# 同济大学 2018 年数学建模竞赛 C 题

# 中超联赛

# 中超联赛结果预测与实力分析

## 摘要

本文在过去两年480场中超比赛数据及球队表现评分的基础上，利用梯度提升回归树（GBRT）[2]建立了中超联赛的比赛结果预测模型，并对第九轮对战结果进行预测。同时结合球员个人场均表现数据建立了球员实力的评估模型，在球员实力评估模型的基础上对球队实力进行了评估排名，利用球队实力排名评估赛事进行过程中利用积分排名的合理性。最后利用胜负预测模型所得到的影响比赛胜负的多项指标，分析外援在比赛时的各项表现指标对球队表现的影响重要程度，并基于所影响较大的表现指标对外援对球队打法的影响进行了分析。具体研究的内容可以分为以下几个方面:

首先，基于梯度提升回归树（GBRT）[2]建立中超联赛比赛结果预测模型，同时得到球队各项表现指标对比赛结果的影响重要程度。利用爬虫获得中超联赛16、17赛季480场及18赛季目前已进行的8轮64场比赛数据[6]，包括主客双方射门次数、射正次数、禁区内射门次数、准确传球次数、控球率、扑救次数、抢断次数等客观真实指标，利用GBRT回归模型进行训练，选定最佳训练参数，并利用中超18赛季前八轮比赛结果进行预测结果检验。最后运用训练好的模型预测中超联赛第九轮的对局结果，其预测对局结果为(只列出主场球队):

大连一方(负)，广州富力(平)，重庆斯威(胜)，华夏幸福(胜)，
上海上港(胜)，北京人和(负)，天津泰达(胜)，江苏苏宁(胜)。

其次，基于线性回归建立球员实力评估模型并评估赛事进行过程中利用积分排名的合理程度。球员实力评估模型将sofaScore[6]中1617年联赛球员场均数据以及其攻击性、创造性、防御性、战术性、技术性五项评分进行线性回归。利用17、18赛季球员场均数据评估分别得到当赛季球员五项实力评分。以球员评分为基础分别评估球队在17年与18年的实力排名，并与当年球队积分排名进行比较。最后结果为17年球队实力与球队积分排名相似，18年前八轮实力与积分会有差异但整体趋势不变。结论为在赛事进行过程中利用积分排名会与真实情况有所差异但随着赛事进行会趋向合理。

最后，利用第一题预测模型中求得的球队比赛表现指标对预测结果影响的重要程度，并利用懂球帝[5]中球员榜单的18赛季个人表现数据，计算外援对这些重要指标的贡献度来分析外援对球队实力的影响。将外援贡献较大的各项指标划分为进攻、防守，分析球队在这两个属性上不同的打法风格和属性值大小的关系，得出外援个人关键指标对球队打法的影响程度。

**关键词**：足球 中超联赛 比赛预测 梯度回归树 线性回归 特征重要度

# 一、 问题重述

足球是当今世界上开展最广、影响最大的体育项目，被人们认为成世界第一运动。而造成足球运动为全世界球迷狂热的重要原因便是足球比赛结果的偶然性,因为在足球的世界里没有绝对的强队也没有绝对的弱队。强如欧冠冠军皇家马德里也有被刚从西乙升级过来的赫罗纳击败的经历。在足球比赛里不到最后一刻，比赛的结果便很难预知。而近年来随着中超联赛的系统化与制度化，中超联赛球迷的人数也与日俱增。球迷对足球比赛的关注度不断提高,对中超比赛结果胜负的预测的需求也越来越多。

中超比赛参赛球队数固定在16支，每赛季一共有30轮的比赛,但由于足球比赛结果的偶然性，在30轮赛事进行的过程中仅用积分来评价球队真实实力的合理性程度就被受到质疑,在联赛进行的过程中有很多不确定的因子。例如近来中超联赛已经完成了一共八个回合的比赛,在八轮对决后球队并没有完成对于其它15个球队的完整对决。此时便出现了有的球队只参加了与较弱球队的对决而没有参加较强球队的对决的情况,这是否导致球队积分虚高也只是一个值得探究的问题。

中超联赛允许每个球队在每场比赛中上场三个外援(外国球员),外援的发挥往往在一定程度上决定了赛季各队的走向，而外援的表现则又跟教练战术、整体磨合、出场时间和本土球员的支持存在很大关系。本赛季有上港外援奥斯卡被评为最佳攻击型中场情况也有中超身价最高球员卡拉斯科表现不佳无法带领大连一方走出降级区的情况。如何评估外援对一个球队实力的影响以及对一个球队打法的影响也是一个问题。

由上可得，待解决的问题分为以下三个部分:

1. 利用赛事数据建立对中超赛事结果的预测模型,并给出第九轮中超赛事的预测结果。
2. 依据前八轮比赛数据建立对16支球队的实力的评估模型，并评估赛事进行中利用积分排名的合理程度。
3. 建立中超外援在球队重要程度的评估模型。

# 二、问题分析

## 2.1 问题1：对于第九轮胜负判断的分析

影响比赛胜负的因素十分复杂,包括比赛双方的实力因素，球队战术，比赛主客场，队员士气，天气因素等等。其中比赛双方的实力与球队战术，比赛主客场可以归纳为决定比赛胜负的主要因素。

而这三个因素都可以从一场比赛中主客场双方的控球率，射门次数，传球数，传球成功率等比赛数据统计中体现。由于题目中未提供比赛的数据统计,转而采用sofa score[6]提供的过去两年中超比赛480场的对局数据,对数据进行预处理后，基于Python的Scikit-learn库[4]，使用梯度提升回归树（GBRT）[2]回归分析得到模型。根据球队前八轮对局数据分析得到球队第九轮对局的数据，建立模型对第九轮不同场合比赛结果进行预测。

### 2.2 问题 2：基于评价体系建立球队的名次并与官方排名相比较

官方排名是依据官方积分标准对每场比赛进行记分统计后得到的，在每个赛季的比赛进行过程中，天气、球员身体健康、心情等大量不确定因素都能影响球队的胜负以及积分排名的结果。客观来讲，比赛胜负的结果不一定能真实代表这支球队的一般实力，但球队由球员组成，球员在球场上的实力表现决定了球队在对决时的实力表现。基于球队上场的队员实力对球队实力进行评估，能够更有效地反映一支球队的真实实力。

利用Sofa score[6]网站上所提供的大部分球员的信息，包括场均射门、过人成功率等能够直接反映其球员水平的技术性指标，以及进攻、防守、创造力、战术、技术这五个综合指标。基于网站提供的数据已经消除了其相关性的假设，利用技术性指标和综合性指标得到球队的完整数据，用回归方程模拟五个综合性指标与球员在场上的表现之间的关系，并计算出球员的综合素养，利用线性回归寻找关系。利用最终得出的18赛季球员综合性指标，计算球队的综合实力并进行排名，与官方排名进行比较，判断赛事进行中，利用积分排名的合理程度

### 2.3 问题 3：外援如何影响球队的综合实力和打法

一个球队的综合实力是由组成这支球队的队员实力决定的，比赛时球员的表现最终影响了球队的综合表现。利用问题1中得到的胜负预测模型，可以求出在取得预测结果时球队各表现指标对预测结果影响的重要程度，从外援对重要指标的贡献程度上就能体现出外援是如何通过影响某些关键表现指标来影响球队的最终表现和综合实力的。

一般来说，一个球队的打法风格是体现在某些关键指标的，如进攻型打法往往会提高单场比赛的射门次数，防守型打法则会提高抢断次数等。根据外援对关键指标的影响，可以进一步总结出外籍球员关键指标所对应的球队比赛风格的影响，并结合whoScore[6]中球队打法的匹配程度来建立模型评估球队外援是如何影响球队打法。

## 三、模型假设与符号说明

### 3.1模型假设：

1. 忽略教练、心情、天气等因素；
2. sofa score[6]中的球员的评价指标是相互独立的
3. 由球员的实力评估球队的实力时不考虑其他因素影响

4. 球员的实力可以完全由5个指标所体现，并且5个指标不相关

**3.2符号说明：**

| 模型 4.1（梯度提升回归树（GBRT）模型） | | |
|---|---|---|
| 符号 | 含义 | 备注 |
| $Tset$ | 训练集 | |
| $(x_i, y_i)$ | 训练集样本 | $i=1,2,3,\cdots,m$ |
| $L$ | 损失函数 | |
| $T$ | 迭代次数 | |
| $r_{ti}$ | 负梯度 | |
| $R_{tj}$ | 叶子节点区域 | $j=1,2,\cdots,J$ |
| $c_{tj}$ | 最佳拟合值 | |
| $f_t(x)$ | 强学习器 | |
| $Home\_val$ | 主场球队实力值 | |
| $Away\_val$ | 客场球队实力值 | |
| $differ$ | 球队实力差值 | 取值 0.35，此值是保证能在过去比赛中找到至少两场类似比赛的最低球队实力差值 |
| $Home\_old\_value$ | 过去比赛中主场球队实力值 | |
| $Away\_old\_value$ | 过去比赛中客场球队实力值 | |
| $N$ | 相似对局数 | |
| $result$ | 对局结果 | 取值-1，0，1。-1 为主场球队负，0 为平，1 为正 |
| $X$ | 相似对局数据矩阵 | $X=\begin{bmatrix} y_1 & \cdots & y_{40} \\ z_1 & \cdots & z_{40} \\ \vdots & \vdots & \vdots \end{bmatrix}$ |
| $Y$ | 新的 64 局对局数据 | |
| $predicting\_set$ | 预测数据集 | |
| $training\_set$ | 训练数据集 | |
| $test\_set$ | 测试数据集 | |

| 模型 4.2（球员能力指标评估与球队实力评估模型） | | |
|---|---|---|
| *Attacking* | 攻击性 | 最高 100 |
| *Creativity* | 创造性 | 最高 100 |
| *Defending* | 防御性 | 最高 100 |
| *Tactical* | 战术性 | 最高 100 |
| *Technical* | 技术性 | 最高 100 |
| *matchdata* | 球员场上表现（射门、犯规） | 1*15 |
| $Exp\_player\_value$ | 期望球员指标(A、C、D、Te、Ta) | 1*5 |
| $y_i$ | 第 i 位球员的 matchdata 中的实际值 | 包含 5 个维度 |
| $\hat{y}_i$ | 根据第 i 位球员的 matchdata 计算出的理论值 | 包含 5 个维度 |
| $x_i$ | 赛场表现，来自第 i 位球员的 matchdata | |
| $r$ | 残差 | |
| $\alpha$ | 置信水平 | 取值 0.05 |
| $\omega$ | 线性回归的系数 | |
| $A$ | 球队 attacking 指标获得的分数 | |
| $C$ | 球队 creativity 指标获得的分数 | |
| $D$ | 球队 defending 指标获得的分数 | |
| $Ta$ | 球队 tactical 指标获得的分数 | |
| $Te$ | 球队 technical 指标获得的分数 | |
| $Gk$ | 球队守门员获得的分数 | |

