计算机科学
COMPUTER SCIENCE

# 融合XGBoost与SHAP模型的足球运动员身价预测及特征分析方法

廖彬，王志宁，李敏，孙瑞娜

引用本文

廖彬，王志宁，李敏，孙瑞娜.融合XGBoost与SHAP模型的足球运动员身价预测及特征分析方法[J]．计算机科学, 2022, 49(12): 195-204.

LIAO Bin, WANG Zhi-ning, LI Min, SUN Rui-na. Integrating XGBoost and SHAP Model for Football Player Value Prediction and Characteristic Analysis [J]. Computer Science, 2022, 49(12): 195-204.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600029

# 融合 XGBoost 与 SHAP 模型的足球运动员身价预测及特征分析方法

廖 彬[1] 王志宁[2] 李 敏[2] 孙瑞娜[2,3,4]
1 贵州财经大学大数据统计学院 贵阳 550025
2 新疆财经大学统计与数据科学学院 乌鲁木齐 830012
3 中国科学院信息工程研究所 北京 100093
4 中国科学院大学网络空间安全学院 北京 100049
(liaobin665@163.com)

**摘 要** 随着足球运动全球化程度的不断提升,全球转会市场愈发庞大,然而针对影响转会交易最关键的因素球员身价的深入模型及应用研究还较为缺乏。以国际足球联合会 FIFA 的官方球员数据库为研究对象,首先,在区分不同球员位置的前提下,运用 Box-Cox 变换、F-Score 特征选择等方法对原始数据集进行特征处理;其次,通过 XGBoost 构建球员身价预测模型,并与 Random Forest,Adaboost,GBDT,SVR 等主流机器学习算法进行 10 折交叉验证实验对比,证明了 XGBoost 模型在 $R^2$,*MAE*,*RMSE* 这 3 项指标上的性能优势;最后,在身价预测模型的基础上,融合 SHAP 框架分析不同位置影响球员身价的重要因素,为球员身价评估、身价对比分析、球员训练策略制定等场景提供决策支持。

**关键词**:机器学习;球员身价预测;训练策略;XGBoost 算法;SHAP 值
**中图法分类号** TP391

## Integrating XGBoost and SHAP Model for Football Player Value Prediction and Characteristic Analysis

LIAO Bin[1],WANG Zhi-ning[2],LI Min[2] and SUN Rui-na[2,3,4]
1 College of Big Data Statistics,Guizhou University of Finance and Economics,Guiyang 550025,China
2 College of Statistics and Data Science,Xinjiang University of Finance and Economics,Urumqi 830012,China
3 Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100093,China
4 School of Networks Security,University of Chinese Academy of Sciences,Beijing 100049,China

**Abstract** With the increasing globalization of football,the global player transfer market is becoming more and more prosperous. However,as the most important factor affecting player transfer transaction,the player's transfer value lacks in-depth model and application research. In this paper,the FIFA's official player database is taken as the research object. Firstly,on the premise of distinguishing different player positions,Box-Cox transformation,F-Score feature selection,etc. are used to perform feature processing on the original data set. Secondly,the player value prediction model is constructed by XGBoost,and compared with the main machine learning algorithms such as random forest,AdaBoost,GBDT and SVR for 10-fold cross validation experiments. Experimental results prove that the XGBoost model has a performance advantage over the existing models on the indicators of $R^2$,*MAE* and *RMSE*. Finally,on the basis of constructing the value prediction model,this paper integrates the SHAP framework to analyze the important factors affecting the players' value score in different positions,and provides decision support for some scenarios,such as player's value score evaluation,comparative analysis,and training strategy formulation,etc.

**Keywords** Machine learning,Player's value prediction,Training strategy,XGBoost algorithm,SHAP value

## 1 引言

自 1863 年 10 月 26 日世界上第一个正式足球组织在英国伦敦成立以来,足球运动以惊人的发展速度遍布全世界,激烈对抗的竞争性、令人叹为观止的观赏性及默契配合的灵巧性,让足球运动在世界范围内收获了巨大的关注与喜爱。据荷兰 KPMG 机构统计[1],综合 Facebook,Ins,Twitter 及 YouTube 四大国际社交平台的粉丝数,欧洲冠军联赛以 14.1

到稿日期:2021-06-03 修日期:2021-10-22
基金项目:国家自然科学基金(61562078);新疆"天山雪松计划"青年拔尖人才后备人选项目:机器学习前沿算法及其应用研究;新疆高校科研计划(XJEDU2021Y037)
This work was supported by the National Natural Science Foundation of China(61562078),Xinjiang "Tianshan Cedar Plan" Young Top Talent Reserve Project:Research on Machine Learning Frontier Algorithm and Its Application and Scientific Research Program of Colleges and Universities in Xinjiang(XJEDU2021Y037).
通信作者:王志宁(740216422@qq.com)

亿粉丝数，位列全球十大体育赛事粉丝数量排名榜单的第一，而榜单的第三至第五位分别为英国足球超级联赛、西班牙足球甲级联赛与 FIFA 世界杯。

足球运动凭借优良的赛事氛围、极高的商业化运作、普及度较高等特点，现已成为世界第一大体育运动；同时，足球在学术界的受关注度也不断攀升。由于职业足球联赛必然涉及球员的交易转会，各大职业足球联赛的不断进步使得当今世界足坛的转会市场有了新的特点：一方面，球员身价普遍呈现几何式的增长；另一方面，世界各大知名财团竞相收购顶级联赛俱乐部。自 2010 年 10 月起，国际足联成员协会俱乐部的所有职业球员转会均需通过国际足联转会系统进行登记，使得足球转会交易市场也愈发系统化和完善。据国际足联的最新报告[2]显示，2020 年全球转会市场共记录了 17 077 次男子职业足球运动员的转会，涉及 4 178 家俱乐部和 187 个足协，并涉及了来自 180 个不同国家的 14 432 名球员。然而，在以机器学习为核心技术的人工智能、无人驾驶、机器翻译、语音识别等领域大放异彩的今天，利用机器学习对足球进行的研究，仅仅集中在比赛结果预测、比赛视频语义分析、技战术动作效用及质量分析这几个方面。特别是在国际转会市场交易如此频繁的背景下，球员身价作为影响转会交易最关键的因素，由于缺乏深入的模型及应用方面的研究，导致基于已有的相关工作无法解答如下几个与球员身价相关的问题。

问题 1：影响球员身价的主要因素是什么？不同位置之间是否存在差异？

问题 2：某球员如何制定针对性的训练项目，以在提高关键技术能力的同时快速提升身价估值？

问题 3：同为亚洲顶级球员的中国球员武磊，怎样才能获得与韩国球员孙兴慜一样在更高水平联赛中踢球的机会？

问题 4：对于不同位置球员能力值的提升，其身价会有怎样的变化？

上述仅仅举例了几个常见的问题，已有关于球员身价预测相关的研究，大多仅简单分析了影响球员身价的因素，无法回答或解决的问题还有很多。为解决以上问题，本文以国际足球联合会全球范围内注册球员的各项技术能力值及身价数据为研究对象，在特征工程的基础上通过 XGBoost 集成算法构建球员身价预测模型，并将模型与已有研究以及其他主流机器学习算法进行对比分析，以验证 XGBoost 模型的有效性。在模型构建的基础上，融合 SHAP 框架分析不同位置影响球员身价的重要因素，并以武磊、孙兴慜等球员为例，展示了模型在球员身价评估、身价对比分析、球员训练策略制定等场景的应用。

