1. 摘要

本文中的交易策略基于 2010 年 1 月 1 日到 2021 年 12 月 31 日 A 股市场的月频交易数据构建，并且基于支持向量机（SVM）方法构建不同选股策略的回测表现。本文中的模型通过下期超额回报数据和当期因子数值进行拟合训练,将相应的预测结果给予当期相应股票,预测结果越大，下期股票上涨概率越大。然后，基于当期月末的预测结果将股票从高到低排序并选取排名前 X 的股票构建投资组合。回测结果表明通过该模型构建的组合显著跑赢深证成指。

2. 策略主要思想

基于股市价格依赖于过去信息的假定，可以利用已有的市场信息来预测未来股市走向。本文进一步假定，每一只股票过去的特征或者构建的因子独立地决定未来价格走向。从而利用多变量分析的方法，利用上一个时间截面的一支股票的因子特征来预测下一个时间截面股票的走势。通过当期因子值以及下期超额回报率训练 SVM 机器学习模型，从而做到通过当期基本面因子数据判断股票下一期的走势，从而提前建仓布局下一期将要上涨的股票。

3. 模型训练

3.1 模型概述

支持向量机（SVM）是一种用于分类和回归的机器学习算法。它允许训练模型中出现一些噪音从而避免过度拟合的情况。SVM 有不同的核函数，本文中使用的核函数为线性核函数。此模型的优点为能够很好的处理高维数据和非线性数据。

3.2 因子构建

本模型中采用的因子主要为估值因子（市盈率倒数和市净率倒数），市值因子（市值对数），波动率因子（过去一个月的波动率）以及财务质量因子（EPS，ROE 和 Operating Profit Margin）。其中，将单日市盈率或者市净率为负的股票剔除，因为在分析上缺少意义。之后，通过去极值的方法，进一步筛选掉一些股票，最后以月为单位，取每月最后一个交易日的因子值为月频因子。

Trade_status	Close	Excess_Return	Monthly_Return	Index_Return	deviation_factor	MarketCap_factor	BP_factor	EP_factor	EPS	ROE	OP_factor
1	11.380000	0.011477	-0.098108	-0.109585	0.016000	4.874205	0.473171	0.041560	0.1370	2.51	19.228961
1	6.290000	0.081813	-0.027772	-0.109585	0.015520	4.834693	0.430645	0.025273	0.0932	3.38	17.365193
1	7.760000	0.038232	-0.071354	-0.109585	0.018587	5.140025	0.287043	0.026210	0.0185	0.83	17.710140
1	7.320000	-0.027013	-0.136598	-0.109585	0.016521	4.799176	0.424719	0.056163	0.1049	3.23	19.351910
1	7.670000	0.020562	-0.089023	-0.109585	0.022022	4.752079	0.358873	0.030008	0.0688	2.47	18.757059
...	...	...	...	...	...	...	...	...	...	...	...
1	121.199997	-0.050606	-0.045413	0.005192	0.045407	6.043579	0.143579	0.015946	0.4074	2.37	18.994395
1	70.449997	-0.382145	-0.376953	0.005192	0.065866	6.072785	0.172479	0.025723	0.3634	3.04	19.279994
1	705.000000	-0.092804	-0.087612	0.005192	0.023220	6.591311	0.038088	0.010000	1.9786	7.65	19.131178
1	254.979996	-0.088048	-0.082855	0.005192	0.040004	6.447750	0.033876	0.010000	0.5044	5.88	18.533801
1	47.470001	-0.153497	-0.148305	0.005192	0.049271	5.250874	0.092085	0.020265	0.2131	5.00	18.332009

此图由 build_factor(df)构建。

3.3 特征和标签

在 Parameter 类里汇总所有模型和策略需要的参数，其中 percent_select 在模型训练中取下期超额回报排名靠前的 30%股票作为正例,取排名靠后的 30%股票作为反例,而非所有的股票。然后通过 label_data(df)函数将正例标为 1，将反例标为 0，以方便 SVM 模型学习这些股票的特征。特征值为所有当期因子，标签值为 label_data(df)函数给定的标签。