# 四、模型建立与分析预测

## 4.1 模型 1：比赛结果预测模型

### 4.1.1 数据准备

从历年比赛的情况来看，两支球队进行比赛受大量因素的影响，最主要因素为比赛时的主客双方球队整体实力，其次为球队在比赛时的表现，包括控球率、射门次数、射正次数、射偏次数、被封堵射门次数、禁区内射门次数、禁区外射门次数、扑救次数、角球数、传球数、准确传球数、越位数、犯规数、黄牌数、一对一对抗成功数、头球数，以及各个球队在主客场同一指标的不同表现情况。

以 2018 年赛季第 8 轮 4 月 29 日长春亚泰对阵上港这一场比赛为例，从实力上来说，虽然比赛进行时，保持当赛季不败且位居积分榜第一的上港球队综合实力要比位居积分榜第九位的长春亚泰强，但是上港是客场作战，且上述 20 项表现指标中，上港占优的只有 8 项，在各种因素综合影响下，最终得到了长春亚泰 2-1 击败上港的结果。

基于 3.1 中的基本假设，利用 Charles 抓取 SofaScore App[6]中的比赛数据请求，分析请求并获取其接口链接，利用网络爬虫抓取 16、17、18 三个赛季中已发生的所有中超比赛技术统计数据，同时抓取 16、17 年经过 30 轮比赛后的各参赛球队评分、18 年经过目前 8 轮比赛后的各参赛球队评分。数据数量见表 4.1-1 单条数据所含的指标类型见表 4.1-2。

表 4.1-1 抓取到的数据条目总数

| 赛季 | 已进行轮数 | 赛季是否结束 | 比赛技术统计数量 | 球队数量 | 球队评分数量 |
|---|---|---|---|---|---|
| 2016 | 30 | 是 | 240 | 16 | 16 |
| 2017 | 30 | 是 | 240 | 16 | 16 |
| 2018 | 8 | 否 | 64 | 16 | 16 |
| 合计 | 68 | NA | 544 | 48 | 48 |

表 4.1-2 抓取到的技术统计各项指标

| Accurate_passes% | Accurate_passes | Aerials_won | Ball_possession | Blocked_shots | Corner_kicks | Counter_attacks | Duels_won | Fouls | Goalkeeper_saves |
|---|---|---|---|---|---|---|---|---|---|
| 准确传球率 | 准确传球数 | 争抢成功 | 控球率 | 被封堵射门 | 角球 | 反攻次数 | 1 对 1 成功 | 犯规次数 | 扑救次数 |
| Offsides | Passes | Red_cards | Shots_inside_box | Shots_off_target | Shots_on_target | Shots_outside_box | Total_shots | Yellow_cards | value |
| 越位次数 | 传球数量 | 红牌数量 | 禁区内射门 | 射偏次数 | 射正次数 | 禁区外射门 | 射门次数 | 黄牌数量 | 球队实力值 |

### 4.1.2 模型分析：基于决策树模型的梯度提升回归树（GBRT）[2]回归算法

**1. 决策树[4]理论基本介绍：**

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。决策树代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。

**2. CART 回归树：**

CART 回归树的目的是通过 CART 算法找出一组基于树的回归方程来预测目标变量。

决策树的构建：不同的算法使用不同的指标来定义“最好”，在这里通常情况下选用 GINI 指数，生成原生的过拟合决策树，之后使用代码中的错误率降低剪枝得到回归树

### 3. GBRT 回归算法[4]模型：

输入是训练集样本 $Tset=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots,(x_m,y_m)\}$，最大迭代次数 $T$ ，损失函数 $L$

输出是经过梯度提升回归树（GBRT）算法优化之后的强学习器 $f(x)$

1） 初始化弱学习器

$$f_0(x)=\arg\min\sum_{i=1}^{m}L(y_i,c)$$

2）对迭代轮数

$$t=1,2,3,\cdots,T \quad \text{有}$$

a) 对样本 $i=1,2,3,\cdots,m$ 计算负梯度

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$$

利用 $(x_i,r_{ti})\quad(i=1,2,3,\cdots,m)$ 拟合一颗 CART 回归树，得到第 $t$ 颗回归树，其对应的叶子节点区域为 $R_{tj}$， $j=1,2,\cdots,J$ 。其中 $J$ 为回归树 $t$ 的叶子节点的个数。

b) 对叶子区域 $j=1,2,\cdots,J$，计算最佳拟合值

$$c_{tj}=\arg\min\sum_{x_i\in R_{tj}}L(y_i,f_{t-1}(x_i)+c)$$

c) 更新强学习器

$$f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$$

2） 得到强学习器 $f(x)$ 的表达式

$$f(x)=f_T(x)=f_0(x)+\sum_{t=1}^{T}\sum_{j=1}^{J}c_{tj}I(x\in R_{tj})$$

这里的强学习器就是使用 16-17 年比赛数据训练得到的回归函数，用于计算预期比赛结果

### 4.1.3 变量选择：

**1. 结合有关文献[1、3]与对中超联赛致胜因素分析,将比赛训练变量划分为主客场双方的数据：**

2016 与 2017 共 480 对局结果显示，受到比赛环境变化，球员心理变化，观众影响与赴客场劳累程度等因素的影响，面对相同的对手而主客场不同每只球队的比赛数据都会有所区别并影响比赛结果,现将主客场因素考虑在内，将变量预处理得到的 4.1-2 中数据划分为主数据与客方数据:

主场方:

表 4.1-3 主场方技术统计各项指标

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Accurate_passes%_home | Accurate_passes_home | Aerials_won_home | Ball_possession_home | Blocked_shots_home | Corner_kicks_home | Counter_attacks_home | Duels_won_home | Fouls_home | Goalkeeper_saves_home |
| Offsides_home | Passes_home | Red_cards_home | Shots_inside_box_home | Shots_off_target_home | Shots_on_target_home | Shots_outside_box_home | Total_shots_home | Yellow_cards_home | value_home |

客场方:

表 4.1-4 客场方技术统计各项指标

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Accurate_passes%_away | Accurate_passes_away | Aerials_won_away | Ball_possession_away | Blocked_shots_away | Corner_kicks_away | Counter_attacks_away | Duels_won_away | Fouls_away | Goalkeeper_saves_away |
| Offsides_away | Passes_away | Red_cards_away | Shots_inside_box_away | Shots_off_target_away | Shots_on_target_away | Shots_outside_box_away | Total_shots_away | Yellow_cards_away | value_away |

**2. 训练集与测试集选择：**

梯度提升回归树（GBRT）的训练数据集是 1617 赛季 60 轮次共 480 场对局的数据统计，如表 4.1-3 与表 4.1-4 所示主客场数据共有 40 个维度。同时每场对局结果用 *result* 表示,其取值为 1，0，-1，其含义分别为主场胜，主场平，主场负。

将 18 赛季前 8 轮共 64 场的对局数据作为测试集，考虑不同实力球队单次对局数据的偶然性，现将 64 场的数据进行均分处理:

将对局双方的 sofa Score[6]评分作为基准,在 1617 赛季 480 场对局中寻找相似对局,设主客双方的评分分别为 $Home\_val$ 与 $Away\_val$,设 480 场中的 sofa Score 主客评分为 $Home\_old\_value$, $Away\_old\_value$,设 $differ$ 为 0.35（此值是保证能在过去比赛中找到至少两场类似比赛的最低球队实力差值）,利用 Python 遍历 480 场的对局数据，其相似对局满足条件公式为:

$$\left|Home\_old\_val - Home\_val\right| \le differ \&\& \left|Away\_old\_val - Away\_val\right| \le differ$$

若 64 场测试数据中出现任意场次找不到对局的情况则令 differ 每次递增 0.01,继续查找相似对局，直至 64 场测试数据都找到相似对局为止。

在寻找到相似对局数据后，将相似对局数据取均值而得到新的 64 场对局数据，设每一场的相似对局有 $N$ 个，

$$X = \begin{bmatrix} y_1 & \cdots & y_{40} \\ z_1 & \cdots & z_{40} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

为多个相似对局 40 个维度的矩阵，设 Y 为新的 64 局对局数据，其处理公式如下所示:

$$Y_i = \frac{(\sum_{j=1}^{N} x_j)}{N} \qquad 1 \le i \le 64$$

以前八轮第 12 场比赛数据为例,均分处理前:

| Accurate_ passes_%_ away | Accurate_ passes_%_ home | Accurate _passes_ away | Accurate _passes_ home | Aerial s_won_ away | Aerial s_won_ home | Ball_pos session_ away | Ball_pos session_ home | ..... ..... ..... | ..... ..... ..... |
|---|---|---|---|---|---|---|---|---|---|
| 0.79 | 0.88 | 293 | 530 | 11 | 15 | 0.39 | 0.61 | ..... | ..... .. |

均分处理后:

| Accurate_ passes_%_ away | Accurate_ passes_%_ home | Accurate _passes_ away | Accurate _passes_ home | Aerial s_won_ away | Aerial s_won_ home | Ball_pos session_ away | Ball_pos session_ home | ..... ..... ..... | ..... ..... ..... |
|---|---|---|---|---|---|---|---|---|---|
| 0.768571429 | 0.837142857 | 213.5714286 | 304.4285714 | 10.71428571 | 9.142857143 | 0.42 | 0.58 | ..... | ..... .. |

