舞于浪尖——
一种应用于 A 股波动上涨情景下的统计套利模型

杨宸源 [1)]

摘　要　通过对 A 股日 K 历史数据的观察，我们发现很多股票的股价都会在某些时期内持续波动上涨，持续时间从数十日至上百日不等，接着往往迎来大幅度的下跌。若在这样的波动上涨区间中，通过一系列数据对股价何时迎来下跌进行相对准确的预测，就能实现套利操作。我们收集了 2019 年年初到 2024 年年末区间内的 5000 余只股票的日 K 数据，并从中截取出符合条件的波动上涨区间，试探究影响波动上涨区间长度的因素，并基于此设计套利模型。

关键词　A 股，LSTM，统计套利，波段套利

Dancing on the Crest —A Statistical Arbitrage Model for Fluctuating Bull Markets in A-Shares

Chenyuan Yang[1)]

[1)](School of Management, Huazhong University of Science and Technology, Wuhan, China)

Abstract　By observing the historical daily K-line data of A-shares, we found that the stock prices of many companies experience sustained fluctuating increases over certain periods, lasting from several dozen to over a hundred days, often followed by significant declines. If we can relatively accurately predict when such a fluctuating upward trend will culminate in a decline using a series of data indicators, arbitrage opportunities could be realized. We collected the daily K-line data of over 5,000 stocks from early 2019 to the end of 2024, extracted the qualified fluctuating upward intervals, and attempted to explore the factors influencing the duration of these intervals. Based on this, we aim to design an arbitrage model.

Keywords　A-share, LSTM, Statistical Arbitrage，Swing Trading

1 问题引出

网络中流传着一个关于 A 股的段子：“问为什么 A 股叫做 A 股，答因为 A 股每次暴涨都会暴跌，K 线形状像字母 A”，忍俊不禁之余也使人反思为何存在这样的现象，借助供给需求模型，我们尝试从上涨动力和下跌动力两个角度对该现象进行分析：从上涨动力视角来看，股价上涨的原因可能是个别企业或整个行业的前景受到市场看好，或者主力资金为达成某些目的大量购入，最终都引起散户相继加仓，需求大于供给使得股价在一段时期内上升；从下跌动力来看，当出现有关于企业或行业的负面消息，或者主力资金出于某些目的大量空头股票时，表现出的股价异常波动会影响散户情绪，导致散户可能相继清仓，另一种可能的情况为股价上升区间中动力不足，增速放缓或横盘震荡持续一段时间时，市场信心降低，上涨区间可能因此结束，上述种种最终都表现为供给大于需求，股价下跌。因此，股价的涨跌是股市中多方博弈的结果，具体来说，是散户与所有主力，主力与其他主力，多头方和空头方的博弈。当一只股票处于上涨区间时，其上涨动力来源可能是多个主力和散户跟投。具体来说，主力造势抬高股价后散户争相入场使得股价进一步拉高，主力会在某个高位撤出，若没有其他主力进一步拉高股价，则股价失去上升动力，在一段时间后便开始持续波动下跌。基于股价上升区间的情景，我们提出研究问题：在一个上升区间过程中，股价即将持续下跌前是否存在某些预兆，从而使我们能够提前撤离，完成套利。接下来我们将围绕该研究问题，通过对股票历史数据的统计研究，试图有所发现。

2 特征工程

我们使用爬虫技术对东方财富网行业涨幅榜中 BK0600 到 BK1099 共 500 个板块的股票数据进行爬取，去重后得到 5443 只 A 股股票从 2019 年到 2024 年的日 K 数据构建数据集，并从中截取出所有 MA20 均线连续 10 天以上持续走高（其间允许最多 3 天的连续走低情况）的区间，得到 100876 段时序数据。接着，我们对这些时序样本的走势进行线性拟合，筛选出 p 值小于 0.01, 斜率为正，R 方值大于 0.9 的样本，以保证所使用的样本均为波动上升区间，从而获得了 6022 条有效数据。

2.1 基础特征

包含所有可直接获取的 K 线数据，如 1所示。

其中，加权均价的计算方式为：

$$WAP_i = 0.25 \cdot HP_i + 0.25 \cdot LP_i + 0.5 \cdot CP_i \tag{1}$$

2.2 空间特征

空间特征在本文中指通过对波动上涨区间内的基础特征时序数据进行统计降维后的结果，如 2所示。

其中，累计涨幅和日均涨幅的计算方法分别为：

$$RC\Delta P_{ij} = \frac{AWP_j - AWP_i}{AWP_i} \tag{2}$$

$$AR\Delta P_{ij} = \frac{RC\Delta P_{ij}}{L_{ij}} \tag{3}$$

Table 1 基础特征

特征名称	符号	解释
开盘价	OP_i	第 i 日开盘股价
收盘价	CP_i	第 i 日收盘股价
最高价	HP_i	第 i 日最高股价
最低价	LP_i	第 i 日最低股价
加权均价	WAP_i	第 i 日加权均价
振幅	A_i	第 i 日振幅
涨跌额	ΔP_i	第 i 日涨跌额
涨跌幅	$R\Delta P_i$	第 i 日涨跌幅
成交量	V_i	第 i 日成交量
成交额	T_i	第 i 日成交额
换手率	RV_i	第 i 日换手率

Table 2 空间特征

特征名称	符号	解释
区间长度	L_{ij}	波动上涨区间中第 i 到 j 日的区间长度（天）
时段均值	$AVG(X)_{ij}$	波动上涨区间中第 i 到 j 日的基础特征 X 的均值
时段标准差	$STD(X)_{ij}$	波动上涨区间中第 i 到 j 日的基础特征 X 的标准差
时段最大值	$MAX(X)_{ij}$	波动上涨区间中第 i 到 j 日的基础特征 X 的最大值
时段最小值	$MIN(X)_{ij}$	波动上涨区间中第 i 到 j 日的基础特征 X 的最小值
累计涨幅	$RC\Delta P_{ij}$	波动上涨区间中第 i 到 j 日的累计涨幅
日均涨幅	$AR\Delta P_{ij}$	波动上涨区间中第 i 到 j 日的日均涨幅

