基于机器学习的多因子 A 股市场量化选股策略研究

队伍编号：2204

摘要：随着计算机与机器学习的兴起，将机器学习应用到选股策略中来成为一大热门。本文着重研究 A 股市场沪深 300 成分股。本文首先会聚焦于因子库的构建。在获取不同类型的各种因子后，对每一类型中选取的因子进行单因子回测，对超额收益、年化收益率、最大回撤率与信息比率进行初筛，筛选出表现较好的因子。随后进行相关系数矩阵判定，排除掉相关性高的因子。经过这一系列的筛选后，每个类型得到 1-2 个优质因子。在这之后，我们会搭建 LSTM 神经网络模型，通过对这些优质因子的学习，最终达到预测股价的目的。结合真实股价，模型即可提供预测第二日股价涨跌的信号。有了这些信号，我们便可以应用在接下来的股票回测中。通过投资策略的构建，更改同期持仓股票的数量，我们发现，在持仓 100 支股票的时候，能获得最大的年化收益。尽管模型与策略仍需进一步的优化，但我们仍可以得出结论，基于 LSTM 模型的多因子 A 股市场量化选股策略是具有一定实操意义的。

关键词：A 股市场；因子筛选；LSTM 神经网络；股票回测

1 研究问题描述

1.1 问题背景

随着量化投资的不断崛起，多因子投资逐渐被广泛应用于股票投资策略的研究。区别于主观投资，基于量化的因子投资更具有系统性与客观性的特点。而在不同类型的因子终发掘有效的因子便是重中之重。而近年来，随着计算机与人工智能的发展，机器学习与量化投资逐渐结合起来，而如何建立机器学习的模型进行选股以至于获得超额收益成了亟待解决的问题。

1.2 问题提出

1) 如何进行因子优选

2) 搭建何种机器学习模型，该模型如何进行股价预测

3）如何搭建回测框架构造多因子选股策略

2 理论和方法

2.1 LSTM 模型

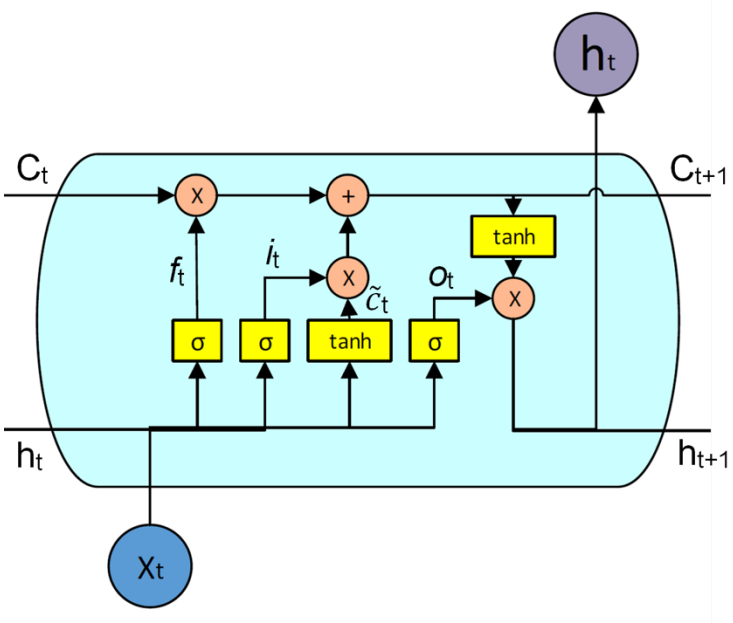
我们最终选择了长短期记忆人工神经网络（Long Short-Term Memory, LSTM）作为股价趋势预测的模型。在做出选择之前，我们比对了大量的模型。普通神经网络因为其“黑箱”的特点，很大概率会给模型的调参与优化造成困难，并且调试与排错的成本高昂，故被放弃。卷积神经网络缺少记忆功能，并且全连模式显得冗余且低效，故也被淘汰。而循环神经网络以其优秀的记忆能力与对非线性特征进行机器学习的能力脱颖而出。对于两种最为常用的循环神经网络，标准循环神经网络（Recurrent Neural Network, RNN）对于长短期记忆人工神经网络（LSTM）有一个重要的功能缺陷，相对严重的梯度消失。“Standard RNN cannot bridge more than 5-10 time steps … Blown-up error signals lead straight to oscillating weights, whereas with a vanishing error”(Staudemeyer & Morris, 2019)。标准循环神经网络会尝试记住所有的信息，A 股市场的股价变化与海量的因子数据会带来沉重的记忆负担，导致最后保留大量的噪音数据，给进一步的分析带来困扰。而长短期记忆人工神经网络拥有记忆细胞，可以对股价变化、因子特征等信息进行筛选，并通过“遗忘门”函数清理噪音数据。“当遗忘门ft被打开时，Ct的梯度可以有效地反向传递给$Ct-1$……通过引入另一个隐藏状态Ct和 3 个门控结构，LSTM 缓解了神经网络训练中的梯度消失问题”（林晓明，2017），这保障了最终数据

