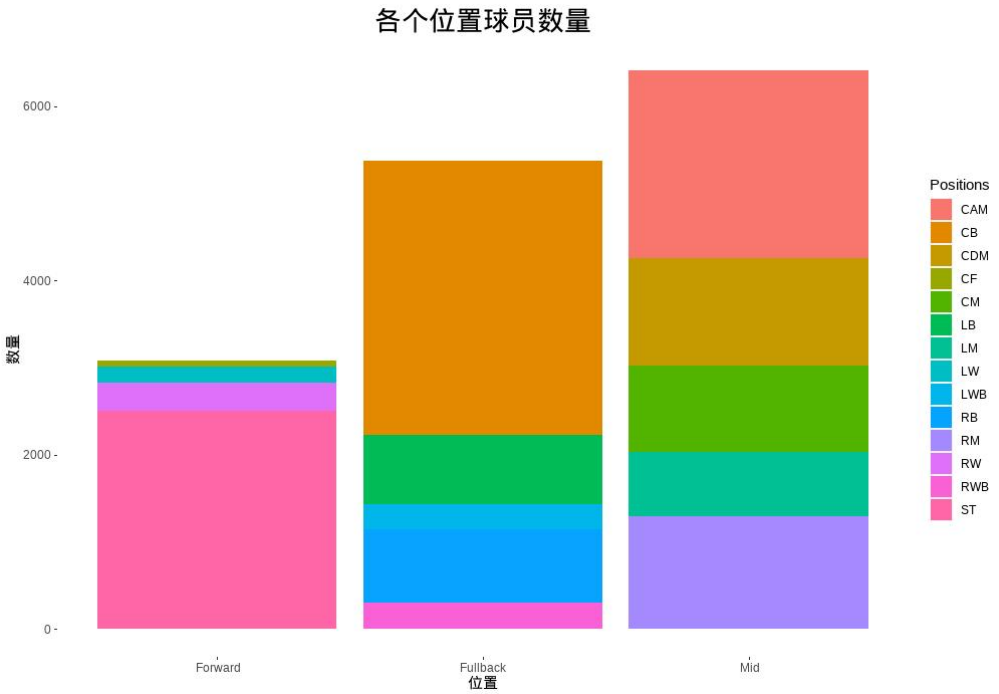


图 9

对于细分位置的平均身价如图 10 所示。从整体来看各个细分位置的平均身价也呈现出前场大于中后场的趋势，和上面分析的超级巨星的状况十分类似，同时也可以看出边路的球员往往身价更高，印证了上面的说法，在当今足坛无论是边锋还是边后卫，都是处于一种物以稀为贵的状态。

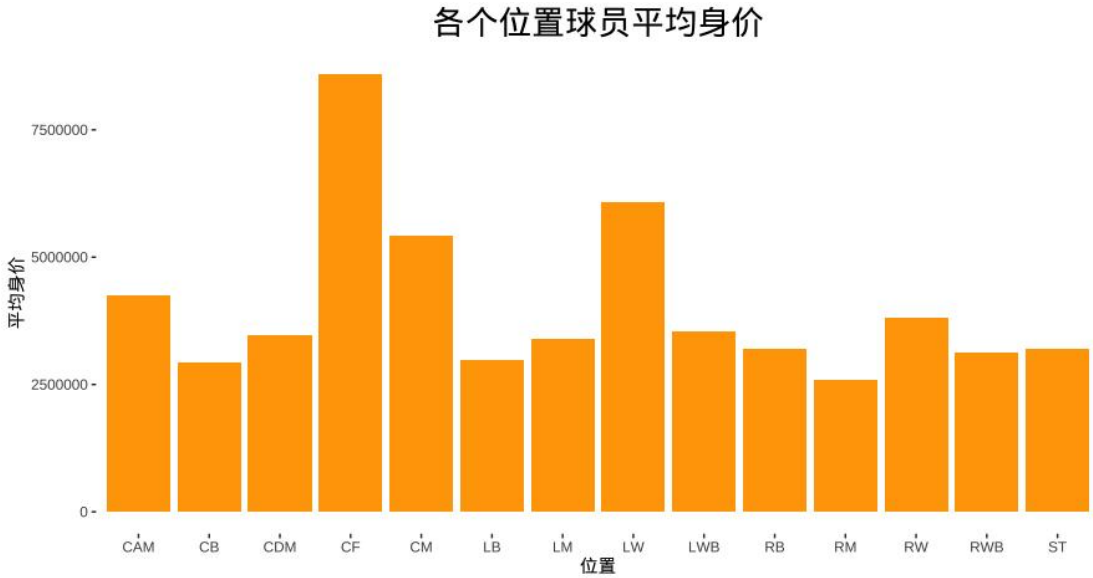


图 10

最后通过绘制各个细分位置上能力值与身价的散点图（图 11）可以发现，对于普通球员来说，其位置对于身价的影响并不明显，但是对于综合能力值在 85 以上的球员来说整体呈现出前场球员高于后场球员的状态，不过差异并不显著与平稳。

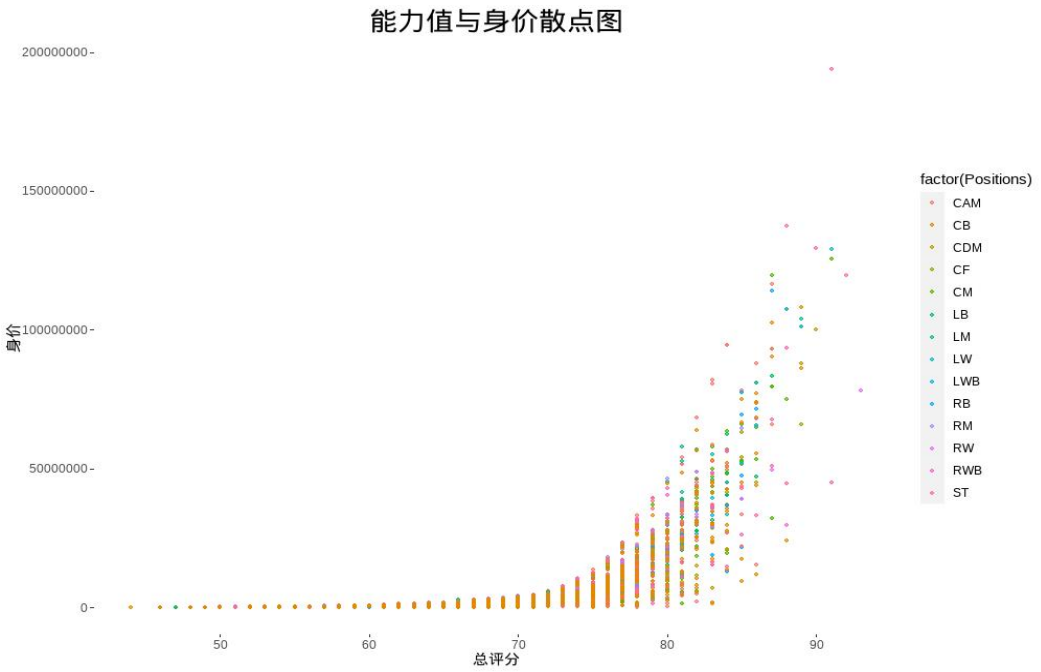


图 11

3、因变量描述分析

根据获取的数据可以发现，球员身价的跨度较大，其分布呈一个极度右偏分布，球员身价的中位数仅为 130 万欧元，而身价最高的姆巴佩达到了 1.94 亿欧元。从数据中可以看出大部分球员的身价低于 500 万欧元；而那些身价高昂的足球天才是足球联赛中的凤毛麟角，他们处于足球金字塔的顶端。为保证回归的合理性，我们决定将球员身价对数化，其对数化后的身价呈大致正态分布，如图 12 所示。本文将分析导致球员身价差异跨度大的因素。

