1. 摘要

本策略参考中州期货撰写的“期货基本面量化系列专题（三）：棉花择时策略探索之 XGBoost”量化专题报告，以郑棉期货合约为交易标的，从估值和驱动两方面选取不同因子，通过 XGBoost 模型构建择时策略并进行回测。结果显示，不同因子在一定程度上捕捉到中短期趋势和短期大幅变动的趋势。在参数进一步调优和组合后，该策略有比较可观的效果。

2. 策略主要思想

a. 基本面择时因子

对于期货品种，可以从估值和驱动两方面分析，从基本面的角度，可以从基差、利润、供给、消费、库存等角度生成较为量化的指标，这些指标与期货合约价格有着非常紧密的联系。然而，这种联系通常是非线性的，并且可能存在较为复杂的逻辑关系。简单的通过基本面数据进行择时，可以管中窥豹，但可获取的信息是非常有限的。

b. 机器学习模型之 XGBoost

XGBoost 模型提供了一个从统计学习的角度处理基本面因子数据（结构化数据）的方法。该模型基于经典机器学习模型决策树，通过梯度算法提升计算速度，是近年来广受欢迎的非线性拟合模型。

将基本面因子与非线性模型相结合，让模型以更加复杂但全面的方式解构基本面因子与价格趋势之间的关系，用模型生成的信号提供择时时机指导交易，是该策略的主要思想。

3. 指标分析与模型训练

a. 单一指标 1：基差因子

基差是指期货合约对应的现货价格与期货合约价格之差。从估值的角度，基差直接反映了市场对该品种未来价格走势的预期。期货价格相对于现货价格折价称为“贴水”，反之则成为“升水”。当期货合约临近实物交割期并处于贴水状态时，一般认为应该做多合约，因为期货价格在临近交割时将收敛于现货价格。反之通过做空合约，可以获得合约升水带来的潜在盈利机会。下文图 1 展示了 2007 年至今郑棉期货主连合约价格与现货价格的走势。可以发现，期货价格时而高于现货，时而低于现货，围绕在其上下反复波动。基于这样的观察，我们尝试通过基差因子来构建数据集，通过非线性模型训练找到**当前基差**与**未来一段时间期货价格走势**的相对关系，从而构建择时策略来买入或做空合约。

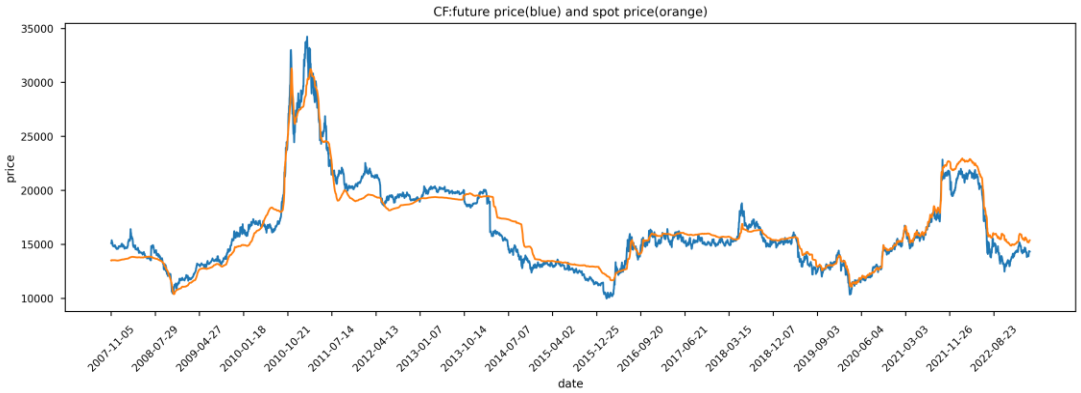


图 1. 郑棉期货与现货价格走势

在实验中，我们选用数据库中的期货收盘价（连续）作为郑棉期货价格走势的表征，选用中国棉花价格指数作为现货价格参考，通过两者差值计算每日基差。考虑到现货指标数据获取的滞后性，将基差因子后移一日作为特征（X）进行训练。选用未来 n 日后期货收盘价相对于今日收盘价的涨跌方向作为标签（y），如下表 1 所示。其中，2007-11-07 的基差因子-1587 为 2007-11-06 当日期限价格的基差值，而 2007-11-07 的 y=1 是 2007-11-08 期货合约收盘价相比 2007-11-06 期货收盘价的涨跌方向，此处 y=1 意味着 2007-11-08 收盘价高于 2007-11-06 期货收盘价。**注意这里假定我们身处 2007-11-07，生成交易信号时并不知道当日收盘价价格，故不能使用当日收盘价。**

date	factor_indicator	y
06/11/2007	-1586	1
07/11/2007	-1587	1
08/11/2007	-1857	0
09/11/2007	-1732	0
12/11/2007	-1662	0
13/11/2007	-1597	0
14/11/2007	-1342	0
15/11/2007	-1379	0
16/11/2007	-1299	0