## 2 相关研究

足球比赛时间长且攻防对阵瞬息万变，跌宕起伏的进程吸引着球迷的狂热关注，同时也吸引了不少学者利用机器学习对其进行深入研究。然而，目前利用机器学习算法对足球进行研究主要集中在比赛结果预测[3-5]、比赛视频语义分析[6-8]、技战术动作效用及质量分析[9-11]这几个方面，针对球员身价预测的相关研究相对较少。在国际转会市场交易如此频繁的背景下，本文将影响转会交易最为关键的球员身价作为本文的研究对象。目前，针对球员身价预测方面的相关研究，按照构建预测模型建立的方法进行分类，主要分为基于传统统计学方法以及基于机器学习方法两大类。

针对球员身价预测及其影响因素分析的早期研究以传统统计学方法和计量经济学方法为主。例如，Herm 等[12]提出了一种基于集体智慧方法(Crowd Wisdom)的职业足球运动员身价估值的方法，该方法为球员的转会交易费用的计算提供了参考；Scelles 等[13]从经济学的角度研究了 2005－2013 赛季期间，影响欧洲足球俱乐部价值的因素，研究结果表明，球员身价估值、俱乐部营业收入和所有权类型是影响俱乐部价值的重要因素；Wan[14]对 2016 赛季中超联赛球员的身价及转会费做了简单的统计分析，其主要研究了国内球员转会市场的现状，但未考虑球员各项技术能力指标对其身价的影响；Rossetti 等[15]运用聚类分析方法，研究了不同俱乐部针对球员转会的不同管理策略，并对球员转会对球队战绩的影响进行了分析；Chen[16]通过选取反映球员能力属性、球员影响力的不同指标，运用相关分析与回归模型构建中超联赛注册球员身价预测的回归方程，预测了中超联赛 275 名前锋球员的身价，但其研究的样本量较小且球员特征选取较少；Ye 等[17]基于基尼系数的构建与计算，研究了 2015 赛季中超球员的身价差距，为中超联赛的管理模式提供了一定的对策建议；Oliver 等[18]将球员特征划分为基础技术属性、临场表现及场外影响力 3 个类别，构建回归模型以预测球员身价；Prabhnoor 等[19]研究了足球运动员的场外受欢迎程度以及赛事临场表现对其市场价值评估的贡献；Kirschstein 等[20]研究了球员的具体技术能力值对其市场价值的影响，选取了 28 个反映球员能力的指标，并运用多元统计分析方法及稳健回归分析了 493 个球员的市场价值，研究结果表明明星球员的身价易被高估。

随着计算机技术及机器学习算法的发展，有研究者开始运用多种机器学习算法构建球员身价预测模型。例如，Zhao[21]应用复杂网络理论构建足球运动员的转会网络图，并在此基础上通过 GBDT 算法构建球员身价的预测模型，其研究结果表明，球员转会网络图对文献[12]中提出的基于集体智慧的足球运动员身价估值方法具有一定的补充作用；Iman 等[22]提出了一种将粒子群优化算法与支持向量机回归算法(Support Vector Regression，SVR)相结合的混合回归方法，用于建立预测模型以估算转会市场球员的价值；Huo[23]将无线传感器网络用于感知球员在场上的表现并记录各种评估指标数据，最终运用贝叶斯算法建立球员的身价预测模型。

本文与已有研究的不同之处在于：1)数据源及预测模型构建方法不同，即在 FIFA 官方权威数据集的基础上，通过 XGBoost 算法建立了相比已有工作性能更加优异的预测模型；2)区分球员位置的身价影响因素分析，即在基于 SHAP 框架提升身价估值模型可解释性的同时，针对不同位置的球员身价影响因素进行分析；3)模型具备落地价值，即通过真实的应用示例，展示了本文模型在球员身价评估、身价对比分析、球员训练策略制定等实际场景的应用价值。

## 3 球员身价预测模型的构建

### 3.1 身价预测模型的构建及其应用总体流程

球员的身价预测模型构建及应用流程如图 1 所示，具体包括以下几个主要步骤。

(1)原始数据获取：本文通过 Python 爬虫从 FIFA(国际足球联合会)的球员数据库网站[1)]获取球员身价值及技术能力值等原始数据。

(2)数据预处理及特征工程：检查并处理原始数据中的缺失值、异常值，进行特征转换、不同位置球员区分等工作，再运用简单相关性分析、相关性热力图及 F-Score 方法进行特征选择。

(3)构建身价预测模型：通过 XGBoost 构建球员身价预测模型，并进行 10 折交叉验证、超参数调优等工作，结合相应算法评价指标获得最优的模型架构。

(4)基于球员身价预测模型的应用：首先，结合 SHAP 模型分析不同位置影响身价估值的主要因素；其次，将模型应用到球员身价评估、身价对比分析、针对性训练策略制定等实际场景。

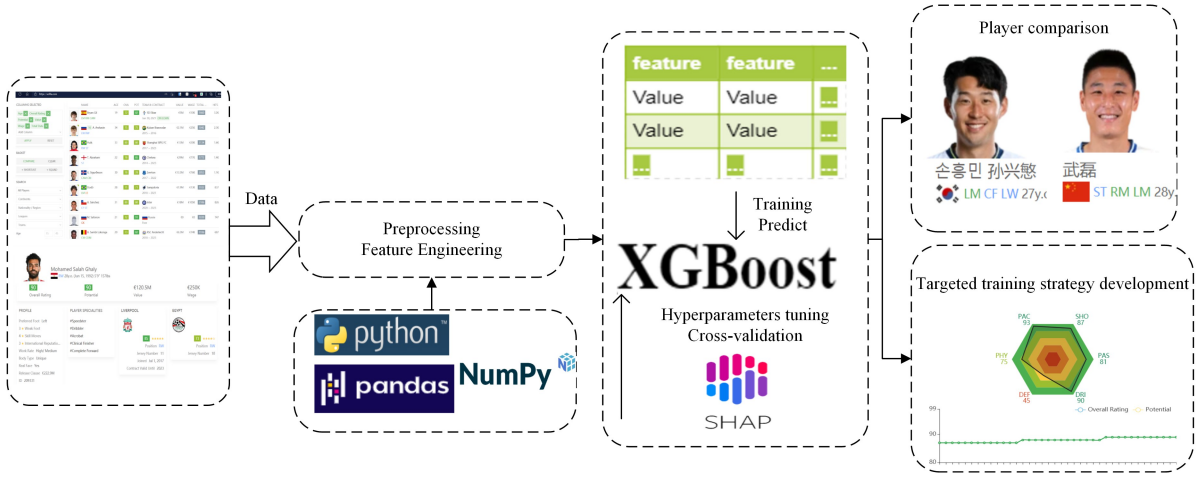



图 1　球员身价预测模型构建及应用流程图

Fig. 1　Flow chart of construction and application of player value prediction model

### 3.2　基于 XGBoost 的身价预测建模过程

基于 XGBoost 算法构建球员身价预测模型，设球员数据集为 $X$(包括球员技术能力值、出生日期、所属俱乐部等特征)，$Y$ 为球员身价值(单位为万欧元)，给定 $n$ 个样本(表示 $n$ 名球员)、$m$ 个特征的训练数据集为 $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中 $x_i=(x_i^{(1)},\cdots,x_i^{(m)})$。将训练集样本数据输入 XGBoost 算法进行训练。XGBoost 是由 Chen 等[24]开源的机器学习项目，其高效实现了梯度提升决策树(Gradient Boosting Decision Tree，GBDT)算法并做了许多改进，在多个领域得到广泛应用并取得了良好的效果[25-26]。XGBoost 算法的本质是 Boosting 迭代方法，涉及加法模型(强评估器由一系列弱评估器线性相加而成)和前向分布算法(下一轮迭代产生的新的评估器是在上一轮的基础上训练得到的)两个关键部分。作为 GBDT 算法的改进，XGBoost 算法将目标函数二阶泰勒展开，保留了更多有关目标函数的信息，并加入正则项以避免过拟合。XGBoost 的核心思想如下。

