

有色金属基差预测模型：基于 stacking 集成模型

-以上期所的金屬為例

近年来，人工智能（AI）的迅猛发展引发了各行各业的广泛关注与应用，而机器学习作为 AI 的重要分支，扮演着不可或缺的角色。当前机器学习模型被广泛地应用在量化投资，尤其是在股票市场和期货市场的探索上，但对于大宗商品现货市场的研究仍然相对不足。探索机器学习在这些领域的应用将会提升现货市场的预测准确性和辅助企业更好的理解未来价格走势。本文尝试在当前有色金属的基差预测上引用机器学习模型。以上期所期货品种：铜、铝、锌、铅、镍等有色金属的日度期现货数据为样本，选取系列影响有色金属价格的特征，采用随机森林模型对未来基差涨跌进行训练和预测。

一、Stacking 集成学习算法

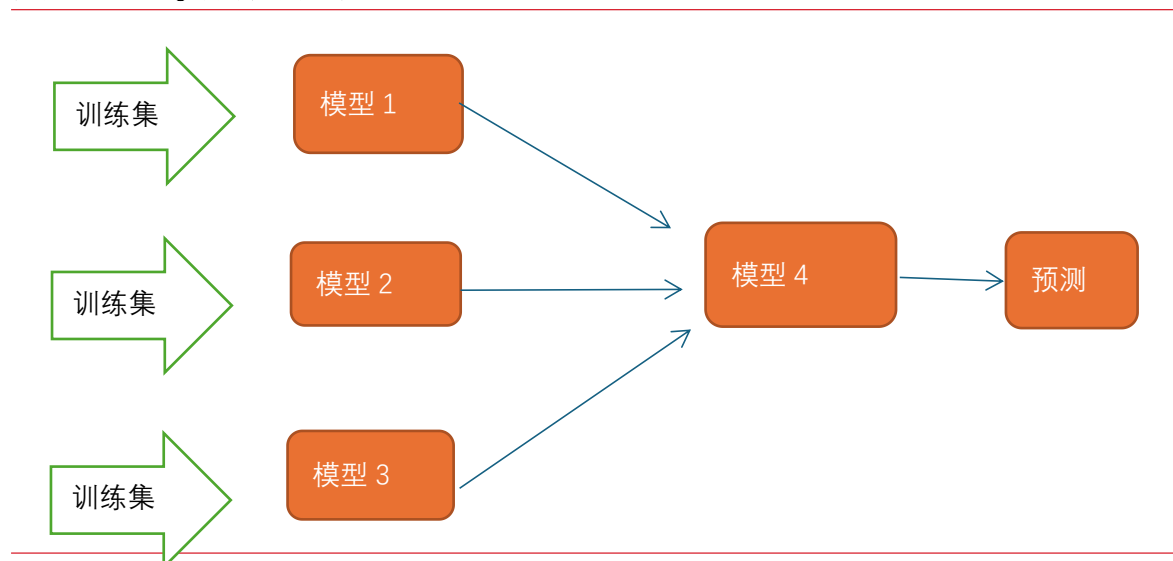
集成学习算法介绍：

Stacking 模型是机器学习中重要的集成算法。本文使用 Stacking 模型将 SVM、随机森林、XGBoost 以及 LR 逻辑回归多种机器学习模型融合。需要注意的是基学习器由于包含不同类型的学习器，例如随机森林、支持向量机等具有异质性，与其他同质性的学习算法有所不同。基学习器为 SVM、随机森林、XGBoost，元学习器使用 Logistic 回归。其中使用 SVM、随机森林、XGBoost 作为第一层基学习器进行特征转换，最后使用 Logistic 回归作为第二层元学习器进行最终分类预测。并且采用带格点搜索的交叉验证法确定模型参数，并以有色金属数据为例进行实证分析，结果表面基于 stacking 模型能够有效的对期货技术面因子进行选择，获取较高的超额收益率。

Stacking 集成学习的优势

由于期货市场中，存在着多种复杂的影响因素，仅仅依靠单个模型很难准确反映因子中的有效信息。Stacking 模型作为一种集成方法，能够融合多种机器学习模型，发挥不同模型的优势。stacking 模型主要是通过给定的数据训练出若干个弱分类器，然后依据这几个弱分类器给出的预测结果作为新的训练数据，训练出新的学习器。本文中采用 SVM、RandomTree、XGBoost 作为个体训练器，并且将输出的结果再次输入到 LogisticRegression 逻辑回归中训练，构建模型，逻辑如图所示：