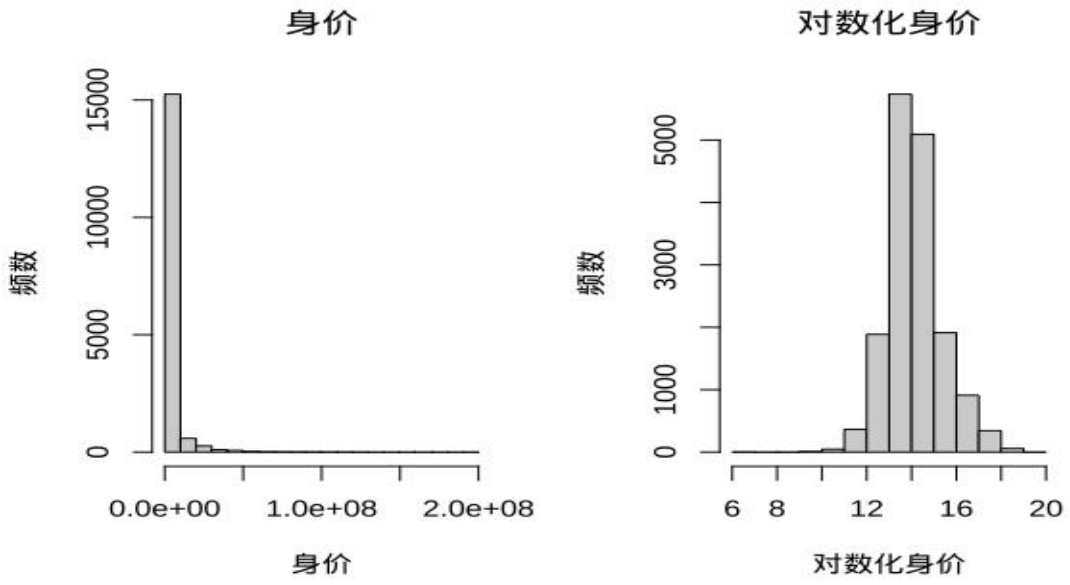


图 12

4、自变量描述分析

对于可能影响球员身价的因素，我们绘制了箱线图进行了初步分析，得到以下结论：

(1)中前场球员身价高于后场球员

由图 13 我们发现不同位置的球员有着不同的身价，其中进攻性位置（例如：中锋 CF）的球员平均身价身价往往高于防守性位置（中后卫 CB）的球员。中锋拥有所有位置中最高的平均身价且身价中位数也最高，中后卫的身价中位数最低。同时我们发现对于顶级球员而言，位置对其身价影响不大。此外，我们还发现位于中场的球员身价中位数高于位于前场和后场的球员，但前场球员身价的下限和上限均高于中后场。

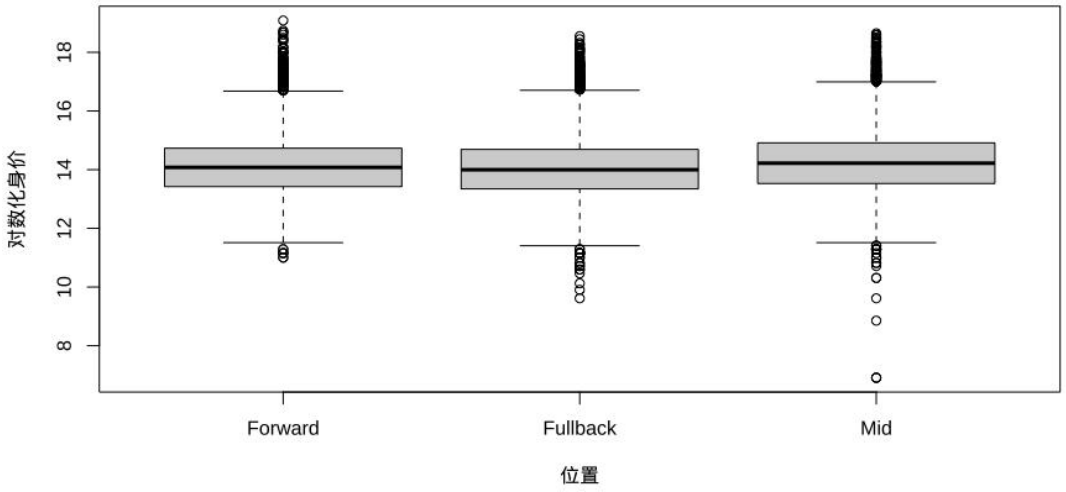


图 13

(2)年轻的球员身价远高于年长的球员

年龄因素也是影响球员身价的一个重要因素。将球员年龄进行划分，从各个年龄段球员身价来看，如图 14 所示，23-28 岁是球员的黄金年龄，在这个阶段的球员平均身价达到了顶峰；而大概将 28 岁作为一个分水岭，在这之前为成长期，球员身价随着年龄的增长而上涨；这以后则开始步入下滑期，球员身价随着年龄增长而下跌。

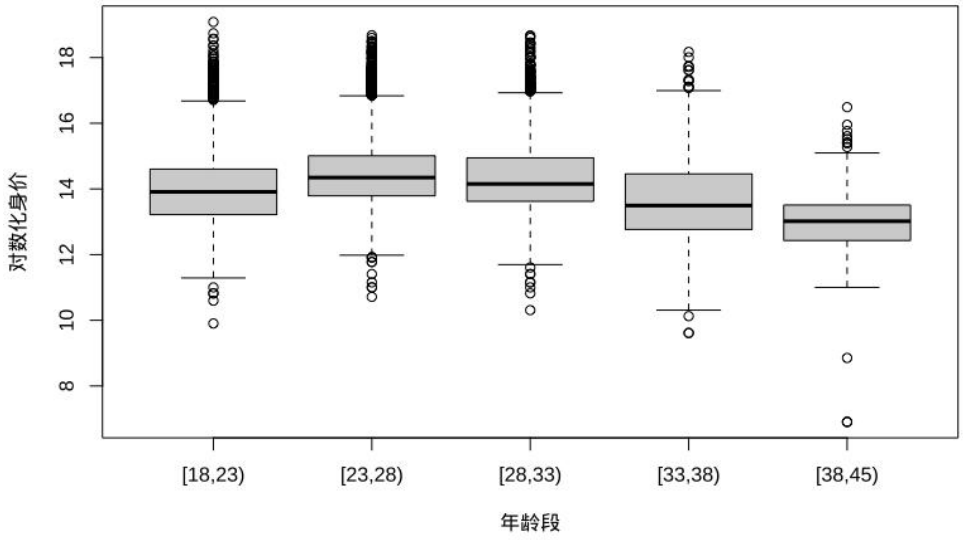


图 14

(3)国际声誉高的球员拥有更高的身价

球员的国际声誉往往关系到他的受欢迎程度，因此也对其身价有着正向的影响。由图可以看出球员的国际声誉与其身价中位数显著正相关，但对于拥有不同国际声誉的球员而言，身价的上限似乎与声誉关系不大。如图 15 所示

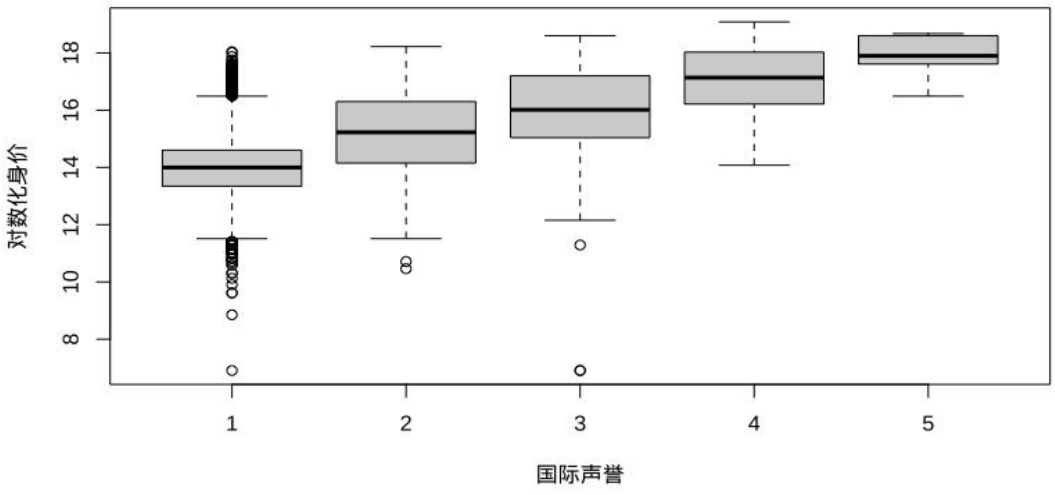


图 15

(4)集体荣誉提升个人价值

球员所效忠的俱乐部对其身价存在一定的影响，我们发现效忠著名球队的球员往往拥有较高的身价，反之则相反。我们对所有的球队进行评分，如图 16 所示，50 分为非足球五大联赛队伍，80 分为五大联赛队伍，100 分为世界知名的一流球队，这些球队在联赛中名列前茅或者打入了本届欧冠，可以看出五大联赛球队尤其是一流球队球员的身价要远高于非五大联赛球员的身价。

