



Top Quant
CHRD前海智库
CHINA HONGKONG RESEARCH DEVELOPMENT



零起点

Python足彩大数据 与机器学习实盘分析

何海群 著

Win Or Home

率先导入机器学习模式
用专业的量化回溯手段，精准分析足彩赔率数据

 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

零起点 Python足彩大数据 与机器学习实盘分析

“零起点Python”系列丛书

本书继续保持了“零起点Python”系列丛书的一贯风格——简单实用。书中配备了大量的图表说明，没有枯燥的数学公式，只要懂Word、Excel，就能够轻松阅读全书。

▼IT零起点：不需要任何电脑编程基础，只要会打字、会使用Excel，就能看懂本书，利用本书配套的Python软件包，轻松学会如何利用Python对股票和足彩数据进行专业分析和量化投资分析。

▼投资零起点：无须购买任何专业软件，本书配套的zwPython软件包，采用开源模式，提供全功能、全免费的工业级数据分析平台。

▼配置零起点：所有软件、数据全部采用苹果“开箱即用”模式，绿色版本，无须安装，解压后即可直接运行系统。

▼理财零起点：不需要任何专业金融背景，采用通俗易懂的语言，配合大量专业的图表和实盘操作案例，轻松掌握各种量化投资策略。

▼数学零起点：全书没有任何复杂的数学公式，只有基本的加、减、乘、除，轻轻松松就能看懂全书。

本书学习资源

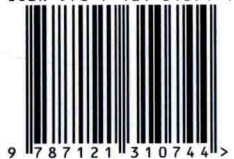
▼本书的读者QQ群是：264880547（Top极宽足彩大数据）。

▼有关的程序和数据下载，请浏览网站<http://TopQuant.vip>，在【下载中心】栏目下载。

▼本书在极宽量化社区设有专栏，读者有任何建议都可以在相关专栏发布信息，我们会第一时间进行反馈和答复。

上架建议：大数据&机器学习

ISBN 978-7-121-31074-4



9 787121 310744 >

定价：99.00元



策划编辑：黄爱萍（QQ：69476637）
责任编辑：徐津平
封面设计：李玲



零起点

Python足彩大数据 与机器学习实盘分析

何海群 著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书使用 Python 编程语言、Pandas 数据分析模块、机器学习和人工智能算法对足彩大数据进行实盘分析，设计并发布了开源大数据项目 tfbDat 足彩数据包，汇总了 2010—2017 年全球近 7 万场足球比赛的赛事和赔率数据。此外，还介绍使用 Python 语言抓取网页数据、下载更新 tfbDat 足彩数据包、预测和分析比赛球队的取胜概率，同时提出了检测人工智能算法优劣的“足彩图灵”法则。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目 (CIP) 数据

零起点 Python 足彩大数据与机器学习实盘分析 / 何海群著. —北京: 电子工业出版社, 2017.5
ISBN 978-7-121-31074-4

I. ①零… II. ①何… III. ①软件工具—程序设计—应用—足球运动—彩票 IV. ①F719.52-39

中国版本图书馆 CIP 数据核字 (2017) 第 050287 号

策划编辑: 黄爱萍

责任编辑: 徐津平

印 刷: 三河市华成印务有限公司

装 订: 三河市华成印务有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编: 100036

开 本: 787×980 1/16 印张: 27.5 字数: 523 千字

版 次: 2017 年 5 月第 1 版

印 次: 2017 年 5 月第 1 次印刷

定 价: 99.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888, 88258888。

质量投诉请发邮件至 zltz@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010) 51260888-819, faq@phei.com.cn。

前 言

从足彩到量化，再从量化到足彩。

生命总是在轮回中，不断完成自我突破与成长壮大。

本书是“Python 量化三部曲”的补充部分。

“Python 量化三部曲”

“Python 量化三部曲”包括：

- 《零起点 Python 大数据与量化交易》（入门课程）；
- 《零起点 Python 量化与机器学习实盘分析》（重点分析 Sklearn）；
- 《零起点 Python 量化与 TensorFlow 深度学习实盘分析》（重点分析 TensorFlow）。

