兴证期货有限公司
CHINA INDUSTRIAL SECURITIES FUTURES CO.,LTD.

专题报告

金融衍生品·金融工程

兴证期货.研发产品系列

基于随机森林的 WTI 原油价格预测研究

2019 年 4 月 8 日　星期一

兴证期货.研发中心

金融工程研究团队

刘文波

从业资格编号：F0286569

投资咨询编号：Z0010856

段宁

从业资格编号：F3048895

孙石

从业资格编号：F3042665

吴耀宏

从业资格编号：F3035075

周英

从业资格编号：F3038963

联系人

吴耀宏

021-20370974

wuyh@xzfutures.com

内容提要

- 随机森林（Random Forest）是由 Breiman 和 Cutler 提出的一种分类学习算法，其本质是基于决策树的分类器集成算法。模型的基本单元是决策树。它的主要特点是处理能力强、结果较为稳定。它的应用主要包括分类和回归两个方面。
- 我们通过原油基本面框架寻找影响其价格的因子，基于随机森林对 WTI 周度平均收盘价进行拟合。实证结果表明，CFTC 非商业持仓量、美国商业原油库存、美国原油出口量、WTI321 裂解价差、贝克休斯原油钻井数这几个指标是 WTI 期货价格的重要影响因素，随机森林对原油价格走势的拟合较好，价格涨跌判断的正确率较理想。

1. 随机森林算法概述

1.1 随机森林的定义

随机森林（Random Forest）是由 Breiman 和 Cutler 提出的一种分类学习算法，从“随机”和“森林”这两个关键词的字面上理解，就是用随机的方式建立一个森林，这个森林中有许多决策树，彼此之间不互相关联。当一个新的样本输入模型时，每一棵树都独立进行判断，最终将所有的判断结果汇总作为模型的输出结果。用易于理解的方式来说，相当于针对某一疑难杂症请教一群医生，每位医生的水平各不相同，他们都独立地给出了自己的评价，最后我们综合所有医生的意见作出判断，因此这样得到的结果更准确和稳定。

随机森林的本质是基于决策树的分类器集成算法。模型的基本单元是决策树，每棵决策树都是一个分类器，对于一个输入样本，每一棵树都会给出一个投票结果，最终将投票结果汇总，将投票次数最多的类别指定为最终输出。随机森林的随机，就是指随机森林通过随机挑选变量和观测样本生成许多决策树。

1.2 随机森林的优点

随机森林算法的主要特点包括：

1. 处理能力强。一方面，随机森林对多元共线性不敏感，只要树的数量足够多，随机森林能够处理极高维度的数据，不需要进行特征选择和降维的数据预处理。另一方面，随机森林体现出了对数据集的强大适应性，在建模前不需要对数据集进行规范化预处理，且在离散型数据与连续型数据的处理上均能够胜任，能够处理非线性和多输出的问题。

2. 结果较为稳定。由于在随机森林建立的过程中有变量选择和样本观测选择两个方面随机性的引入，随机森林不容易陷入过拟合，且具有良好的抗噪声能力，受个别数据缺失的影响不大。

1.3 随机森林的应用

概括来说，随机森林的应用主要包括分类和回归两个方面。当因变量是离散型变量时，随机森林处理的就是分类问题；当因变量为连续型变量时，随机森林处理的就是回归问题。具体来说，主要包括几种应用情形：一是判别分析，即在因变量被明确划分为几类的情况下，根据几个自变量判别每个样本类型归属问题的多变量统计分析方法；二是二元回归分析，即因变量为有或无、发生或不发生等二元变量的情形，相比 logistic 回归法，随机森林对自变量多元共线性不敏感，且不要求自变量之间相互独立；三

是多元回归分析，即通过一组自变量对取值连续的因变量进行解释分析。

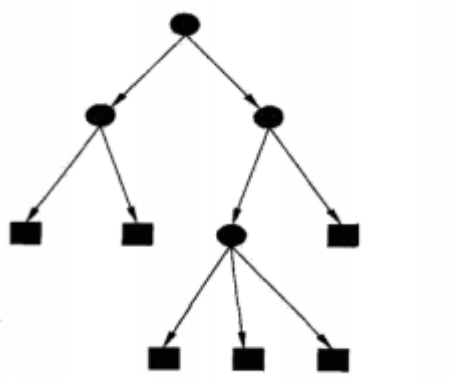
因此，随机森林算法在许多领域都有着广泛应用，例如：在金融领域，它可用来对资产未来价格变化趋势进行预测，也可被用来构建信用评价体系；在生物领域，它可用于物种分类或研究环境对生物行为的影响；在化学领域，它可用于分析物质的成分组成；在市场领域，它可用于分析消费者偏好。