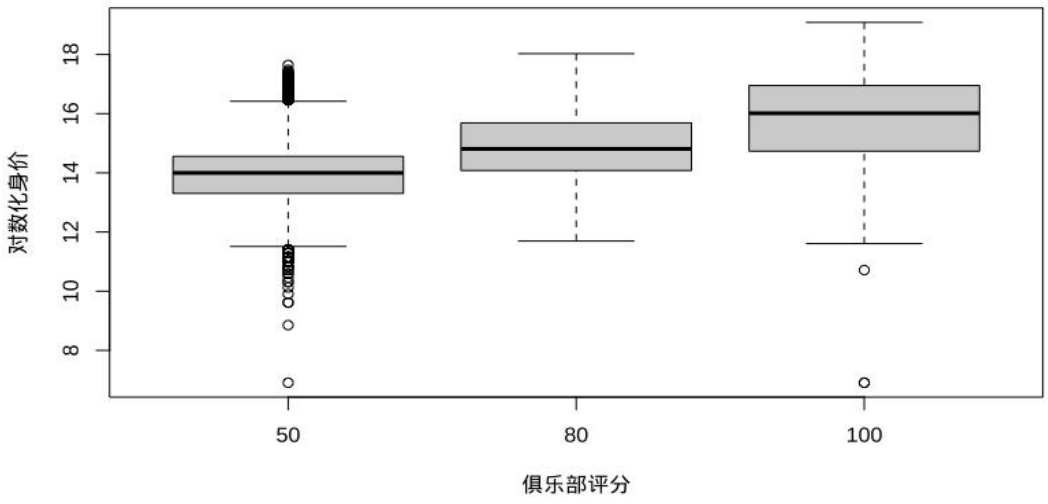


图 16

5、变量处理

（1）正如因变量分析中看到的，球员身价具有很强的后尾性，为保证回归的合理性，我们决定将球员身价对数化，其对数化后的身价呈现大致正态分布。（具体内容参考因变量分析）

（2）由于球员身价受到年龄影响，我们选择 28 岁这一公认的球员巅峰年龄，通过公式 adjust_Age=(Age-28)^2 来构建调整后的年变量。

（3）此外为了统一数据的原始数据中的国际声誉一项与其他项的分布范围，我们对国际声誉数据进行乘 20 处理，此步骤在数据导入前已经完成。

（4）球员的各项指标反应了他们在比赛时的表现，这与他们的身价有关。我们获取了球员进攻、防守两个维度的指标根据相关系数图可以发现，这些变量之间具有较强的相关关系，因此考虑对这些指标进行主成分降维处理。而在这些指标当中，与身价的正相关性较强的指标有射门能力、射门力量、远射能力和点球能力这些指标。

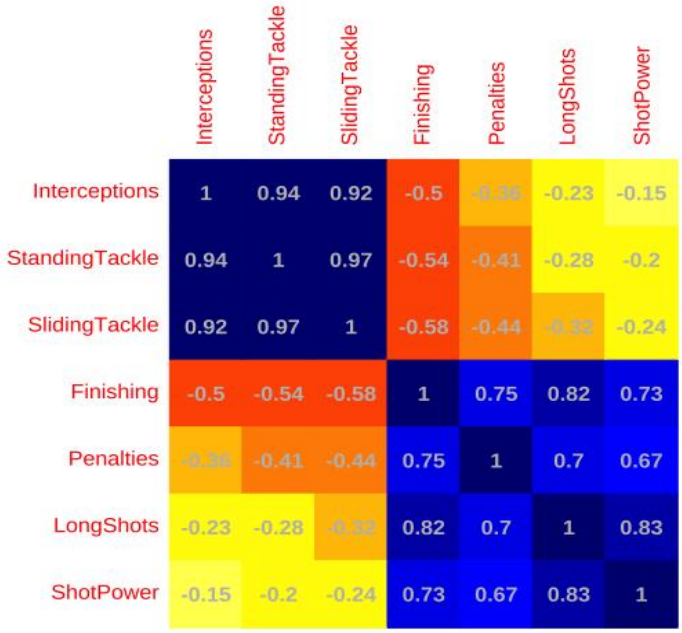


图 17

从图 17 可以看出，射门能力、射门力量、远射能力和点球能力之间存在极强的线性关系，抢断能力、逼抢能力和铲球能力之间存在极强的线性关系，为了减小回归时多重共线性对结果的影响我们绝对采用主成分的方法实现降维。

通过绘制出的碎石图和方差最大正交旋转图（如图 18 和 19 所示）初步分析看出这些变量可以构造两大主成分，分别是包括了门前终结能力，点球，射门力量，远射能力的主成分 1 和包括了拦截，抢断和铲球的主成分 2，这两大主成分在现实分别体现了球员的进攻和防守能力，具有现实意义。