此外还有两部补充作品：

- 《零起点 Python 足彩大数据与机器学习实盘分析》；
- 《零起点 Python 机器学习快速入门》。

较好的 Python 机器学习入门教程

本书中的机器学习算法章节，是目前较系统的 Python 机器学习入门教程，其特点如下。

- 独创的黑箱教学模式，全书无任何抽象理论和深奥的数学公式。
- 首次系统化融合 Sklearn 人工智能软件和 Pandas 数据分析软件，无需直接使用复杂的 Numpy 数学矩阵模块。
- 三位一体的课件模式：图书+开发平台+成套的教学案例。系统讲解，逐步深入。
- 业内第一个系统化的 Sklearn 函数和 API 中文文档，可作为案头工具书随时查阅。
- 基于 Sklearn+Pandas 模式，无需任何理论基础，全程采用 MBA 案例模式，懂 Excel 就可看懂。

这些内容采用独创的黑箱模式和 MBA 案例教学机制，结合一线实战案例，介绍 Sklearn 人工智能模块库和常用的机器学习算法。

进一步学习

本书的读者如果有兴趣可以进一步学习“Python 量化三部曲”的内容，虽然“Python 量化三部曲”的内容是以金融量化分析为主，但基本原理都是相通的，本质上都是数据分析，只是数据源不同，一个是金融数据，另一个是足彩赔率数据。

对于“Python 量化三部曲”的读者而言，本书也有很大的价值，特别是对于入门的读者。

本书有多章关于网络爬虫的内容和具体案例，讲解了提取网络数据的方法，为希望自己编写自动交易程序的读者提供了一种基于 Web 的操作接口。

网络资源

本书的读者 QQ 群是：264880547（Top 极宽足彩大数据）。

本书有关的程序和数据下载，请浏览网站：TopQuant.vip 极宽量化社区。网站【下载中心】有最新的程序和数据下载地址。

本书在 TopQuant.vip 极宽量化社区设有专栏，对本书和足彩有任何建议的读者，请在社区相关专栏发布信息，笔者会在第一时间进行反馈和答复。

“零起点 Python”系列丛书

本书继续保持了“零起点 Python”系列丛书的一贯风格，简单实用，书中配有大量图表说明，没有一条数学公式，普通读者只要懂 Word 和 Excel，就能够轻松阅读全书。

- IT 零起点，不需要任何电脑编程基础，会打字、会 Excel 就能看懂本书，利用本书配套的 Python 软件包，轻松学会利用 Python 对股票、足彩数据进行专业分析、量化投资分析。
- 投资零起点，无须购买任何专业软件，本书配套的 zwPython 软件包采用开源模式，提供 100% 全功能、全免费的工业级数据分析平台。
- 配置零起点，所有软件、数据全部采用苹果“开箱即用”模式，绿色版本，无须安装，解压即可，直接运行系统。
- 理财零起点，不需要任何专业金融背景，使用通俗易懂的语言，配合大量专业图表和实盘操作案例，轻松掌握各种量化投资策略。
- 数学零起点，全书没有任何复杂的数学公式，只有最基本的加、减、乘、除，易于理解。

关于足彩的几个误区

近年来，大数据产业、人工智能行业风生水起，可国内足彩大数据的专业研究还处于冷门和偏门，这其中有很多关于足彩领域的误区。

很多主流学者，到现在都看不起足彩，认为是赌博，这个我们不讨论。最近国家级别的彩票大数据研究中心已经正式成立了，资本的力量是无穷的。

为此，笔者做了一个简单的小结，以正视听。

- 足彩虽然容易与赌球混淆，但却是最好的大数据研究对象，没有之一。
- 微软、百度和谷歌等公司目前都有专业团队在做足彩大数据研究，并定期发布结果。
- 足彩相当于十倍配资的股票。
- 国内足彩赔率的确很低，差不多是全球最低的，比欧洲平均低 10% 左右。

- 彩票和股票的发明人据说都是同一个英国爵士。
- 必胜足彩交易所成立当年获得了英国 MBA 商业创新大奖。
- 高盛公司多年前就开始进行足彩套利业务，维基百科中有介绍。
- 极宽黑天鹅（红牛吧）足彩是业内首家公开进行实盘测试的足彩大数据模型。
- 黑天鹅算法在业内率先以“盈利率”而不是“胜率”测试足彩算法模型。