2. 随机森林算法原理

2.1 决策树

决策树是随机森林的基本处理单位，能够处理离散型和连续型的输入及输出变量，根据输入变量中最显著的分裂点将总体或样本划分为几个类别。它的基本结构如图所示：

图 1：决策树基本结构图



图中，原点代表决策节点，指的是将会进一步分裂出子节点的节点，其中最顶端的节点是根节点，代表整个总体或样本。方点代表叶节点，即不再分裂的终端节点。决策树学习的本质就是从训练集中总结出它的概率分布情况，以损失函数为准则来寻找最优化结果。对于分类树，其损失函数以基尼系数或者熵来度量；对于回归树，则是以方差来度量。

决策树的学习，是从根节点开始，选择一个最优特征进行测试，根据测试结果将样本分配到子节点，每个子节点对应着该特征的一个取值，如果这些子集已经能够被基本正确分类，则子节点就是叶节点；如果还有子集不能被正确分类，则对这些子集选择新的最优特征，继续将它们分配到下一层子节点中，直至所有子集都被分到叶节点上，则生成了一棵决策树。

因此，可以看出树模型与线性模型最大的区别在于，树模型是对特征逐一处理的，而线性模型是将所有的特征按照某个权重加权相加。

2.2 决策树的算法

决策树通过将一个节点分裂为多个子节点使得分裂后各子节点的纯度增加，那么决定分裂方式的标准便是生成决策树的一个关键因素。决定分裂方式的标准主要包括：

1. 基尼系数

基尼系数是一个衡量纯度的统计量，指的是从当前集合中随机挑选两个样本，这两个样本为同类别的概率。如果总体是纯的，基尼系数就是 1，基尼系数越高，纯度越高。将子节点的基尼系数按照子集样本数占父节点总样本数的比例加权平均可以得到父节点的基尼系数，选择特征时按照基尼系数最大的原则进行。基尼系数一般用于处理二元分裂的情况。

2. 卡方系数

卡方系数可以衡量子节点和父节点之间是否存在显著差异，通过目标变量的观测频率和期望频率之间的标准离差平方和计算而得。计算公式为：

$$\text{Chi}-\text{square} = \left(\frac{(Actual - Expected)^2}{Expected}\right)^{\frac{1}{2}}$$

将每个节点下所有情况的卡方值加总可得到该节点的卡方值，父节点的卡方值为所有子节点卡方值的加总。卡方系数可以处理多元分裂的情形。

3. 信息熵

熵指的是信息的混乱程度，可以理解成是与基尼系数完全相反的一个统计指标。计算公式为：

$$\text{Entropy} = -\text{p} \times log_2(p) - q \times log_2(p)$$

p 和 q 分别指分类完的子集中两种类别的概率，如果子集中的个体是完全同类的，那么系统的熵为 0，如果子集中两种类别各占 50%，则系统的熵为 1。熵可以处理多元分裂的情形，值越低越纯。

4. 均方误差

均方误差是处理回归问题时使用的指标，用于衡量连续型输出变量学习效果，计算出每个节点的均方误差后将它们加权平均，从而得到一个分裂方式的总方差，根据方差最小原则来确定分裂方式。计算公式为：

$$\text{Variance} = \frac{\sum(\text{X} - \bar{X})^2}{n}$$

2.3 随机森林的回归算法

在 WTI 原油价格的研究中，涉及到的变量都是连续型变量，因此我们将构建的是由回归树组成的随机森林。

随机森林回归的基本思想是使用自助法（bootstrap）的重抽样技术，从原始样本中抽取一定数量的样本，允许重复抽样，根据抽出的样本计算给定的统计量，重复多次，得到该统计量的多个结果。

具体步骤为：

1. 假设原始样本容量为 n，应用 bootstrap 方法有放回地随机抽取 k 个新的自助样本集，并由此构建 k 棵回归树，每个样本集生成一棵树，未被抽取到的数据则成为 k 组袋外数据作为模型的测试样本。

2. 假设原始样本的特征个数为 M，则在树的每个节点处从 M 个特征中随机挑选 m 个特征，然后根据决策树算法中的均方误差标准来挑选特征进行节点分裂。所挑选的特征应该要使得这一分裂的均方误差下降最大，即按照子节点均方误差最小的标准来进行分裂。