XGBoost 利用前向分布算法学习到包含 $k$ 个基评估器(模型)的加法模型，$f_k$ 为第 $k$ 个基评估器，$\hat{y}_i$ 为第 $i$ 个球员样本 $x_i$ 的身价预测值，其表达式如下：

$$\hat{y}_i=\sum_{t=1}^{k}f_t(x_i) \tag{1}$$

损失函数反映真实值和预测值之间的差异，$n$ 为样本数量，目标函数由传统损失函数 $l$ 及抑制模型复杂度的正则项 $\Omega$ 组成，$y_i$ 为球员身价真实值，$\hat{y}_i$ 为模型的预测输出。球员身价预测模型训练过程中的损失函数如式(2)所示：

$$Obj=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{t=1}^{k}\Omega(f_i) \tag{2}$$

作为前向分布算法，XGBoost 通过贪心算法来寻找局部最优解，通过每一轮迭代来寻找使损失函数降低最多的树结构，在第 $t$ 轮迭代中，对第 $i$ 个球员样本 $x_i$ 的身价预测值为：

$$\hat{y}_i^t=\hat{y}_i^{t-1}+f_t(x_i) \tag{3}$$

其中，$f_t(x_i)$为迭代需要新加入的基评估器，因此，将目标函数改写为：

$$Obj=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{t-1}+f_t(x_i))+\Omega(f_t) \tag{4}$$

此时最优化目标函数即求得 $f_t(x_i)$，对目标函数进行二阶泰勒展开，$g_i$，$h_i$ 分别为损失函数对前 $t-1$ 轮迭代得到的球员身价预测值 $\hat{y}_i^{t-1}$ 所求的一阶及二阶导数，并将第 $t$ 轮迭代的常数项 $l(y_i,\hat{y}_i^{t-1})$移除得到式(5)：

$$Obj=\sum_{i=1}^{n}(g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i))+\Omega(f_t) \tag{5}$$

XGBoost 算法引入叶子节点分数 $\omega_q$ 来代表每个叶子节点上的预测值，$q(x)$表示样本落入的具体叶子节点，则 $\omega_{q(x)}$ 代表每个样本的预测值，即可以将 $f_t(x)$转换为 $\omega_{q(x)}$。定义树模型的复杂度 $\Omega(f_t)$与叶子节点 $j$ 的样本集合 $I_j$，其表达式为：

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{6}$$

$$I_j=\{i|q(x_i)=j\} \tag{7}$$

将样本集合改写为叶子节点集合，设 $G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$ 得到新的目标函数，如式(8)所示：

$$Obj=\sum_{j=1}^{T}[G_j\omega_j+\frac{1}{2}(H_j+\lambda)\omega_j^2]+\gamma T \tag{8}$$

在式(9)中，需要预测的值 $\omega$ 不确定，令目标函数一阶导数值为 0，求得叶子节点 $j$ 对应的值，并得到目标函数的最小值，如式(10)所示：

1) https://sofifa.com/

$$\omega_j^* = -\frac{G_j}{H_j+\lambda} \tag{9}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{10}$$

### 3.3 融合 SHAP 的模型可解释性

SHAP 是由 Lundberg 等[27]于 2017 年在合作博弈论启发下构建的一个加性的解释模型，其核心是计算每个特征的 Shap Values，以此来反映特征对整个模型的预测能力的贡献程度。SHAP 将模型的预测值解释为每个输入特征的归因值(Shap Values)之和，即：

$$\hat{y} = f_0 + \sum_{i=1}^{M} f_i \tag{11}$$

其中，$\hat{y}$ 为模型预测值，$f_i$ 为每个特征所对应的归因值，$f_0$ 是所有训练样本的预测均值。SHAP 针对 XGBoost 算法做了 Shap Values 与模型预测值的关系转换，如式(12)所示：

$$\ln\frac{\hat{y}}{1-\hat{y}} = f_o + \sum_{i=1}^{M} f_i \tag{12}$$

传统方法 Feature Importance 可以直观地反映出特征的重要性，但无法判断特征与最终预测结果的关系。SHAP 使用特征归因方法计算特征的归因值，可以反映出每个特征对最终预测值的影响，并且可以表现出影响的正负性，增加了模型的可解释性。SHAP 提供了强大的数据可视化功能，以展示模型及预测的解释结果，广泛适用于解释较为复杂的算法模型。

## 4 模型评估及其对比分析

### 4.1 数据预处理

在原始训练数据集方面，本文从 FIFA 官方球员数据库中，通过 Python 爬虫获取到 2020－2021 赛季全球各大职业足球联赛中所有注册在籍的球员的最新信息。数据集包括球员的身价及违约金、所属俱乐部及联赛、个人身体属性及技术能力值等多项基本信息。而在球员属性信息中，共包括基础、进攻、技巧、移动、力量、心理、防守、守门及特殊 9 个类别。将球员的身价作为标签值，数据集所含部分特征及具体含义如表 1 所列。

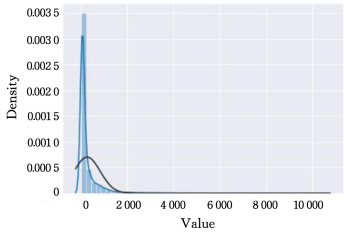
表 1 数据集特征列表

Table 1 Data set feature

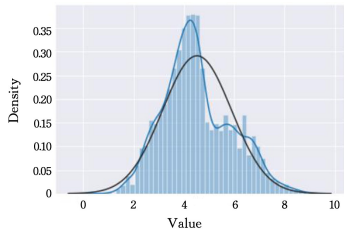
| 变量名称 | 均值 | 方差 | 变量含义 | 变量类型 |
|---|---|---|---|---|
| age | 25.1033 | 20.3954 | 年龄 | 数值型 |
| potential | 71.7425 | 36.9600 | 潜力 | 数值型 |
| reputation | 1.1043 | 0.1444 | 国际知名度 | 数值型 |
| skill | 2.5389 | 0.3942 | 技巧动作(能力值) | 数值型 |
| pace | 67.7195 | 127.3319 | 速度(能力值) | 数值型 |
| shoot | 52.2776 | 196.4576 | 射门(能力值) | 数值型 |
| pass | 57.2054 | 108.4184 | 传球(能力值) | 数值型 |
| dribble | 62.5284 | 105.6990 | 带球(能力值) | 数值型 |
| defend | 51.5041 | 269.2228 | 防守(能力值) | 数值型 |
| physical | 64.8485 | 95.3145 | 身体对抗(能力值) | 数值型 |
| crossing | 54.1476 | 197.5129 | 传中(能力值) | 数值型 |
| finishing | 49.8343 | 266.5895 | 完成射门(能力值) | 数值型 |
| head_accuracy | 56.9422 | 135.3220 | 头球精度(能力值) | 数值型 |
| short_pass | 62.7034 | 92.5498 | 短传(能力值) | 数值型 |
| volleys | 46.6435 | 215.3452 | 凌空球(能力值) | 数值型 |
| dribbling | 60.8569 | 152.2892 | 盘带(能力值) | 数值型 |
| curve | 51.4206 | 227.6680 | 弧线(能力值) | 数值型 |
| fk_accuracy | 46.2388 | 224.6517 | 定位球精度(能力值) | 数值型 |
| long_pass | 56.0853 | 151.4690 | 长传(能力值) | 数值型 |
| ball_control | 63.3053 | 99.0212 | 控制球(能力值) | 数值型 |
| agility | 66.5132 | 150.1958 | 灵活性(能力值) | 数值型 |
| reactions | 62.1088 | 77.3594 | 反应(能力值) | 数值型 |
| shot_power | 59.6291 | 176.1997 | 射门力量(能力值) | 数值型 |
| stamina | 66.9626 | 129.8300 | 体能(能力值) | 数值型 |
| long_shots | 51.1139 | 248.5976 | 远射(能力值) | 数值型 |
| aggression | 59.4197 | 204.7401 | 侵略性(能力值) | 数值型 |
| positon | 54.9871 | 212.0886 | 位置感(能力值) | 数值型 |
| vision | 55.5443 | 164.0652 | 视野(能力值) | 数值型 |
| penalties | 51.9638 | 154.3437 | 罚点球(能力值) | 数值型 |
| composure | 60.4110 | 104.2331 | 心理素质(能力值) | 数值型 |
| ST | 59.8425 | 84.5472 | 球员在影锋位置的能力值 | 数值型 |
| LR_W | 61.0936 | 97.6191 | 球员在边锋位置的能力值 | 数值型 |
| CF | 60.7821 | 97.1934 | 球员在中锋位置的能力值 | 数值型 |
| CAM | 61.0549 | 95.3070 | 球员在前腰位置的能力值 | 数值型 |
| LR_M | 61.7689 | 85.6269 | 球员在左右中场位置的能力值 | 数值型 |
| CM | 60.2346 | 78.1323 | 球员在中场位置的能力值 | 数值型 |
| CDM | 58.7128 | 103.7724 | 球员在后腰位置的能力值 | 数值型 |
| LR_WB | 59.3841 | 82.0599 | 球员在进攻型边后卫位置的能力值 | 数值型 |
| LR_B | 58.6577 | 92.3980 | 球员在边后卫位置的能力值 | 数值型 |
| CB | 57.4532 | 139.0840 | 球员在中后卫位置的能力值 | 数值型 |
| value | 261.2587 | 322585.2629 | 市场价值(单位为万欧元) | 数值型 |