致谢

虽然很多网友在笔者博客中留言，建议笔者早日完成这本书的写作，但本书的创作和正式出版还是经历了许多波折。

如今本书终于出版，在此，要特别感谢电子工业出版社的黄爱萍和戴新编辑，感谢她们在选题策划和稿件整理方面所做的大量工作。

同时，在本书创作过程中，极宽开源量化团队和培训班的全体成员提出很多宝贵的意见，并对部分课件程序做了中文注解。

特别是吴娜、余勤、王硕三位同学，为极宽开源量化文库和 zwQuant 开源量化软件编写文档，并在团队成员管理方面做了大量工作，对他们的付出表示感谢。

何海群（字王）

北京极宽科技有限公司 CTO

2017 年 3 月 25 日

目 录

第 1 章 足彩与数据分析	1
1.1 “阿尔法狗”与足彩	1
1.2 案例 1-1: 可怕的英国足球	3
1.3 关于足彩的几个误区	7
1.4 足彩·大事件	8
1.5 大数据图灵(足彩)原则	10
1.6 主要在线彩票资源	11
1.7 主要在线足彩数据源	15
1.8 足彩基础知识	17
1.9 学习路线图	18
第 2 章 开发环境	19
2.1 数据分析首选 Python	19
2.1.1 大数据, why Python	19
2.1.2 入门简单, 功能强大	21
2.1.3 难度降低 90%, 性能提高 10 倍	23
2.1.4 “零对象”编程模式	24
2.2 用户运行平台	25
2.3 程序目录结构	26
2.4 tfbDat 足彩数据包	27

2.5	Spyder 编辑器界面设置	28
2.5.1	开发环境界面设置	28
2.5.2	代码配色技巧	29
2.5.3	图像显示配置	31
2.5.4	重剑无锋	32
2.6	Notebook 模式	34
2.7	模块库控制面板	36
2.7.1	模块库资源	37
2.7.2	模块库维护更新	37
2.7.3	系统关联	38
2.8	使用 pip 命令更新模块库	39
2.8.1	pip 常用命令	39
2.8.2	进入 Python 命令行模式	41
2.8.3	pip 安装模板	41
2.8.4	pip 参数解释	42
2.8.5	pip-install 参数选项	43
第 3 章	入门案例套餐	45
3.1	案例 3-1: 第一次编程, “hello,ziwang”	45
3.1.1	简单调试	46
3.1.2	控制台复位	47
3.2	案例 3-2: 增强版 “hello,ziwang”	47
3.3	案例 3-3: 列举系统模块库清单	49
3.4	案例 3-4: 常用绘图风格	50
3.5	案例 3-5: Pandas 常用绘图风格	52
3.6	案例 3-6: 常用颜色表 cors	53
第 4 章	足彩量化分析系统	55
4.1	功能简介	55
4.1.1	目录结构	56
4.1.2	TFB 安装与更新	56

4.2	TFB 主体框架	57
4.2.1	模块构成	57
4.2.2	Top-Base 极宽基础模块库	57
4.2.3	Top-Football 极宽足彩专业模块库	58
4.2.4	tfbDat 极宽足彩数据包	59
4.2.5	量化系统模块构成	60
4.2.6	案例 4-1: 赔率文件切割	61
4.2.7	案例 4-2: 批量切割数据文件	64
4.3	tfbDat 数据结构	66
4.3.1	案例 4-3: tfb 数据格式	67
4.3.2	gid 基本比赛数据格式	67
4.3.3	xdat 赔率数据格式	69
4.4	足彩基本数据分析	73
4.4.1	案例 4-4: 比赛数据基本图表分析	73
4.4.2	案例 4-5: 比赛数据进阶图表分析	77
4.4.3	案例 4-6: 比赛数据年度图表分析	80
4.4.4	案例 4-7: 比赛数据时间细分图表分析	81
4.5	胜、平、负数据分析	88
4.5.1	案例 4-8: 胜、平、负数据分析	88
4.5.2	@修饰符	88
4.5.3	胜、平、负分析	90
4.6	赔率数据分析	91
4.6.1	案例 4-9: 赔率分析	91
4.6.2	扩充 dr_gid_top10 绘图函数	92
4.6.3	赔率对比	93
第 5 章	常用数据分析工具	96
5.1	Pandas 数据分析软件	96
5.1.1	Pandas 简介	96
5.1.2	案例 5-1: Pandas 常用统计功能	99
5.2	科学计算	104