2.3 时间特征

空间特征记录了一段时间内股价的整体状态，而时间特征则刻画该段时间内股价的运动变化情况，如3所示。

其中，我们首先使用三次多项式对波动上涨区间的股价走势进行拟合，得到效果类似 1所示，并记录下其回归系数、拟合优度。

对于某个特定极值点 k，设该点处回归曲线值为 $CURV_k$，则其残差极值乖离率 REB_k 和绝对残差极值乖离率 $RAEB_k$ 的计算方法分别为：

$$REB_k = \frac{WAP_k - CURV_k}{CURV_k} \tag{4}$$

$$RAEB_k = \frac{|WAP_k - CURV_k|}{CURV_k} \tag{5}$$

Table 3 时间特征

特征名称	符号	解释
趋势系数	TC_{ij}	波动上涨区间中第 i 到 j 日价格走势的多项式回归系数
一阶增长率	TFD_j	波动上涨区间中第 i 到 j 日价格走势的多项式回归系数在点 j 处的一阶导数值
二阶增长率	TSD_j	波动上涨区间中第 i 到 j 日价格走势的多项式回归系数在点 j 处的二阶导数值
拟合优度	R^2	波动上涨区间中第 i 到 j 日价格走势的多项式回归的拟合优度
绝对残差极值乖离率	$RAEB$	波动上升过程中股价围绕回归曲线波动的程度
绝对残差极值乖离率均值	$AVG(RAEB)_{ij}$	略
绝对残差极值乖离率标准差	$STD(RAEB)_{ij}$	略
残差极值乖离率最大值	$MAX(REB)_{ij}$	波动上涨区间中第 i 到 j 日的股价高于回归曲线对应值的最大幅度
残差极值乖离率最小值	$MIN(REB)_{ij}$	波动上涨区间中第 i 到 j 日的股价低于回归曲线对应值的最大幅度（负）

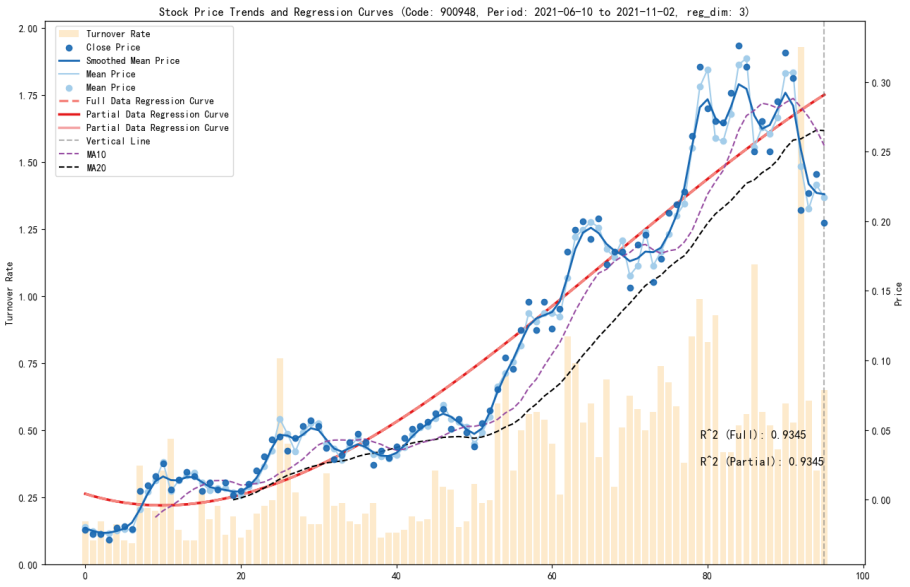

Fig. 1 三次多项式拟合股价波动上涨趋势示例

3 因子分析

3.1 上涨区间长度的概率分布

我们想要首先探究股票波动上涨区间的长度的概率分布。设随机变量 L 为上涨区间长度，我们将样本中所有大于等于 10 的上涨区间长度进行了统计，得到条件样本概率分布 $P\{L = l|L \geqslant 10\}$ 如 2所示。由该图可见，当上涨区间长度大于等于 10 时，其条件分布为一个双峰分布，第一个概率峰值出现在 010 区间内，第二个概率峰值出现 59 左右，这说明，当上涨区间长度大于等于 10 时，该上涨区间长度在 10 左右或 59 左右的条件概率最大，即对于在第十天仍波动上涨的股票，其在之后两天或者五十天之后面临大幅下跌的概率较在其他天下跌的概率都要大。

接下来我们要确定的问题是，给定一个整数 m，若要最大化股价在未来 m 天内仍保持上涨趋势的概率，则应选择至少已持续上涨多少天的股票。设 m 为任意正整数，则上述问题变为找到正整数 n 以解得条件概

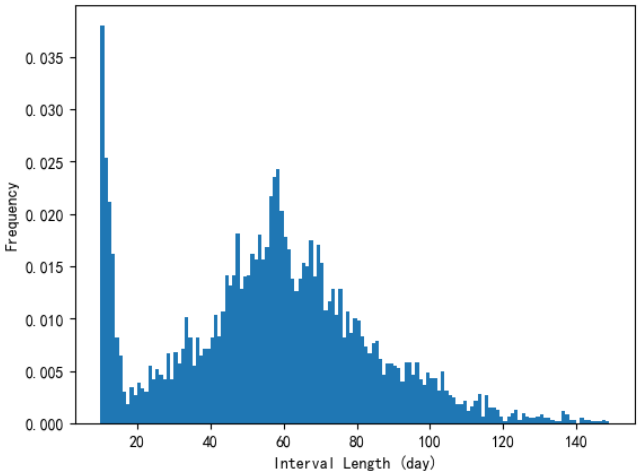

Fig. 2 $L = l|L \geqslant 10$ 样本概率分布

率最大值:

$$max\ P\{L > n+m|L > n\} \tag{6}$$

对于任意 $n \in \{n \in Z|n \geqslant 10\}$，可将上式进行改写:

$$\begin{aligned} P\{L > n+m|L > n\} &= \frac{P\{L > m+n\}}{P\{L > n\}} \\ &= \frac{P\{L > n\} - P\{n < L \leqslant n+m\}}{P\{L > n\}} \\ &= 1 - \frac{P\{n < L \leqslant n+m\}}{P\{L > n\}} \\ &= 1 - \frac{P\{n < L \leqslant n+m|L \geqslant 10\}P\{L \geqslant 10\}}{P\{L > n|L \geqslant 10\}P\{L \geqslant 10\}} \\ &= 1 - \frac{P\{n < L \leqslant n+m|L \geqslant 10\}}{P\{L > n|L \geqslant 10\}} \end{aligned} \tag{7}$$