的可靠性。

2.1.1 LSTM 模型概述

长短期记忆网络（Long Short-Term Memory, LSTM）是一种时间循环神经网络，与所有的 RNN 类似，都具有一种重复神经网络模块的链式形式。LSTM 模型是为了解决传统 RNN 模型容易产生梯度消失，难以处理长序列数据的问题而设计的。作为非线性模型，LSTM 可作为复杂的非线性单元用于构造更大型深度神经网络（陈亮, 王震, 王刚，2017）。

LSTM 通过精心设计的隐藏层神经元来缓解传统 RNN 的梯度消失问题。 LSTM 模型中，每个序列索引位置 t 时刻除了向前传播隐藏状态 h_t，还有另一个隐藏状态 C_t，该状态被称为细胞状态(Cell State)。实际上，C_t 在 LSTM 中起到了 RNN 中 h_t 的作用。除了细胞状态，LSTM 神经元中还包括三种 RNN 所不具备的门控结构(Gate)，分别遗忘门，输入门和输出门。



2.1.1.1 LSTM 神经网络模型示意图

2.1.2 遗忘门

遗忘门是图中标识为f_t的一个 sigmoid 激活函数：

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

函数输出值介于[0,1]之间，代表对上一层序列索引位置 t-1 时刻细胞状态进行遗忘处理的概率。当模型遇到新的主语x_t并希望对其进行新预测时，可以通过遗忘门消除旧主语 h_{t-1} 的一些无关特征。

2.1.3 输入门

输入门是图中标识为 i_t 的一个 sigmoid 激活函数和标识为 $\tilde{c}t$ 的一个双曲正切函数（tanh）：

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

$$t = \tanh(W_C h_{t-1} + U_C x_t + b_C)$$

输入门负责控制是否将当前时刻的输入值x_t融入细胞状态。i_t的值介于[0,1]之间，代表记住这一层输入信息的概率。i_t与C_t的乘积表示当前细胞状态所需要添加的信息。

2.1.4 输出门

输出门是图中标识为 o_t 的一个 sigmoid 激活函数：

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

函数的值在[0,1]之间，决定了当前时刻细胞状态的输出部分，其目的在于过滤细胞状态，并且从C_t中产生隐藏状态 h_t，输出给下一个序列索引位置 t+1 时刻。其数学表达式为：

$$h_t = o_t \odot \tanh(C_t)$$

2.2 总体思路

首先通过因子筛选搭建因子库，其次通过 LSTM 神经网络模型对于这些优质因子的学习，输出股价涨跌的信号，最后通过回测构建投资策略，并对策略进行总结与分析。

3 模型建立

3.1 因子库构建

我们选择万得作为数据源，在行情数据、交易衍生数据、财务数据等表格中提取所需初始数据。随后通过数据清洗与计算得到因子数据。

3.1.1 数据选择

选取 2017-01-01 至 2022-07-01 沪深 300 成分股数据，根据聚宽(JoinQuant)的因子数据字典，我们选择了价值类、基础类、风险类、情绪类以及财务与质量类五大类因子，从每一类型中选取十至二十个因子。具体各类型所包含因子及部分因子的解释/计算方式详见附录 1。