表 1. 基差因子模型 dataset

在模型参数上，通过 5-fold 交叉验证调参，选择出最优的超参数如下。使用最优的超参数训练模型，得到训练集内和测试集内对未来期货价格走势的预测值。

参数名	参数意义	调参结果
max_depth	Boosting 过程中每棵树的最大深度	7
learning_rate	较小的 learning rate 有助于缓解 over fitting	0.3
subsample	每次选取训练集中数据的比重	0.6

num_round	决策树的数量。选取较大的数值意味着使用比较复杂的模型	30

在仓位管理上，当模型给出 1 时，满仓做多合约（如果本身正在做多合约，则保持持有状态）；当模型给出 0 时，满仓做空合约（如果本身正在做多合约，则保持持有状态）。下文图 2 给出了该策略的仓位持有以及做多做空情况。

从回测表现来看，模型在训练集上的交易胜率达到 68.98%，净值从 1 增长到 33.08。在测试集上的交易胜率达到 50%。策略具体表现在第四部分的表格中列出。

另外，由策略信号图可以看出，基差因子给出的交易信号比较频繁，考虑到我们在做的是日频策略并使用每日 CLOSE 价格作为收益率因子，以及频繁交易会增加手续费的问题，可以筛掉一部分持仓时间较短（如 1-3 天）的交易信号来优化策略。在筛选后，年均交易频率减少约一半，收益率表现也较之更优。上图以及第四部分给出的回测表现是去除掉持仓时间等于 1 天的交易信号的得出的结果。