### 4.1.4 模型训练与效果评估

#### 1. 训练参数选择

利用 4.1.3 得到的*training _ set*作为输入用于训练梯度提升回归树（GBRT）模型的训练集，得到的*test _ set*作为测试模型效果的测试集。利用 Sklearn[2]的 GridSearchCV 函数，在 4.1-5 参数列表中，得出表中训练效果最好的各项参数，见表 4.1-6。

表 4.1-5 参数列表

| 参数名 | 意义 | 值 | | | | |
|---|---|---|---|---|---|---|
| n_estimators | 迭代次数 | 500 | 1000 | 2000 | 4000 | |
| max_depth | 单个回归估计器的最大深度 | 2 | 3 | 5 | 8 | 12 |
| min_samples_split | 用来划分一个内部节点的最小抽样数 | 2 | 4 | 6 | 8 | |
| learning_rate | 每次迭代对模型训练的贡献率 | 0.01 | 0.05 | | | |

表 4.1-6 最佳参数及评价

| learning_rate | max_depth | min_samples_split | n_estimators | mean | std |
|---|---|---|---|---|---|
| 0.01 | 2 | 2 | 4000 | 0.45726 | 0.08590 |

其余参数使用 Sklearn 的 ensemble.GradientBoostingRegressor 默认参数，如参数表 4.1-7 所示。

表 4.1-7 最终参数表

| 参数名 | 值 | 参数名 | 值 |
|---|---|---|---|
| alpha | 0.9 | min_impurity_decrease | 0.0 |
| criterion | 'friedman_mse' | min_impurity_split | None |
| init | None | min_samples_leaf | 1 |
| learning_rate | 0.01 | min_samples_split | 2 |
| loss | 'ls' | min_weight_fraction_leaf | 0.0 |
| max_depth | 2 | n_estimators | 4000 |
| max_features | None | presort | ‘auto’ |
| max_leaf_nodes | None | random_state | None |
| subsample | 1.0 | verbose | 0 |
| warm_start | False | | |

## 2. 模型表现

利用 4.1-7 最佳参数表训练得到的梯度提升回归树（GBRT）模型，绘制随迭代次数不断增加时结果估计值与结果真值之差平方的期望值(MSE)，如图 4.1-1 所示。

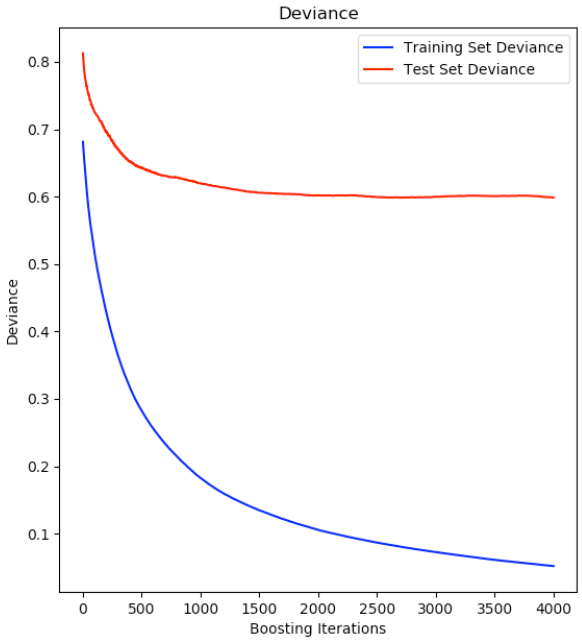


图 4.1-1 MSE 随迭代次数变化图

可以看出，随着迭代次数的增加，模型基本收敛，由于足球比赛的不确定因素较多，故测试集收敛后的 MSE 要比训练集收敛后的 MSE 高。

对 2018 赛季前八轮发生的 64 轮比赛结果进行预测，结果如表 4.1-8 所示。

表 4.1-8 2018 前八轮预测结果比对

| 回归结果 | 胜负判定 | 真实结果 | 回归结果 | 胜负判定 | 真实结果 | 回归结果 | 胜负判定 | 真实结果 |
|---|---|---|---|---|---|---|---|---|
| 0.369094852 | -1 | -1 | 0.272929866 | 1 | 1 | 0.186065642 | 1 | 1 |
| 0.606663982 | 1 | 0 | 1.004469826 | 1 | 1 | 0.995667232 | 1 | 1 |
| 0.964009619 | 1 | -1 | -0.200360552 | -1 | -1 | 0.554717274 | 1 | 1 |
| 0.566564452 | 1 | 1 | 0.631764956 | 1 | 1 | 0.338854861 | 1 | 0 |
| 1.144812938 | 1 | 1 | 0.138095747 | 1 | -1 | 0.318358466 | -1 | -1 |
| 0.272929866 | 1 | 0 | 0.525556967 | 1 | 0 | 0.458029966 | 1 | 0 |
| 0.242527171 | -1 | -1 | 0.762693387 | 1 | 1 | 0.120085887 | -1 | -1 |
| 0.859552922 | 1 | 1 | 0.630326862 | 1 | 1 | 0.371253168 | -1 | -1 |
| 0.77510689 | 1 | 1 | 0.639003089 | 1 | -1 | 0.369094852 | -1 | 0 |
| 0.458029966 | 1 | -1 | 0.745679191 | 1 | 1 | 0.799851139 | 1 | 1 |
| 0.834292925 | 1 | 1 | 0.136776346 | 1 | -1 | 0.704659951 | 1 | 0 |
| 0.847368588 | 1 | 1 | 0.77438865 | 1 | 0 | 0.463076748 | 1 | 0 |
| 0.851643778 | 1 | -1 | -0.269548078 | -1 | 0 | 0.138095747 | 1 | 1 |
| 0.218075465 | -1 | -1 | 0.458029966 | 1 | 1 | 0.102037384 | 1 | -1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.584466265 | 1 | -1 | -0.567672317 | -1 | -1 | 0.649430653 | 1 | 1 |
| 0.132507632 | 1 | 1 | 0.617093997 | 1 | 1 | 0.019440235 | 0 | 1 |
| 0.785403698 | -1 | -1 | -0.218075465 | -1 | -1 | 1.004469826 | 1 | 1 |
| 0.278611356 | 1 | -1 | -0.982382731 | -1 | -1 | 1.074945578 | 1 | 1 |
| -0.07912118 | 0 | 1 | 0.334334886 | 1 | 0 | 0.639003089 | 1 | 1 |
| 0.012368244 | 0 | 0 | 0.03121996 | 0 | 0 | 0.575009899 | 1 | -1 |
| 0.473321359 | -1 | -1 | 0.611476215 | 1 | 1 | | | |
| 0.02094987 | 0 | -1 | 0.316222585 | 1 | 1 | | | |

$$\text{总预测准确率}=\frac{\text{预测成功数}}{\text{总比赛场数}}=0.625$$

对于胜负平三个分类指标来看，每个类的精确度，召回率，F1 值如表 4.1-9 所示。

表 4.1-9 分类预测评价指标

| 类名 | 准确度 (Precision) | 召回率 (recall) | F1 值(f1-score) | 出现次数 (support) |
|---|---|---|---|---|
| 负（主场球队） | 0.86 | 0.52 | 0.65 | 23 |
| 平（主场球队） | 0.40 | 0.15 | 0.22 | 13 |
| 胜（主场球队） | 0.58 | 0.62 | 0.59 | 28 |
| 加权平均 | 0.64 | 0.62 | 0.59 | 64（合计） |

由此可知：

- **从预测结果来看：**

目前本模型对于一场比赛中预测主场球队输掉比赛的准确度较高，即当模型给出预测为输的结果时，该球队在真实对决中输掉的可能性较高。但当模型给出预测为平的结果时，该球队在真实对决中确实打平的概率只有 40%，也就是说，当球队表现、球队实力相当时，比赛的结果受运气在内的多种复杂因素的影响因而难以进行预测。

- **从实际结果来看：**

所有对决结果为胜或负的比赛中，本模型能正确判断胜负的比例大于 50%，其中对胜的判断能力较好，比例超过 60%。但对于所有真实结果为平的比赛，能够判断正确的比例较低，也就是说，对于所有对决结果为平的比赛，其结果在赛前是难以预料的。

### 4.1.5 中超 2018 赛季第 9 轮比赛结果预测

基于 3.1 中的基本假设，不考虑主球员心情、天气等因素的影响，利用 4.1.3 得到的第 9 轮比赛 *predicting _ set* 预测数据集和 4.1.4 得到的最佳预测模型，对第 9 轮比赛情况进行预测，预测结果如表 4.1-10 所示。

表 4.1-10 中超 18 赛季第 9 轮比赛结果预测表

| 预测结果 | 预测胜负(主场) | 主场 | 客场 |
|---|---|---|---|
| -0.73996552 | 负 | 大连一方 | 广州恒大 |
| 0.03762937 | 平 | 广州富力 | 上海绿地申花 |
| 0.1764996 | 胜 | 重庆斯威 | 天津权健 |
| 0.67013998 | 胜 | 华夏幸福 | 河南建业 |
| 0.25242681 | 胜 | 上海上港 | 北京中赫国安 |
| -0.55402812 | 负 | 北京人和 | 山东鲁能 |
| 0.17469777 | 胜 | 天津泰达 | 贵州恒丰 |
| 0.68672041 | 胜 | 江苏苏宁 | 长春亚泰 |

## 4.2 模型 2:球员能力指标评估与球队实力评估模型：

### 4.2.1 数据准备：

球员每次进行一场比赛，不仅球队会产生一组数据，球员自己也会拥有自己的当场比赛的赛场上的表现，包括：传球成功率、助攻、创造射门机会、错失射门机会，单场过人成功率、犯规、被犯规、进球平均数、断球、成功带球、阻挡、总射门次数，等等数据，但是因为其量纲并不相同，并且相关性难以判断，直接选取 sofa score[6]所计算得出的球员 5 维评价指标建立回归。