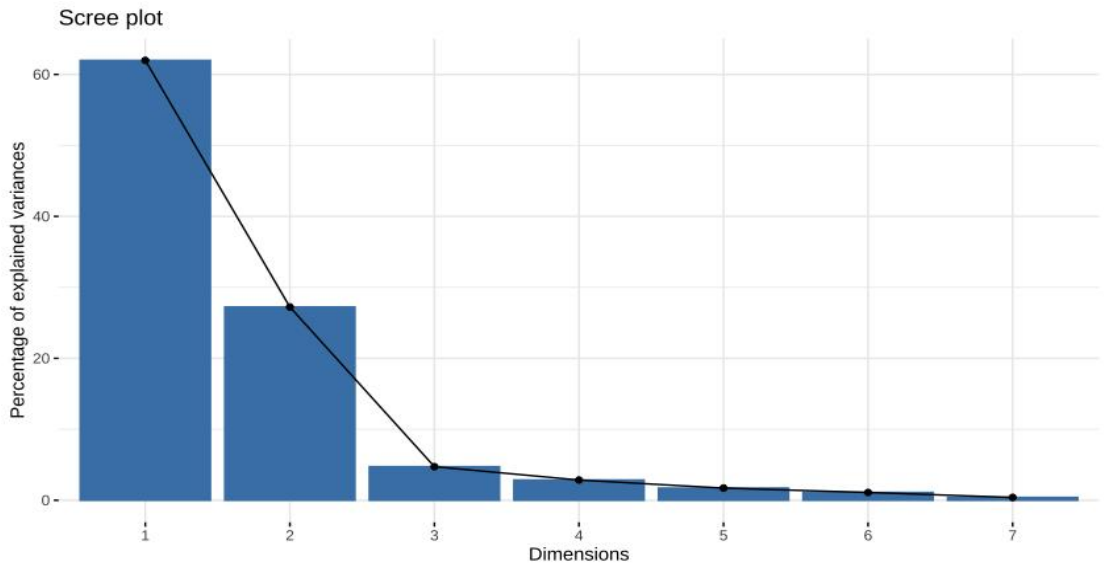


图 18

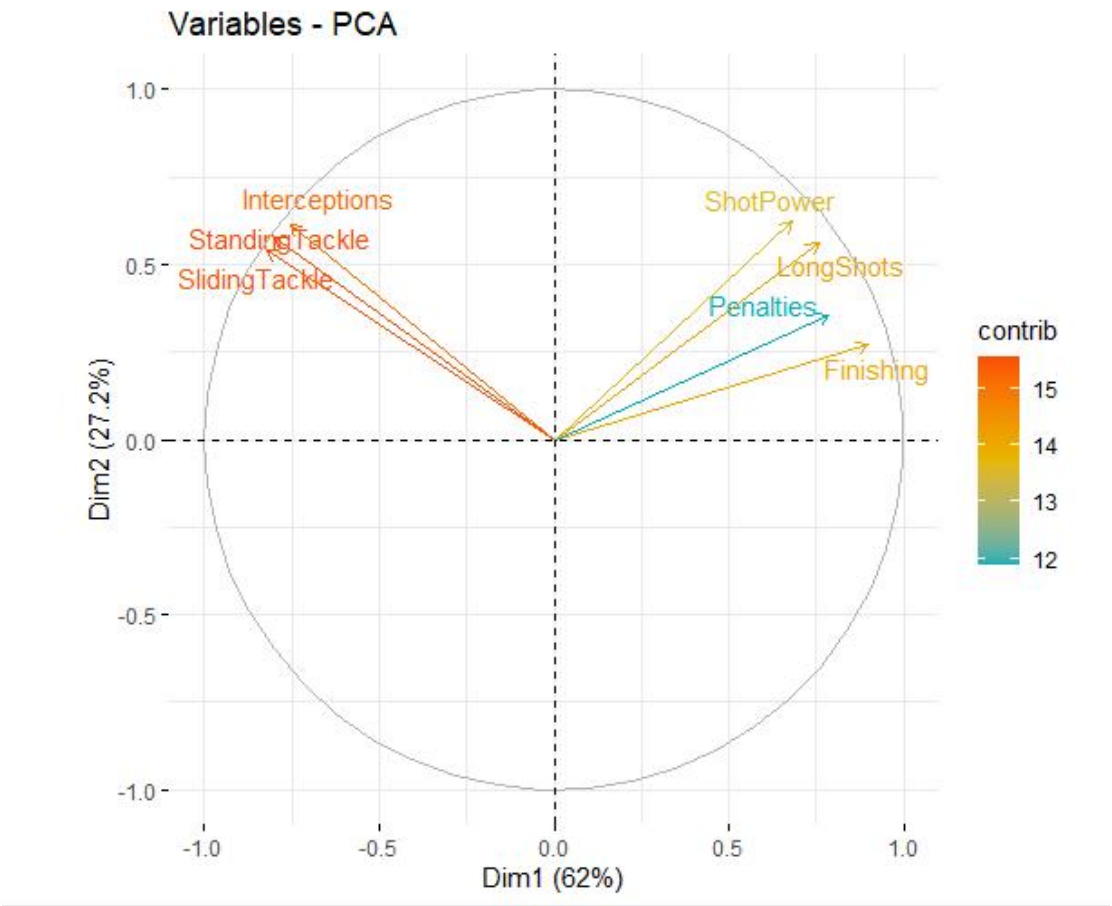


图 19

利用主成分的方法，提取前两个主成分，解释比达到了 89%，得到主成分的因子载荷矩阵如表 2 所示

指标	主成分 1	主成分 2
射门能力	0.431	0.196
射门力量	0.327	0.450
远射能力	0.363	0.407
点球能力	0.377	0.258
抢断能力	-0.362	0.445
逼抢能力	-0.383	0.419
铲球能力	-0.394	0.392

表 2

主成分 1 主要包括了射门能力、远射能力和点球能力等指标，在上面的叙述中我们已经将它命名为球员进攻性因子；主成分 2 主要包括抢断和逼抢能力等指标，我们也已经将它命名为球员防守性因子。下图展示球员身价随着主成分得分上升的均值柱状图。自然，能力越高的球员身价越高，球员身价随着主成分 1 和主成分 2 的增加而增加，且身价随主成分 1 上升的趋势非线性，主成分 1 得分越高，身价上涨越快。同时我们还发现主成分 1 在为负的时候，平均身价随主成分 1 的增加而减少，我们认为这与主成分 1 存在负系数有关。具体结果如图 20 所示。

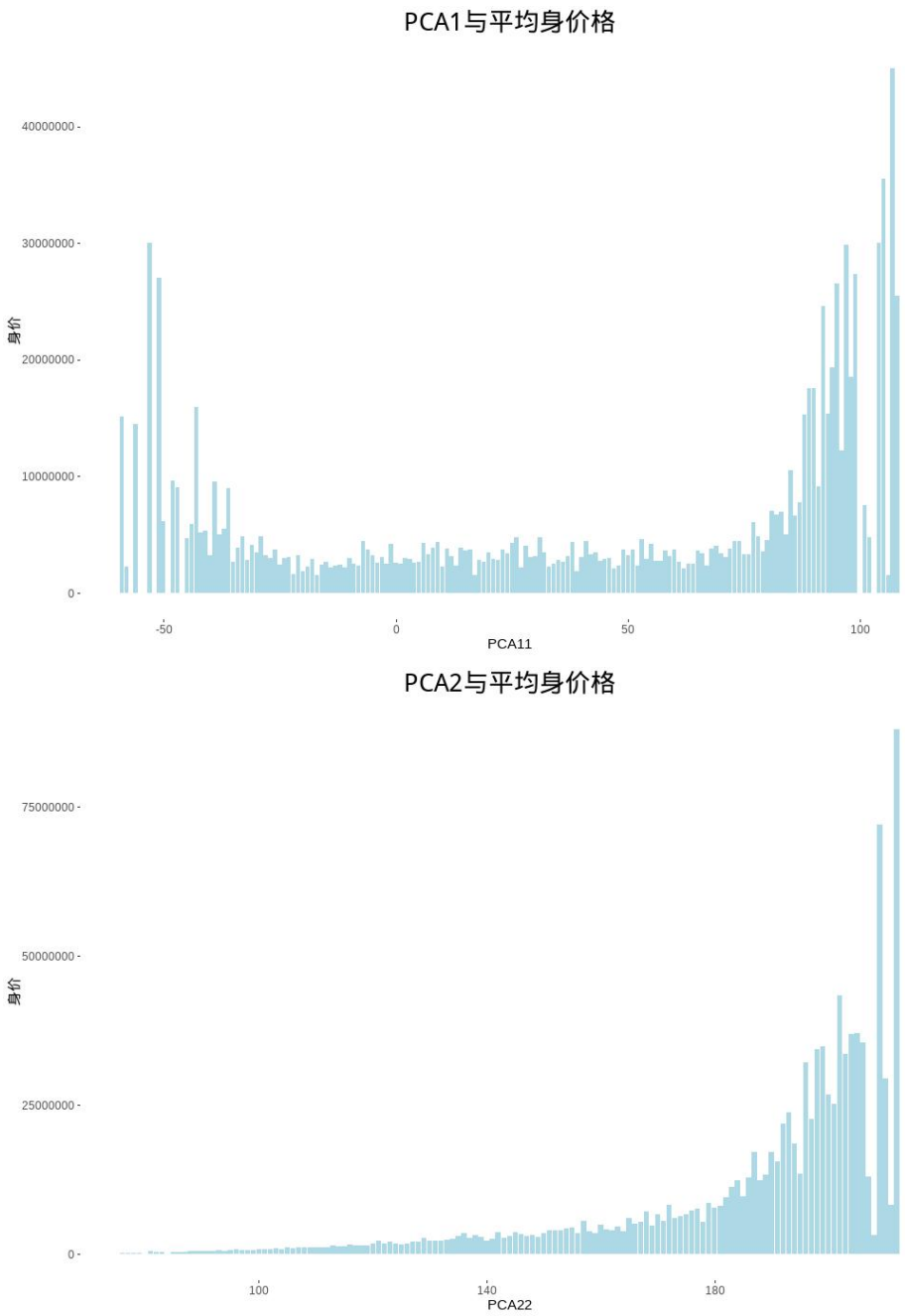


图 20

图 21 展示了各位置球员的两个主成分的分布的关系。可以看出，对于前场球员，球员 PCA1 即进攻因子的得分分布明显高于中场球员，中场球员又明显高于后场球员（从图中可以看出，前场球员 Forward 的最高点的 PCA1 得分靠右，而中场 Mid 球员位于中间，后场球员 Fullback 的 PCA1 得分位于最左边），这与现实中各位置球员的进攻能力相符合。而对于防守因子 PCA2 而言，前场球员则位于最左边，中场球员与后场球员不分上下，这也是与 FIFA 中将部分多位置球员的推荐位置设置为中场所致，其中就不乏在俱乐部与国家队的优秀后卫球员，但总的的来说，球员的综合能力因子还是与现实相符合，具有较强的现实意义。

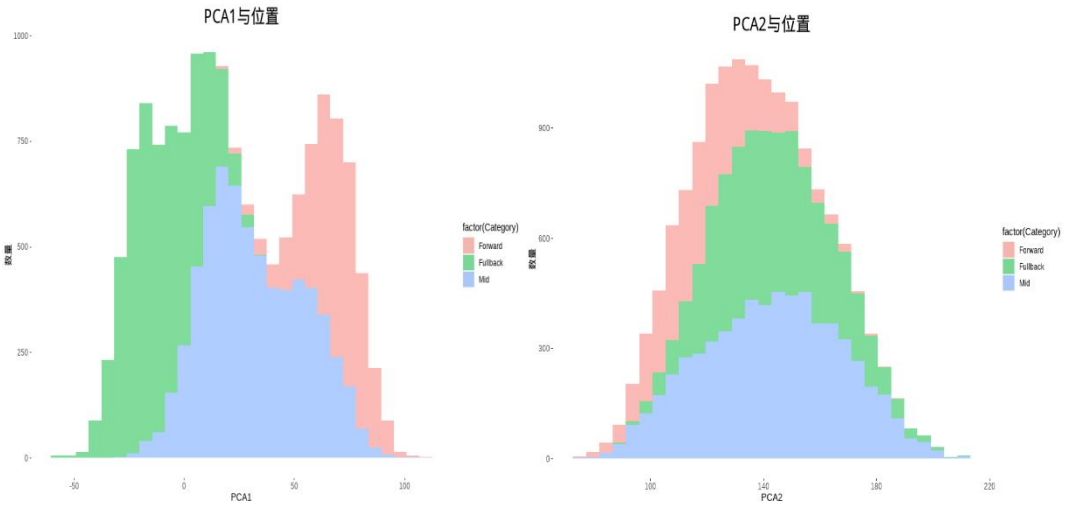


图 21

接下来我们将进一步建立回归模型，衡量这些因素的影响效果，为球队甄别哪些球员可能是水货，而哪些球员拥有巨大的潜力，为球队转会操作提供建设性意见。

四、线性回归

1.全模型分析

我们将计算得到的两个主成分 PCA1、PCA2 与球员的国际声誉、年龄、所在俱乐部、所属国籍共计六个因素作为回归自变量，对因变量——球员的身价进行回归建模，希望以此找出球员身价与这六个因素之间存在的内在关系。但根据对于数据的描述性分析，我们有必要对部分变量进行如下调整，保证回归模型更加准确：