5.3	人工智能	105
5.4	NLTK 语义分析	107
5.5	数据清洗统计分析	109
5.6	数据可视化	109
第 6 章	辅助工具	114
6.1	性能优化	114
6.1.1	Numexpr 矢量加速库	115
6.1.2	Numba 支持 GPU 的加速模块库	115
6.1.3	Blaze 大数据优化模块库	115
6.1.4	Pyston 加速模块	116
6.1.5	PyPy 加速模块	116
6.1.6	Cython	116
6.1.7	其他优化技巧	117
6.2	网页信息抓取	117
6.2.1	Requests 人性化的网络模块	118
6.2.2	Scrapy 网页爬虫框架	118
6.2.3	Beautiful Soup 4	119
6.3	其他工具模块	120
6.3.1	Logging 日志模块	120
6.3.2	Debug 调试工具	121
6.3.3	re 正则表达式	121
6.3.4	并行编程	122
6.4	网络辅助资源	123
6.5	arrow 优雅简捷的时间模块库	125
6.5.1	案例 6-1: arrow 入门案例	126
6.5.2	创建 arrow 时间对象	128
6.5.3	创建时间戳	128
6.5.4	arrow 属性	129
6.5.5	replace 替换和 shift 位移	130
6.5.6	format 格式化参数	130

6.5.7	时间转换	131
6.5.8	短命令	131
6.5.9	人性化	131
6.5.10	范围和跨度	132
6.5.11	工厂模式	133
6.5.12	Token 特殊字符	133
第 7 章	网络足彩数据抓取	135
7.1	500 彩票网站数据接口的优势	135
7.1.1	案例 7-1: 抓取赔率数据网页	136
7.1.2	网页数据实战操作技巧	139
7.2	网页解析的心灵鸡汤	141
7.2.1	BS4 四大要素三缺一	142
7.2.2	Tag 标签对象	142
7.2.3	案例 7-2: Tag 标签对象	142
7.2.4	案例 7-3: Tag 标签对象数据类型	145
7.2.5	NavigableString 导航字符串	149
7.2.6	BeautifulSoup 复合对象	149
7.2.7	Comment 注释对象	150
7.2.8	案例 7-4: BS4 查找匹配功能	150
7.2.9	BS4 节点遍历功能	154
7.3	足彩基本数据抓取	155
7.3.1	案例 7-5: 分析网页比赛数据	155
7.3.2	案例 7-6: 提取网页比赛数据	157
7.3.3	gid 比赛基本数据结构	159
7.3.4	案例 7-7: 提取比赛得分	161
7.3.5	案例 7-8: 提取球队 id 编码	164
7.3.6	案例 7-9: 抓取历年比赛数据	167
7.3.7	案例 7-10: 流程图工具与 Python	171
7.3.8	实盘技巧	172
7.3.9	案例 7-11: 进程池并发运行	174

7.4	批量抓取足彩网页数据实盘教程	177
7.4.1	案例 7-12: 批量抓取赔率数据	177
7.4.2	fb_gid_getExt 扩展网页下载函数	178
7.4.3	bars 节点数据包与 pools 彩票池	178
7.4.4	抓取扩展网页	180
7.5	足彩赔率数据抓取	181
7.5.1	gid 与赔率数据网页	181
7.5.2	案例 7-13: 提取赔率数据	184
7.5.3	赔率数据与结构化数据	186
7.5.4	瀑布流数据网页与小数据理论	189
第 8 章	足彩数据回溯测试	191
8.1	TFB 系统构成	192
8.1.1	TFB 系统模块结构	192
8.1.2	Top-Base 极宽基础模块库	192
8.1.3	Top-Football 极宽足彩专业模块库	193
8.2	实盘数据更新	194
8.2.1	案例 8-1: 实盘数据更新	194
8.2.2	实盘要点: 冗余	195
8.2.3	实盘要点: 耐心	196
8.2.4	实盘要点: 数据文件	197
8.2.5	main_get 函数	197
8.3	变量初始化	199
8.3.1	全局变量与类定义	201
8.3.2	彩票池内存数据库	202
8.3.3	案例 8-2: 内存数据库&数据包	204
8.4	回溯测试	205
8.4.1	案例 8-3: 回溯	206
8.4.2	main_bt 回溯主入口	207
8.4.3	案例 8-4: 实盘回溯	209
8.4.4	彩票池与统计池	211