Close	Excess_Return	Monthly_Return	Index_Return	deviation_factor	MarketCap_factor	BP_factor	EP_factor	EPS	ROE	OP_factor	Selected
34.540001	0.443856	0.334271	-0.109585	0.023515	4.107432	0.124740	0.015386	0.6251	13.53	18.995159	1.0
6.740000	0.426768	0.317183	-0.109585	0.018392	3.668631	0.388863	0.010000	0.0812	3.05	17.602537	1.0
13.720000	0.374818	0.265233	-0.109585	0.018913	3.234040	0.192256	0.010000	0.0245	0.93	15.353163	1.0
20.709999	0.373970	0.264385	-0.109585	0.028675	3.841547	0.219370	0.025819	0.2194	4.71	17.960590	1.0
10.540000	0.369629	0.260043	-0.109585	0.015438	4.662001	0.258806	0.032953	0.1158	4.16	18.846384	1.0
...	...	...	...	...	...	...	...	...	...	...	...
480.000000	-0.227486	-0.222294	0.005192	0.033153	6.707973	0.059344	0.010000	0.7034	8.50	18.651563	0.0
23.500000	-0.232514	-0.227322	0.005192	0.045484	5.484376	0.104587	0.010000	0.0218	0.89	17.257042	0.0
10.900000	-0.276951	-0.271759	0.005192	0.050749	3.925701	0.274379	0.020574	0.1192	2.85	17.642767	0.0
21.790001	-0.307303	-0.302111	0.005192	0.044632	5.718555	0.123122	0.011275	0.0860	3.26	18.569680	0.0
70.449997	-0.382145	-0.376953	0.005192	0.065866	6.072785	0.172479	0.025723	0.3634	3.04	19.279994	0.0

3.4 特征预处理

根据华泰证券人工智能研究报告，将数据分为训练集和交叉检验集。在进行样本内和样本外数据拆分的时候，本文采用分配更多数据作为样本内数据，不同于研报中的分配，因为通常样本内数据需要大于样本外数据。之后，使用主成分分析函数（PCA）降低数据维度，以便 SVM 模型可以更准确的进行判断

3.5 训练结果

	StatDate	Code	Close	Monthly_Return	Index_Return	y_score
99734	2018-01-31	S.CN.SSE.600507	15.120000	0.271485	-0.027003	2.766975
99841	2018-01-31	S.CN.SZSE.300166	12.500000	0.258264	-0.027003	0.018609
100607	2018-01-31	S.CN.SSE.600673	7.470000	0.256131	-0.027003	-0.270136
100376	2018-01-31	S.CN.SSE.600845	19.350000	0.249865	-0.027003	-0.613886
100045	2018-01-31	S.CN.SSE.600789	9.330000	0.239210	-0.027003	-0.200777
...	...	...	...	...	...	...
153864	2021-11-30	S.CN.SSE.603290	480.000000	-0.222294	0.005192	0.369245
153254	2021-11-30	S.CN.SZSE.000829	23.500000	-0.227322	0.005192	-0.757764
153573	2021-11-30	S.CN.SSE.600215	10.900000	-0.271759	0.005192	-0.360670
153534	2021-11-30	S.CN.SSE.600110	21.790001	-0.302111	0.005192	-0.503758
153562	2021-11-30	S.CN.SZSE.300741	70.449997	-0.376953	0.005192	-0.649027

26544 rows × 6 columns

此图由 svm_train(df)函数构建。首先，通过训练集数据训练模型，之后用训练好的模型在训练集特征值上做预测。y_score 为预测结果，代表涨跌的概率，如果其数值越大，证明下一期该股票上涨的概率较大。

3.6 模型评价

```
training set, accuracy = 0.52
training set, AUC = 0.55
test set, accuracy = 0.52
test set, AUC = 0.53
```