（1）球员身价对数化。原因在于球员身价分布呈现出极度右偏分布，直接建立回归模型会导致结果的不准确。

（2）将球员年龄进行调整，用（实际年龄-28）^{2}来代替年龄。进行如此变换的原因也易于理解，正如箱线图所表达的，球员的年龄对于身价的影响并不是一致的，28 岁为球员年龄的分水岭，在 28 岁前身价会随年龄递增，28 岁后则随年龄递减。具体如表 3 所示

变量名	参数估计值	P 值	显著水平
Intercept	9.899	0	***
Adjust_age	-0.006	0	***
Nationality_score	0.001	0.00142	**
club_score	0.024	0	***
IntReputation	0.018	0	***
PCA1	0.007	0	***
PCA2	0.018	0	***
F 检验	P<0.001		
判决系数(R-square) 0.4489	调整后的判决系数(R-square)0.4487		

表 3

全模型的 F 检验拒绝了原假设 H_0，即构建的模型中至少存在一个自变量，使得其参数估计不为 0，可见建立的全模型是显著的，全模型调整后 R^2 达到 0.4487，模型的拟合程度

较好。各自变量的方差膨胀因子值如表 4 所示

方差膨胀因子 VIF					
adjust_Age	Nationality_score	club_score	IntReputation	PCA1	PCA2
1.20	1.02	1.20	1.31	1.19	1.57

表 4

各自变量方差膨胀因子均为 1 附近的值，并没有出现远离 1 的情况，可见自变量之间没有出现多重共线性的情况。

2.模型诊断

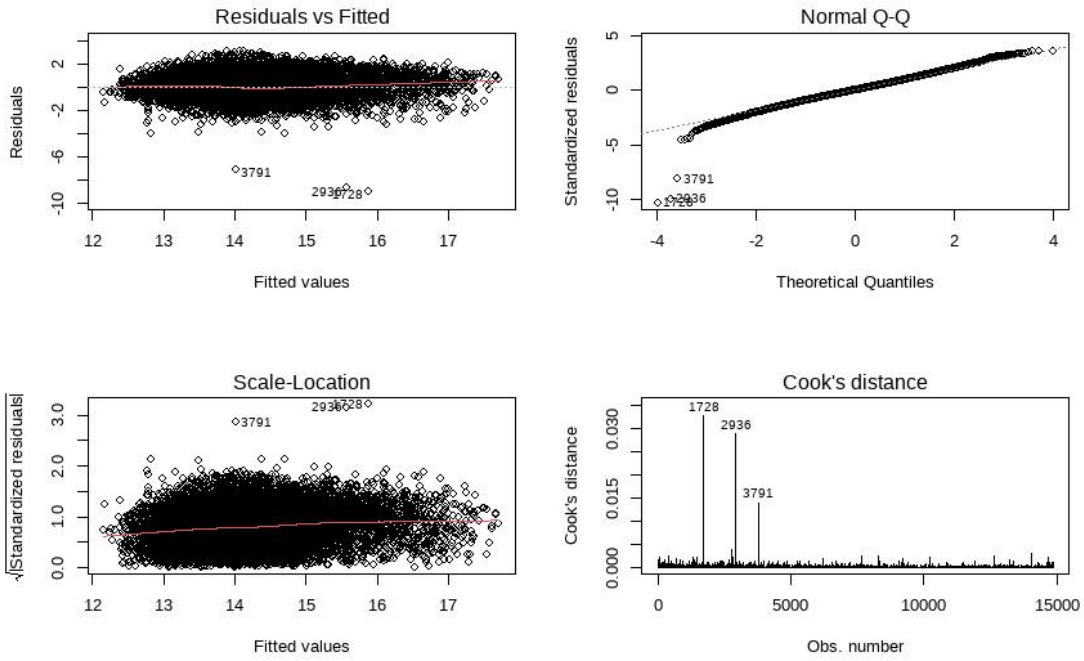


图 22

通过初步观察图 22 四张图我们可以发现，1728、2936、3791 三个数据在四张图中均为离群值，我们认为这三个数据是异常值数据。需要对异常值数据进行剔除。

在删除了 1728、2936、3791 三个异常值数据后，我们重新建立了回归模型，各自变量的参数如表 5 所示，图 22 为删除异常值后的模型诊断。

变量名	参数估计值	P 值	显著水平
Intercept	9.866	0	***
Adjust_age	-0.006	0	***
Nationality_score	0.001	0.00142	**
club_score	0.024	0	***
IntReputation	0.018	0	***
PCA1	0.007	0	***
PCA2	0.018	0	***

F 检验	P<0.001
判决系数(R-square) 0.4548	调整后的判决系数(R-square)0.4546

表 5

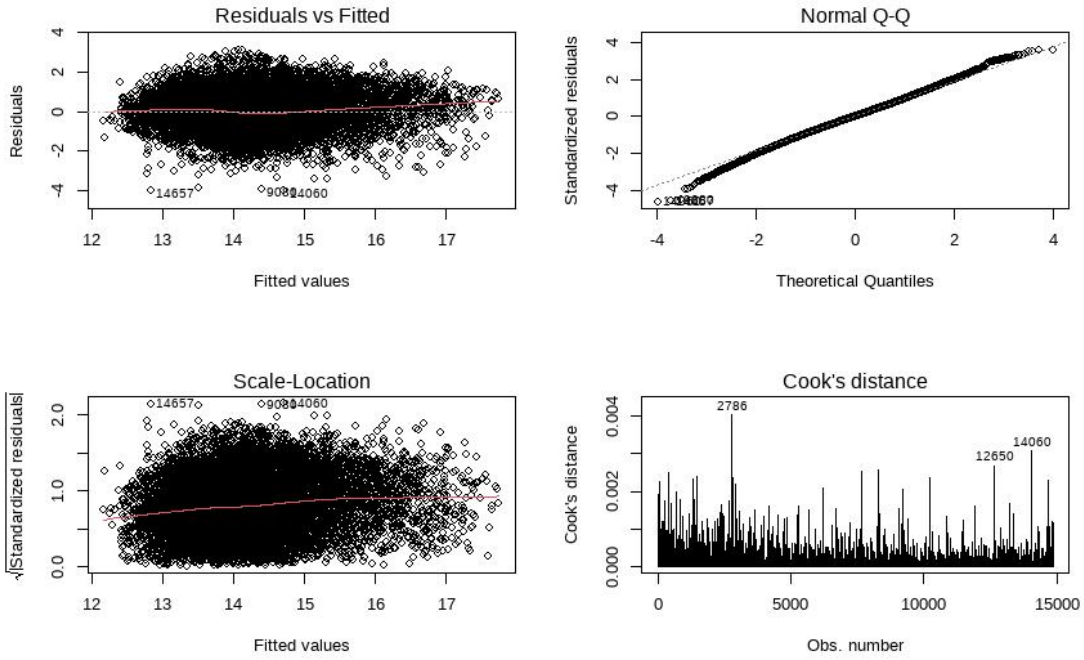


图 23

（1）残差的独立性检验

对于残差的独立性检验，我们着眼于图 23 中左上角的图，即 Residual vs fitted 图。其中取 fitted values 为横坐标 x，residual 为纵坐标 y。可以看出图中的点较为均匀的分布于 y=0 的上下两侧，并没有呈现出一侧点多于另一侧且没有呈现出某种有序的趋势，因此我们认为残差之间是相互独立的。

（2）残差的正态性检验

对于残差的正态性检验，我们着眼于图 23 中右上角的图，即 Normal Q-Q 图。其中取理论分位数 Theoretical Quantiles 为横坐标 x，标准化残差 standardize residual 为纵坐标 y。可以看到图中的点近似位于一条直线附近，说明数据满足正态性。

（3）同方差检验

对于数据的同方差检验，我们着眼于图 23 中左下角的图，即 Scale-Location 图。其中取 Fitted values 为横坐标 x，标准化残差 standardize residual 的平方根 $\sqrt{\text{standardize residual}}$ 为纵坐标 y。可以看到数据并没有呈现出某种趋势，而是均匀的分布于近似水平的直线的两端，这也说明了数据满足同方差的假设。

（4）cock 距离检验

对于 cock 距离检验，我们着眼于图 23 中右下角的图，即 Cock’s distance 图。其中取数据序号 Obs.number 为横坐标 x，Cock 距离为纵坐标 y。可以看到图中数据并没有大量的高杠杆值点，因此我们认为数据通过 cock 距离检验。