其中球员真实能力评价指标如表 4.2-1 所示：

表 4.2-1 球员真实能力评价指标表

| Actual_player_value 能力指标评价(16\17\18) | | |
|---|---|---|
| 符号 | 含义 | 备注与行数列数 |
| Attacking | 攻击性 | 最高 100 |
| Creativity | 创造性 | 最高 100 |
| Defending | 防御性 | 最高 100 |
| Tactical | 战术性 | 最高 100 |
| Technical | 技术性 | 最高 100 |
| | | |
| matchdata | 球员场上表现（射门、犯规） | 1*15 |
| r | 残差 | |
| Exp_player_value | 期望球员指标(ACDTeTa) | 1*5 |

数据均来自 sofa_score[6]中球员能力表中的数据，部分球员的数据在网络中无法找到，暂时不考虑在内，删去缺省数据，因为问题仅仅涉及到 18 年的比赛，仅仅考虑在 18 年中超赛事中出现的球队的球员的能力指标与场上表现的数据，使用的数据包括：

1. 16 年上场的所有有比赛记录的球员（并且在 18 年上场）的信息一共 152 名，1*21 列数据（其中 5 维指标，15 列场上的表现，以及球员对应的球队编号）
2. 17 年上场的所有有比赛记录的球员（并且在 18 年上场）的信息一共 187 名，1*21 列数据（其中 5 维指标，15 列场上的表现，以及球员对应的球队编号）
3. 18 年上场的所有有比赛记录的球员的信息一共 183 名，1*21 列数据（其中 5 维指标，15 列场上的表现，以及球员对应的球队编号）

例：表 4.2-2

表 4.2-2 单个球员评价指标示例表

| Attacking | Creativity | Defending | Tactical | Technical |
|---|---|---|---|---|
| 84 | 69 | 21 | 54 | 82 |

| accuratePassesPerGame | assists | bigChanceCreated | bigChanceMissed | duelsWonPerGame | fouls | goalsAverage | goalsFrequency |
|---|---|---|---|---|---|---|---|
| 31.54 | 0.467 | 4.53 | 2.6 | 10.99 | 1.9 | 0.313 | 369 |
| interceptionsPerGame | lostBallsPerGame | successfulDribblesPerGame | tacklesPerGame | team_id | totalClearancesPerGame | totalShotsPerGame | wasFouled |
| 0.24 | 3.313 | 5.58 | 0.67 | 41537 | 0.526 | 3.84 | 3.52 |

### 4.2.2 模型分析：

SofaScore[6]上的评分可以体现球员的综合素养，尝试通过回归的方式复现评分的结果，能够得出球队所有球员的水平，进而考虑球队的客观实力：先使用 matlab 的 regressionlearner 工具箱中的 4 种线性回归进行尝试：

表 4.2-3 中超 18 赛季第 9 轮比赛结果预测表

| RMSE | Attacking-matchdata | Creativity-matchdata | Defending-matchdata | Tactical -matchdata | Technical matchdata |
|---|---|---|---|---|---|
| Linear Regression | **4.96** | **4.46** | **5.17** | **6.03** | **5.00** |
| SVM | 4.99 | 4.50 | 7.49 | 6.07 | 5.04 |
| Tree | 5.95 | 5.67 | 5.23 | 7.06 | 6.02 |

以上结果分别选用了线性回归，支持向量机回归与决策树回归，RMSE（均方根误差）在可接受的范围，可知线性回归的效果已经足够优秀，因此我们选用线性回归模型连接球员评价指标与球员赛场表现。

### 4.2.3 理论准备

表 4.2-4 模型变量说明表

| | |
|---|---|
| $y_i$ | 分别取 Attacking、Creativity、Defending、Tactical、Technical |
| $\hat{y}_i$ | 根据第 i 位球员的 matchdata 计算出的理论值 |
| $x_i$ | 来自第 i 位球员的 matchdata |

在使用表 4.2-4 说明变量的模型中，对于数据集$\{(x_1,y_1),...,(x_m,y_m)\}$，尝试找到使得残差平方和

$$\min_{\omega} \| X\omega - y \|_2^2$$

最小时的$\omega=(\omega_1,\omega_2,\omega_3,...,\omega_p)$，得到线性模型

$$\hat{y}(\omega,x)=\omega_0+\omega_1x_1+...+\omega_px_p \quad ,$$

在 matlab 中使用 regress 函数，取$\alpha$=0.05（置信水平）并使用 rcoplot 命令画出下图：

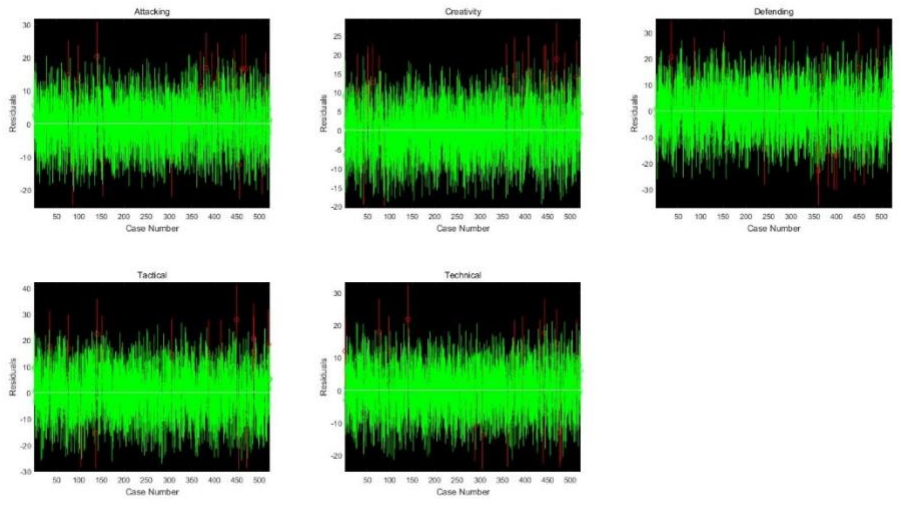


图 4.2-1 线性拟合结果图

图中亮色部分代表落在置信区间内的期望值，线性拟合的情况很好，其中 $\hat{y}(\omega,x)=\omega_0+\omega_1x_1+...+\omega_px_p$，中的线性回归系数 $\omega$ 分别对应 5 项指标有 $\omega_k^{(1)}$，$\omega_k^{(2)}$ $\omega_k^{(3)}$ $\omega_k^{(4)}$ $\omega_k^{(5)}$，$k=0,1,2,3...15$ 其值为：

表 4.2-5 五维评价指标系数表

| $\omega_k^{(1)}$ | $\omega_k^{(2)}$ | $\omega_k^{(3)}$ | $\omega_k^{(4)}$ | $\omega_k^{(5)}$ |
|---|---|---|---|---|
| attacking | Defending | Creativity | Tactical | Techinical |
| 45.79 | 43.22 | 45.38 | 34.11 | 44.15 |
| 0.03 | 0.2 | 0.16 | 0.31 | 0.26 |
| 9.73 | 9.27 | -9.75 | -2.18 | -0.61 |
| 0.14 | 0.55 | 0.15 | -0.16 | 0.36 |
| -0.02 | -0.49 | -0.48 | -0.09 | -0.33 |
| 1.32 | -0.46 | -0.67 | 1.16 | -0.01 |
| -0.27 | -0.25 | 0.41 | 0.59 | -0.25 |
| 9.32 | -5.86 | -7.59 | 4.13 | 4.28 |
| 0 | 0 | 0 | 0 | 0 |
| -2.02 | -1.63 | 3.08 | 2.33 | -1.42 |
| 2.44 | 1.2 | -4.58 | -2.36 | -0.19 |
| -0.41 | 2.23 | 0.99 | -1.08 | 5.14 |
| -1.82 | 0.97 | 2.35 | -1.8 | -0.8 |
| -2.69 | -1.88 | 3.37 | 1.71 | -1.11 |
| 3.9 | 2.56 | -1.05 | 2.92 | 1.49 |
| -0.64 | 0.11 | -0.59 | -0.78 | 0.12 |

对于其他缺省的数据（例如没有 5 维评价指标的队员）可以使用公式

$$\hat{y}(\omega,x)=\omega_0+\omega_1x_1+...+\omega_px_p$$

来分别计算其 5 项能力值。

### 4.2.4 球队能力的评价

由上一步可以得到球队中所有队员的 5 维评价指标，接下来需要据此对球队整体实力进行评估与横向对比。

一个球队的球队实力体现在各个方面，球员的实力直接影响球队的实力，据假设不考虑其他因素，包括教练、战术安排等等，而且此处计算的是球队的绝对实力评价而非比赛中显示的相对实力，基于假设默认 5 维指标是独立的体现在不同的方面并且是相同重要的，因此可以写出公式，将其与守门员实力相加得到总成绩其中：

$$球队实力评价 = A + C + D + Ta + Te + Gk$$

其中 A 代表球队 attacking 指标获得的分数，C 代表球队 creativity 指标获得的分数,D 代表球队 defending 指标获得的分数,Ta 代表球队 tactical 指标获得的分数

Te 代表球队 technical 指标获得的分数,Gk 代表球队守门员获得的分数。

其中 X 代表某一个指标：计算方法为：

$$X\text{指标的得分} = \frac{\text{所有队员}X\text{指标5位最高得分之和}}{25}$$

$$Gk\text{指标的得分} = \frac{\text{守门员指标的5项指标总分}}{25}$$

因为守门员与其他球员的评价方法不同，守门员的 5 项指标分别为：SAV（扑救）ANT（预判）TAC（战术）BAL（处理球）AER（空中能力），每一项对于比赛均十分重要，因此也将其相加。

下表为 17 年部分参与了 18 年比赛的球队（14 支，16 支中有两支被淘汰）的指标评估情况：（因为积分规则的制定积分有先后顺序这里用 0.1 加以区分 52.1 与 52 的积分）