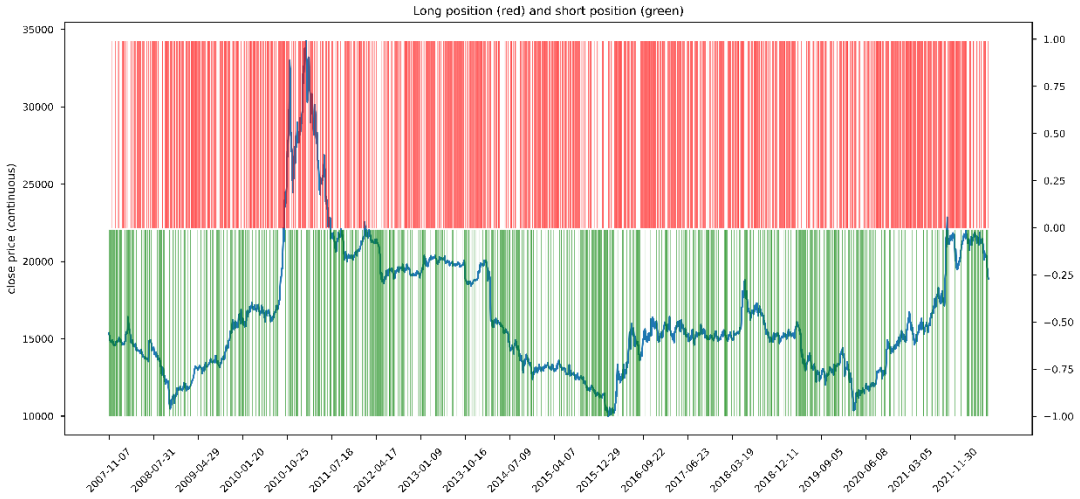


图 2. 郑棉主连合约走势与基差的 XGBoost 策略信号

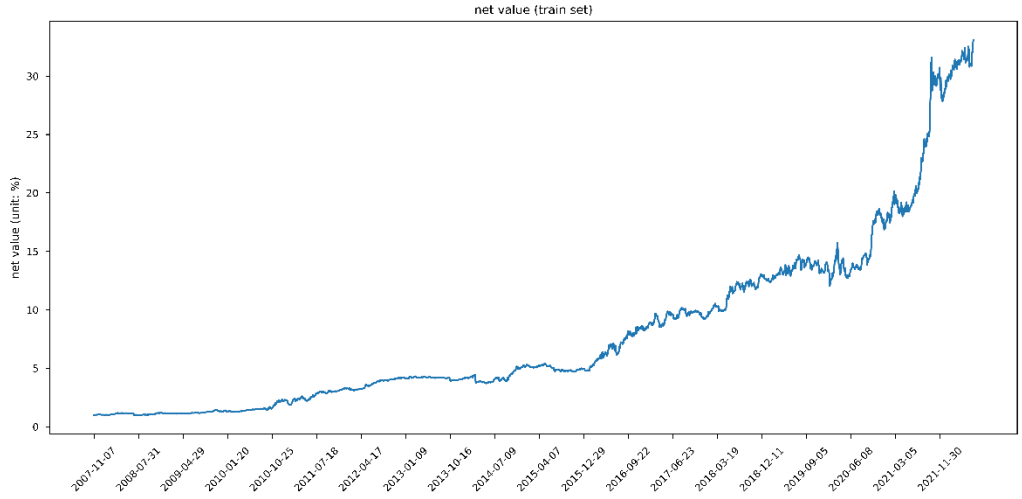


图 3. 棉花基差策略训练集净值走势

需要注意的是，我们根据研报中的结果选择两天作为时间间隔来检验基差因子对合约价格走势的预测效果。在后续的实验中，应该选取不同的时长（如从 1 天到 7 天）来分别检验因子的预测效果。本周因时间限制暂且选择 2 天作为时间间隔。（**在代码设计上，indicator_processing.py 中将 gap_year 设置为可调变量，可以方便地选取不同时间间隔来进行测试**）

仔细观察图 2，我们得到与中州期货研报相类似的结论，即基差因子比较善于捕捉中短期趋势。图 4 黄色箭头标注出了该策略明确给出的持仓一段时间的时段，这几个时段内，策略较好的捕捉到中短期策略。思考其背后原因，可能是这些时段接近交割期限，故基差因子较好地反映了一段时间的期货价格走势。

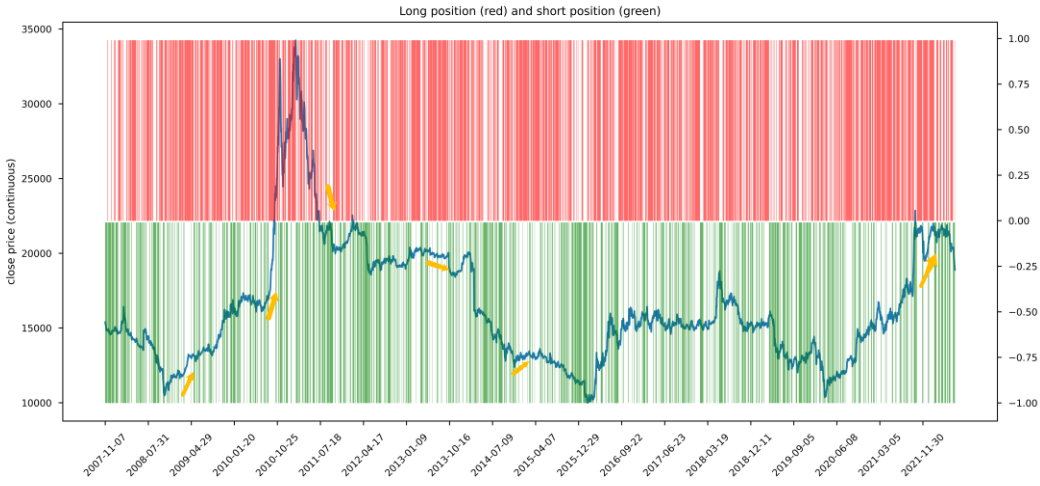


图 4. 棉花基差策略对中短期趋势的捕捉

b. **单一指标 2：利润因子**

利润指标在一定程度上反映了合约的相对价格。我们选取数据库中进口棉利润作为利润因子指标。观察下图合约价格（蓝色）与进口利润（黄色）的相对走势，我们发现两者走势时而趋同，时而相反，简单的线性模型难以刻画这样的关系，XGBoost 作为非线性集成模型在此派上用场。**回想决策树的判断逻辑，我们猜想，当利润水平处于不同分位时，利润的走势与期货价格的走势可能有着不同关系。**