3.模型选择

基于全模型的回归结果我们不难发现，用于回归的自变量对球员的身价都有较为显著的影响，但是否进行回归自变量的选择仍是一个完整的回归模型构建过程中必不可少的一环。因此我们仍对全模型进行基于 AIC（表 7）与 BIC 准则（表 8）的回归自变量选择，查看应用 AIC 准则或 BIC 准则后，选出的自变量是否会与全模型有所不同。

AIC 准则

变量名	参数估计值	P 值	显著水平
Intercept	9.866	0	***
Adjust_age	-0.006	0	***
Nationality_score	0.001	0.00142	**
club_score	0.024	0	***
IntReputation	0.018	0	***
PCA1	0.007	0	***
PCA2	0.018	0	***
F 检验 P<0.001			
判决系数(R-square) 0.4548 调整后的判决系数(R-square)0.4546			

表 7

BIC 准则

变量名	参数估计值	P 值	显著水平
Intercept	9.899	0	***
Adjust_age	-0.006	0	***
Nationality_score	0.001	0.00142	**
club_score	0.024	0	***
IntReputation	0.018	0	***
PCA1	0.007	0	***
PCA2	0.018	0	***
F 检验 P<0.001			
判决系数(R-square) 0.4544 调整后的判决系数(R-square)0.4543			

表 8

通过表 7 与表 8 可见，全模型与 AIC 模型在本项目中完全相同，而 BIC 模型对于自变量参数的估计值与全模型、AIC 模型完全相同，仅仅在截距项的估计上存在很小偏差。综合考虑后我们决定选用 AIC 模型作为本项目的最终模型。且在自变量选择过程中我们发现，本项目在对球员身价分析过程中，所选择的自变量对于球员身价均具有显著性影响，无需剔除任何自变量。

五、模型解读

在确定了本项目的最终模型为 AIC 准则选取的模型后，我们需要对模型进行分析解读。

首先图 24 呈现了 AIC 模型回归系数的条形图。

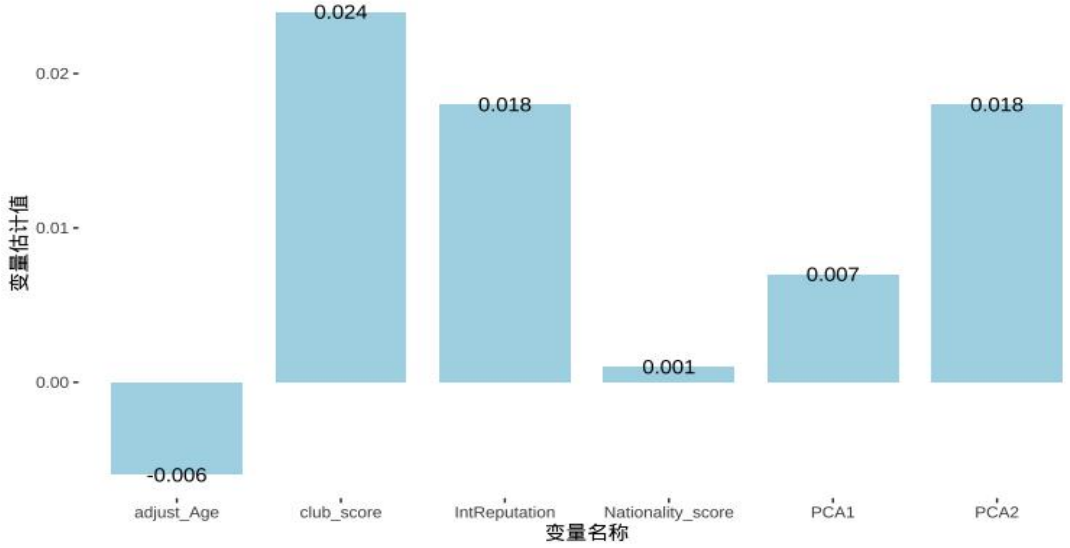


图 24

1.正如实际所表现出的，28 岁为足球运动员的巅峰年龄，在 28 岁之前身价一般会随着年龄的增加而升高，而过了 28 岁之后，身价会随年龄的增加而降低。因此当我们用 adjust_Age=(实际年龄-28)² 来表示年龄的影响时，年龄距离 28 岁越近，adjust_Age 越小，球员身价会越高，因此 adjust_Age 的系数应当为负值，在回归结果中 adjust_Age 的系数确实为-0.006，与实际预期相符合。且 adjust_Age 偏离原点 1 个单位，球员对数身价下降 0.6%。

2.球员所在的俱乐部往往会很大程度上影响到球员的身价，一方面优秀的俱乐部往往有较强的购买力，可以负担更高身价的球员；另一方面在这些俱乐部球员有更多的成长空间与展示自己的机会，身价自然容易水涨船高。俱乐部评分每提高 1 个单位，球员对数身价能够增加 2.4%。

3.球员的国际声誉 InReputation 对球员的身价存在正向的影响，其内在原因在于当球员具有较高的国际影响力时，其在场外能为球队带来更多的商业价值，因此该球员的身价也会因为这个原因得到提高。在回归中我们发现，国际声誉不仅对球员身价有显著影响，而且影响相对较大，国际声誉每提高 1 个单位，球员对数身价能够增加 1.8%。

4.受联赛户口规则和球员出身国家在足球世界知名度与地位的影响，球员的身价往往会不同。现实中英超联赛对本土球员总是愿意支付更高的价格，此外一些球探也更愿意去巴西，英格兰，意大利，德国，阿根廷等等传统足球强国去挖掘球员，不过可以看到与其他因素相比国籍这一因素的影响并不明显，一是因为传统的足球强国本身数量就不多，二是因为这些国家本身足球球员的数量更多，右偏更明显，三是因为这些足球强国本身的国情和经济情况差异较大，这些因素都制约着国籍这一变量在影响球员身价时发挥作用。

5.PCA1 是我们之前定义的进攻能力因子，一般来说球员的进攻能力越强，球员的身价也会越高。在回归中该因子对于身价影响显著，且进攻能力因子每提高 1 个单位，球员对数升价提高 0.7%。

6.PCA2 是我们之前定义的防守能力因子，一般来说球员的进攻能力越强，球员的身价也会越高。在回归中该因子对于身价影响显著，且发觉感受能力因子每提高 1 个单位，球员

对数升价提高 1.8%。可见对于大多数球员，提升自身的防守能力会显得更加重要一些。这一点在现实中也有所体现，因为相比于提高进攻，更需要球员防守做好才能保证球队不输球，这也是现代足球教练强调的理念。

六、预测

在建立了 AIC 模型之后，我们对一万多名球员的身价进行了预测，其身价分布如图 25 所示

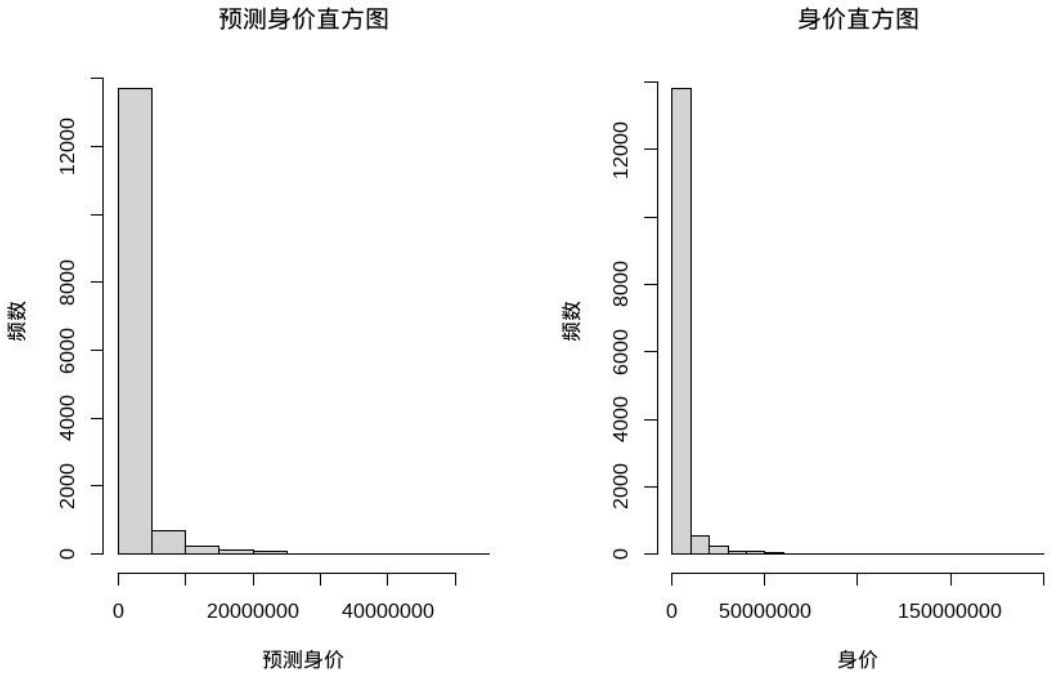


图 25

可以发现球员的身价仍然是呈现出一个极右分布，与实际相符合，但是预测的结果和原来相比总体上相对偏低了一些，但数据更为集中，这正是因为现实中大部分球员并不是金字塔顶端的顶级球员，本项目建立的模型是用于分析大部分球员的身价情况，因此例如姆巴佩这类较为极端的情况会被得到中和。