3.1.2 数据清洗

1） 由于部分因子是季度数据（如财务类因子），首先对他们进行升采样，以季度数据填充其上三个月的数据，如此可转为日频数据。

2） 将所有因子数据按日期合并，合并后，数据日期从 2015-02 至 2022-03。将所有数据按照日期进行分组。

3） 对于每组中的数据，以每列的均值填充缺失值。然后使用绝对值差中位数法（参数 n=5）去除极值。接着使用线性回归，对因子进行市值中性化，去除市值对因子的影响。

4） 上述操作全部完成后，将所有组的数据合并，得到清洗完毕的数据。

3.1.3 因子筛选

对每一类型中选取的因子进行单因子回测，超额收益、年化收益率、最大回撤率与信息比率进行初筛，筛选出表现较好的因子。随后进行相关系数矩阵判定，排除掉相关性高的因子，以避免机器学习过程中的多重共线性问题。

1） 单因子回测

我们使用 alphalens 包对因子进行回测，根据因子表现筛选出表现较好的股票并排除表现较差的股票，最终得到测试报告。

由于篇幅限制，这里将呈现价值类因子的回测结果。

因子代码	因子名称	超额收益 Alpha	信息比率	年化收益率
S_VAL_PE	市盈率	0.038	82.99%	5.21%
S_VAL_PB_NEW	市净率	0.025	76.34%	4.37%
S_VAL_PS	市销率	0.037	87.45%	8.63%
NET_ASSETS_TODAY	当日净资产	0.028	91.07%	5.61%
S_DQ_TURN	换手率	-0.023	62.01%	4.29%
S_DQ_MV	流通市值	0.009	79.23%	3.28%

表 3.1.3.1 价值类因子的回测结果

2) 相关系数选择

a) 在特定类别中的相关性选择

如图所示，在价值类因子中，市盈率 PE、市净率 PB 和市销率 PS 的相关性较高，所以在该三个因子中我们选择保留表现较好的 PS 因子。同理，根据相同的相关性判断逻辑，我们在该类中选择了当日净资产（Net Asset Today）作为另一因子。