则问题进一步转化为求 $\frac{P\{n<L\leqslant n+m|L\geqslant 10\}}{P\{L>n|L\geqslant 10\}}$ 最小值问题，结合图像可知 n 取值点应该在**??**横坐标靠左区域。接下来，使用样本分布似总体概率分布，如下式所示 (其中 q_k 为 $L = k$ 的样本频率)

$$\frac{P\{n < L \leqslant n+m|L \geqslant 10\}}{P\{L > n|L \geqslant 10\}} \approx \frac{\sum_{k=n+1}^{n+m} q_k}{\sum_{k=n+1}^{\infty} q_k} \tag{8}$$

接着，通过计算机程序对 n 进行遍历求解，得到结果如 4所示。由上述计算，我们此处选用 $L > 15$ 的计算结果，将 $L = 16$ 作为股票买入决策制定的最早时点，即只有当一只股票持续波动上升至少 16 天，我们才有较大的把握认为其有可能在未来数天内保持增长趋势，进而才会对其进行进一步分析以决定是否买入。同时，我们使用条件样本分布数据拟合计算了 $L = l|L \geqslant 16$ 的分布，结果如 3所示。

3.2 影响上涨区间长度的因素

为了探究是否能对股票时序数据的某些特征进行分析，从而对其波动上涨区间长度进行预测，我们首先通过统计方法对上涨区间长度的影响因素进行回归分析。因上涨区间长度 $L = l|L \geqslant 10$ 不符合正态分布，且具体分布不易得出，故考虑使用非参数回归进行分析。

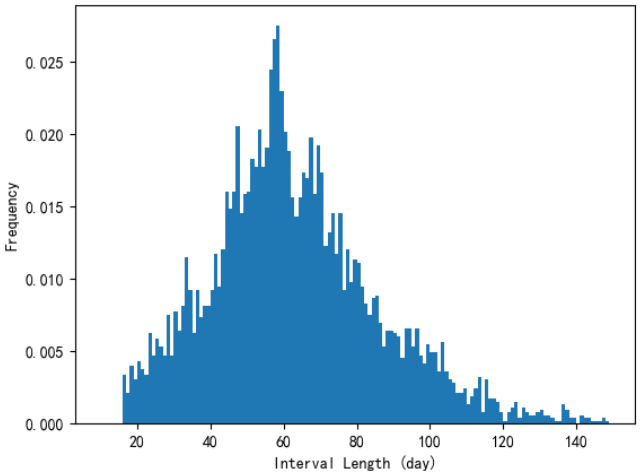

Fig. 3 $L = l|L \geqslant 16$ 样本概率分布

Table 4 条件概率最优化结果

m	n	$\frac{P\{n<L\leqslant n+m\|L\geqslant 10\}}{P\{L>n\|L\geqslant 10\}}$	$P\{L>n+m\|L>n\}$
1	16	0.002071	0.997929
2	15	0.005442	0.994558
3	16	0.009038	0.990962
4	15	0.012385	0.987615
5	15	0.016701	0.983299
6	15	0.020454	0.979546
7	15	0.023832	0.976168
8	15	0.030024	0.969976
9	15	0.034716	0.965284
10	15	0.040533	0.959467
11	15	0.045787	0.954213
12	15	0.050479	0.949521
13	14	0.057377	0.942623
14	15	0.062676	0.937324
15	14	0.069486	0.930514
16	15	0.076750	0.923250
17	14	0.083458	0.916542
18	14	0.091468	0.908532
19	13	0.099686	0.900314
20	13	0.110947	0.889053

我们使用随机森林回归模型对整个波动上涨区间数据进行分析，区间长度（L）作为被解释变量，本文第 2 部分提及的所有其他特征作为解释变量，构建出回归模型如下式所示：

$$\begin{aligned} L = f[&L_{ij}, AVG(X)_{ij}, STD(X)_{ij}, MAX(X)_{ij}, MIN(X)_{ij}, \\ &RC\Delta P_{ij}, AR\Delta P_{ij}, TC_{ij}, TFD_j, TSD_j, R^2, RAEB, \\ &AVG(RAEB)_{ij}, STD(RAEB)_{ij}, MAX(REB)_{ij}, MIN(REB)_{ij}] \end{aligned} \tag{9}$$

我们使用 2022 至 2024 年的波动上涨区间数据作为数据集，其中 80% 用于训练，20% 用于测试，得到模型均方误差为 84.30412706333973 ，R2 为 0.8928558864531428。接着，我们根据模型得出的因子重要性对各个因子进行排序，同时显示因子在模型中的系数，得到结果如 4所示。