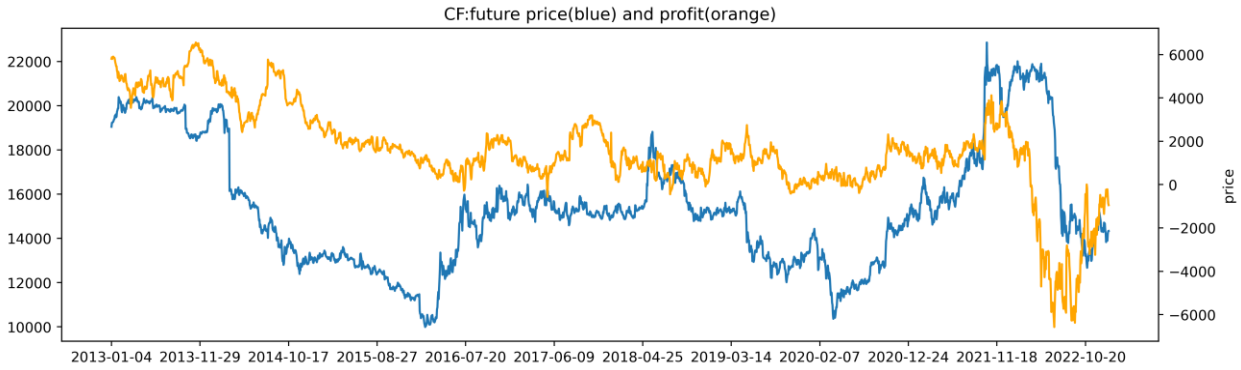


图 5. 棉花期货价格与进口利润走势

在模型建立、参数调优以及持仓管理上，我们沿用上文中描述的方法。超参数调优后的结果如下

参数名	参数意义	调参结果
max_depth	Boosting 过程中每棵树的最大深度	8
learning_rate	较小的 learning rate 有助于缓解 over fitting	0.3
subsample	每次选取训练集中数据的比重	0.6
num_round	决策树的数量。选取较大的数值意味着使用比较复杂的模型	30

从回测表现来看，模型在训练集上的交易胜率达到 75.49%，净值从 1 增长到 3.83。在测试集上的交易胜率高达 67.80%。策略具体表现在第四部分的表格中列出。