同时作为本项目最重要的结果，我们希望通过建立的 AIC 模型绘制出各个位置中身价最具有潜力（也就是预计身价与现实身价差值较大）一些球员。以下是我们基于 AIC 模型绘制出的前场、中场、后场身价最具有潜力的球员。

前场最具潜力的球员（图 26）

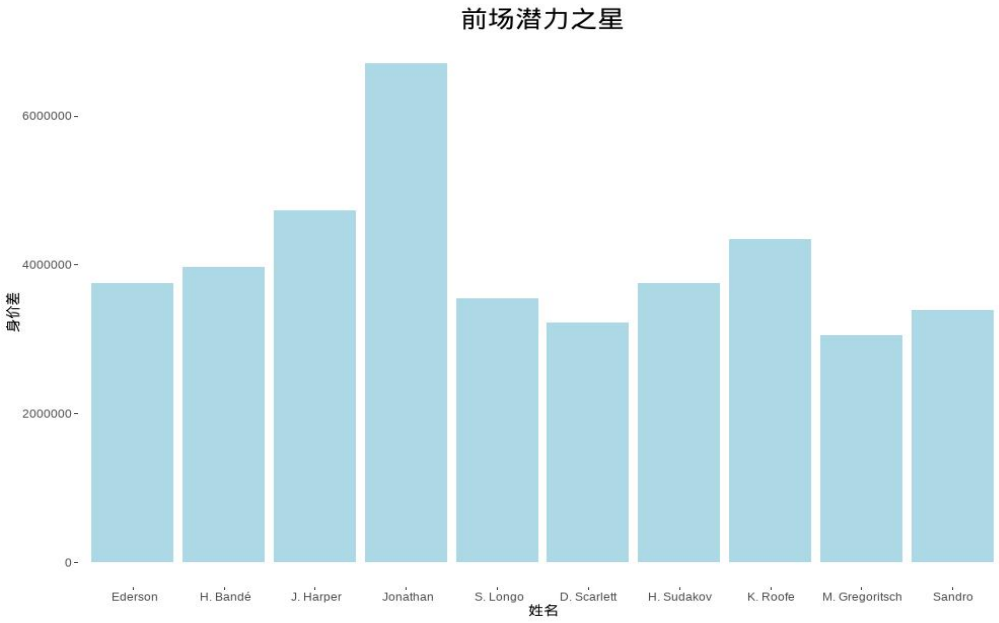


图 26

在前场球员中，预测中的潜力球员之一例如奥地利的米夏埃尔·格雷戈里奇（M.Gregoritsch），其目前效力于德甲球队弗赖堡，该球员在 8 月开始的本赛季德甲联赛中就稳坐首发前锋位置，贡献了 15 场 6 球 3 助的数据，其所在的球队目前在德甲也排名第 2，仅次于德甲传统豪门拜仁，但是在此之前改球员表现相对并没有那么亮眼，基本很少受到球探等关注，因此可以看到我们构建的平价体系是可以做到预测高潜力球员的，该球员本赛季发挥出了原超身价的水平，兑现了自己的天赋。

中场最具潜力的球员（图 27）

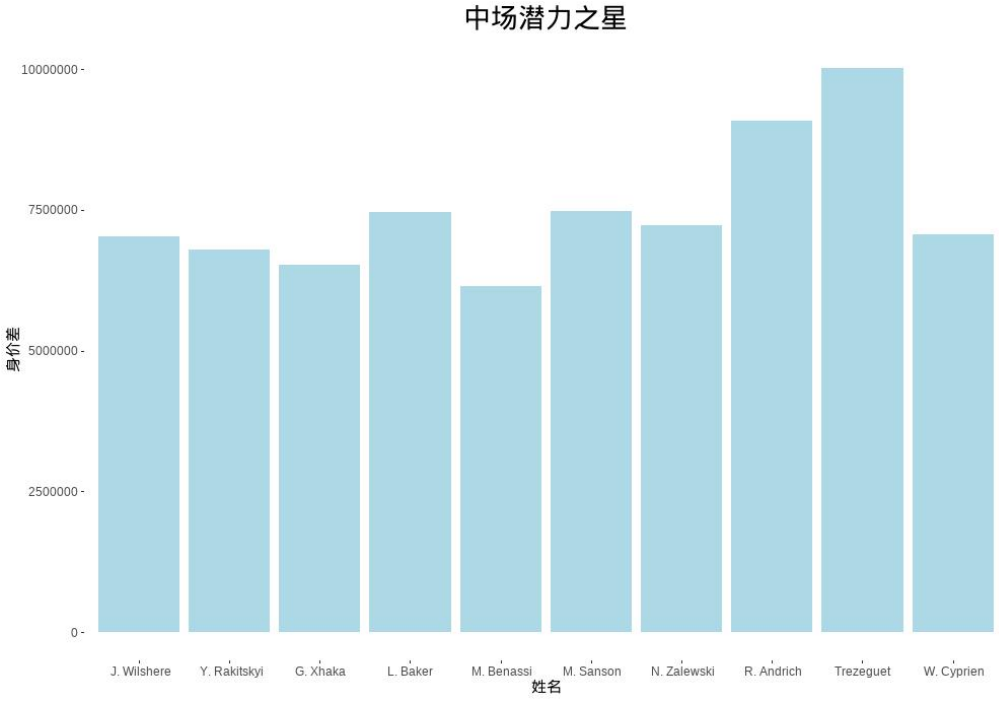


图 27

而在中场的预测球员中，扎卡（G.Xhaka）目前效力于本赛季英超排名第一传统豪门阿

森纳，其同时也是瑞士国家队的核心中场，在上一赛季 27 场只有 1 球 2 助的前提下，本赛季仅 14 场就贡献了 3 球 3 助，场均 1.4 个关键传球，其表现肉眼可见的提升，状态火热，尽管我们使用的 FIFA22 是近两年前的相关数据，当时其场上表现并不尽如人意，但是通过对球员本身能力的评估依旧成功寻找到了这一球员，可见在当时其身价的确被低估。

后场最具潜力的球员（图 28）

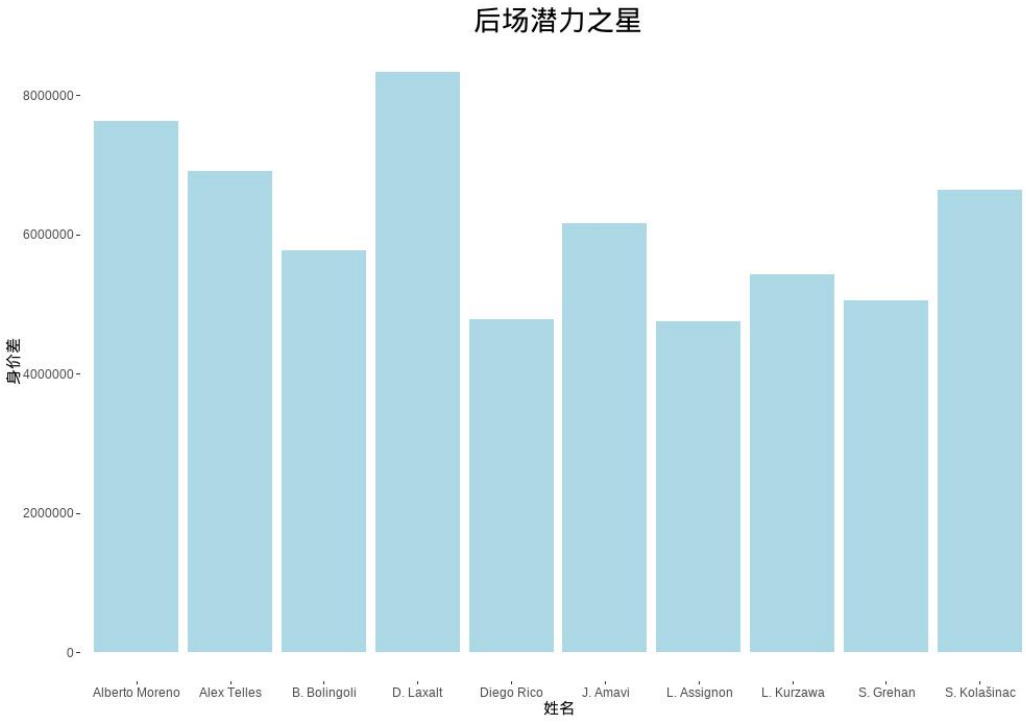


图 28

在后场球员的预测中，我们发现了一名来自爱尔兰超级联赛的 17 岁小将肖恩·格雷汉（S.Grehan），未来可以进一步观察这名球员的状况，看其是否有机会进入五大联赛。预测球员中也包括目前法甲排名第四的球队——马赛的后卫科拉希纳茨（L.Kurzawa），其在本赛季至今场均能贡献出 1.2 次拦截，首发六次作为后卫便贡献了一球一助，表现较上赛季也有很大提升。