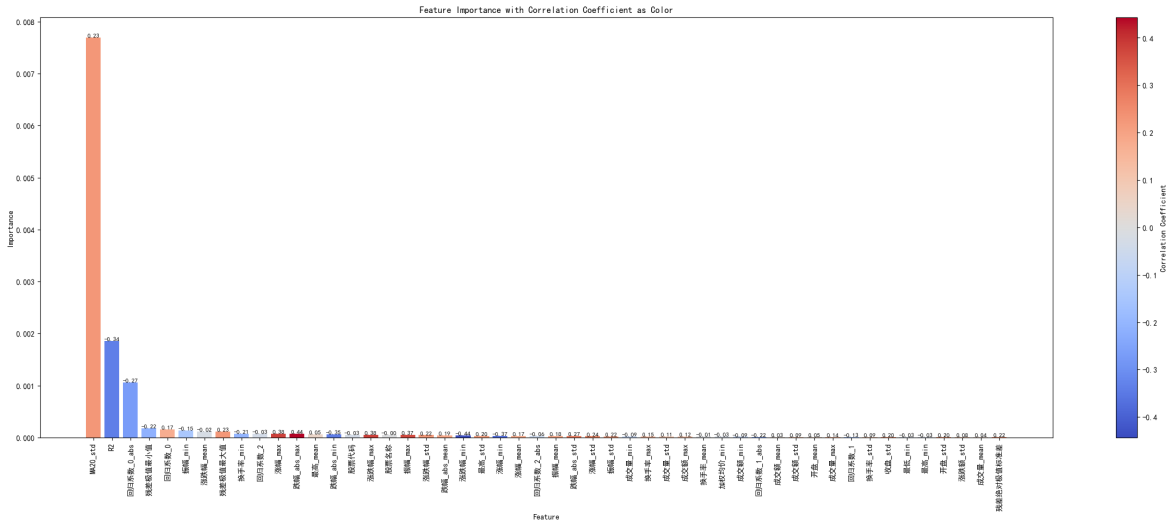

Fig. 4　随机森林回归因子重要性排序即模型系数

由 4我们得出以下发现：

- **M20_std**: 反映出整个上升区间的累计涨额（+），正相关说明上升区间越长，累计涨额越大，反过来又增强市场信心，进一步增长上升区间
- **R2**: 反映出回归曲线的拟合效果（+），负相关说明拟合度越好说明上升区间越短，毕竟长的上升区间一般都有一些中途波动，降低曲线的拟合度
- **回归系数 _0_abs**: 反映出曲线的弯曲程度（+），负相关说明上升区间越长，拟合曲线越像直线
- **残差极值最小值（一般为负值）**：反应了绕回归线下跌波动最大幅度
- **回归系数**: 同上
- **振幅 _min**: 上升区间中日内价格波动情况的最小值（+），负相关表示这玩意儿很大的话会使市场信心降低，导致上升区间较小
- **涨跌幅 _mean（一般为正值）**：整个上升区间内平均每天涨幅（+），负相关表示上升区间内如果上涨速度太快股民容易拿不住股票，没几天就空头，然后股价大跌

- **残差极值最大值（一般为正值）**：反应了绕回归线上涨波动最大幅度，正相关说明股价突破回归线上涨越多，上涨动力越强劲，增强市场信心，上升区间容易更长
- 有一个很有意思的地方，7. 说明股价整体不能涨得太快，8. 说明更大的单次上涨波动使得上升区间容易更长，所以这个上升速度需要把握一个度，太快了容易大量空头，但一次较大的上涨幅度能增强市场信心
- **涨幅 _max（正值）**：上升区间中最大的日际涨幅，正相关说明的情况和 8. 类似
- **跌幅 _abs_max（正值）**：正相关可能说明的是一些股票暴涨前会先暴跌一下，但是有救不回来的风险，所以不要轻易去接盘
- **最高 _mean**: 上升区间内股票的一种均价，似乎股票均价越高，上涨区间越长，但是系数太小了，感觉没啥用
- **跌幅 _abs_min（正值）**：负相关说明跌幅太大股票上涨区间不容易长
- **股票代码**: 略
- **涨跌幅 _max**: 和涨幅 _max 作用一样
- **股票名称**: 略
- **振幅 _max**: 日内波动涨幅大小，我觉得是因为会有一天日内暴涨所以会使得当天振幅很大，所以本质和 8. 和 9. 一样
- **跌幅 _abs_mean（正值）**：和 12 一样
- **最高 _std**: 和 1. 类似
- **涨幅 _min**: 负相关表明上升区间内涨得太快容易很快被大量空头，然后下跌，和 7 类似
- **涨幅 _mean**: 作用和 9. 类似，而不是和 19. 类似
- **成交量 _min**: 负相关表明
- **换手率 _max**: 正相关表明最大市场关注度越高，上升区间越长，一般这种 max 值都出现在暴涨的时候，如果不是这个时候需要非常留意，所以留意异常换手率上升才是选这个指标的目的
- **残差绝对值标准差**: 这个描述了股价绕回归曲线波动情况，正相关的原因是一般暴涨的几天都会使得这个指标很高，所以本质还是暴涨