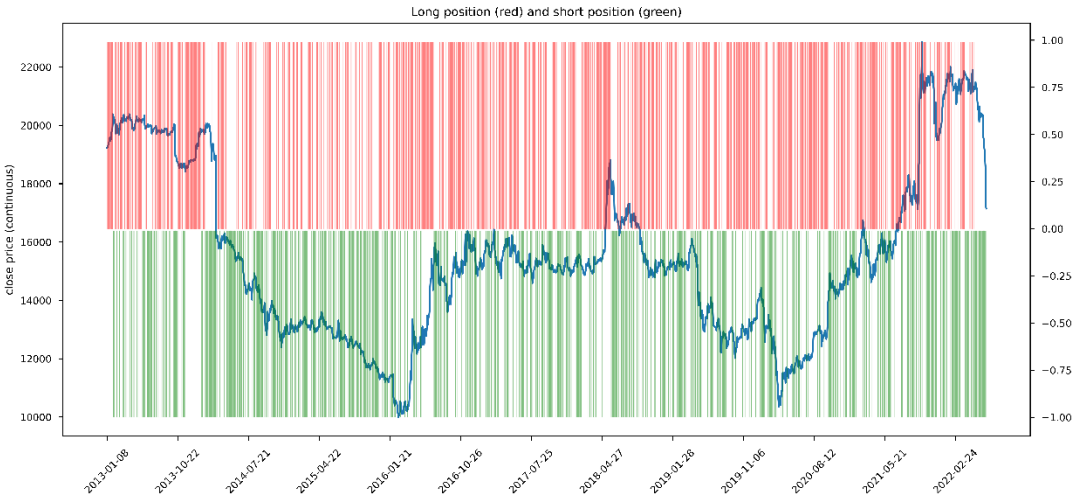


图 6. 郑棉主连合约走势与利润因子的 XGBoost 策略信号

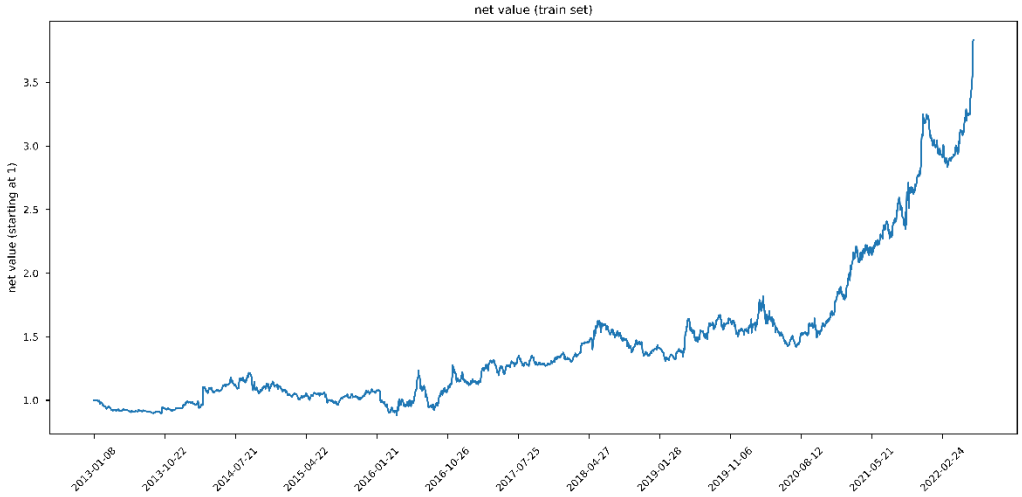


图 7. 棉花利润策略训练集净值走势

仔细观察图 6，发现利润因子比较善于捕捉大趋势。在几次期货合约价格出现短期较大价格浮动时都给出了正确的持仓方向。思考原因，利润因子指标存在滞后性，需要经历生产-消费的周期回馈到生产端，进而对价格走势产生影响。

c. 单一指标 3：消费因子

在本部分，我们以数据库中的“棉花：抛储：成交量”作为需求指标，从消费端的角度构建因子进行测试。

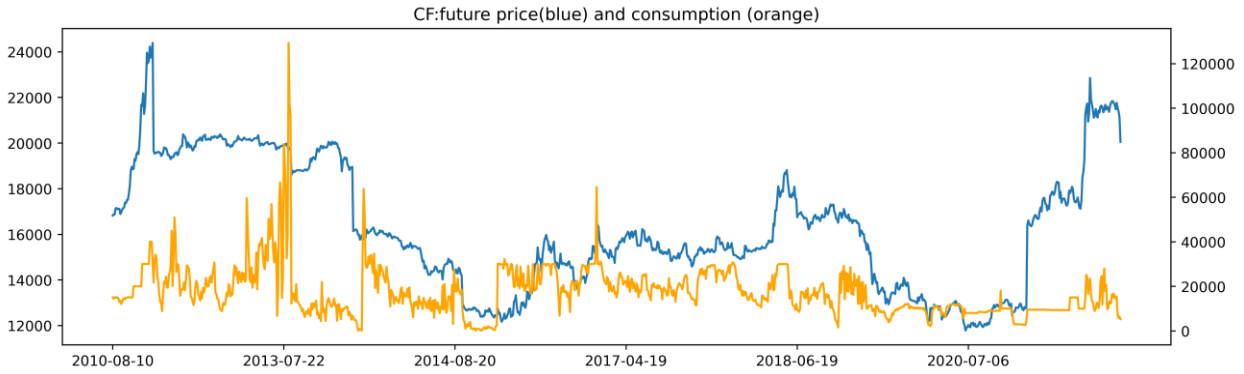


图 7. 郑棉主连合约与棉花抛储成交量走势

在模型建立、参数调优以及持仓管理上，我们沿用上文中描述的方法。超参数调优后的结果如下

参数名	参数意义	调参结果
max_depth	Boosting 过程中每棵树的最大深度	9
learning_rate	较小的 learning rate 有助于缓解 over fitting	0.3
subsample	每次选取训练集中数据的比重	0.6
num_round	决策树的数量。选取较大的数值意味着使用比较复杂的模型	30

从回测表现来看，模型在训练集上的交易胜率达到 84.43%，净值从 1 增长到 1.6。在测试集上的交易胜率为 47.17%。**该结果似乎反映模型出现过拟合的情况，泛化能力较差**。由于时间关系，没有在超参数调优上作进一步的计算和分析，后续应进行补充。策略具体表现在第四部分的表格中列出。