此模型在训练集上的正确率为 52%，在训练集上分类正确的概率为 55%，在交叉检验集上的正确率为 52%，在交叉检验集上分类正确的概率为 53%

```
test set, 2018-01 accuracy = 0.54
test set, 2018-01 AUC = 0.61
test set, 2018-02 accuracy = 0.52
test set, 2018-02 AUC = 0.56
test set, 2018-03 accuracy = 0.53
test set, 2018-03 AUC = 0.59
test set, 2018-04 accuracy = 0.54
test set, 2018-04 AUC = 0.68
test set, 2018-05 accuracy = 0.54
test set, 2018-05 AUC = 0.62
test set, 2018-06 accuracy = 0.52
test set, 2018-06 AUC = 0.48
test set, 2018-07 accuracy = 0.51
test set, 2018-07 AUC = 0.53
test set, 2018-08 accuracy = 0.48
test set, 2018-08 AUC = 0.49
test set, 2018-09 accuracy = 0.51
test set, 2018-09 AUC = 0.54
test set, 2018-10 accuracy = 0.47
test set, 2018-10 AUC = 0.47
test set, 2018-11 accuracy = 0.50
test set, 2018-11 AUC = 0.57
test set, 2018-12 accuracy = 0.54
test set, 2018-12 AUC = 0.64
test set, 2019-01 accuracy = 0.52
test set, 2019-01 AUC = 0.48
```

```
test set, 2019-02 accuracy = 0.52
test set, 2019-02 AUC = 0.58
test set, 2019-03 accuracy = 0.53
test set, 2019-03 AUC = 0.62
test set, 2019-04 accuracy = 0.50
test set, 2019-04 AUC = 0.50
test set, 2019-05 accuracy = 0.54
test set, 2019-05 AUC = 0.59
test set, 2019-06 accuracy = 0.50
test set, 2019-06 AUC = 0.58
test set, 2019-07 accuracy = 0.53
test set, 2019-07 AUC = 0.65
test set, 2019-08 accuracy = 0.51
test set, 2019-08 AUC = 0.59
test set, 2019-09 accuracy = 0.56
test set, 2019-09 AUC = 0.64
test set, 2019-10 accuracy = 0.54
test set, 2019-10 AUC = 0.59
test set, 2019-11 accuracy = 0.51
test set, 2019-11 AUC = 0.57
test set, 2019-12 accuracy = 0.51
test set, 2019-12 AUC = 0.59
test set, 2020-01 accuracy = 0.54
test set, 2020-01 AUC = 0.61
test set, 2020-02 accuracy = 0.51
test set, 2020-02 AUC = 0.51
test set, 2020-03 accuracy = 0.53
test set, 2020-03 AUC = 0.66
test set, 2020-04 accuracy = 0.51
test set, 2020-04 AUC = 0.62
```

```
test set, 2020-04 accuracy = 0.51
test set, 2020-04 AUC = 0.62
test set, 2020-05 accuracy = 0.52
test set, 2020-05 AUC = 0.65
test set, 2020-06 accuracy = 0.51
test set, 2020-06 AUC = 0.57
test set, 2020-07 accuracy = 0.53
test set, 2020-07 AUC = 0.63
test set, 2020-08 accuracy = 0.50
test set, 2020-08 AUC = 0.49
test set, 2020-09 accuracy = 0.57
test set, 2020-09 AUC = 0.70
test set, 2020-10 accuracy = 0.47
test set, 2020-10 AUC = 0.43
test set, 2020-11 accuracy = 0.55
test set, 2020-11 AUC = 0.59
test set, 2020-12 accuracy = 0.56
test set, 2020-12 AUC = 0.69
test set, 2021-01 accuracy = 0.46
test set, 2021-01 AUC = 0.44
test set, 2021-02 accuracy = 0.48
test set, 2021-02 AUC = 0.46
test set, 2021-03 accuracy = 0.57
test set, 2021-03 AUC = 0.74
test set, 2021-04 accuracy = 0.51
test set, 2021-04 AUC = 0.51
test set, 2021-05 accuracy = 0.51
test set, 2021-05 AUC = 0.55
test set, 2021-06 accuracy = 0.54
test set, 2021-06 AUC = 0.57
test set, 2021-07 accuracy = 0.51
test set, 2021-07 AUC = 0.55
test set, 2021-08 accuracy = 0.47
test set, 2021-08 AUC = 0.43
test set, 2021-09 accuracy = 0.53
test set, 2021-09 AUC = 0.61
test set, 2021-10 accuracy = 0.45
test set, 2021-10 AUC = 0.50
test set, 2021-11 accuracy = 0.46
test set, 2021-11 AUC = 0.37
```