3.3 上涨区间结束的迹象

3.3.1 乖离率异常

当围绕回归曲线的波动幅度异常增大，股价偏离之前的增长速度持续大幅上涨时，到一定高度后部分主力可能撤离，散户跟出，股价开始持续下跌。

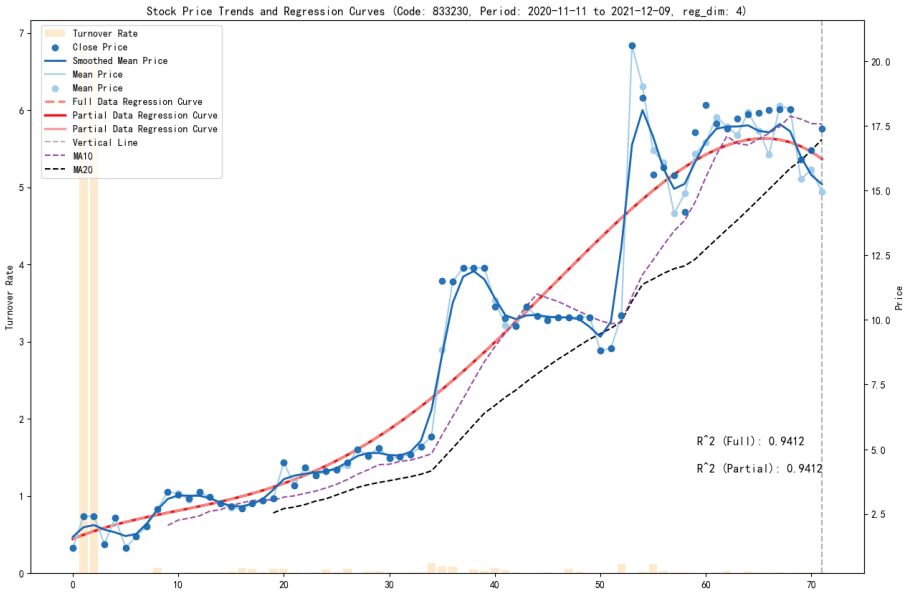

Fig. 5　乖离率异常举例

3.3.2　上升动力减弱

股票在持续上升一段时间后上升速度逐渐减缓，到一个高位后动力缺失开始横盘震荡或者直接开始下跌。

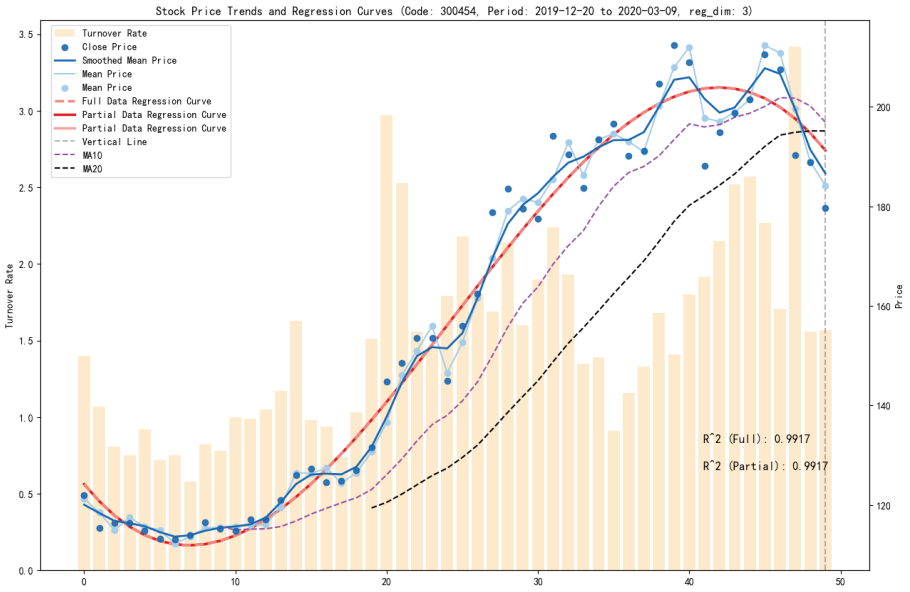

Fig. 6　上升动力缺失举例

3.3.3　尚无法分析预兆的情况

也有部分情况是我们目前无法找到预兆的，此类波动上升一直保持较好的增长动力和相对稳定的乖离率，然而却在某个点戛然而止。

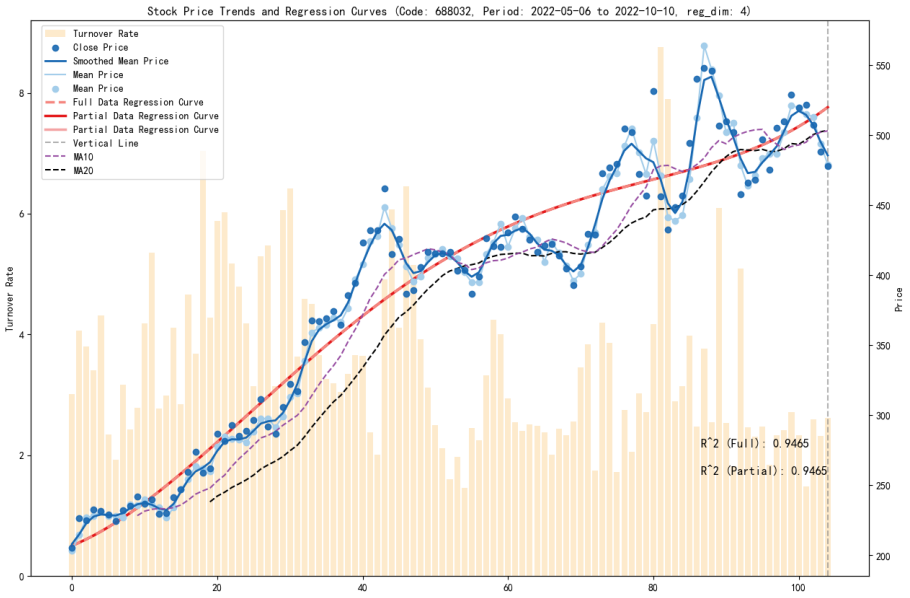

Fig. 7 上升动力缺失举例

4 决策模型构建

4.1 模型框架

如 8所示，决策模型分为买入决策和卖出决策两个部分。首先，每日开盘前向买入模型依次输入所有待选 A 股最近 16 天的日 K 数据序列，若经检测某支股票近 16 日内不处于波动上升区间，则不买入，若处于波动上升区间，则进一步使用预测模型对该支股票未来 5 日内是否会出现连续下跌进行预测，若是，则为规避买入即跌的风险选择暂时观望不买入。对于已持有的股票，每日依次输入持仓股票自买入至当日的日 K 数据序列，若当日已出现异常下跌或者相关征兆，则立即卖出止损，若没有出现上述情况，则进一步使用预测模型对该支股票未来 5 日内是否会出现连续下跌进行预测，若是，则卖出，否则继续持有该股并实时监控股价走势。

4.2 股价波动上升区间检测

首先，我们判断股价在最近 16 日的 MA20 均线是否为持续上升状态，并设置了允许上升期间出现不超过 3 日的均线连续下降情况。接着，我们对筛选出的股价序列进行一元线性回归，筛选出 p 值小于 0.01，斜率为正，R 方值大于 0.9 的样本，以保证所使用的样本均为波动上升区间。

4.3 股价 5 日内异常下跌预测

如果能使用一些方法预测出一只处于波动上升区间的股票何时迎来该次波动上升的结束，则能在临近预测点时再提前卖出股票，减少过早卖出或者过晚卖出的操作失误，从而提升该波动上升区间的套利收益率。基于此想法，我们通过逻辑回归等方法对股价波动上升区间结束点进行预测。

在特征工程中，我们将特征分为两类，第一类是时点特征，即仅反映股票当天情况的数据，包括开盘价、收盘价、成交量等 1中提及的所有特征，此处不再加以赘述；第二类是累计特征，即反映股票从某日起始至当日变化情况的特征，如趋势系数、R2、累计涨幅、日均涨幅等。