球员的身价值范围在 1～10 550 万欧元之间，平均身价为 261.259 万欧元，通过概率密度分布图查看球员身价的分布，如图 2(a)所示，呈现较为明显的右偏。为使模型能够更好地处理数据，本文运用 Box-Cox 变换使目标值正态化，变换

前后的球员身价值分布如图 2 所示。




(a)球员身价值分布图




(b)Box-Cox 变换后球员身价值分布图

图 2　球员身价分布图

Fig. 2　Distribution map of players' value

绘制特征相关性热力图对数据集特征进行探索及分析，热力图中颜色越浅表示两个特征之间相关性越高，颜色越深，表示两特征之间相关性越高。部分特征与身价标签值的相关性热力图如图 3 所示。

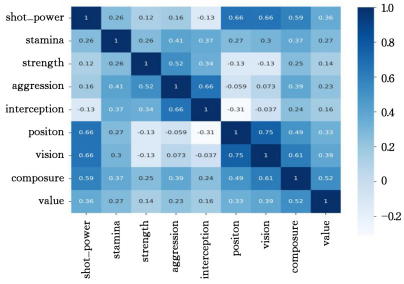



图 3　部分变量与目标变量(value)热力图

Fig. 3　Heat diagram of partial variables and target variables (value)

运用 f_regression 得分函数进行特征选择。F-Score 最初由台湾国立大学的 Chen 等[28]提出，是区别于分类问题算法评价指标 F1_score 的一种衡量特征在两类之间分辨能力的方法，基于此方法可以实现最有效的特征选择。F-Score 越大说明该特征的辨别能力越强，计算所有特征的 F-Score 并进行特征选择，部分特征的 F-Score 如表 2 所列。

表 2　特征 F-Score 重要性值

Table 2　Feature F-Score importance value

| 变量名称 | 变量解释 | F-Score 值 |
|---|---|---|
| potential | 潜力 | 8373.844 950 56 |
| reputation | 国际知名度 | 10 488.886 486 74 |
| skill | 技巧动作能力值 | 2 604.923 352 44 |
| pass | 传球能力值 | 4 135.527 277 97 |
| dribble | 带球能力值 | 3 665.604 810 40 |
| volleys | 凌空球能力值 | 2 107.328 523 98 |
| curve | 弧线能力值 | 2 192.394 523 95 |
| acceleration | 加速度能力值 | 440.491 244 96 |
| sprint_speed | 冲刺速度能力值 | 474.422 306 89 |
| reactions | 反应能力值 | 8 114.493 037 20 |
| shot_power | 射门力量能力值 | 2 425.764 646 98 |
| jumping | 弹跳能力值 | 205.928 017 88 |
| stamina | 体能能力值 | 1 270.809 006 25 |

通过特征相关性热力图及计算特征的 F-Score 重要性值进行特征选择。球员的国际知名度于球员身价的影响最大，但该特征受多种因素影响，综合性较强，球员难以通过提高个人技术能力值来直接改变其知名度，因此，将该项特征剔除。此外，将球员在数据库中的 id、球员姓名、所属俱乐部及联赛、国籍、球衣号码、违约金、合同到期日、周工资等无关特征剔除，并将球员的出生日期转换为新特征年龄。由于门将位置球员具有特定的门将扑球、门将开球、门将位置感及门将反应等特有属性值，与其他位置球员的技术能力属性有较大的区别，因此将其他位置球员与门将位置球员进行区分。经特征选择及处理所得到的数据集的样本总量为 16 032，特征总数为 40。

### 4.2　算法评价指标

本文选用决定系数($R^2$)、平均绝对误差($MAE$)、均方根误差($RMSE$)作为衡量模型效果的评价指标。计算公式分别如式(13)－式(15)所示。

$$R^2(y,\hat{y})=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{13}$$

$$MAE(y,\hat{y})=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{14}$$

$$RMSE(y,\hat{y})=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{15}$$

其中，$y_i$ 为样本真实值，$\hat{y}_i$ 为预测值，$n$ 为样本数。

决定系数反映因变量的全部变异能通过回归关系被自变量所解释的比例，$R^2$ 取值在[0,1]之间，决定系数值越大，模型的效果越好。MAE 与 RMSE 衡量模型预测值与真实值之间的差异，两者具有相同的量纲，但 RMSE 先对误差平方进行累加再开方，放大了较大误差之间的差距，相比 MAE，RMSE 的值更大，较为直接地反应了真实误差。MAE 与 RMSE 的值越小，表明模型预测越准确。

### 4.3　模型训练及对比实验

对原始数据集进行数据预处理及特征工程之后，将数据集中的特征及标签输入模型，划分为训练集与测试集数据，建立 XGBoost 集成算法模型进行训练，并对测试集数据进行预测，根据算法评价指标评估模型的表现，再进行超参数调优。XGBoost 算法的重要参数有 $n\_estimators$(基评估器的数量)、$learning\_rate$(学习率)、$max\_depth$(树的最大深度)、$objective$(损失函数的类型)、$subsample$(对样本所采样的比例)、$reg\_alpha$($L1$ 正则项的权重)、$reg\_lambda$($L2$ 正则项的权重)。运用网格搜索方法及绘制学习曲线，获得的最佳超参数组合为：$n\_estimators=500$，$learning\_rate=0.1$，$objective=reg:gamma$，$max\_depth=5$，$subsample=1$，$reg\_alpha=0$，$reg\_lambda=1$。

运用 10 折交叉验证方法进一步对模型超参数调优后的效果进行评估，结果表明，XGBoost 算法的性能表现较好，相应的算法评价指标结果如表 3 所列。将本文所采用的 XGBoost 算法与文献[21]和文献[22]中采用的 GBDT 算法、支持向量机回归算法以及主流机器学习算法如 Random Forest 和 Adaboost 等 9 种算法进行对比分析，对比时采用决定