表 4.2-6 17 年球队实力评分与排名表

| 队名 | 评估得分 | 排名 | 球队积分 |
|---|---|---|---|
| 上海上港 | 63.13943 | 1 | 58 |
| 广州恒大淘宝 | 62.47839 | 2 | 64 |
| 广州富力 | 61.54754 | 3 | 52.1 |
| 北京中赫国安 | 61.40816 | 4 | 40 |
| 天津权健 | 61.1854 | 5 | 54 |
| 山东鲁能泰山 | 60.56966 | 6 | 49 |
| 河北华夏幸福 | 60.31437 | 7 | 52 |
| 上海绿地申花 | 60.2604 | 8 | 35 |
| 河南建业 | 59.99105 | 9 | 30 |
| 贵州恒丰 | 58.57496 | 10 | 42 |
| 长春亚太 | 58.47396 | 11 | 44 |
| 天津泰达/亿利 | 56.09616 | 12 | 31 |
| 重庆斯威/力帆 | 55.75588 | 13 | 36 |
| 江苏苏宁易购 | 55.39352 | 14 | 32 |

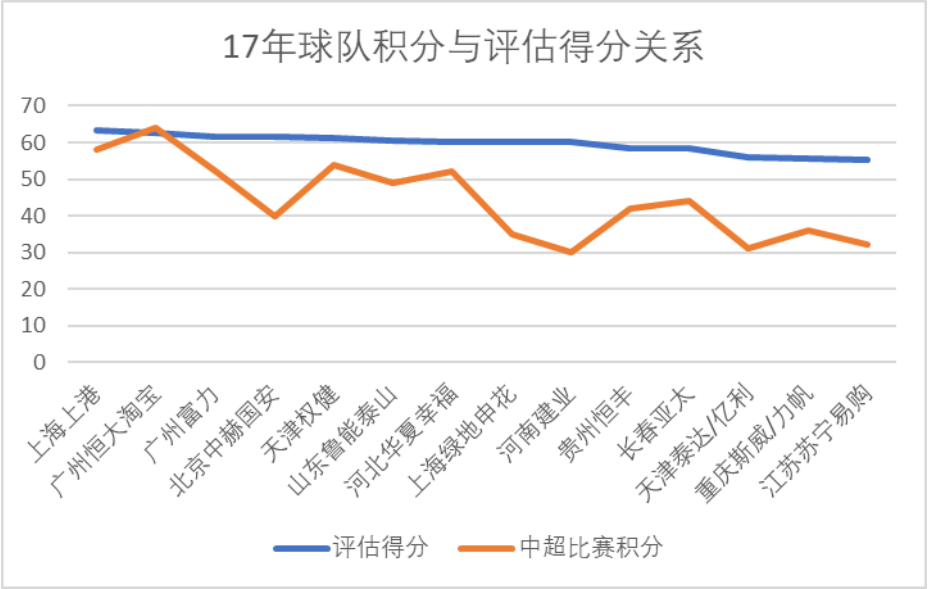


图 4.2-2 17 赛季球队积分与评估得分关系图

从表 4.2-6 及图 4.2-2 可以看出，经过实力评估后得到的评分排名与球队经过 30 轮比赛后得到的积分排名是具有近似的趋势，当前模型使用的实力评估方法能够较为准确地反映球队的真实实力水平。用此模型计算 18 赛季的评估得分（其中大连一方与贵州恒丰的球员数据由 4.1 建立的模型得出，并据此计算球队评估得分）表 4.2-7 。并且绘制关系图如图 4.2-3 所示：

表 4.2-7 18 年球队实力评分与排名表

| 队名 | 评估得分 | 排名 | 球队积分 |
|---|---|---|---|
| 北京中赫国安 | 64.59444 | 1 | 15 |
| 上海上港 | 64.24223 | 2 | 19 |
| 山东鲁能泰山 | 61.99884 | 3 | 18 |
| 上海绿地申花 | 61.87227 | 4 | 14 |
| 江苏苏宁易购 | 61.08144 | 5 | 13 |
| 鲁能泰山 | 61.07654 | 6 | 18 |
| 广州富力 | 61.05918 | 7 | 12 |
| 广州恒大淘宝 | 61.05782 | 8 | 17 |
| 天津权健 | 58.77661 | 9 | 8 |
| 天津泰达 | 58.15069 | 10 | 9 |
| 河南建业 | 57.73379 | 11 | 4 |
| 长春亚泰 | 57.1122 | 12 | 11.3 |
| 重庆斯威 | 56.92601 | 13 | 11.4 |
| 河北华夏幸福 | 54.68245 | 14 | 11.2 |
| 大连一方 | 50.07853 | 15 | 3.1 |
| 贵州恒丰 | 49.87352 | 16 | 3 |

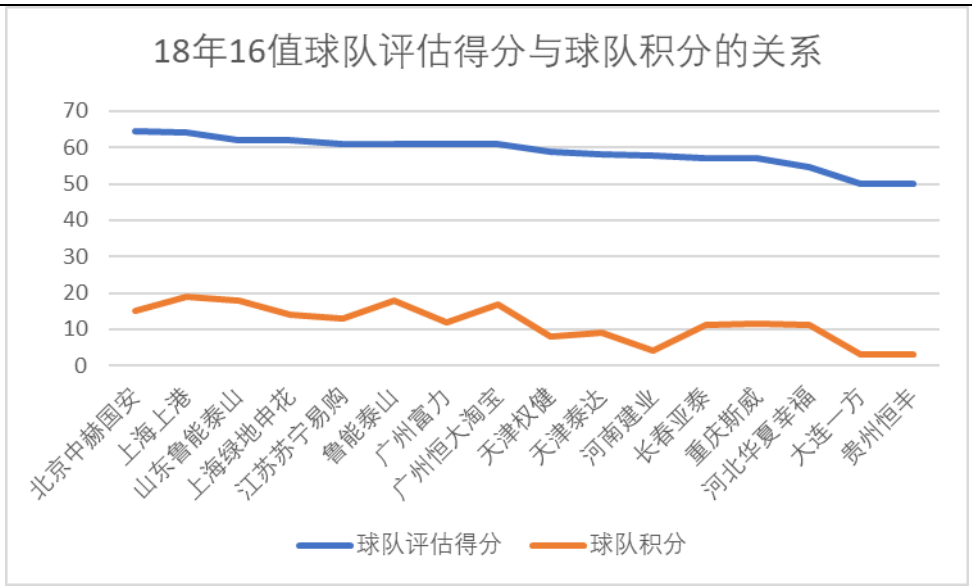


图 4.2-3 18 赛季球队积分与评估得分关系图

球队评估得分的计算方法，兼顾了球队在 5 个方向的素质，并且考量了守门员的能力值，图 4.2-1 中将评估得分自低到高排列，同时将对应的中超比赛排名以相同的球队顺序在图中画出，可以看出两者的趋势相似，但是球队的评估得分不能完全反映出球队的比赛结果，将同一模型套用在 18 年目前进行过的 8 轮比赛，得到的结果也较为明显，随着球队评估得分的上升，球队积分也有上升的趋势，但是因为没有进行所有的比赛，不排除有的球队目前仅仅遇到的是水平较低的球队所以会有积分较高的情况，但是总体上可以体现积分的标准是合理的。具体排名如下：

## 4.3 模型 3：外援影响模型

### 4.3.1 数据准备

对于一场比赛来说，球员的综合表现直接影响球队对决的胜负情况。即便在传统强队与刚晋级的弱队之间，也可能因为双方球员发挥失常或超常发挥而影响比赛胜负。从历年比赛的情况来看，两支球队在比赛时与单一球员有较大关系的表现指标包括射门次数、射正次数、射偏次数、被封堵射门次数、禁区内射门次数、禁区外射门次数、扑救次数、准确传球数、头球数等。
基于 3.1 中的基本假设，利用 Charles 抓取懂球帝 App[5]中的球员榜单请求，分析请求并获取其接口链接，利用网络爬虫抓取 16、17、18 赛季中的球员表现榜单，选取与上述比赛表现指标有关联的指标见表 4.3-1

表 4.3-1 2018 中超球员表现榜各单项指标

| Name | goals | assists | key_passes | shots | shots_on_target | passes | tackles | interception | clearances | saves |
|---|---|---|---|---|---|---|---|---|---|---|

| 指标名 | 射手榜 | 助攻榜 | 关键传球榜 | 射门榜 | 射正榜 | 成功传球榜 | 抢断榜 | 拦截榜 | 解围榜 | 扑救榜 |
|---|---|---|---|---|---|---|---|---|---|---|
| 人数 | 91 | 90 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 20 |