图表 1:Stacking 集成学习示意图



资料来源：华安期货研究院

总体来看，利用历史数据训练机器学习模型，可以相对准确的给出未来一段时间的基差相对走势。这可以辅助拥有定价主动权的一方在期货交易中把握基差的变化规律，在点价期内获取寻找较优势的价格，从而制定更有效的交易策略，降低成本，减少市场价格波动带来的不确定性。

二、特征因子的创建和处理

本文使用 2011 年 1 月-2024 年 9 月有色金属，包含铜、铝、锌、铅、镍等有色金属的日度量价数据。通过特征因子的组建，计算多个与价格和成交量有关的技术因子，如均线、角度、乖离率、趋势线和布林带等因子，试图寻找到最能够影响基差的因子数据。对有色金属的日度数据计算出来多个技术面因子，使用多个模型算法集成进行训练和预测，目标是预测下一个交易日的日基差。本次计算中共有 29 个量价因子。具体的特征因子包括

- 1、趋势因子：移动平均线的变化率
- 2、高低价差因子：日内高点和收盘价的百分比
- 3、成交量因子：成交量均线及其差分
- 4、简单移动均线和指数移动均线
- 5、布林带:上下轨
- 6、变化率：价格均线的变化率
- 7、乖离率：多日乖离率及其差分
- 8、收益率因子
- 9、均线因子：价格和成交量均线
- 10、角度因子：即反三角函数的计算价格价差。

2.2 因子数据的处理：

样本数量总体有 14661 行数据，29 列因子。通过使用 scikit-learn 中的常用的函数，将数据集划分为训练集和测试集。按照 25%的数据作为验证集，75%数据作为训练集。在划分训练集和验证集中，通过打乱数据集中样本，随机地划分数据样本。为了保障模型拥有足够的泛化能力，使用过去 10 年的数据作为训练集。由于较长时间维度的数据涵盖了多个牛熊周期的变化，使得模型总体拥有较好的泛化能力。使用最近 4 年的数据作为验证集，通过测试模型在新的数据上的表现，来避免过拟合。

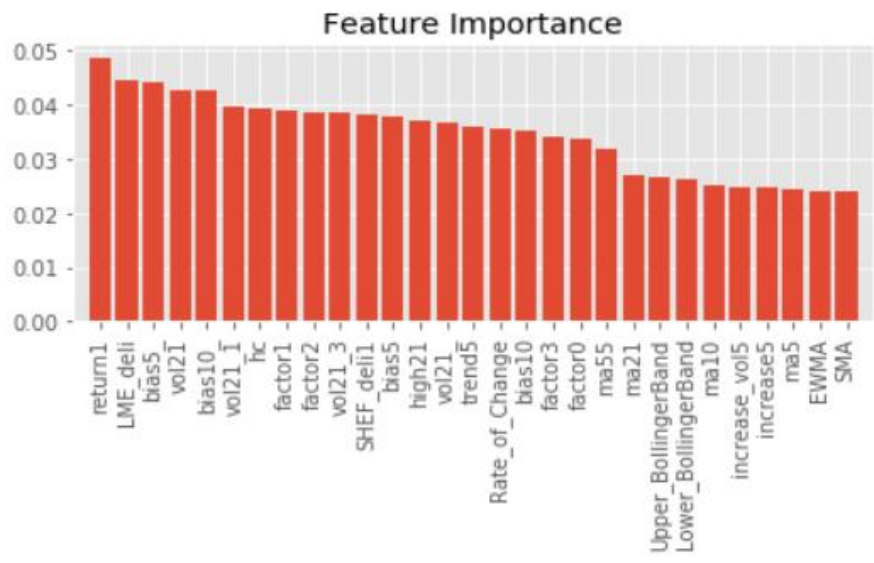
由于模型中使用梯度下降法训练数据，因此需要对于因子进行标准化处理。原始的量价因子的取值范围往往相差加大，而通过标准化处理后，能够确保各个因子贡献度一致，加速参数收敛，使损失函数更对称。本文使用 Z-score 标准化（零均值标准化），减去均值后除以标准差，得到均值为 0，标准差为 1 的特征值。

2.3 因子的贡献度衡量

本文使用随机森林中的因子贡献度来衡量特征因子的重要性，即每个特征对于模型预测的贡献程度。该模型通过计算每个特征在决策树中分裂质量来计算特征的重要性。主要有两种计算方式，1、基于每颗树的分裂增益值，通过加总所有树种的增益，归一化所有特征的重心性。当一个特征用于分裂节点时，计算分裂前后的基尼不纯度差值。记录每棵树每个特征在所有分裂种所产生的增益，加总增益，进行归一化，即可得到每个特征的总重要性。2、通过打乱某个特征的值（例如，将该特征的所有值随机排序），观察模型性能的变化。如果打乱特征后模型性能大幅下降，说明该特征对模型预测的重要性较高。