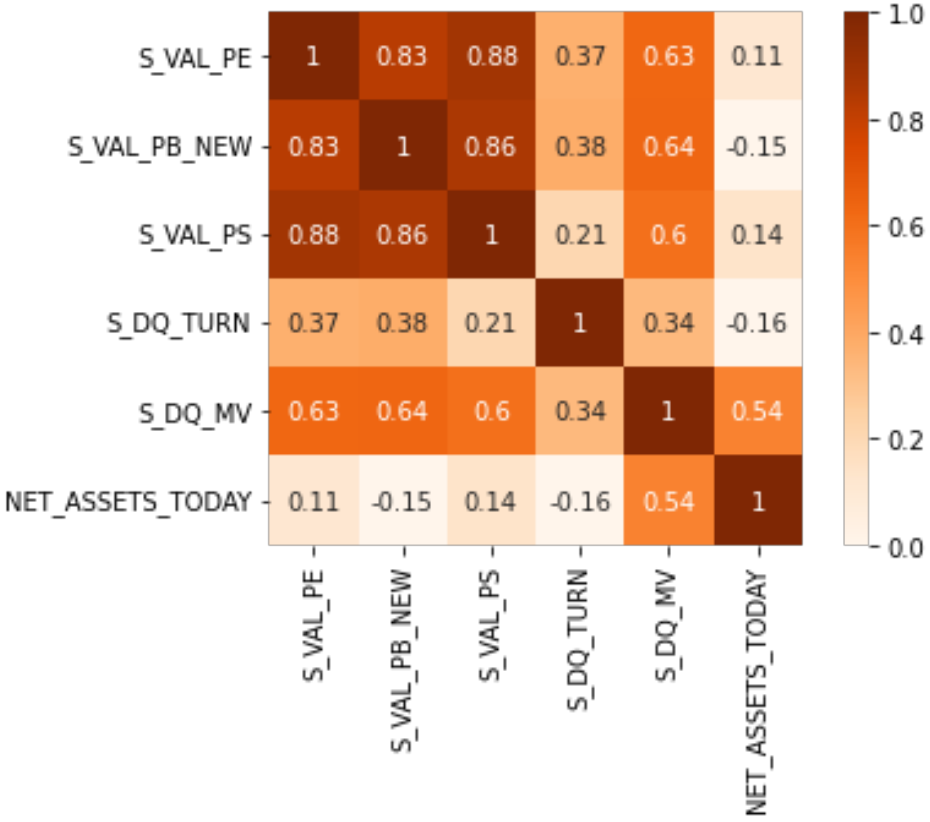


图 3.1.3.2 价值类因子热力图

同时，对于其他类型的因子，也进行上述的筛选工作。并对所有筛选出来的因子绘制相关性矩阵，进行最后一步筛选。

b) 所有因子的相关性判断

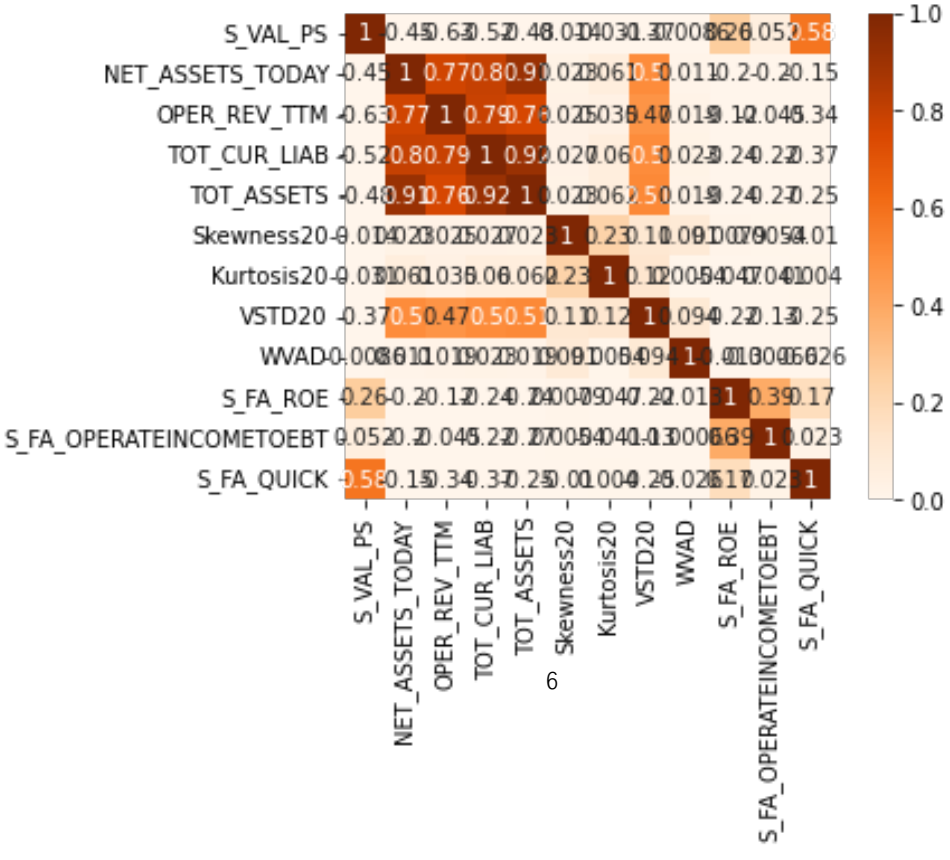


图 3.1.3.3 筛选后的因子热力图

如上图所示，我们选择去除营业收入 TTM、总资产、20 日成交量标准差、经营活动净收益/利润总额、固定资产周转率这几个因子。虽然从回测结果来看，这些因子的表现都较为优异，但由于和其他因子的相关性较高，我们选择去除这些因子。