### 4.3.2 外援在比赛表现中的影响力分析

根据 4.1.4 中训练得到的胜负预测模型，可得到预测胜负时各指标在预测结果中的重要程度如图 4.3-1 所示：

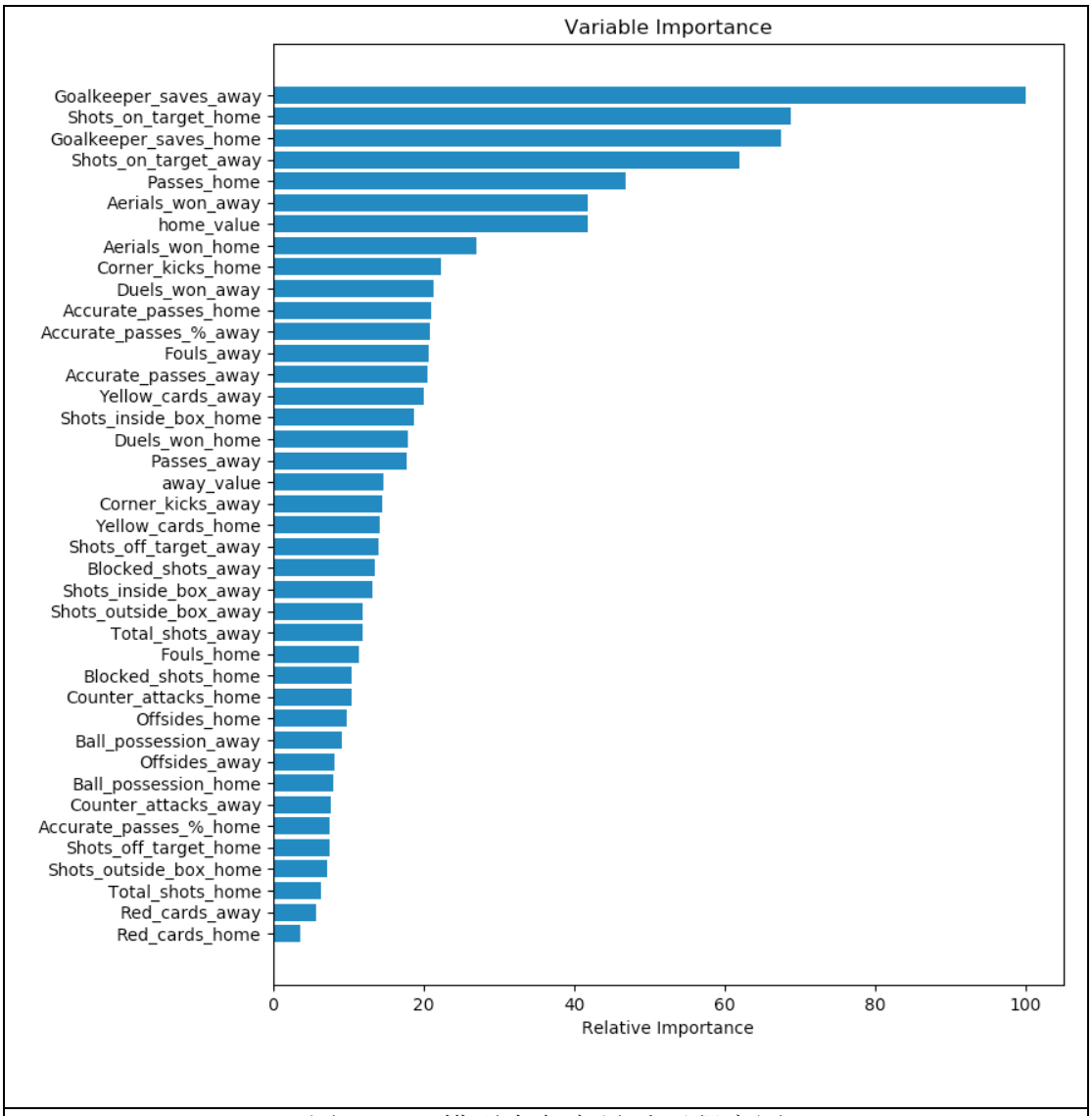


图 4.3-1 模型中各变量重要程度图

从中可以看出，除去主客场球队综合实力值外，排在前 10 的指标分别如表 4.3-2 所示：

表 4.3-2 胜负预测模型中重要程度指标

| Name | Goalkeeper_saves | Shots_on_target | Passes | Aerials_won | Corner_kicks |
|---|---|---|---|---|---|
| 指标名 | 扑救次数 | 射正 | 传球 | 争抢 | 角球 |
| 重要程度 | 1 | 2 | 3 | 4 | 5 |
| Name | Duels_win | Accurate_passes | Fouls | Shots_inside_box | Passes |
| 指标名 | 1 对 1 成功 | 精确传球 | 犯规 | 禁区内射门 | 被封堵射门 |
| 重要程度 | 6 | 7 | 8 | 9 | 10 |

与这 10 个指标对应的榜单如表 4.3-3 所示：

表 4.3-3 胜负预测模型中重要程度指标

| Name | goals | assists | key_passes | shots | shots_on_target |
|---|---|---|---|---|---|
| 指标名 | 射手榜 | 助攻榜 | 关键传球榜 | 射门榜 | 射正榜 |
| 外籍球员人数 | 41 | 32 | 41 | 50 | 47 |
| 外籍球员在榜单所有人数中的占比 | 45% | 35.6% | 41% | 50% | 47% |
| Name | passes | tackles | interception | clearances | saves |
| 指标名 | 成功传球榜 | 抢断榜 | 拦截榜 | 解围榜 | 扑救榜 |
| 外籍球员人数 | 35 | 22 | 19 | 17 | 0 |
| 外籍球员在榜单所有人数中的占比 | 35% | 22% | 19% | 17% | 0% |

表 4.3-4 中超 18 赛季球队中外援数量与占比表

| 外援人数 | 总人数 | 外援占比 |
|---|---|---|
| 69 | 397 | 17.4% |

表 4.3-5 中超外援进球数量占比表

| 外援进球 | 总进球 | 外援进球数量占比 |
|---|---|---|
| 118 | 197 | 59.9% |

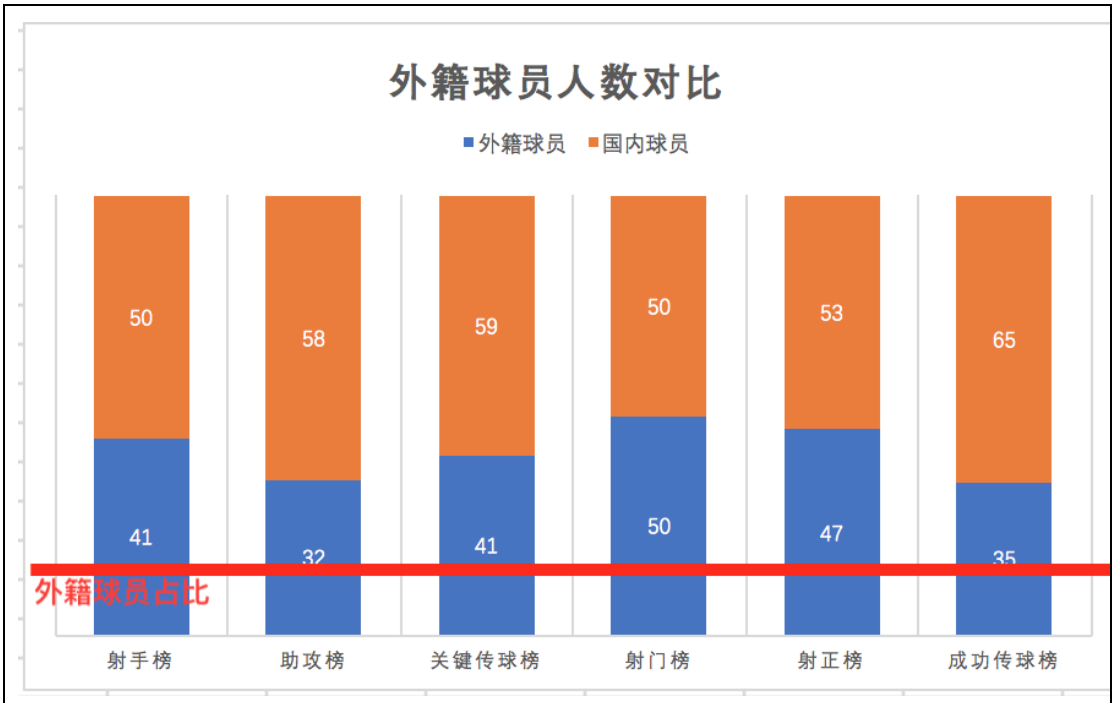


图 4.3-2 外援数量占比图

根据表 4.3-3 所示的结果，在与胜负预测模型中最为重要的前 10 个指标有关的榜单中，有 7 个榜单的外援数量比例都大于外援在所有球员中所占的比例。在与进攻相关的榜单中，外籍球员占比都超过了百分之三十五。

目前中超制定了“3 加 1（亚洲外援）”的政策，每支球队最多拥有四位外援，每次上场不超过三位外援，根据外籍球员占比较高的榜单都与进攻表现相关，且占比均超过了每次上场时外籍球员在上场球员中的最高占比这一现象，且 59.9%的进球都是由外援打入的，可以判断目前外籍球员是通过影响整支球队的进攻能力来提升球队的实力的，外籍球员是球队得分的中坚力量。

### 4.3.3 外援对球队打法影响的分析

基于 whoscored[10]中对球队进攻打法的统一分类，列打法表如表 4.3-6 所示：

表 4.3-6 进攻型打法表

| | |
|---|---|
| 1. 尝试从边路进攻 | 2. 尝试采用传控打法 |
| 3. 进行大量射门尝试 | 4. 尝试使用反越位 |
| 5. 尝试将球控制在对方半场 | 6. 尝试中路渗透进攻 |
| 7. 尝试大范围转移球 | 8. 尝试使用短传进攻 |

判断球队打法风格基于如上简表来总结，但网上数据不够充足，如在每场比赛中没有足够的球队短传统计，对于打法总结简表中的第 8 条便无法概括。

而利用射门次数则可判断球队是否进行大量射门尝试的打法，下表采用了懂球帝统计数据中超球队射门次数的前 7 名队伍：

表 4.3-7 不同球队进攻表现

| 球队 | 射门次数 | 射正 | 射正率 |
|---|---|---|---|
| 上海上港 | 111 | 58 | 52% |
| 上海绿地申花 | 92 | 46 | 50% |
| 广州恒大淘宝 | 89 | 37 | 41% |
| 江苏苏宁易购 | 82 | 35 | 43% |
| 北京中赫国安 | 79 | 38 | 48% |
| 重庆斯威 | 76 | 33 | 43% |
| 广州富力 | 75 | 36 | 48% |

以广州恒大淘宝为例，球队的射门次数高居中超第三而射正率则是前七名最低的，而由第一个模型比赛胜负预测模型可知,决定比赛结果的最重要因素便是球队的射正次数，这表示广州恒大淘宝队的打法风格是进行大量的射门来弥补其射门转化率低的劣势。

同时，对于球员打法同样给出一些简单的个人打法描述 ：

表 4.3-8 进攻型打法表

| | |
|---|---|
| 1.将球控制在对方半场 | 2.尝试大范围转移球 |
| 3.得到大量射门机会 | 4.尝试反越位 |
| 5.尝试短传进球 | 6.喜欢直塞进球 |
| 7.尝试远距离传中 | 8.喜欢尝试过人 |