系数、平均绝对误差、均方根误差3组性能指标在10折交叉验证集上进行实验，对比结果如表4所列。各算法的重要超参数配置如表5所列。

表3 交叉验证下XGBoost算法的预测评价指标结果

Table 3 Prediction and evaluation indicators of XGBoost algorithm with cross validation

| 评价指标 | $R^2$ | *MAE* | *RMSE* |
|---|---|---|---|
| 10折交叉验证次数一 | 0.979 | 25.579 | 85.037 |
| 10折交叉验证次数二 | 0.976 | 25.141 | 68.927 |
| 10折交叉验证次数三 | 0.983 | 24.239 | 65.733 |
| 10折交叉验证次数四 | 0.977 | 26.587 | 78.861 |
| 10折交叉验证次数五 | 0.974 | 28.019 | 94.204 |
| 10折交叉验证次数六 | 0.957 | 28.243 | 127.474 |
| 10折交叉验证次数七 | 0.979 | 24.037 | 71.267 |
| 10折交叉验证次数八 | 0.981 | 24.585 | 79.601 |
| 10折交叉验证次数九 | 0.944 | 28.911 | 127.074 |
| 10折交叉验证次数十 | 0.978 | 28.443 | 96.799 |
| 均值 | 0.973 | 26.378 | 89.498 |

表4 算法预测评价指标对比结果

Table 4 Comparison of algorithm prediction and evaluation indicators

| 算法 | $R^2$ | *MAE* | *RMSE* |
|---|---|---|---|
| XGBoost | 0.973 | 26.378 | 89.498 |
| Decision Tree Regression | 0.907 | 48.629 | 168.229 |
| Extra Tree Regression | 0.876 | 56.469 | 188.521 |
| Random Forest | 0.958 | 30.663 | 111.631 |
| Bagging | 0.942 | 33.385 | 122.463 |
| Adaboost | 0.502 | 331.625 | 352.759 |
| GBDT[21] | 0.963 | 40.478 | 96.547 |
| Linear Regression | 0.595 | 195.465 | 353.272 |
| Support Vector Regression(SVR)[22] | 0.192 | 158.236 | 499.368 |
| K-Nearest Neighbors(KNN) | 0.832 | 77.289 | 227.132 |

表5 各算法重要超参数配置

Table 5 Important hyperparameter configuration for each algorithm

| 算法 | 算法重要超参数配置 |
|---|---|
| XGBoost | *n_estimators*=500, *learning_rate*=0.1, *max_depth*=5 |
| Decision Tree | *criterion*='*mse*', *min_samples_split*=2 |
| Extra Tree | *criterion*='*mse*' |
| Random Forest | *n_estimators*=100 |
| Bagging | *n_estimators*=30 |
| Adaboost | *n_estimators*=50 |
| GBDT | *n_estimators*=50 |
| SVM | *kernel*='*rbf*' |

分析表4中的数据结果可知，本文所选用的XGBoost算法具有最优的效果，3项算法评价指标$R^2$，*MAE*，*RMSE*的取值分别为0.973，26.378，89.498，在10种机器学习算法的对比结果中拥有最佳的数据集拟合效果且预测最准确。与已有同类研究中的算法进行对比，文献[21]中运用的GBDT算法不含球员身价预估值，该项特征的模型的$R^2$为0.55。将转会市场网站的球员身价预估值加入球员特征，所得到的球员身价预测模型的$R^2$为0.84，相比之下，本文的特征工程方法及机器学习算法拥有更好的性能表现。与主流机器学习算法模型相比，XGBoost在3项评价指标上的表现同样最佳。随机森林、Bagging算法、决策树回归对于数据集也具有不错的拟合优度，与XGBoost的差距甚微，而线性回归和支持向量机算法的$R^2$值较低。在指标*MAE*上，XGBoost算法的值最低，为26.378，随机森林为30.663，线性回归与支持向量机及Adaboost的效果最差，与XGBoost的差距较大。对于指标*RMSE*，XGBoost同样拥有最好的实际表现，除GBDT算法与其较为接近外，其余算法在该项指标上的表现都与XGBoost有较大的差距。

综上所述，XGBoost在3项评价指标上的表现效果都优于其他主流算法，预测效果最理想，可以很好地反映出球员属性信息与其身价之间的关系。球员的属性特征与身价之间表现出一种复杂的非线性关系，基于集成方法的XGBoost，GBDT和随机森林算法预测的效果较好。XGBoost算法使用二阶泰勒展开优化目标函数，保留了更多有关目标函数的信息，并加入了正则项来控制模型的复杂度，以防止过拟合，具有较好的预测效果。研究结果表明，在进行特征选择及处理的基础上，相比已有工作，本文模型可以更好地预测球员的身价。

## 5 球员身价预测模型的应用及分析

### 5.1 算法特征重要性对比分析

图4给出了运用SHAP模型绘制的球员特征SHAP摘要图，选取部分重要特征分析其贡献度，可以看出，随机森林算法在数据集上的表现与XGBoost算法的差距较小，尤其在指标*MAE*上，XGBoost为26.952，随机森林为30.663，优于其他算法。分别计算两个算法的特征重要性度，并与SHAP value进行比较分析，结果如表6所列。

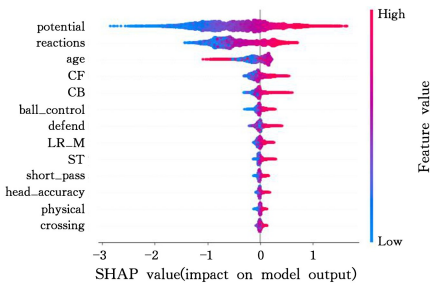



图4 SHAP特征分析摘要图

Fig. 4 Summary of SHAP feature analysis

结合图4及表5的结果可以发现，3种方法对于特征的重要度排序有略微差异。综合分析可以得出，reactions(反应能力值)、potential(潜力)、CF(球员在中锋位置能力值)、CB(球员在中后卫位置能力值)、ball_control(控制球能力值)、LR_M(球员在左右中场位置能力值)是影响球员身价的关键因素。reactions(反应能力值)和potential(潜力)在3种算法进行重要度排序时都排在前两位。具体而言，reactions(反应能力值)是球员在赛场上的技术动作基础，影响进攻时的传接球配合、跑位、抢点以及把握进球机会的能力；对于防守，反应能力会影响球员的卡位、造越位战术、抢断球及拦截能力。球员训练计划中有专业、系统的针对反应的专项训练，提高反应能力对于攻防技术动作的提升有极大帮助，也有利于提高球员的身价。potential(潜力)对于球员身价也具有重要的影响，这是因为潜力代表球员的天赋及能力提升前景，球员的潜力值越高，球员本身具有更大的技术能力提升空间与更广阔的发展前景，球员身价也相应越高。世界足坛大部分俱乐部都拥有完善的青训系统，以及专业的球探发掘有潜力的年轻球员，再通过系统的训练让潜力转化为能力以兑现天赋，潜力对球员身价的影响与足球转会

市场的实际情况相符。CF(球员在中锋位置能力值)、CB(球员在中后卫位置能力值)、LR_M(球员在左右中场位置能力值)的重要程度也反映了球员在场上进攻、组织、防守的能力对其身价的影响。