根据上述步骤，最终所选因子如下表所示：

因子类型	因子代码	因子名称
价值类因子(value_factor)	S_VAL_PS	市销率
	NET_ASSETS_TODAY	当日净资产
基础类因子(basic_factor)	TOT_CUR_LIAB	总流通负债
风险类因子(risk_factor)	Kurtosis20	个股收益的 20 日峰度
	Skewness20	个股收益的 20 日偏度
情绪类因子(trade_factor)	WVAD	威廉变异离散量
财务与质量类因子(fa_factor)	S_FA_ROE	净资产收益率

表 3.1.3.4 终选因子

至此，因子库搭建完成。

3.2 LSTM 模型搭建

3.2.1 搭建框架

在决定了以长短期记忆人工神经网络为基础构建模型之后，Python 语言扩展包丰富，编程自由度高，数据处理能力好的优点我们选择使用其进行编程来构建模型以及之后的模型优化与数据处理。模型以交易日的各因子数据为自变量，当日收盘价为因变量作回归分析。股票的因子数据与收盘价均从“new_factor_data”数据表直接读取。自变量数据表与因变量数据列受参数“mem_day”调节。自变量与因变量搭建的步骤被封装进“LSTM_stock”函数中，输入参数为数据表“stock”与记忆天数“mem_day”为了避免自己搭建 LSTM 神经网络，降低时间成本，我们从 tensorflow 扩展包中的 keras 附属包中直接调用 LSTM 模块。

3.2.2 模型调参与优化

LSTM 模型的调参与优化工作由代码文件“LSTM_Test.py”执行。模型循环运行的步骤被封装进“opt_model”函数中，输入参数为记忆天数列表“mem_days”，神经网络层数列表“lstm_layers”，全连接层层数列表“dense_layers”，神经元个数列表“units”。该函数通过 for 循环更改参数组合，使用预设的训练集（分割自“new_factor_data”数据表）运行 LSTM 模型，每一种参数组合进行 50 次学习，输出模型文件并标注拟合误差率。考虑数据缺失和拟合失败等问题，一轮测试将生成不超过 2,700 个模型，其中拟合度最高，误差率最小的模型会被保存并进入回测阶段。

3.2.3 模型效果检验

通过调参获取的最优模型将会被读取进入代码文件“LSTM_Formal.py”，在这里代码会展示模型的具体信息。利用 sklearn 扩展包的“train_test_split()”函数，我们对检验股票的因子数据与收盘价数据划分训练集与测试集，参数 test_size=0.1。调用读取的模型使用“.evaluate()”函数对测试集进行评估，获得模型准确率与损失数据。将这两项数据与模型拟合度标签进行对比，检验模型的可靠性。

同时，模型对因子数据测试集使用“.predict()”函数，预测检验股票的股价变化，并与真实股价变化一并绘制折线图，将拟合度数据可视化。绘图工具由 matplotlib 扩展包导入。

3.2.4 预测数据输出

通过运行代码文件“LSTM_Run.py”，我们对“沪深 300”股票池中的每一支股票进行模型预测。预测所得的收盘价数据与股票实际收盘价数据被录入表格中，以“file_name.gz”的压缩包格式进行数据储存。所有数据表格将用于下一步的回测分析工作。

4 投资策略分析

4.1 策略构建

前文提到，我们通过相关系数因子选择器与机器学习模型（长短期记忆人工神经网络）进行了因子优选。该模型可以在每日产生所有成分股下一日上涨或下跌的预测值，故也可以将该模型看作因子合成模型，即将所有因子通过机器学习方法进行合成。接下来，我们将对该“合成因子”进行回测，回测具体细节如下：

- 成分股：沪深 300 所有成分股

- 回测周期：2017 年 2 月 10 日----2022 年 3 月 31 日
- 换仓期：该策略为日频策略，采取每日换仓的操作。每日核算因子值，并在下一交易日按照当日收盘价进行换仓
- 回测细节：印花税率为 0.001，最低印花税为 1，佣金率为 0.001，最低佣金费为 5，按比例进行收费。单位为人民币
- 评价方法：回测年化收益率，夏普比率，最大回撤等