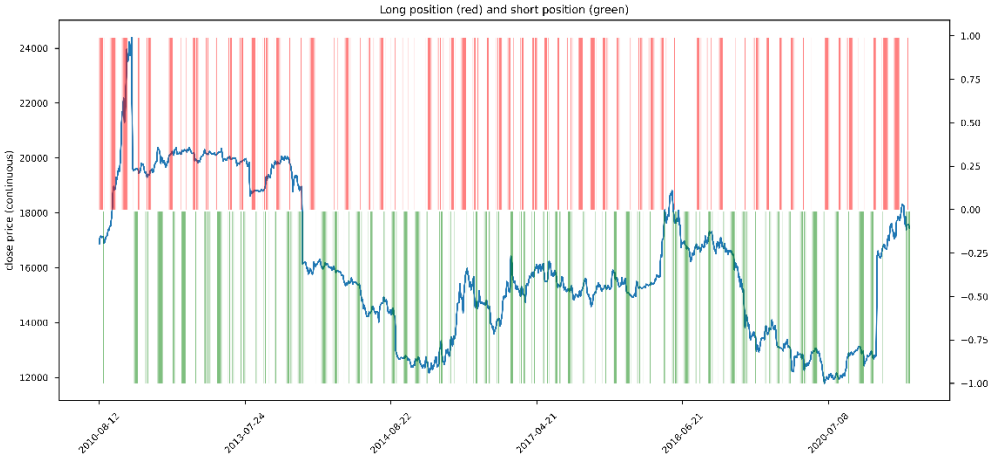


图 8. 郑棉主连合约走势与消费因子的 XGBoost 策略信号

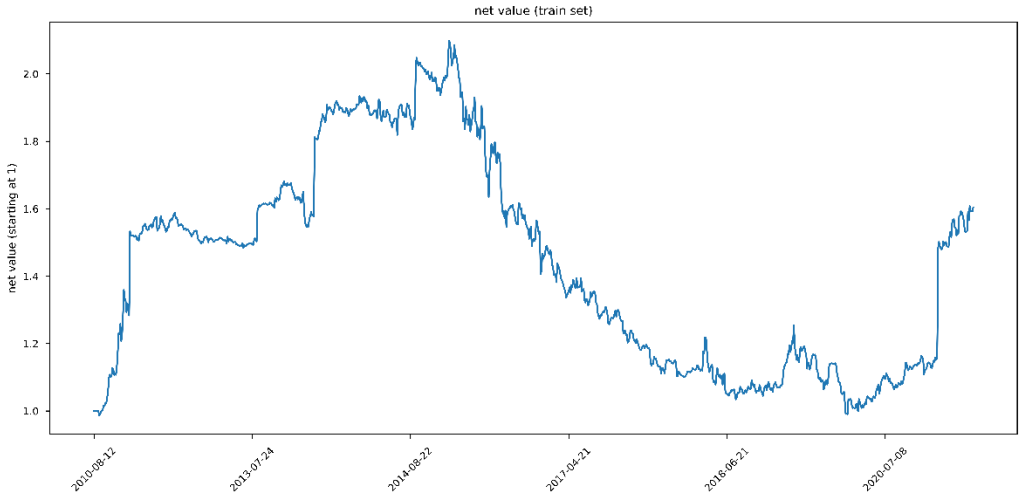


图 9. 棉花需求因子策略训练集净值走势

d. **复合指标：**

通过上述对单因子指标的分析，我们发现不同因子善于把握不同时间维度的增长趋势，同时，它们是从估值和驱动的不同角度来解构价格走势。本部分尝试将上述三个因子结合起来构建多因子择时策略。这里采用的方法是构建多维的 features，生成由基差因子、利润因子以及消费因子组成的多变量数据集，使用 XGBoost 训练模型，得到的超参数优化效果如下所示。**注意这里由于使用多个变量作为 X，我们在参数调优中增加 colsample_bytree。**

参数名	参数意义	调参结果

max_depth	Boosting 过程中每棵树的最大深度	4
learning_rate	较小的 learning rate 有助于缓解 over fitting	0.3
subsample	每次选取训练集中数据的比重	0.6
num_round	决策树的数量。选取较大的数值意味着使用比较复杂的模型	30
colsample_bytree	**在 random sample 时，控制使用 features 的比例。每次随机选取 features 比例可以避免某个 feature 影响过大**	0.9

从回测表现来看，模型在训练集上的交易胜率达到 87.49%，净值从 1 增长到 1.6。在测试集上的交易胜率为 34.69%。**该结果似乎同样存在过拟合，泛化能力较差**。由于时间关系，没有在超参数调优上作进一步的计算和分析，后续应进行补充。策略具体表现在第四部分的表格中列出。