再次以恒大淘宝队为例，其当家外援高拉特射门次数为 22 次,占全队射门次数的 25%,其得到大量射门机会但进球转化率只有 4 个,进球转化率不高，球队需要给予高拉特更多的射门次数来提高球队的总进球数。这是球员影响球队打法的一个举例。

最后总结，外援在中超联赛对于每个球队来说都是球队的核心，在评估外援对球队打法的影响时需要考虑到：

1.对球队打法的定义。

2.对于球员与球队特定统计数据的分析。

以上第二点可以使用统计学的思想进行分析但由于对于球队打法的定义过于宽泛很难建立一个宽泛的模型去统计球员对于所有打法的影响。但通过高拉特与广州恒大淘宝特定数据的分析还是可以得到外援对于一个队伍的打法有很大影响的。

# 五、模型评价

## 5.1 模型的优点

1. 采用梯度提升回归树（GBRT）回归训练得到的比赛胜负预测模型对于比赛胜负预测的准确率较高，整体预测成功率达到了62.5%，同时预测模型对于一场比赛中预测主场球队输掉比赛的准确度更达到了87%。
2. 球员五维实力评估模型与球员近期的表现相符，并且以此为基础对球队实力的评价也是符合实际情况，并合理的评估了赛事进行过程中利用积分评估球队实力的不足与合理性。
3. 使用Python爬虫爬取的比赛数据十分广泛，以第一个模型的训练为例，其采用的训练数据即为两年480场共40多个维度的数据，这也在一定程度上保证了模型训练的准确性。

## 5.2 模型的局限性

1. 采用梯度提升回归树（GBRT）回归训练得到的比赛胜负预测模型对于比赛胜负预测的准确率较高，整体预测成功率达到了62.5%，同时预测模型对于一场比赛中预测主场球队输掉比赛的准确度更达到了87%。
2. 第二个球队实力评估模型在进行数据处理时在17赛季没有添加被降级的两个队伍，对于导致17年分析结果会与真实情况存在一定差异。
3. 由于时间限制，梯度提升回归树（GBRT）的模型没有对训练维度进行更细致的挑选，这意味着预测模型在精准度上还能有进一步提高。

# 六、参考文献

[1] 王溪源，黄迎乒，袁玉玲，中超球队引进外援对俱乐部战绩的影响研究[J]，竞技体育，1674-151X（2016）07-034-03:1-5, 2016 年 7 月 141 期.
[2] Gradient Boosting regression，http://sklearn.apachecn.org/cn/0.19.0/auto_examples/ensemble/plot_gradient_boosting_regression.html#sphx-glr-auto-examples-ensemble-plot-gradient-boosting-regression-py，2018-5-1
[3] 夏飞. BP 神经网络在足球比赛胜负预测中的应用研究[D].重庆师范大学,2017.
[4] 集成方法，http://sklearn.apachecn.org/cn/0.19.0/modules/ensemble.html，2018-5-1
[5] 懂球帝，http://www.dongqiudi.com ，2018-4-30
[6] Football LiveScore-sofaScore.com，https://www.sofascore.com ，2018-4-30
[7] 可能是最专业的足球数据中文网站，http://www.tzuqiu.cc/teams/42/show.do，2018-5-3
[8] 足球，https://zh.wikipedia.org/wiki/足球，2018-5-2
[9] 中超联赛，https://zh.wikipedia.org/wiki/中超联赛，2018-5-2
[10] Football Statistics，https://www.whoscored.com，2018-5-1

# 附录:

## 1. 爬虫数据来源:

### (1)懂球帝球员榜单(2018 年):

https://api.dongqiudi.com/data/v1/person_ranking/0?type=goals&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=assists&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=key_passes&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=shots&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=shots_on_target&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=offsides&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=passes&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=success_passes&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=interceptions&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=tackles&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=clearances&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=fouls&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=fouled&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=red_cards&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=yellow_cards&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=saves&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=appearances&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=time_played&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=high_speed_running_distance&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=foreign_time_played&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=u23_time_played&version=149&refer=person_ranking&season_id=10219

https://api.dongqiudi.com/data/v1/person_ranking/0?type=total_running_distance&version=149&refer=person_ranking&season_id=10219

### (2)sofaScore 球员数据:

上海上港: https://mobile.sofascore.com/mobile/v4/team/41537/players

鲁能泰山:https://mobile.sofascore.com/mobile/v4/team/3371/players

广州恒大淘宝:https://mobile.sofascore.com/mobile/v4/team/24156/players

江苏苏宁易购:https://mobile.sofascore.com/mobile/v4/team/34693/players

北京中赫国安:https://mobile.sofascore.com/mobile/v4/team/3376/players

重庆斯威:https://mobile.sofascore.com/mobile/v4/team/3362/players

上海绿地申花:https://mobile.sofascore.com/mobile/v4/team/3373/players

天津泰达:https://mobile.sofascore.com/mobile/v4/team/3367/players

广州富力:https://mobile.sofascore.com/mobile/v4/team/3375/players

天津权健:https://mobile.sofascore.com/mobile/v4/team/50018/players

长春亚泰:https://mobile.sofascore.com/mobile/v4/team/34694/players

河北华夏幸福:https://mobile.sofascore.com/mobile/v4/team/71564/players

北京人和:https://mobile.sofascore.com/mobile/v4/team/34696/players

河南建业:https://mobile.sofascore.com/mobile/v4/team/34692/players

贵州恒丰:https://mobile.sofascore.com/mobile/v4/team/50017/players

大连一方:https://mobile.sofascore.com/mobile/v4/team/49936/players

(3)其他数据:

每一个球员的数据

https://mobile.sofascore.com/mobile/v4/player/34705/statistics

每一场比赛的 event

https://mobile.sofascore.com/mobile/v4/event/7723723/lineups

2018 赛季所有比赛

https://mobile.sofascore.com/mobile/v4/tournament/652/season/16186/event

2017 赛季所有比赛

https://mobile.sofascore.com/mobile/v4/tournament/652/season/12938/events

2016 赛季所有比赛

https://mobile.sofascore.com/mobile/v4/tournament/652/season/11342/events

## 2. GradientBoostingRegressor 参数选择表:

| learning_rate | max_depth | min_samples_split | n_estimators | std | mean |
|---|---|---|---|---|---|
| 0.01 | 2 | 2 | 500 | 0.07761 | 0.32991 |
| 0.01 | 2 | 2 | 1000 | 0.07306 | 0.39967 |
| 0.01 | 2 | 2 | 2000 | 0.08843 | 0.44409 |
| 0.01 | 2 | 2 | 4000 | 0.0859 | 0.45726 |
| 0.01 | 2 | 4 | 500 | 0.07817 | 0.32955 |
| 0.01 | 2 | 4 | 1000 | 0.07419 | 0.40017 |
| 0.01 | 2 | 4 | 2000 | 0.08807 | 0.44424 |
| 0.01 | 2 | 4 | 4000 | 0.08791 | 0.45259 |
| 0.01 | 2 | 6 | 500 | 0.07633 | 0.32883 |
| 0.01 | 2 | 6 | 1000 | 0.07315 | 0.39974 |
| 0.01 | 2 | 6 | 2000 | 0.08807 | 0.44123 |
| 0.01 | 2 | 6 | 4000 | 0.09286 | 0.4473 |
| 0.01 | 2 | 8 | 500 | 0.0769 | 0.32875 |
| 0.01 | 2 | 8 | 1000 | 0.07418 | 0.4007 |
| 0.01 | 2 | 8 | 2000 | 0.08864 | 0.44212 |
| 0.01 | 2 | 8 | 4000 | 0.09286 | 0.44917 |
| 0.01 | 3 | 2 | 500 | 0.07985 | 0.35232 |
| 0.01 | 3 | 2 | 1000 | 0.07587 | 0.38161 |
| 0.01 | 3 | 2 | 2000 | 0.07857 | 0.39436 |
| 0.01 | 3 | 2 | 4000 | 0.07364 | 0.39357 |
| 0.01 | 3 | 4 | 500 | 0.07925 | 0.3513 |
| 0.01 | 3 | 4 | 1000 | 0.07676 | 0.38471 |
| 0.01 | 3 | 4 | 2000 | 0.07293 | 0.40118 |
| 0.01 | 3 | 4 | 4000 | 0.06923 | 0.40338 |
| 0.01 | 3 | 6 | 500 | 0.07945 | 0.35223 |