8.4.5	poolTrd 下单交易数据	212
8.4.6	poolRet 回报记录数据	213
8.4.7	实盘足彩推荐分析	214
8.4.8	实盘回报分析	214
8.4.9	全数据分析与足彩数据集	215
8.5	bt_main 回溯主函数	216
8.5.1	bt_1dayMain 单日回溯函数	218
8.5.2	赔率数据合并函数	219
8.5.3	单日回报分析函数	220
8.5.4	单日回报分析	221
8.5.5	单场比赛回报分析	223
8.6	sta01 策略的大数据分析	224
8.6.1	一号策略函数	226
8.6.2	超过 100%的盈利策略与秘诀	227
8.6.3	统计分析	228
8.6.4	回溯时间测试	229
8.6.5	bt_main_ret 总回报分析	230
第 9 章	参数智能寻优	232
9.1	一元参数寻优	233
9.1.1	案例 9-1: 一号策略参数寻优	233
9.1.2	一元测试函数	234
9.1.3	测试结果数据格式	236
9.1.4	案例 9-2: 一元参数图表分析	237
9.2	策略函数扩展	241
9.2.1	扩展一号策略函数	241
9.2.2	案例 9-3: 一号扩展策略	242
9.2.3	案例 9-4: sta10 策略	244
9.3	二元参数寻优	246
9.3.1	案例 9-5: sta10 参数寻优	246
9.3.2	案例 9-6: 二元参数图表分析	248

9.4	策略 310 准多因子策略	252
9.4.1	案例 9-7: 数据预处理	254
9.4.2	案例 9-8: 策略 310 参数寻优	257
9.4.3	案例 9-9: 策略 310 图表分析	259
9.4.4	案例 9-10: 策略 310	264
第 10 章	Python 人工智能入门与实践	266
10.1	从忘却开始	266
10.2	Iris 经典爱丽丝	269
10.2.1	案例 10-1: 经典爱丽丝	270
10.2.2	案例 10-2: 爱丽丝进化与矢量化文本	272
10.3	AI 操作流程	273
10.3.1	机器学习与测试数据集	274
10.3.2	机器学习运行流程	274
10.3.3	经典机器学习算法	275
10.3.4	黑箱大法	275
10.3.5	数据切割函数	276
10.3.6	案例 10-3: 爱丽丝分解	277
10.3.7	案例 10-4: 线性回归算法	281
第 11 章	机器学习经典算法案例 (上)	286
11.1	线性回归	286
11.2	逻辑回归算法	293
11.2.1	案例 11-1: 逻辑回归算法	294
11.3	朴素贝叶斯算法	296
11.3.1	案例 11-2: 贝叶斯算法	297
11.4	KNN 近邻算法	299
11.4.1	案例 11-3: KNN 近邻算法	301
11.5	随机森林算法	302
11.5.1	案例 11-4: 随机森林算法	306
第 12 章	机器学习经典算法案例 (下)	308
12.1	决策树算法	308

12.1.1	案例 12-1: 决策树算法	310
12.2	GBDT 迭代决策树算法	311
12.2.1	案例 12-2: GBDT 迭代决策树算法	312
12.3	SVM 向量机	313
12.3.1	案例 12-3: SVM 向量机算法	315
12.4	SVM-cross 向量机交叉算法	316
12.4.1	案例 12-4: SVM-cross 向量机交叉算法	317
12.5	神经网络算法	318
12.5.1	经典神经网络算法	319
12.5.2	Sklearn 神经网络算法	320
12.5.3	人工智能学习路线图	320
12.5.4	案例 12-5: MLP 神经网络算法	321
12.5.5	案例 12-6: MLP_reg 神经网络回归算法	323
第 13 章	机器学习组合算法	326
13.1	CCPP 数据集	326
13.1.1	案例 13-1: CCPP 数据集	327
13.1.2	案例 13-2: CCPP 数据切割	328
13.1.3	数据切割函数	330
13.1.4	案例 13-3: 读取 CCPP 数据集	331
13.1.5	数据读取函数	333
13.2	机器学习统一接口函数	334
13.2.1	案例 13-4: 机器学习统一接口	334
13.2.2	统一接口函数	336
13.2.3	机器学习算法代码	338
13.2.4	效果评估函数	339
13.2.5	常用评测指标	340
13.3	批量调用机器学习算法	341
13.3.1	案例 13-5: 批量调用	341
13.3.2	批量调用算法模型	344
13.4	一体化调用	345