下图展示了策略仓位以及净值变化。

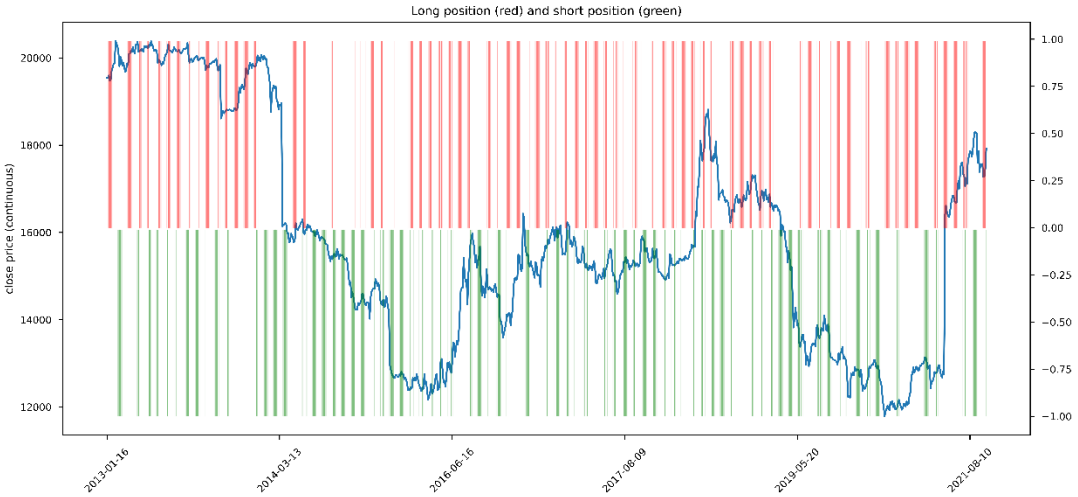


图 9. 郑棉主连合约走势与多因子的 XGBoost 策略信号

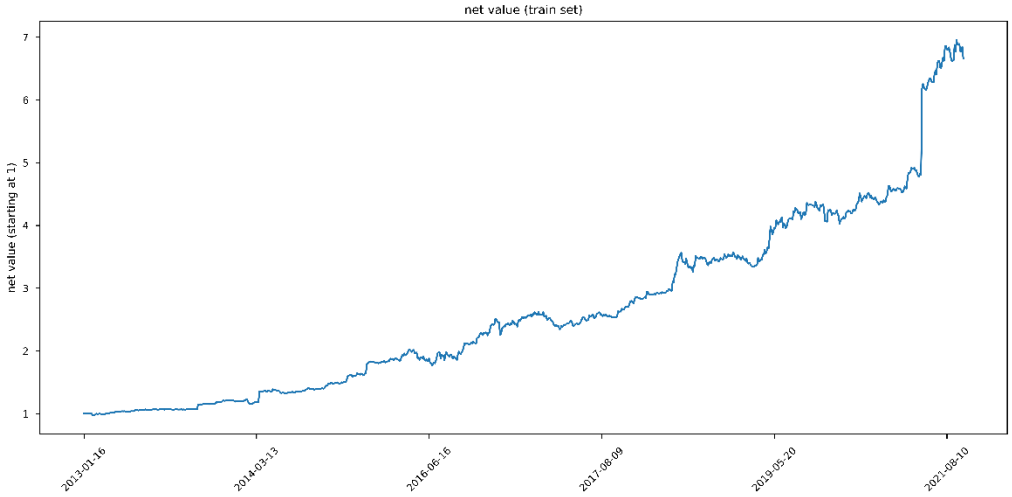


图 10. 棉花多因子策略训练集净值走势

e. 仓位管理方法

在前文单一指标 1 中，我们已经对仓位管理进行了说明。这里给出 pandas 表头进一步阐述仓位管理和计算方法。当模型给出信号 0 时，我们满仓做空合约（如本身处于做空状态则保持），当模型给出信号 1 时，如果之前做空合约，则平仓并满仓做多合约。由此进一步计算每一日的收益率、累计收益以及净值变化。

	date	close_price	signal	action	position	daily_return	\
2007-11-07		15390.0	0.0	0.0	-1.0	-0.016848	
2007-11-08		15265.0	0.0	0.0	-1.0	0.008122	
2007-11-09		15195.0	0.0	0.0	-1.0	0.004586	
2007-11-12		15140.0	0.0	0.0	-1.0	0.003620	
2007-11-13		14885.0	0.0	0.0	-1.0	0.016843	
...		...	...	...	...	...	
2022-06-16		19560.0	0.0	0.0	-1.0	0.015106	
2022-06-17		19515.0	0.0	0.0	-1.0	0.002301	
2022-06-20		19240.0	0.0	0.0	-1.0	0.014092	
2022-06-21		19065.0	1.0	1.0	1.0	-0.009096	
2022-06-22		18890.0	0.0	-1.0	-1.0	0.009179	

cumulative_return	daily_net_return	net_value
-0.016848	-0.016848	0.983152
-0.008863	-0.008863	0.991137
-0.004318	-0.004318	0.995682
-0.000714	-0.000714	0.999286
0.016117	0.016117	1.016117
...	...	...
618.766122	618.766122	619.766122
620.191964	620.191964	621.191964
628.945630	628.945630	629.945630
623.215875	623.215875	624.215875
628.945630	628.945630	629.945630