下面介绍第二类特征的计算。在对波动上升区间中未来 5 日内连续下跌起始点的预测过程中，我们将

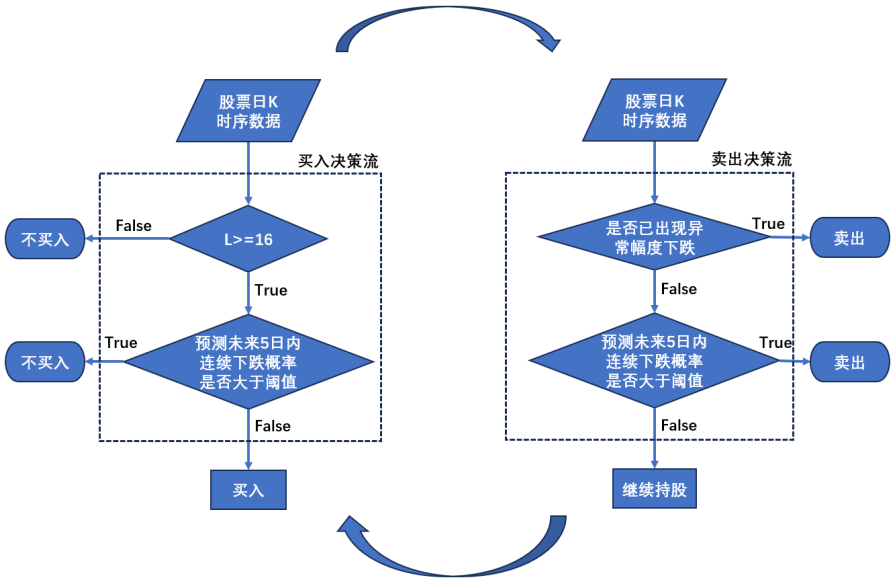

Fig. 8 决策模型框架示意图

获取到一支股票近 16 日内的日 K 数据序列，为了保证用于拟合趋势回归曲线的数据量，我们从第 10 日开始计算累计特征，10 日以前的累计数据列保持空值。接着，设 16 日数据序列中第一天序号为 0，开始分别计算所有 $j \geqslant 10$ 区间长度（$L_{0,j}$）、累计涨幅（$RC\Delta P_{0,j}$）、日均涨幅（$AR\Delta P_{0,j}$）、趋势系数（$TC_{0,j}$）即其符号以及绝对值、一阶增长率（TFD_j）、二阶增长率（TSD_j）、拟合优度（$R^2_{0,j}$）、MSE（$MSE_{0,j}$）、绝对残差极值乖离率（$RAEB_{0,j}$）即其统计量，具体计算方法已在第二部分介绍过，此处不再赘述。

另外，我们需要添加用于训练的标签 $Label_j$。由于我们的样本数据截取了完整的波动上升区间，即区间右端点之后股价便会连续下跌，因此我们将每个样本区间的倒数 5 日设为未来 5 日内连续下跌出现点，该 5 天的标签值为 1，区间上其他点标签值为 0。

我们通过上述方法对每个独立的波动上升区间样本进行了特征提取，接着将每个每个区间中前十日累计统计量为空的日数据去除，然后将所有区间数据表合并为一个整体，并对其中除标签外的其他所有数值特征进行归一化处理，得到预测模型的训练集，该训练集描述性统计如 8、9、10所示。

得到了特征之后，我们首先构建出预测模型的概念公式。设一从高维空间向实数闭区间 [0,1] 的映射 f，Y_j 为 [0,1] 上任意实数，则有预测模型

$$Y_j = f(WAP_j, L_{0,j}, RC\Delta P_{0,j}, AR\Delta P_{0,j}, TC_{0,j}, TFD_j, TSD_j, R^2_{0,j}, MSE_{0,j}, RAEB_{0,j}) \tag{10}$$

预测模型的具体构建过程包括以下步骤：首先提取特征和标签，并对特征进行标准化处理，以确保数据的一致性。接着，应用 SMOTE 技术处理类别不平衡问题，以平衡数据集。然后将数据划分为训练集和测试集，并通过 GridSearchCV 对 XGBoost 模型进行超参数调优，以获得最佳模型。调优后的模型通过调整决策阈值，优化其分类性能。随后，将 Logistic 回归、随机森林和 XGBoost 模型结合，采用集成学习方法（VotingClassifier）进行训练，旨在发挥各模型的优势。最后，通过准确率、混淆矩阵和分类报告对最终模型进行评估，以全面检验其预测能力。

具体来说，首先提取特征和标签，并对特征进行标准化处理，以确保数据的一致性。标准化是通过'StandardScaler'

完成的，它将数据缩放至均值为 0，标准差为 1 的标准正态分布，这对于很多机器学习算法的性能是至关重要的。

接着，使用 SMOTE（Synthetic Minority Over-sampling Technique）技术处理类别不平衡问题，以平衡数据集。SMOTE 通过生成少数类的合成样本，解决了数据集中类别不平衡的问题，使得模型在训练过程中能够更好地学习到少数类的特征。

然后，将数据划分为训练集和测试集，以便在模型训练后对其进行验证。我们将数据集按 80% 和 20% 的比例划分为训练集和测试集。

接下来，通过'GridSearchCV' 对 XGBoost(Extreme Gradient Boosting)模型进行超参数调优。XGBoost 是一种提升树算法，以其高效的计算能力和出色的预测性能而著称。超参数调优可以帮助找到模型的最佳参数组合，从而提高模型的性能。

调优后的模型通过调整决策阈值，优化其分类性能。使用最佳模型的概率预测值，尝试不同的决策阈值，选择最佳的阈值以提高模型的分类效果。

然后，将 Logistic 回归、随机森林和 XGBoost 模型结合，采用集成学习方法（VotingClassifier）进行训练。Logistic 回归是一种简单而高效的线性分类模型，适用于二元分类问题。随机森林是一种集成学习方法，通过构建多个决策树，进行投票表决来输出最终的分类结果，具有很好的泛化性能。VotingClassifier 通过结合多个模型的预测结果，进一步提高了整体模型的性能和稳定性。