从本文中可以发现所有技术指标中，收益率、成交量和乖离率的因子重要性更大。

图表 1: 因子贡献度排行



资料来源：华安期货研究院

3.4 使用回归模拟基差值，而非进行分类

3.5 确定评价定价

三、模型表现

3.1 逻辑回归

逻辑回归主要是利用 sigmoid 函数将线性组合的输入特征映射到概率值上进行分类：

公式表达：

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

$$p(y = 0|x) = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

其中 p 是预测分类标签的概率， x 是输入特征， β_n 是模型的参数。

损失函数：
$$L(\beta) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

其中 y_i 是真实标签， p_i 是模型预测的概率

3.2 支持向量机

支持向量机主要是寻找最优超平面，将不同类别的数据点分开，最大化两类样本点之间的间隔。其中超平面表达式 $f(x) = w^T x + b$ ，寻找多数数据在一个 ϵ 区间内

$$|y_i - f(x_i)| \leq \epsilon \text{ 损失函数: } L(w, b) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$$

其中 ξ_i 和 ξ_i^* 是松弛变量，反映了预测值与真实值的偏差。

3.3 随机森林：

通过构建多棵决策树并将其结果进行汇总，利用投票或平均的方式来提高模型的准确性和鲁棒性。

回归问题：

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N \hat{y}_i$$

其中，针对回归问题主要计算所有树的预测值的平均值。

3.4 GBR 模型：

是一种常用的回归算法，通过逐步构建多个弱学习器（通常是决策树）来提高预测精度。其基本思想是每一步都根据前一步的残差进行优化。

$$\text{损失函数: } \hat{y}_i = \sum_{m=1}^M f_m(x_i) \quad L(\theta) = \sum_{i=1}^N l(y_i, \hat{y}_i) \quad r_i^{(m)} = -\frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i^{(m-1)}}$$

其中， \hat{y}_i 是第 i 个样本的预测值， M 是树的数量，第 m 颗树的预测。 $L(\theta)$ 是损失函数（如均方误差）。在每一轮中新的树是根据前一轮模型的残差进行训练的。残差定义为： $r_i^{(m)}$

因此我们将基差未来走势简化为涨跌问题，减少预测难度。将未来基差的预测当成分类问题，根据时间序列中的基差值进行分类，判断当前值相对于前一个时间点的基差是否上涨或下跌，如果后一个时间点的基差值比前一个时间点高，则标记为上涨(1)，反之则标记为下跌(0)。

模型选择：随机森林模型的简介。该算法是一种非线性模型，之所以叫森林，是因为将多个决策树的结果集成为森林。通过对于多个决策树的预测结果进行投票，可以避免单个决策树受到极端值影响。每一个决策树根据样本特征（库存、动量、经济好坏等）对未来基差的涨跌进行判断。最终对于 N 个决策树的结果采取投票形势确定最终的基差涨跌判断。该模型的决策思路即将每个决策树的预测结果进行投票，未来涨跌的判断根据获得投票数最多的预测结果。

3.3 评价指标

在基差预测中，随机森林模型的最中准确率 = 57.66%。

1. 准确率 (Accuracy = 0.5766)

准确率表示模型预测正确的样本占总样本的比例。在随机森林模型中，交叉验证后的准确率为 57.66%。这意味着在整个数据集中，大约 57.66%的样本被正确分类。

2. AUC (AUC = 0.5998)

AUC 是另一个衡量模型性能的指标，它关注模型在区分不同类别（上涨/下跌）时的能力。AUC 取值范围为 0 到 1，数值越大，模型的区分能力越好。AUC = 0.5998 这两个指标表明，当前模型的预测效果一般，特别是区分上涨和下跌有一定能力。

后期改进：为了提高模型性能，后期可以尝试以下几种方式：1、调整模型超参数：你可以尝试其他参数组合进行调优。2、增加特征处理：更多基于经济和市场规律的因子可能帮助提高模型的预测能力。3、数据预处理：进一步检查数据质量，比如平衡数据集或处理异常值。

四、本文总结

结合机器学习，在基差点价策略中引入随机森林算法不仅能够有效识别市场风险，还能提高策略优化的能力。通过精确的预测和动态调整，企业能够更好地应对价格波动，从而在竞争中保持优势。通过随机选择特征和样本构建多棵决策树，能有效处理过拟合问题，适用于复杂的基差预测任务。同时该模型可以通过特征重要性分析，找出哪些因素对基差变动影响最大，从而优化决策，提高基差预测的准确度。