图 11. 仓位管理示例（Pandas 表头）

4. 策略表现

a. **回测表现**

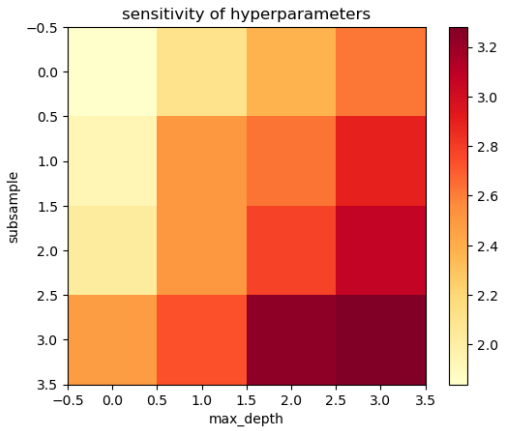
因子	预测间隔天数	回测表现（训练集）								
		回测范围	回测天数	最终净值	交易胜率	年化收益率	夏普比率	Calmar率	最大回撤	年均交易频率
基差因子	2	2007-11-06 到 2022-06-22	3556	629.94	68.98%	42.97%	2.89	2.15	20%	50
利润因子	7	2013-01-07 到 2022-06-27	2240	3.83	75.49%	29.90%	1.02	0.01	27.46%	85
消费因子	5	2010-08-11 到 2021-08-30	1002	1.60	84.43%	5.46%	0.67	0.001	52.80%	30
多因子	5	2013-01-15 到 2021-09-03	919	6.66	87.49%	65.57%	2．56	0.05	12.78%	24

因子	预测间隔天数	回测表现（测试集）								
		回测范围	回测天数	最终净值	交易胜率	年化收益率	夏普比率	Calmar率	最大回撤	年均交易频率
基差因子	2	2022-06-23 到 2023-03-30	188	0.83	50.00%	-13.06%	-0.87	-0.005	26.76%	66
利润因子	7	2022-06-28 到 2023-03-14	118	1.50	67.80%	71.21%	3.41	0.06	11.66%	33
消费因子	5	2021-08-31 到 2021-11-23	53	1.06	47.17%	25.18%	1.10	0.01	17.13%	95
多因子	5	2021-09-06 到 2021-11-23	49	0.71	34.69%	-137.25%	-5.52	0.05	31.40%	80

从回测表现来看，利润因子在训练集和测试集上均给出了不错的表现。对于消费因子和多因子策略，由于时间上部分数据存在缺失值的原因，数据长度较小，回测范围较短，后续还需要进一步研究观察。

b. **超参数敏感性分析**

此处对 XGBoost 所使用的重要参数“max_depth”和“subsample”进行超参数敏感性分析，以基差因子的策略模型为例，下图给出的是不同超参数组合下策略在训练集上夏普指数的变化情况。



5. 总结

总结来看，通过基本面因子以及 XGBoost 模型构建择时策略信号有着不错的效果。基本面因子从估值、驱动等不同角度解构驱动价格走势的因素，拥有非线性拟合能力的 XGBoost 模型能从这些因子数据中学到较为复杂的相互影响关系，进而生成交易信号指导交易。

反思来看，目前在该策略的设计中存在一些不足，同时也有一些改进方向可以在未来实践。

- **策略的可交易性**：为简化研究过程，目前选用郑棉主连合约作为投资标的，并且未考虑手续费、滑点的问题，导致策略的可交易性较弱。未来应选取不同时段的主力合约作为投资标的，并考虑手续费、滑点等实际交易过程的问题。
- **模型训练**：在消费因子和多因子模型上存在较为明显的过拟合现象。在超参数调参过程中使用 gridSearch 进行参数调优，受到算力以及计算速度的制约，没有找到更好的超参数。未来可使用 randomSearch 以及 hyperopt（使用贝叶斯统计的方法进行超参数调参）等方法寻找更优化的参数。
- **模型设计：**使用 XGBoost 模型有很多优点，在前文中已经提及。但是，目前设计的训练方法，破坏了金融数据中包含的时序信息。具体而言，以利润因子为例，当前利润因子值与未来走势的关系可能与此前一段时间的利润走势非常相关，但是在我们目前设计的数据集上，这种时序关系被破坏了。在未来，应该着手考虑加入时序信息。

6. 附注：策略代码说明

本部分对策略代码架构进行简要说明。

- codes
    - data_processing
        - indicator_processing.py （用以生成因子数据以及策略指标）
    - fine_tune_train
        - train_fine_tune_gridSearch.py （用以调参）
        - train.py （用以训练模型并获得结果）
    - strategy
        - strategy.py （用以构建策略以及进行回测）
- indicator_data
- raw_data
- signal_data
- result

各部分代码通过简单的修改即可测试更多因子和品种，具有一定的可扩展性。