最后，通过准确率、混淆矩阵和分类报告对最终模型进行评估，以全面检验其预测能力。这些评估指标可以帮助了解模型在不同类别上的表现，从而进一步优化模型。

我们使用 2019 至 2023 年波动上升区间数据对模型进行了训练，得到训练结果如 5、6、7所示。基于实验结果，该模型在区分两个类别时表现良好，达到了 96% 的总体准确率。分类报告显示模型在类别 0 和类别 1 上的精度分别为 0.98 和 0.94，召回率分别为 0.94 和 0.99，F1 分数均为 0.96，表明模型在检测正负样本时都较为可靠。混淆矩阵进一步证实了模型的有效性，只有少量假阳性（2711 个）和假阴性（615 个）。综上所述，模型在处理分类任务时具有较高的精度和召回率，适用于实际应用场景。

类别	精度（precision）	召回率（recall）	F1 分数（f1-score）
0	0.98	0.94	0.96
1	0.94	0.99	0.96

Table 5 分类报告

指标	值
准确率	0.96
宏平均	0.96
加权平均	0.96

Table 6 总体指标

	预测为 0	预测为 1
实际为 0	39850	2711
实际为 1	615	42299

Table 7 混淆矩阵

	日期	开盘	收盘	最高	最低	成交量	成交额	振幅	涨跌幅
count	280728	280728	280728	280728	280728	280728	280728	280728	280728
mean	2021-03-25 6:43	23.03717336	23.20085788	23.77853781	22.50138758	189376.0571	325623079.1	4.998011955	0.737851728
min	2019-01-29 0:00	-0.067	-0.066	-0.058	-0.069	0	10	-700	-700
25%	2020-02-27 0:00	6.53	6.57	6.7	6.41	27471	45081780.03	2.89	-1.23
50%	2021-04-19 0:00	11.68	11.75	12.03	11.43	72173	117204998	4.26	0.38
75%	2022-06-09 0:00	23.1	23.26	23.82	22.58	186661.5	304504664.2	6.32	2.34
max	2023-12-29 0:00	1619.34	1619.34	1624.35	1585.75	41144530	20262833664	266.67	342.86
std		43.5348045	43.83331271	44.85832866	42.54394798	499924.2425	710718293.9	3.985708047	4.16956273

Table 8 训练集描述性统计（I）

	涨跌额	换手率	MA20	加权均价	区间长度	累计涨幅	日均涨幅	回归系数 __3	回归系数 __2
count	280728	280728	280728	280728	280728	280728	280728	235706	235706
mean	0.159767957	3.216771715	21.50928229	23.17041029	36.38314311	0.385597222	0.009740918	5.92E-05	-0.000176522
min	-138.53	0	-0.080325	-0.06425	1	-16.38554217	-0.207009858	-0.141511613	-6.041680922
25%	-0.13	0.95	6.18775	6.56	16	0.092930709	0.004452773	-9.46E-05	-0.004061878
50%	0.03	1.92	10.97325	11.74	32	0.233040612	0.007624138	-1.75E-06	0.000346393
75%	0.28	3.89	21.5950625	23.215	51	0.480096871	0.012280644	8.62E-05	0.00572359
max	88	82.18	1512.491	1603.765	220	17.59550562	0.516949153	0.323849415	2.076343593
std	1.890263443	4.011395869	40.56079431	43.76167424	25.86480965	0.573519188	0.010424357	0.003169707	0.070615108

Table 9 训练集描述性统计（II）

	回归系数 __1	回归系数 __0	R2	绝对极值残差 __mean	绝对极值残差 __std	极值残差 __max	极值残差 __min	label	一阶导值	二阶导值
count	235706	235706	235706	235706	235706	235706	235706	235706	235706	235706
mean	0.129508849	16.8872091	0.917245642	0.036133004	0.027914919	0.085422312	-0.074099455	0.078438916	0.204243354	33.77760645
min	-12.58075361	-0.063763006	0.018923885	0.000970861	0.000150498	0.001795123	-442.2181918	0	-17.84933007	-0.124305525
25%	-0.014161148	5.295092269	0.901116974	0.018539111	0.010777041	0.033740761	-0.082368067	0	-0.003265431	10.58886946
50%	0.038096698	9.167648227	0.939682609	0.028666657	0.017980308	0.057141272	-0.053267022	0	0.059086796	18.3338531
75%	0.140553119	17.16074424	0.964634794	0.042964688	0.028145945	0.091341611	-0.032484171	0	0.214151191	34.3337042
max	41.35354338	984.213072	0.999766134	85.24375837	321.7794775	1331.273581	-0.001272932	0	31.14003585	1968.414486
std	0.649132223	31.09530177	0.082436495	0.251412016	0.850054413	3.026105433	1.471246498	0.268862	0.78799505	62.20946451

Table 10 训练集描述性统计（III）

4.4 策略回测结果

回测（Backtesting）是通过使用历史市场数据来模拟和评估交易策略在过去市场条件下的表现，从而推测其在未来的潜在表现。这个过程包括获取准确的历史数据、定义明确的交易规则、逐日模拟交易并记录每次交易的结果。通过回测，可以计算出策略的各种绩效指标，如总收益率、年化收益率、最大回撤、夏普比率和胜率等。回测的优点在于可以提前识别交易策略的潜在风险和弱点，验证策略的有效性，并通过优化参数来提高策略的表现。然而，回测也需要注意数据质量、交易成本和过拟合等问题，以确保回测结果的可靠性和可操作性。通过回测，投资者可以更科学地制定投资决策，提升交易策略的稳定性和鲁棒性。由于我们的交易策略属于中短期波段交易，需要较长的时间以检验策略效果，因此我们根据决策模型搭建了自动交易回测系统如，并使用 2024 年 A 股波动上升区间数据进行测试。得到回测结果如 9所示。

```
def get_strategy_res():
    '''16天连续波动上涨后当参考预测结果买卖（买卖价均按当日加权均价计）'''
    trade_log_list = []
    for i in tqdm(range(len(feature_df_list_24[:]))):
        n = len(feature_df_list_24[i][:])
        if n < 16:  # 没有连续上涨撑过16天的直接不要
            continue
        new_data = feature_df_list_24[i][:]
        features = new_data[new_data['R2'].notna()].drop(columns=['股票名称
            ', '股票代码', '日期', 'label'])
        scaler = StandardScaler()
        features_scaled = scaler.fit_transform(features)
        trade_log = {
            'buy_index': None,
            'sale_index': None,
            'cost': None,
            'revenue': None,
            'profit': None,
            'profit_rate': None,
            'upward_period': None
        }
        for j in range(16, len(features_scaled)):
            # 预测时重塑特征数组
            predict = ensemble_model.predict(features_scaled[j].reshape(1,
                -1))[0]
            if predict == 0 and trade_log['buy_index'] is None:
                trade_log['cost'] = list(features['加权均价'])[j]
                trade_log['buy_index'] = j
```

```
            elif predict == 1 and trade_log['buy_index'] is not None:
                trade_log['revenue'] = list(features['加权均价'])[j]
                trade_log['sale_index'] = j
                trade_log['profit'] = trade_log['revenue'] - trade_log['
                    cost']
                trade_log['profit_rate'] = trade_log['profit'] / trade_log[
                    'cost']
                trade_log['upward_period'] = trade_log['sale_index'] -
                    trade_log['buy_index']
                break

        # 如果在最后一天还未卖出，强制卖出(模拟在股票开始下跌前没能准确预
            测，于是卖出止损)
        if trade_log['buy_index'] is not None and trade_log['sale_index']
            is None:
            trade_log['revenue'] = list(features['加权均价'])[-1] * (1 -
                max(features['涨跌幅'])*0.01*1.5)
            trade_log['sale_index'] = len(features_scaled) - 1
            trade_log['profit'] = trade_log['revenue'] - trade_log['cost']
            trade_log['profit_rate'] = trade_log['profit'] / trade_log['
                cost']
            trade_log['upward_period'] = trade_log['sale_index'] -
                trade_log['buy_index']
        if trade_log['buy_index'] is not None:
            trade_log_list.append(trade_log)
    return trade_log_list


res = get_strategy_res()
```