上述 3 个图为每月模型评估，从 2018 年 1 月到 2021 年 11 月。

4. 策略构建

4.1 信号描述

本文中的策略对应的是 A 股市场，所以策略为只看多，不看空。通过 SVM 模型的预测结果，前 n（选股数量）个预测结果较高的股票会被标记为 1（买入），其他均为 0（不参与交易）。

4.2 仓位管理

初始资金为：1000 万

个股仓位：等权分配资金

	Bought_on	Code	Close	Monthly_Return	Index_Return	y_score	Signal	Trade Volume	Monthly_Result	Monthly_Profit	Sell_Date
99561	2018-01-31	S.CN.SZSE.000537	17.040001	-0.193236	-0.027003	4.606274	1	74.0	0.0	-24366.234520	2018-02-28
99734	2018-01-31	S.CN.SSE.600507	15.120000	0.271485	-0.027003	2.766975	1	83.0	0.0	34070.240268	2018-02-28
100293	2018-01-31	S.CN.SSE.601003	7.340000	0.063633	-0.027003	2.461343	1	171.0	0.0	7986.843109	2018-02-28
100392	2018-01-31	S.CN.SSE.601155	40.220001	-0.149824	-0.027003	2.349961	1	32.0	0.0	-19282.982359	2018-02-28
99507	2018-01-31	S.CN.SSE.600338	34.150002	0.087682	-0.027003	2.083199	1	37.0	0.0	11079.060710	2018-02-28
...	...	...	...	...	...	...	...	...	...	...	...
153879	2021-11-30	S.CN.SSE.688188	415.640015	-0.068195	0.005192	-0.080947	1	4.0	0.0	-11337.766433	2021-12-31
153170	2021-11-30	S.CN.SZSE.000799	224.580002	-0.049848	0.005192	-0.085527	1	6.0	0.0	-6716.905001	2021-12-31
154026	2021-11-30	S.CN.SZSE.002557	53.580002	0.143861	0.005192	-0.087617	1	24.0	0.0	18499.373800	2021-12-31
153380	2021-11-30	S.CN.SSE.600563	261.200012	-0.109377	0.005192	-0.098657	1	5.0	0.0	-14284.694827	2021-12-31
153596	2021-11-30	S.CN.SZSE.300459	4.550000	0.226142	0.005192	-0.107304	1	275.0	0.0	28296.010131	2021-12-31