4.2 回测结果

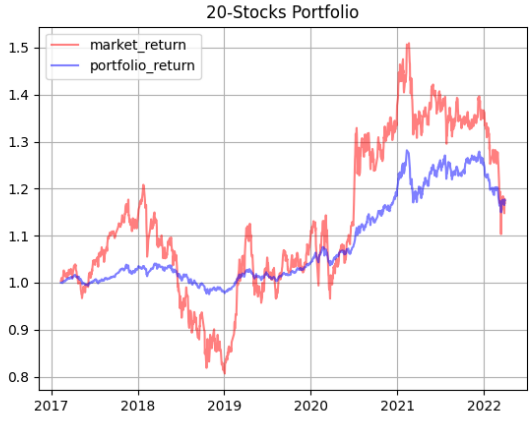


图 4.2.1 每期持仓 20 支股票

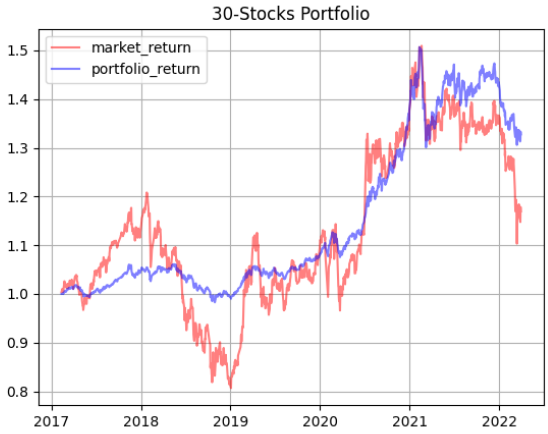


图 4.2.2 每期持仓 30 支股票

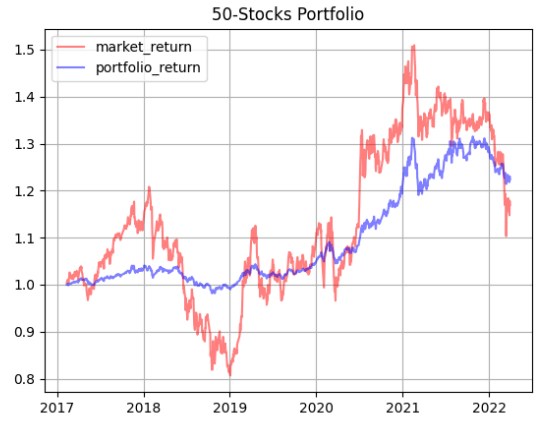


图 4.2.3 每期持仓 50 支股票

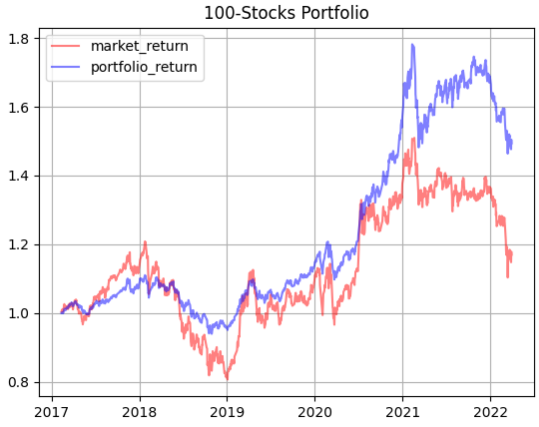


图 4.2.4 每期持仓 100 支股票

持股数量	年化收益率（%）	夏普比率	最大回撤（%）
20	3.31	0.60	10.23
30	5.87	0.77	13.62
50	4.15	0.70	9.28
100	8.41	0.74	17.84