4.4.1 回测结果分析

- **总收益率**: 52.38
  - 这个数字表示在回测期间，策略的总回报率为 52.38%。这个结果表明策略总体上是盈利的。
- **平均收益率**: 0.07
  - 平均收益率为 0.07，意味着每次交易的平均回报率为 7%。这表明尽管个别交易可能波动较大，但平均来说，策略能够产生正收益。
- **胜率**: 0.64

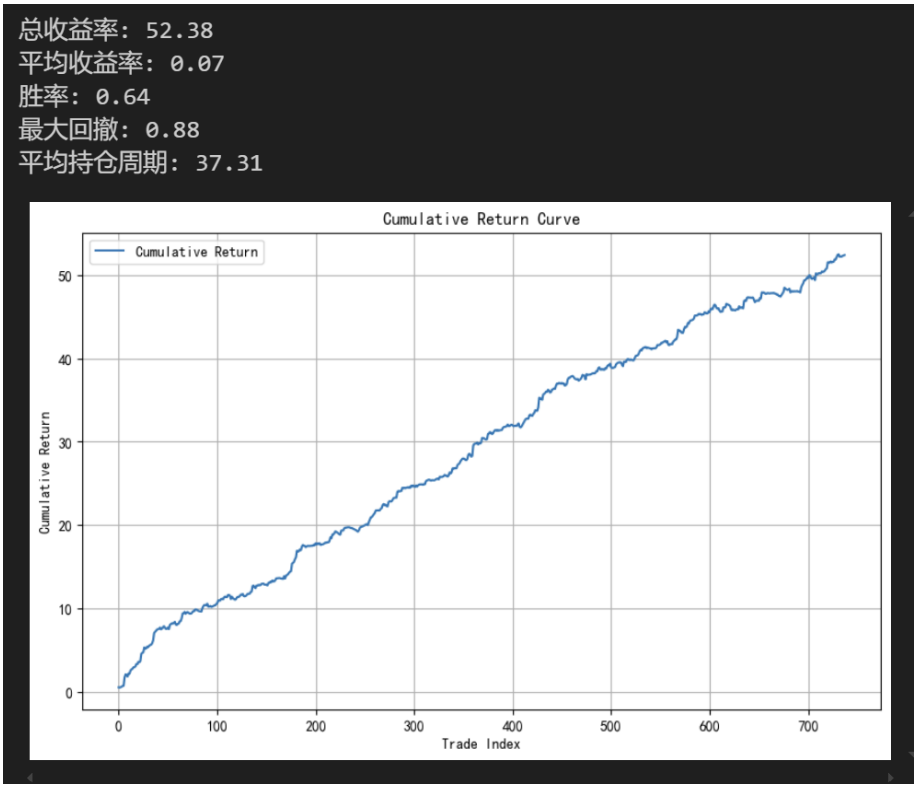

Fig. 9 回测结果

- 胜率为 64%，即在所有交易中，有 64% 的交易是盈利的。这是一个较高的胜率，表明策略在多数情况下是有效的。

- **最大回撤**: 0.88
  - 最大回撤为 0.88，这表明在回测期间，策略的最大亏损为 88%。这个数值相对较高，显示策略在某些情况下会面临较大的下行风险。

- **平均持仓周期**: 37.31
  - 平均持仓周期为 37.31 天，这表明每笔交易的平均持有时间为 37 天左右。这个指标有助于了解策略的交易频率和持仓周期。

4.4.2 综合分析

- **收益与风险**: 总收益率和平均收益率表明策略在回测期间具有良好的盈利能力。然而，最大回撤值较高，提示策略在某些时候面临较大的风险。应该关注如何降低回撤，或调整策略以减少波动。

- **胜率与稳定性**: 64% 的胜率显示策略在大多数情况下是盈利的，证明其具备一定的稳定性。然而，较高的最大回撤值表明在个别情况下可能会出现较大的亏损。

- **交易频率**: 平均持仓周期为 37.31 天，表明策略不是短期频繁交易，而是中期持有策略。需要根据市场环境和交易目标确定这种持仓周期是否合适。

4.4.3 未来改进

- **优化策略参数**: 通过调整策略参数（例如买卖条件、止损止盈点）来优化收益与风险的平衡。
- **分散投资**: 通过分散投资标的和增加多样化资产组合来降低风险。
- **持续监控与调整**: 持续监控策略在不同市场环境中的表现，并根据实际情况进行调整，以提高策略的鲁棒性和适应性。

总的来说，策略表现良好，但仍有优化空间，尤其是降低回撤和提高收益稳定性方面。