3760 rows × 11 columns

此表格为选股数量为 80 的仓位管理图，由 trading_book(df,buy)函数生成。因为策略为只看多，不看空，所以信号在这里并没有太大意义。

4.3 策略构建

股票池：为 stock_price_standard 文件里的所有股票，通过 ST 股以及停盘时间等因素，层层筛选可供策略交易股票。

回测区间：2018 年 1 月 1 日至 2021 年 11 月 31 日

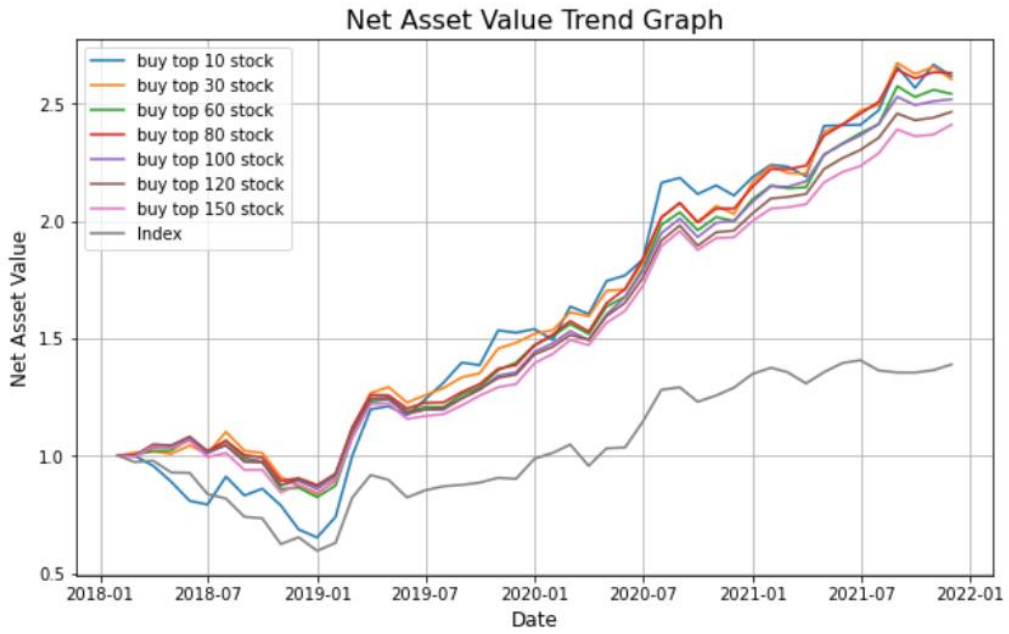
换仓期：月频调仓

数据处理方式：通过合并数据集以及因子去极值构建更高质量的数据集，以供 SVM 机器学习模型更高效做出判断。

评价指标：本策略选用年化回报率，年化波动率，信息比率以及最大回撤率来判断组合表现的优良等级。

5. 策略表现

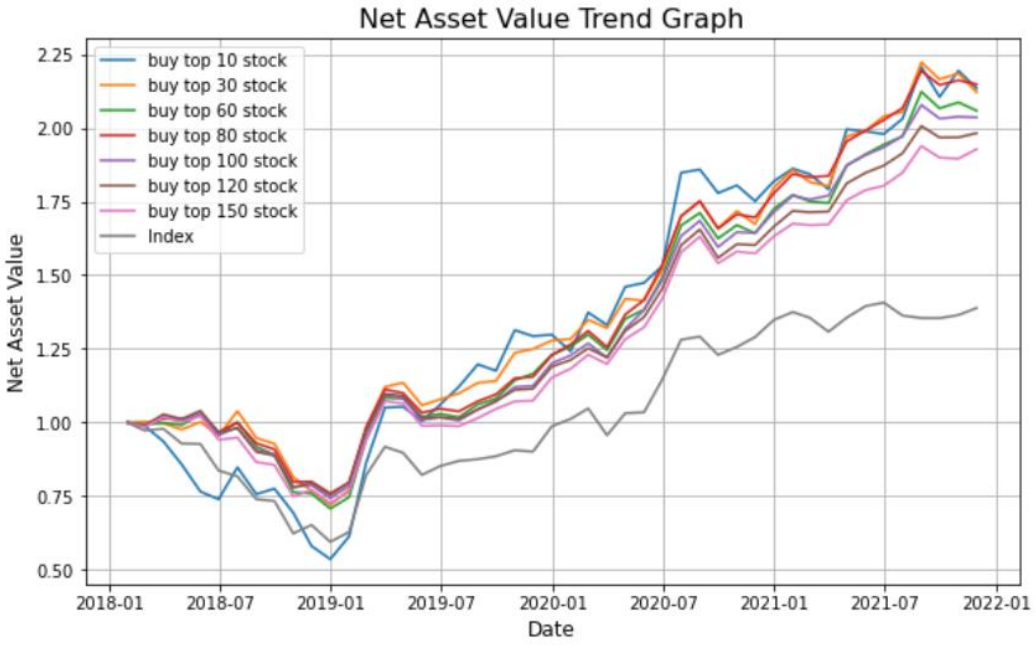
5.1 回测结果



	Strategy	Annual Return	Annual Volatility	Info Ratio	Max Drawdown
1	Strategy1	0.412771	0.331803	1.244023	34.831388
2	Strategy2	0.409484	0.256419	1.596935	24.165230
3	Strategy3	0.393569	0.234591	1.677684	23.233962
4	Strategy4	0.416377	0.219680	1.895375	18.028031
5	Strategy5	0.387703	0.209729	1.848591	19.855299
6	Strategy6	0.374030	0.201872	1.852806	19.645582
7	Strategy7	0.360196	0.203457	1.770378	20.907924

这两个图通过 back_testing(df,buy),strategy_evaluation(df)和 graph(df)构建。从净值图趋势可以看出，通过 SVM 机器学习模型所创造的 7 个股票组合均在 2018 年 1 月 1 日至 2021 年 11 月 31 日跑赢深证成指。从信息比率和最大回撤来看，当选股数量为 80 的时候，每单位风险收益能力最高并且回撤最低。