除了我们发现的这些球员外，我们还可以通过设定更多限制来有针对性的挖掘更多具有潜力的球员，上面的展示可以初步证明我们小组模型的有效性，这些球员有很多本赛季以及贡献出了不错的表现，并且很多拥有这些潜力球员的俱乐部本赛季在联赛中也有着相当的进步。

足球不只有俱乐部，还有国家队。我们还可以通过所建立的模型来预测 23 人国家队的累计身价前十名（如图 29 和 30 所示），看看哪些国家的配置在身价层面最为豪华。分别给出现实中累计身价前十名的国家与累计预测身价前十名的国家的直方图。

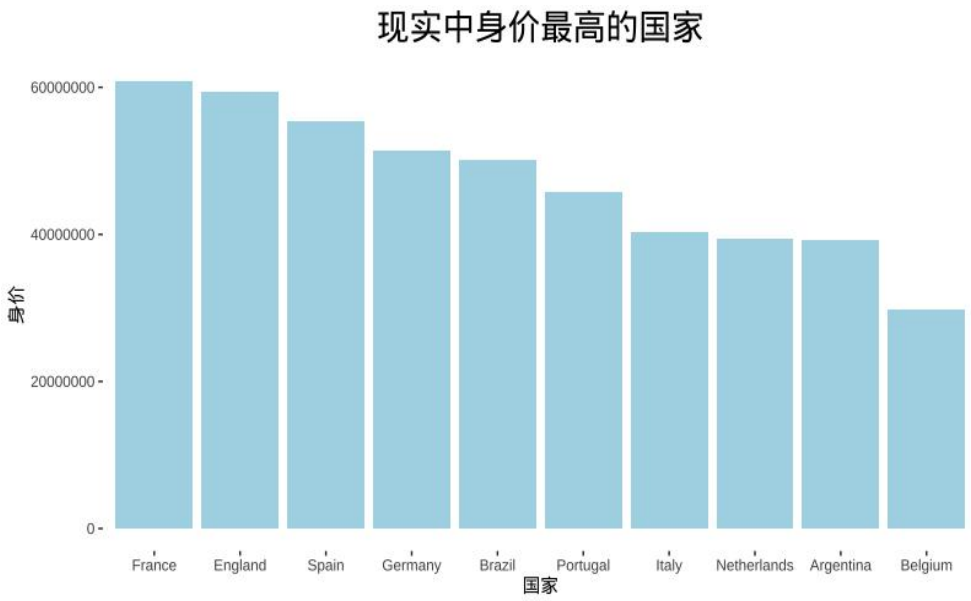


图 29

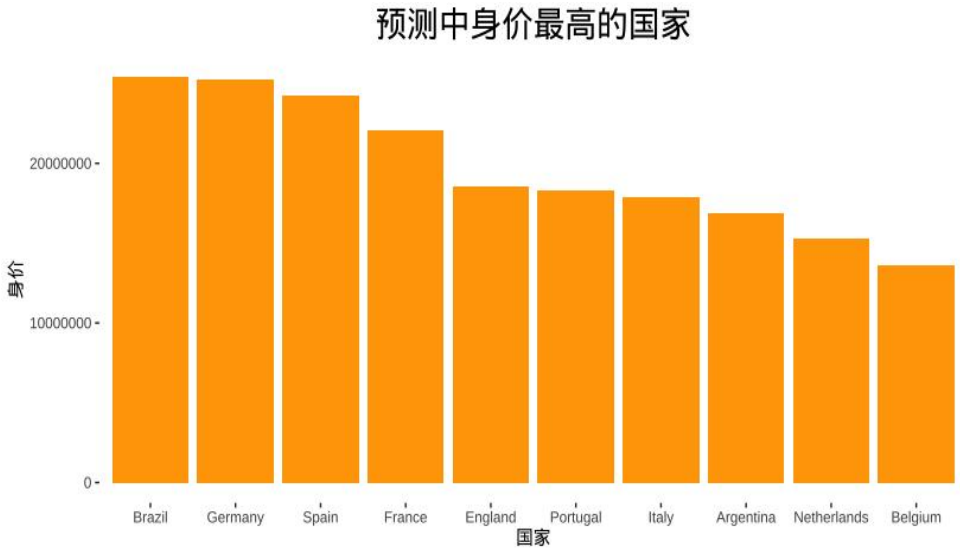


图 30

可以看到，预测身价与现实身价排名前十名的国家均为巴西、德国、西班牙、法国、英格兰、葡萄牙、意大利、阿根廷、荷兰、比利时，只是排名方面存在一定差异，这很大一部分是当今足球部分球员的溢价严重的现象导致的。这一点体现出对于高价值球员来讲国籍因素对身价具有较大的贡献，但是前面的回归过程也说明了对于所有球员来看国籍这一因素影响并不显著，譬如一些顶尖的英国球员可能就受到其“户口本”在英超产生了溢价，导致其国家队的预测身价较真实值有所下降，但是大多数英国球员并没有机会进入英超效力，其国籍溢价自然不明显；此外也可以看出通过回归预测后阿根廷和法国的身价排名差距显著缩小了一倍，也为我们在世界杯决赛前看好阿根廷提供了数据支撑。

七、总结与不足

从上述的分析结果可知，足球运动员的身价确实如我们在项目开展之前所预料的，受到

包括年龄、进攻因子、防守因子在内的很多因素影响，本项目也通过多元回归分析的理论知识与 R 语言成功构建出了球员身价与预期的各个相关变量之间的多元回归模型，模型顺利通过了各种统计检验且预测的自变量对身价均有显著影响，自变量的回归系数也符合现实预期，能够很好的解释该自变量对于球员身价的影响（具体解释参考模型解读部分）。

项目构建出的模型具有很强的现实意义。针对球员方面，球员可以通过本模型了解到提升自己身价的主要方向并在相关方面加强训练、提高，以此实现身价的提升；对于俱乐部而言，俱乐部管理者可以通过本模型来寻找身价存在潜力的球员，无论是对俱乐部转手球员获得差额资金或是培养新的优秀球员都具有很强的指导意义。

虽然模型具有较强的现实指导意义，能够应用于大部分球员的分析结果，但是我们小组成员在实际应用分析的过程中仍然发现了模型存在的不足。这并不是由于模型的构建或者检验存在问题，而是一种必然的，内生的问题。模型的主要不足点就是在对身价远高于普通球员的超级巨星（例如哈兰德、姆巴佩等）进行身价预测时，预测的结果往往会比实际结果低出很多，我们认为这是自然的，首先是因为这些球员本身的溢价现象就十分严重，例如巨星姆巴佩一人的身价就可以抵上两三支足球队，这是近年来足球领域出现大量溢价现象的一个缩影，超级巨星的溢价现象是当代足坛出现的一大问题；其次就是能力值边际效益递减的缘故，因为大部分职业足球运动员的能力值都会在 70-80 之间，但是能力值 90 以上的少之又少，往往几百名职业球员中才会有一名球员，这就导致了其实当一名球员的能力值达到 90 以上时，该能力对于身价解释的权重应当要适量增加，因为能达到 90 以上代表着他在这领域是百里挑一的水平，因此 90 分以上的能力值理应对球员的身价有大幅度的提升，但在我们建立的多元回归分析模型中并没有考虑到这一因素，没有很好地解决这个问题，由此导致了对于一些超级巨星的身价预测值往往比实际值低的情况。

模型第二个不足点是在于能够影响球员身价的因素实在太多了，即使我们尽可能多的考虑对球员身价存在影响的变量，对很多可能存在相关性的变量进行了主成分处理，我们仍无法做到将所有可能影响球员身价的自变量都一一考虑进来，这就导致了模型的判决系数 R^2 并没有预期的那么大，也算是模型存在的另一个小小的不足。

总的来说，我们还是认为项目构建出的针对足球运动员身价的多元回归预测模型还是满足线性回归理论并且具有很强的现实指导意义的。