| | | | | | |
|---|---|---|---|---|---|
| **0.01** | 3 | 6 | 1000 | 0.07711 | 0.38591 |
| **0.01** | 3 | 6 | 2000 | 0.07769 | 0.40405 |
| **0.01** | 3 | 6 | 4000 | 0.07558 | 0.40379 |
| **0.01** | 3 | 8 | 500 | 0.07918 | 0.35096 |
| **0.01** | 3 | 8 | 1000 | 0.07594 | 0.38492 |
| **0.01** | 3 | 8 | 2000 | 0.07599 | 0.40286 |
| **0.01** | 3 | 8 | 4000 | 0.07782 | 0.40144 |
| **0.01** | 5 | 2 | 500 | 0.08824 | 0.33058 |
| **0.01** | 5 | 2 | 1000 | 0.08652 | 0.33869 |
| **0.01** | 5 | 2 | 2000 | 0.08252 | 0.3415 |
| **0.01** | 5 | 2 | 4000 | 0.08561 | 0.34091 |
| **0.01** | 5 | 4 | 500 | 0.08482 | 0.33448 |
| **0.01** | 5 | 4 | 1000 | 0.08231 | 0.34673 |
| **0.01** | 5 | 4 | 2000 | 0.08869 | 0.34195 |
| **0.01** | 5 | 4 | 4000 | 0.08242 | 0.34758 |
| **0.01** | 5 | 6 | 500 | 0.08614 | 0.3385 |
| **0.01** | 5 | 6 | 1000 | 0.08623 | 0.34922 |
| **0.01** | 5 | 6 | 2000 | 0.0869 | 0.34658 |
| **0.01** | 5 | 6 | 4000 | 0.08307 | 0.34645 |
| **0.01** | 5 | 8 | 500 | 0.08524 | 0.32962 |
| **0.01** | 5 | 8 | 1000 | 0.08482 | 0.3375 |
| **0.01** | 5 | 8 | 2000 | 0.0825 | 0.33939 |
| **0.01** | 5 | 8 | 4000 | 0.08393 | 0.33709 |
| **0.01** | 8 | 2 | 500 | 0.06469 | 0.04886 |
| **0.01** | 8 | 2 | 1000 | 0.0731 | 0.042 |
| **0.01** | 8 | 2 | 2000 | 0.07104 | 0.04886 |
| **0.01** | 8 | 2 | 4000 | 0.06775 | 0.04229 |
| **0.01** | 8 | 4 | 500 | 0.09713 | 0.07404 |
| **0.01** | 8 | 4 | 1000 | 0.10264 | 0.08177 |
| **0.01** | 8 | 4 | 2000 | 0.10574 | 0.08797 |
| **0.01** | 8 | 4 | 4000 | 0.10035 | 0.07821 |
| **0.01** | 8 | 6 | 500 | 0.10323 | 0.17249 |
| **0.01** | 8 | 6 | 1000 | 0.10472 | 0.17699 |
| **0.01** | 8 | 6 | 2000 | 0.10615 | 0.17201 |
| **0.01** | 8 | 6 | 4000 | 0.10219 | 0.16641 |
| **0.01** | 8 | 8 | 500 | 0.03196 | 0.24051 |
| **0.01** | 8 | 8 | 1000 | 0.04296 | 0.25108 |
| **0.01** | 8 | 8 | 2000 | 0.03726 | 0.25043 |
| **0.01** | 8 | 8 | 4000 | 0.0403 | 0.24365 |
| **0.01** | 12 | 2 | 500 | 0.08581 | -0.2739 |
| **0.01** | 12 | 2 | 1000 | 0.08677 | -0.2752 |

| | | | | | |
|---|---|---|---|---|---|
| **0.01** | 12 | 2 | 2000 | 0.08552 | -0.2763 |
| **0.01** | 12 | 2 | 4000 | 0.08781 | -0.28197 |
| **0.01** | 12 | 4 | 500 | 0.17083 | -0.18944 |
| **0.01** | 12 | 4 | 1000 | 0.16813 | -0.1965 |
| **0.01** | 12 | 4 | 2000 | 0.17045 | -0.19089 |
| **0.01** | 12 | 4 | 4000 | 0.16845 | -0.18496 |
| **0.01** | 12 | 6 | 500 | 0.20445 | 0.03034 |
| **0.01** | 12 | 6 | 1000 | 0.1751 | 0.00507 |
| **0.01** | 12 | 6 | 2000 | 0.20703 | -0.00713 |
| **0.01** | 12 | 6 | 4000 | 0.19827 | 0.0191 |
| **0.01** | 12 | 8 | 500 | 0.1003 | 0.1523 |
| **0.01** | 12 | 8 | 1000 | 0.0932 | 0.14392 |
| **0.01** | 12 | 8 | 2000 | 0.09665 | 0.15077 |
| **0.01** | 12 | 8 | 4000 | 0.10412 | 0.13968 |
| **0.05** | 2 | 2 | 500 | 0.08779 | 0.45582 |
| **0.05** | 2 | 2 | 1000 | 0.09186 | 0.44991 |
| **0.05** | 2 | 2 | 2000 | 0.08678 | 0.44019 |
| **0.05** | 2 | 2 | 4000 | 0.08878 | 0.42943 |
| **0.05** | 2 | 4 | 500 | 0.08667 | 0.44889 |
| **0.05** | 2 | 4 | 1000 | 0.07979 | 0.45074 |
| **0.05** | 2 | 4 | 2000 | 0.08186 | 0.44432 |
| **0.05** | 2 | 4 | 4000 | 0.0788 | 0.44038 |
| **0.05** | 2 | 6 | 500 | 0.08783 | 0.44495 |
| **0.05** | 2 | 6 | 1000 | 0.08621 | 0.44762 |
| **0.05** | 2 | 6 | 2000 | 0.08972 | 0.43537 |
| **0.05** | 2 | 6 | 4000 | 0.09025 | 0.42925 |
| **0.05** | 2 | 8 | 500 | 0.09091 | 0.4477 |
| **0.05** | 2 | 8 | 1000 | 0.08946 | 0.44643 |
| **0.05** | 2 | 8 | 2000 | 0.09185 | 0.43513 |
| **0.05** | 2 | 8 | 4000 | 0.0922 | 0.43192 |
| **0.05** | 3 | 2 | 500 | 0.08598 | 0.40728 |
| **0.05** | 3 | 2 | 1000 | 0.07466 | 0.40802 |
| **0.05** | 3 | 2 | 2000 | 0.07777 | 0.40639 |
| **0.05** | 3 | 2 | 4000 | 0.07851 | 0.40609 |
| **0.05** | 3 | 4 | 500 | 0.08123 | 0.39751 |
| **0.05** | 3 | 4 | 1000 | 0.07676 | 0.3899 |
| **0.05** | 3 | 4 | 2000 | 0.07644 | 0.39257 |
| **0.05** | 3 | 4 | 4000 | 0.07547 | 0.39076 |
| **0.05** | 3 | 6 | 500 | 0.07306 | 0.40641 |
| **0.05** | 3 | 6 | 1000 | 0.06876 | 0.40405 |
| **0.05** | 3 | 6 | 2000 | 0.06861 | 0.40292 |

| | | | | | |
|---|---|---|---|---|---|
| **0.05** | 3 | 6 | 4000 | 0.06714 | 0.40186 |
| **0.05** | 3 | 8 | 500 | 0.07543 | 0.41588 |
| **0.05** | 3 | 8 | 1000 | 0.07275 | 0.41668 |
| **0.05** | 3 | 8 | 2000 | 0.0687 | 0.4172 |
| **0.05** | 3 | 8 | 4000 | 0.06981 | 0.41802 |
| **0.05** | 5 | 2 | 500 | 0.08086 | 0.32784 |
| **0.05** | 5 | 2 | 1000 | 0.08513 | 0.33322 |
| **0.05** | 5 | 2 | 2000 | 0.08821 | 0.34362 |
| **0.05** | 5 | 2 | 4000 | 0.0868 | 0.34665 |
| **0.05** | 5 | 4 | 500 | 0.09285 | 0.33219 |
| **0.05** | 5 | 4 | 1000 | 0.10022 | 0.33924 |
| **0.05** | 5 | 4 | 2000 | 0.09374 | 0.33897 |
| **0.05** | 5 | 4 | 4000 | 0.09433 | 0.33675 |
| **0.05** | 5 | 6 | 500 | 0.08075 | 0.33162 |
| **0.05** | 5 | 6 | 1000 | 0.08699 | 0.346 |
| **0.05** | 5 | 6 | 2000 | 0.08539 | 0.33414 |
| **0.05** | 5 | 6 | 4000 | 0.09498 | 0.33905 |
| **0.05** | 5 | 8 | 500 | 0.07577 | 0.34163 |
| **0.05** | 5 | 8 | 1000 | 0.09107 | 0.3368 |
| **0.05** | 5 | 8 | 2000 | 0.09125 | 0.34276 |
| **0.05** | 5 | 8 | 4000 | 0.09554 | 0.34047 |
| **0.05** | 8 | 2 | 500 | 0.06857 | 0.06424 |
| **0.05** | 8 | 2 | 1000 | 0.06401 | 0.03968 |
| **0.05** | 8 | 2 | 2000 | 0.0672 | 0.03627 |
| **0.05** | 8 | 2 | 4000 | 0.0695 | 0.0455 |
| **0.05** | 8 | 4 | 500 | 0.09167 | 0.0823 |
| **0.05** | 8 | 4 | 1000 | 0.08529 | 0.07921 |
| **0.05** | 8 | 4 | 2000 | 0.08322 | 0.09118 |
| **0.05** | 8 | 4 | 4000 | 0.08263 | 0.09284 |
| **0.05** | 8 | 6 | 500 | 0.10529 | 0.17268 |
| **0.05** | 8 | 6 | 1000 | 0.09102 | 0.17009 |
| **0.05** | 8 | 6 | 2000 | 0.09505 | 0.17818 |
| **0.05** | 8 | 6 | 4000 | 0.09813 | 0.17032 |
| **0.05** | 8 | 8 | 500 | 0.04426 | 0.23754 |
| **0.05** | 8 | 8 | 1000 | 0.05307 | 0.24594 |
| **0.05** | 8 | 8 | 2000 | 0.04892 | 0.24488 |
| **0.05** | 8 | 8 | 4000 | 0.04894 | 0.24715 |
| **0.05** | 12 | 2 | 500 | 0.08105 | -0.28904 |
| **0.05** | 12 | 2 | 1000 | 0.06246 | -0.2662 |
| **0.05** | 12 | 2 | 2000 | 0.0638 | -0.28034 |
| **0.05** | 12 | 2 | 4000 | 0.07159 | -0.29916 |

| | | | | | |
|---|---|---|---|---|---|
| **0.05** | 12 | 4 | 500 | 0.1714 | -0.18012 |
| **0.05** | 12 | 4 | 1000 | 0.17523 | -0.15506 |
| **0.05** | 12 | 4 | 2000 | 0.17679 | -0.17852 |
| **0.05** | 12 | 4 | 4000 | 0.18398 | -0.17342 |
| **0.05** | 12 | 6 | 500 | 0.16485 | 0.01412 |
| **0.05** | 12 | 6 | 1000 | 0.18345 | 0.00233 |
| **0.05** | 12 | 6 | 2000 | 0.16586 | 0.03166 |
| **0.05** | 12 | 6 | 4000 | 0.17012 | 0.01326 |
| **0.05** | 12 | 8 | 500 | 0.10372 | 0.1536 |
| **0.05** | 12 | 8 | 1000 | 0.10392 | 0.14188 |
| **0.05** | 12 | 8 | 2000 | 0.09992 | 0.16448 |
| **0.05** | 12 | 8 | 4000 | 0.10394 | 0.15397 |