表 6　XGBoost,Random Forest,SHAP 算法的特征重要度对比

Table 6　Feature importance comparison of XGBoost,Random Forest and SHAP algorithms

| 排名 | XGB Feature Importance | | RF Feature Importance | | SHAP Feature Importance | |
|---|---|---|---|---|---|---|
| | feature | value | feature | value | feature | value |
| 1 | reactions(反应能力值) | 0.451 | reactions(反应能力值) | 0.457 | potential(潜力) | 0.655 |
| 2 | potential(潜力) | 0.129 | potential(潜力) | 0.268 | reactions(反应能力值) | 0.522 |
| 3 | CF(球员在中锋位置能力值) | 0.076 | CF(球员在中锋位置能力值) | 0.082 | age(年龄) | 0.166 |
| 4 | ball_control (控制球能力值) | 0.075 | LR_M(球员在左右中场位置能力值) | 0.044 | CF(球员在中锋位置能力值) | 0.123 |
| 5 | CB(球员在中后卫位置能力值) | 0.056 | CB(球员在中后卫位置能力值) | 0.036 | CB(球员在中后卫位置能力值) | 0.077 |
| 6 | reputation(国际知名度) | 0.042 | CAM(球员在前腰位置能力值) | 0.025 | ball_control (控制球能力值) | 0.062 |
| 7 | LR_M(球员在左右中场位置能力值) | 0.038 | ball_control (控制球能力值) | 0.022 | defend(防守能力值) | 0.051 |
| 8 | defend(防守能力值) | 0.021 | age(年龄) | 0.015 | LR_M(球员在左右中场位置能力值) | 0.036 |
| 9 | age(年龄) | 0.020 | ST(球员在影锋位置能力值) | 0.005 | short_pass(短传能力值) | 0.027 |
| 10 | composure(心理素质能力值) | 0.018 | LR_W(球员在左右边锋位置能力值) | 0.003 | ST(球员在影锋位置能力值) | 0.026 |

## 5.2　不同位置的球员身价影响因素分析

足球赛场有具体的位置区分,不同位置的球员有相应的技术能力值区别,在各个位置也具有特定的战术倾向及比赛贡献,如后场球员主要负责防守与拦截,中场球员倾向于进行组织串联,锋线球员把握进球得分机会。分析不同位置球员的具体能力值对于其身价的影响,能够为不同球员的身价评估、身价对比分析、训练策略的制定提供一定的支持。根据球员位置构建后场位置球员、中场位置球员、锋线位置球员和门将位置球员 4 个数据集。其中,门将位置球员与其余球员区别较大,具有特定的属性,具体特征如表 7 所列。

通过绘制特征相关性热力图以及计算特征的 F-Score 重要性度进行特征选择,并基于 XGBoost 算法分别建立球员身价预测模型,对不同位置球员的身价进行预测,再与随机森林、GBDT 和 SVM 算法在决定系数、平均绝对误差、均方根误差 3 组性能指标上进行对比实验.算法评价指标对比结果如表 8 所列。

表 7　门将位置球员特定特征列表

Table 7　Specific features of goalkeeper position players

| 变量名称 | 均值 | 方差 | 变量含义 | 变量类型 |
|---|---|---|---|---|
| diving | 65.3407 | 60.0342 | 门将扑救(能力值) | 数值型 |
| handling | 63.0721 | 52.5301 | 门将控球(能力值) | 数值型 |
| kicking | 61.7876 | 56.6055 | 门将开球(能力值) | 数值型 |
| reflexes | 66.3196 | 66.8602 | 门将反应(能力值) | 数值型 |
| speed | 37.7595 | 112.8965 | 门将速度(能力值) | 数值型 |
| position | 63.2866 | 71.7033 | 门将位置感(能力值) | 数值型 |

表 8　不同位置算法模型预测评价指标对比结果

Table 8　Comparison of prediction and evaluation indicators of different location algorithm models

| 位置 | XGBoost | | | Random Forest | | | GBDT | | | SVM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | *MAE* | *RMSE* | $R^2$ | *MAE* | *RMSE* | $R^2$ | *MAE* | *RMSE* | $R^2$ | *MAE* | *RMSE* |
| 门将 | 0.971 | 18.047 | 84.382 | 0.933 | 29.045 | 122.633 | 0.938 | 29.219 | 118.101 | 0.013 | 129.963 | 495.123 |
| 后场 | 0.965 | 24.609 | 78.233 | 0.943 | 29.663 | 104.453 | 0.951 | 35.381 | 95.313 | 0.193 | 123.178 | 403.597 |
| 中场 | 0.966 | 33.435 | 87.656 | 0.942 | 45.964 | 143.321 | 0.954 | 51.604 | 126.726 | 0.176 | 175.283 | 541.601 |
| 锋线 | 0.964 | 38.089 | 107.392 | 0.943 | 50.043 | 159.726 | 0.943 | 55.208 | 152.431 | 0.125 | 204.319 | 224.007 |

表 8 中的结果显示,无论是在门将、后场、中场还是锋线位置上,相比随机森林、GBDT 以及 SVM,XGBoost 算法在 3 组评价指标上都具有最佳表现。结合 SHAP 模型应用已建立的球员身价预测模型,绘制 SHAP 摘要图,如图 5 所示,分析不同位置球员具体特征对其身价的影响。不同位置球员 SHAP 特征重要度的排序及对比结果如表 9 所列。

结合图 5 及表 9 可以看出,除 potential(潜力)和 age(年龄)外,不同位置球员的具体技术能力值对其身价预测值的影响有所不同。

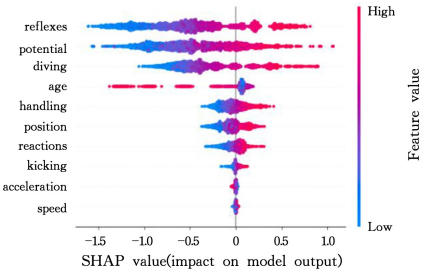
(1)对于门将位置球员,reflexes(门将反应能力值)、diving(门将扑救能力值)、hangling(门将控球能力值)、position(门将位置感能力值)对其身价的影响较为显著。具体而言,门将是球队的最后一道防线,需要快速反应以进行扑救,且门将的控球和位置感不仅是门将与后卫球员在防守中进行配合的基础,也能为发掘良好的防守反击机会提供帮助,门将可以利用开球机会精准寻找到中前场球员,以把握最佳的进攻时机。因此,门将需要提高反应、扑救、控球及位置感等能力值来提升自己的身价估值。

(2)对于后场球员,defend(防守能力值)、standing_tackle(断球能力值)、physical(身体)是较为重要的技术能力属性,后场球员主要进行防守,需要良好的身体素质以进行对抗、铲抢、争顶、卡位等防守技术动作,破坏对手的进攻并断球以获得球权。

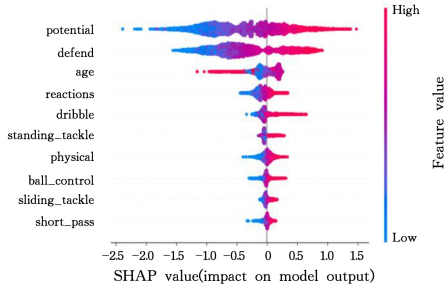
(3)对于中场球员,short_pass(短传能力值)、cross(传中能力值)、positioning(位置感能力值)对其身价预测值有重要的贡献,中场球员主要进行组织配合,为锋线球员创造良好的射门机会,同时也帮助后卫球员进行一定程度的防守和拦截。短传配合、传中球的准确落点以及发现对手防守漏洞的位置感及视野是中场球员提升身价的必备技术能力。

(4)对于后场球员,ball_control(控制球能力值)、shoot(射门能力值)、dribble(带球能力值)对其身价影响较为重要,

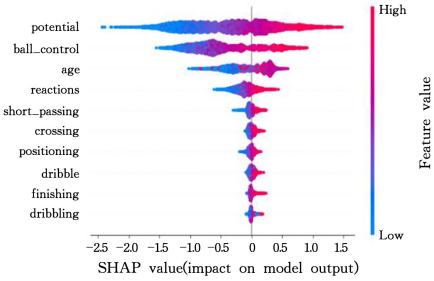
锋线球员的职责是射门得分，在前场的传控配合、门前把握得分机会都需要控制足球以及盘带的技术基础。作为一名锋线球员，若期望获得更高的身价估值，需要不断提升射门、控制球、盘带技术能力。