3. 每棵树都做最大限度生长，不进行剪枝。将生成的 k 棵回归树组成随机森林回归模型，根据每棵树的回归结果汇总成随机森林回归的结果，根据袋外数据计算而得的残差均方误差和拟合优度 R 方值来评价回归效果。

3. 随机森林的训练与原油价格预测

3.1 数据选取

首先，美国是目前世界上原油产量前三的国家，上市的原油期货是 WTI 轻质原油，这一合约有良好的流动性，价格公开透明且交易连续，是国际原油定价的基准之一，而其产量、库存等指标更新频率较高，数据也是最为透明和权威的。因此，我们以 WTI 原油期货周度平均收盘价格作为因变量。

从数据更新频率的角度考虑，我们分别从原油的供应和需求等方面选取了几个指标作为自变量：

1. 库存是供需博弈后的结果，可以反映当前供给是过剩还是紧缺。当需求较弱时，库存便会积累；当需求超过供应时，便会消耗库存。因此，库存对未来价格比较敏感。这里采用对短期油价影响较大的美国商业原油库存。

2. 出口量是衡量供给端的重要指标。即使产油国产量足够，但出口量下滑，仍会造成供给紧张。美国出口量会受到管输能力限制。这里采用美国原油出口量指标。

3. 钻井数也是供给方面对未来产量的一个预判指标，若产油商对未来油价预期较好，则会增加钻井数。这里采用贝克休斯原油钻井数，每周六公布数据，对短期油价有较强影响。

4. 在短期需求方面，我们采用美国炼厂总输入量、炼厂开工率、原油

加工量、原油进口量等指标。

5. 裂解价差代表炼厂原材料和产品的价差，是衡量炼厂利润的指标，价差走强将提振炼厂的开工意愿，从而增加需求。我们采用 WTI321 裂解价差，321 代表原油:汽油:取暖油=3:2:1。

6. CFTC 非商业净持仓量代表投机者对油价的预判，因此我们也将其纳入自变量。

7. 除了这些指标外，我们将滞后一期的期货价格也作为一个自变量。

3.2 随机森林训练与测试结果

我们使用了从 2012 年开始到 2017 年结束的数据作为训练样本。随机森林的一个重要参数是单棵决策树的特征数，即一棵树从所有特征集中随机抽取用来分裂的备选特征数。由于我们的特征数量不大，因此采用循环遍历的方式尝试这一参数各种取值下的结果。

表 1：遍历单棵数特征数的结果

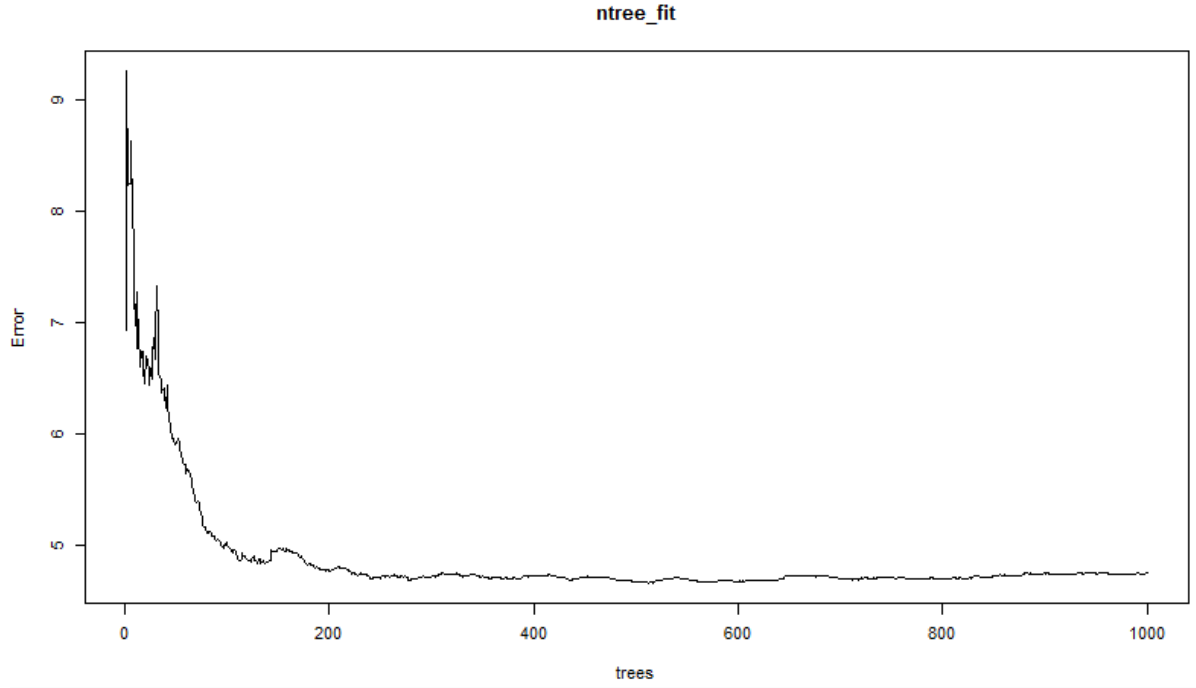
单棵决策树特征数	均方误差
2	8.6449
3	6.3387
4	5.1762
5	4.9170
6	4.9503
7	4.9982
8	4.9953
9	5.0550
10	5.1831