表 4.2.5 每期不同数量持股数数据比较

4.3 结果分析与讨论

从结果来看，所有的投资组合收益都超过了所选择的基准指数（沪深 300）。将持仓范围扩大到 100 股带来的年化收益率最高，但同时面临的问题是难以避免的大回撤。在 2021 年，所有投资组合都面临了一个较大的回撤，目前该策略并没有找到合适的办法来规避 2021 年出现的波动。同时，所有的投资组合都有较好的夏普比率，这使得承担单位风险的同时能够获得较为理想的超额回报。

后续的策略改进需要从如下几个方面进行改善：

1. 改变持仓期，每日调仓虽然比较灵活，但会有手续费较高的问题，改变持仓周期后每周、或每月为单位进行因子测算有可能可以带来更大的收益。
2. 找到控制波动的方法，以规避较大的回撤风险。

5 参考文献

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. 17-18. https://doi.org/10.48550/arXiv.1909.09586

林晓明.（2017）. 华泰人工智能系列之九：人工智能选股之循环神经网络模型. *金工研究/深度研究*. 华泰证券.

陈亮, 王震, 王刚. （2017）. 深度学习框架下 LSTM 网络在短期电力负荷预测中的应用. *电力信息与通信技术*. 8-11.

6 附录

附录 1：各类型因子汇总

因子类型	因子代码	因子名称	计算方式（/因子解释）
价值类因子 (value_factor)	S_VAL_PE	市盈率	
	S_VAL_PB_NEW	市净率	
	S_VAL_PS	市销率	市销率 TTM=（股票在指定交易日期的收盘价 * 当日人民币外汇牌价 * 截至当日公司总股本）/营业总收入 TTM
	S_DQ_TURN	换手率	
	S_DQ_MV	流通市值	
	NET_ASSETS_TODAY	当日净资产	为公布的母公司净资本及相关风险控制指标之一。
基础类因子 (basic_factor)	OPER_REV	营业收入	
	OPER_PROFIT	营业利润	
	TOT_PROFIT	总利润	
	LESS_SELLING_DIST_EXP	销售费用	
	EBIT	息税前利润	
	EBITDA	息税折旧摊销前利润	
	TOT_CUR_ASSETS	总流通资本	
	TOT_CUR_LIAB	总流通负债	流动负债合计是指企业在一年内或超过一年的一个营业周期内需要偿还的债务，包括短期借款、应付帐款、其他应付款、应付工资、应付福利费、未交税金和未付利润、其他应付款、预提费用等。
	TOT_ASSETS	总资产	

	FIX_ASSETS	固定资产	
	NET_PROFIT_PARENT_COMP_TTM	归属母公司净利润 TTM	
	NET_CASH_FLOWS_OPER_ACT_TTM	经营活动产生的现金流量净额 TTM	
	OPER_REV_TTM	营业收入 TTM	
风险类因子 (risk_factor)	Variance20	20 日年化收益方差	
	Kurtosis20	个股收益的 20 日峰度	
	Skewness20	个股收益的 20 日偏度	
	SharpeRatio20	20 日夏普比率	
情绪类因子 (trade_factor)	VOL20	20 日平均换手率	
	VSTD20	20 日成交量标准差	
	TVMA20	20 日成交金额的移动平均值	
	WVAD	威廉变异离散量	(收盘价－开盘价)/(最高价－最低价)×成交量，再做加和，使用过去 6 个交易日的数据
财务与质量类因子 (fa_factor)	S_FA_FCFF	企业自由现金流量	
	S_FA_EPS_BASIC	基本每股收益	
	S_FA_BPS	每股净资产	
	S_FA_ORPS	每股营业收入	
	S_FA_NETPROFITMARGIN	销售净利率	
	S_FA_GCTOGR	营业总成本/营业总收入	
	S_FA_ROE	净资产收益率	净利润/期末股东权益
	S_FA_OPERATEINCOMETOEBT	经营活动净收益/利润总额	
	S_FA_CATOASSETS	流动资产/总资产	

	S_FA_CURRENT	流动比率	
	S_FA_QUICK	速动比率	
	S_FA_FATURN	固定资产周转率	
	S_FA_OPTOLIQDEBT	营业利润/流动负债	
	S_FA_PROFITTOOP	利润总额/营业收入	