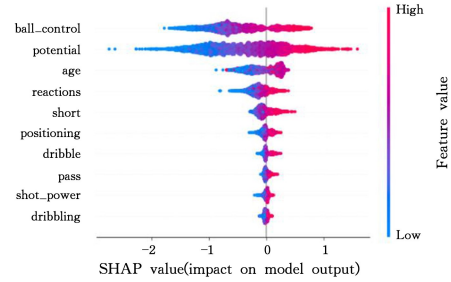

(a)门将位置球员 SHAP 特征分析摘要图




(b)后场球员 SHAP 特征分析摘要图




(c)中场球员 SHAP 特征分析摘要图




(d)锋线球员 SHAP 特征分析摘要图

图 5 球员不同位置 SHAP 特征分析摘要图

Fig. 5 Summary of SHAP feature analysis of players in different positions

表 9 不同位置球员特征 SHAP 重要度对比

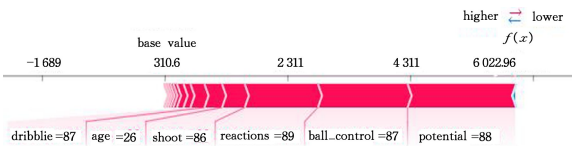
Table 9 Feature SHAP importance comparison of players in different positions

| 排名 | 门将位置 | | 后场位置 | | 中场位置 | | 锋线位置 | |
|---|---|---|---|---|---|---|---|---|
| | feature | value | feature | value | feature | value | feature | value |
| 1 | reflexes(门将反应能力值) | 0.622 | potential(潜力) | 0.601 | potential(潜力) | 0.638 | ball_control(控制球能力值) | 0.625 |
| 2 | potential(潜力) | 0.514 | defend(防守能力值) | 0.579 | ball_control(控制球能力值) | 0.623 | potential(潜力) | 0.605 |
| 3 | diving(门将扑救能力值) | 0.445 | age(年龄) | 0.147 | age(年龄) | 0.271 | age(年龄) | 0.231 |
| 4 | age(年龄) | 0.149 | reactions(反应能力值) | 0.016 | reactions(反应能力值) | 0.141 | reactions(反应能力值) | 0.155 |
| 5 | hangling(门将控球能力值) | 0.109 | dribble(带球能力值) | 0.088 | short_pass(短传能力值) | 0.048 | shoot(射门能力值) | 0.106 |
| 6 | position(门将位置感能力值) | 0.103 | standing_tackle(断球能力值) | 0.059 | cross(传中能力值) | 0.036 | positioning(位置感能力值) | 0.062 |
| 7 | reactions(反应能力值) | 0.071 | physical(身体) | 0.058 | positioning(位置感能力值) | 0.034 | dribble(带球能力值) | 0.046 |
| 8 | kicking(门将开球能力值) | 0.022 | ball_control(控制球能力值) | 0.042 | dribble(带球能力值) | 0.031 | pass(传球能力值) | 0.029 |

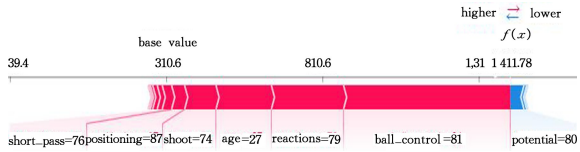
### 5.3 球员身价评估及对比分析(以武磊与孙兴慜为例)

韩国足球运动员孙兴慜与中国足球运动员武磊都是亚洲领域内最优秀的锋线球员，孙兴慜在英国足球超级联赛中的托特纳姆热刺队效力，其身价为 6000 万欧元；武磊在西班牙足球乙级联赛中的皇家西班牙人俱乐部效力，其身价为1450 万欧元。两名球员都司职前锋，同为亚洲人，具有相近的身体素质，且年龄仅相差一岁，但两人的身价差距较大。此外，孙兴慜所处的英超联赛是世界足坛最顶级的联赛之一，孙兴慜是球队首发主力，但武磊效力的俱乐部属于西班牙的二级联赛，武磊也难以占据球队阵容中的主力位置，获得的出场机会与时间较少。将球员身价预测模型应用于两名具体球员的身价评估中，对中国球员武磊制定针对性训练策略，为提高身价估值提供一定的决策支持。

如图 6 所示，在 SHAP 特征分析效应图中，base_value 为球员身价预测模型的目标变量拟合值的均值，即锋线球员身价的平均值——310.6 万欧元。红色区域表示该特征对目标值的贡献为正数，蓝色区域表示该特征对目标值的贡献为负数。球员孙兴慜的实际身价为 6000 万欧元，模型预测值为 6022.96 万欧元，除去潜力值与年龄的影响，孙兴慜的 ball_control(控制球能力值)、reactions(反应能力值)、shoot(射门能力值)是对其身价预测值贡献最大的前 3 项特征，且在 3 项技术上孙兴慜都具有极高的能力水平，其高达 87 的控制球能力值使其身价增加了 1461.557 万欧元。而球员武磊的实际身价值为 1450 万欧元，预测值为 1411.78 万欧元，ball_control(控制球能力值)、reactions(反应能力值)对武磊的身价预测值贡献较大，武磊 81 的控制球能力值使其身价增加了 536.009 万欧元。两名球员的足球运动风格、技术动作倾向有所不同，导致两者的重要技术能力值及排序略有差异。若球员武磊想要获得更高的身价估值，继而在转会市场中有更多选择，增加去高水平联赛踢球的机会与可能性，需要保持现有技术能力中的控制球能力、赛场视野以及反应能力，并通过针对性和系统化的训练提高射门能力和短传能力。




(a)球员孙兴慜 SHAP 特征分析效应图




(b)球员武磊 SHAP 特征分析效应图

图 6　不同球员 SHAP 特征分析效应图(电子版为彩图)

Fig. 6　Effect diagram of SHAP feature analysis of different players

### 5.4　球员训练策略及其身价变化分析

应用身价预测模型为球员进行针对性训练提供决策参考，在锋线、中场、后场、门将位置上分别选取 2 名年龄在 27 岁及以下的不同球员，获得其身价真实值、特征真实值，选择一项对其身价影响较为重要的技术能力值，增加 5 个单位并由身价预测模型得到新的身价预测值，身价变化及特征由表 10 所列。

表 10　不同位置球员能力变化及身价评估

Table 10　Players' ability changes and valuations in different positions

| | | 身价真实值/万欧元 | feature 真实值 | feature 预测值 | 身价预测值/万欧元 | 身价变化/万欧元 |
|---|---|---|---|---|---|---|
| 锋线位置 | 武磊 | 1 450 | shoot(射门能力值):74 | shoot(射门能力值):79 | 1 511.64 | +61.64 |
| | 拉什福德 | 3 550 | composure 心理素质(能力值):73 | composure 心理素质(能力值):78 | 3 608.78 | +58.78 |
| 中场位置 | 贝尔温 | 3 000 | reactions(反应能力值):81 | reactions(反应能力值):86 | 3 052.64 | +52.64 |
| | 班巴 | 1 450 | ball_control(控制球能力值):82 | ball_control(控制球能力值):87 | 1 753.66 | +303.66 |
| 后场位置 | 阿诺德 | 3 200 | standing_tackle(断球能力值):83 | standing_tackle(断球能力值):88 | 3 217.83 | +17.83 |
| | 德利赫特 | 5 000 | reactions(反应能力值):83 | reactions(反应能力值):88 | 5 026.27 | +26.27 |
| 门将位置 | 奥纳纳 | 2 200 | position(门将位置感能力值):76 | position(门将位置感能力值):81 | 2275.29 | +75.29 |
| | 卢宁 | 1 100 | handling(门将控球能力值):71 | handling 门将控球(能力值):76 | 1 123.39 | +23.39 |