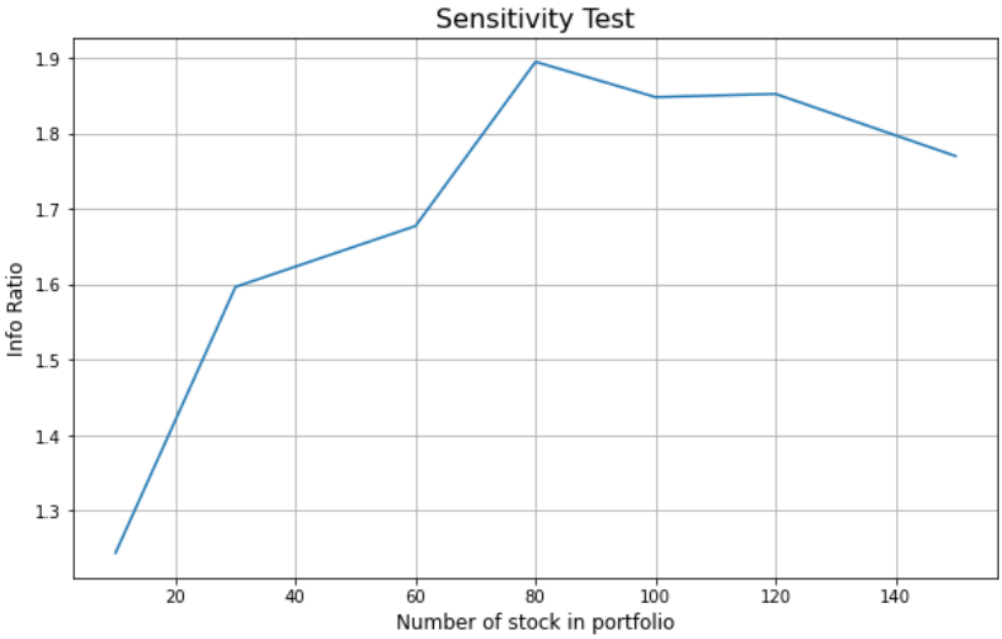
5.2 滑点分析



	Strategy	Annual Return	Annual Volatility	Info Ratio	Max Drawdown
1	Strategy1	0.289452	0.331567	0.872981	46.381388
2	Strategy2	0.286165	0.256116	1.117328	30.684955
3	Strategy3	0.270250	0.234274	1.153561	31.314179
4	Strategy4	0.293058	0.219319	1.336214	25.933024
5	Strategy5	0.264384	0.209381	1.262691	27.808617
6	Strategy6	0.250711	0.201526	1.244060	27.506153
7	Strategy7	0.236877	0.203129	1.166140	28.957455

这两个图通过 trading_cost(df,buy),strategy_evaluation(df)和 graph(df)构建。从净值图趋势可以看出，通过 SVM 机器学习模型所创造的 7 个股票组合仍然可以在 2018 年 1 月 1 日至 2021 年 11 月 31 日跑赢深证成指，但是所有评估指标均有所恶化，交易成本对该投资策略影响略大。

5.3 参数敏感性分析



此图通过 Sensitivity（df）函数构建。从参数敏感性分析图标可以看出当选股数量为 80 的时候，信息比率最高，因此我们可以判定在该模型下，最优月频交易组合选股数量为 80。

6. 总结

通过 PCA 降维数据来提高 SVM 模型预测准确性是很有意义的，当选股数量增多时，组合会更加多样化，但是过度添加，可能会购买到模型判定为下跌的股票而且提高组合里标的间的相关性，从而加大组合风险，并降低组合回报。本文发现当选股数量为 80 的时候，回撤率最低并且信息比率最高。

下一步研究思路：本文中复现的模型比较简单，需要考虑更多因素，例如：实际交易成本，最佳换仓频率（本文中为月频）等等，总的来说，该模型还有待进一步提升。未来可以将此模型的核函数进行改变，尝试不同训练拟合结果，还可以使用不同的模型来探究更优解。最后，由于时间关系，因子挑选较少，应当加大因子数量，同时消除因子间的共线性。根据之前海通证券的量化研究表明，多因子模型中因子间的共线性可能会导致模型训练结果不够好，当因子通过正交处理后，线性相关性将不会存在，从而可以得到更好的训练结果。