对于锋线球员武磊，由 5.3.1 节的内容分析可知，若期望获得更高的身价估值，需要通过针对性的训练来提高射门能力、短传能力。若其射门能力值由 74 提升为 79，武磊把握进球机会的能力会更加优异，其身价也能增加 61.64 万欧元。同为锋线球员的拉什福德可通过提升 5 个单位的心理素质能力值，获得身价 58.78 万欧元的增长。中场球员班巴的控制球能力值为 82，在该项技术上已经表现不错，但对于中场球员而言，需要长时间的传接球配合寻找进攻机会以及化解对手的紧逼、压迫式防守，控制球能力极为重要。如果班巴通过训练将该项技术能力提升至 87，其身价可增加 303.66 万欧元。年仅 19 岁的欧洲金童奖得主中后卫德利赫特有极佳的成长空间，若其反应能力值提升为 88，其身价可增加 26.27 万欧元。若门将奥纳纳提升 5 个单位的门将位置感能力值，则其门前站位选择以及对出击时机的把握会得到提升，对于门将的防守及扑救都有一定程度的提高，其身价估值也会相应增加 75.29 万欧元。

**结束语**　在国际转会市场交易如此频繁的背景下，球员身价作为影响转会交易最关键的因素，由于缺乏深入的模型及应用方面的研究，致使基于已有研究无法回答诸如影响球员身价的主要因素是什么，且不同位置之间是否存在差异等与球员身价预测相关的类似问题。本文以国际足球联合会的官方球员数据为研究对象，首先，在区分不同位置球员的前提下，运用 Box-Cox 变换、F-Score 特征选择等方法进行特征处理；其次，将特征提取后的数据输入 XGBoost 模型进行训练及预测，并与 Random Forest，GBDT，SVR 等主流机器学习算法进行对比分析，对比时采用 $R^2$，$MAE$，$RMS$ 3 组性能指标在交叉验证集上进行实验，验证了本文模型的有效性；最后，结合 SHAP 模型分析了不同位置影响身价估值的主要因素，并以武磊、孙兴慜为例，将模型应用到球员身价评估、身价对比分析、针对性训练策略制定等实际场景中。

由于本文建模数据来自官方组织——国际足球联合会(FIFA)的球员数据，数据更新周期较长(通常每年更新一次)，使得模型的时效性受到巨大的影响。因此，为了更好地解决这一问题，下一步工作考虑将球员阶段性表现、赛场外因素等动态因素引入模型，进一步提高球员身价预测模型的效果及其应用价值。

## 参 考 文 献

[1] Football Clubs' Valuation: The European Elite 2020[EB/OL]. (2020-05-28)[2020-10-13]. http://www.footballbenchmark.com/library/football_clubs_valuation_the_european_elite_2020.

[2] Global Transfer Market Report 2020 [EB/OL]. (2020-01-18)[2020-10-13]. http://img.fifa.com/image/upload/ijiz9rtpkfnbhxwbqr70.pdf.

[3] AO X Q, GONG Y J, LI J. Prediction of soccer match results based on handicapdata[J]. Journal of Chongqing Technology Business University(Natural Science), 2016, 33(6): 86-89.

[4] NAZIM R, AIDA M, ROSHIDI D, et al. A Review on football match outcome prediction using bayesian networks [J]. Journal of Physics: Conference Series, 2018, 1020(1): 1-9.

[5] LEONARDO E, FRANCESCO P, NICOLA T. Combining historical data and bookmakers' odds in modelling football scores [J]. Statistical Modelling, 2018, 18(6): 1-24.

[6] XIA Z C, YANG G B, ZHANG Z Y, et al. Video adaptation scheme for football sports video on mobile terminals[J]. Journal of Chinese Computer Systems, 2011, 32(8): 1660-1664.

[7] TONG M, DING L W, JI C L. Fusion of HCRF and AAM highlight events detection in soccer videos[J]. Journal of Computer Research and Development, 2014, 51(1): 225-236.

[8] YU J Q, ZHANG Q, WANG Z K, et al. Soccer highlight detec-

tion based on replay and affection arousal model[J]. Chinese Journal of Computers, 2014, 37(6): 1268-1280.
[9] CHAWLA S, ESTEPHAN J, GUDMUNDSSON J, et al. Classification of passes in football matches using spatiotemporal data [J]. ACM Transactions on Spatial Algorithms and Systems, 2017, 3(6): 11-25.
[10] GOES F R, KEMPE M, MEERHOFF L A, et al. Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches[J]. Big Data, 2018, 7(1): 57-70.
[11] REIN R, RAABE D, MEMMERT D. 'Which pass is better?' Novel approaches to assess passing effectiveness in elite soccer [J]. Hum Movement Science, 2017, 55(10): 172-181.
[12] HERM S, CALLSEN-BRACKER H M, KREIS H. When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community[J]. Sport Management Review, 2014, 17(4): 484-492.
[13] SCELLES N, HELLEU B, DURAND C, et al. Professional sports firm values: Bringing new determinants to the foreground? A study of European soccer, 2005—2013[J]. Journal of Sports Economics, 2014, 17(7): 1-18.
[14] WAN B. Study on the transfer of the super league players in winter of the 2016 Season[J]. Bulletinof Sport Science & Technology, 2016, 24(9): 107-109.
[15] ROSSETTI G, CAPRONI V. Football Market Strategies: Think Locally, Trade Globally [C]// IEEE 16th International Conference on Data Mining Workshops (ICDMW). Barcelona, Spain: IEEE, 2016: 152-159.
[16] CHEN C. The model construction of transfer price about football forward players in China football association super league [D]. Beijing: Beijing Sport University, 2017.
[17] YE X S, MA L, CHEN J T, et al. Study on the inter-team gap of players' market value in Chinese football association super league[J]. China Sport Science and Technology, 2017, 53(3): 63-70.
[18] OLIVER M, ALEXANDER S, MARKUS W. Beyond crowd judgments: data-driven estimation of market value in association football[J]. European Journal of Operational Research, 2017, 263(2): 611-624.
[19] PRABHNOOR S, PUNEET S L. Influence of crowd-sourcing, popularity and previous year statistics in market value estimation of football players[J]. Journal of Discrete Mathematical Sciences & Cryptography, 2019, 22(2): 113-126.
[20] KIRSCHSTEIN T, STEFFEN L. Assessing the market values of soccer players-a robust analysis of data from German 1. and 2. Bundesliga[J]. Journal of Applied Statistics, 2019, 46(7): 1336-1349.
[21] ZHAO Y. Analysis of professional soccer player transfer market based on complex network theory[D]. Nanjing: Southeast University, 2018.
[22] IMAN B, SEYED M R. A novel machine learning method for estimating football players' value in the transfer market[J]. Soft Computing, 2020, 25(10): 2499-2511.
[23] HUO D. Evaluation of the value of basketball players based on wireless network and improved Bayesian algorithm[J]. EURASIP Journal on Wireless Communications and Networking, 2020, 236(9): 1-11.
[24] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.
[25] SONG L L, WANG S H, YANG C, et al. Application research of improved XGBoost in imbalanced data processing[J]. Computer Science, 2020, 47(6): 98-103.
[26] LI B S, LI L Z, SUN Y, et al. Intranet defense algorithm based on pseudo boosting decision tree[J]. Computer Science, 2018, 45(4): 157-162.
[27] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 4765-4774.
[28] CHEN Y W, LIN C J. Combining SVMs with various selection strategies[J]. Studies in Fuzziness and Soft Computing, Berlin: Springer, 2008: 315-324.

**LIAO Bin**, born in 1986, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include deep learning, data mining and big data computing model.

**WANG Zhi-ning**, born in 1994, postgraduate. His main research interests include machine learning and big data.

(责任编辑:何杨)