资料来源：兴证期货研发部

从结果中我们可以看出，当单棵数特征数为 5 时，对应的误差最小，因此我们将取这一参数为 5。

在特征数为 5 的情况下，我们取决策树数目为 1000。从图中可见，在树的数目取 1000 时，误差已经稳定。

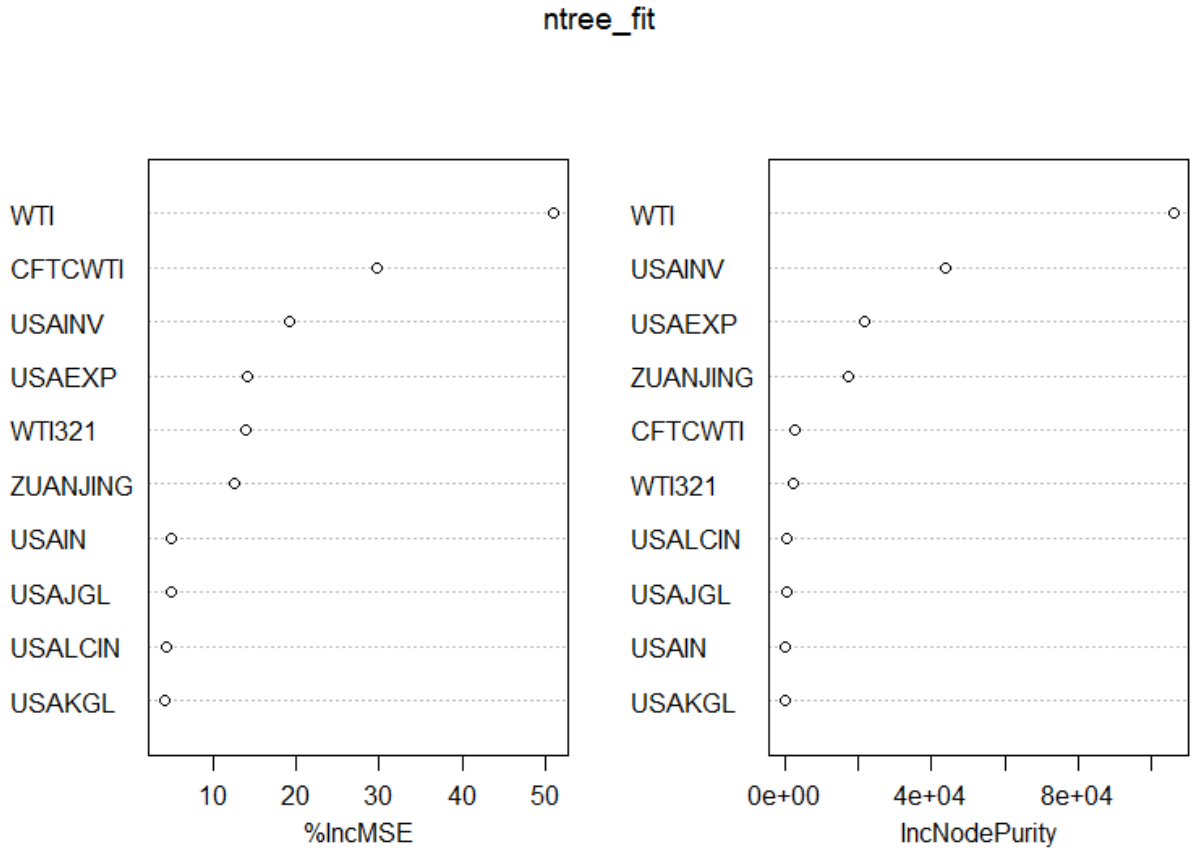
图 2：误差与树数目关系图



资料来源：兴证期货研发部

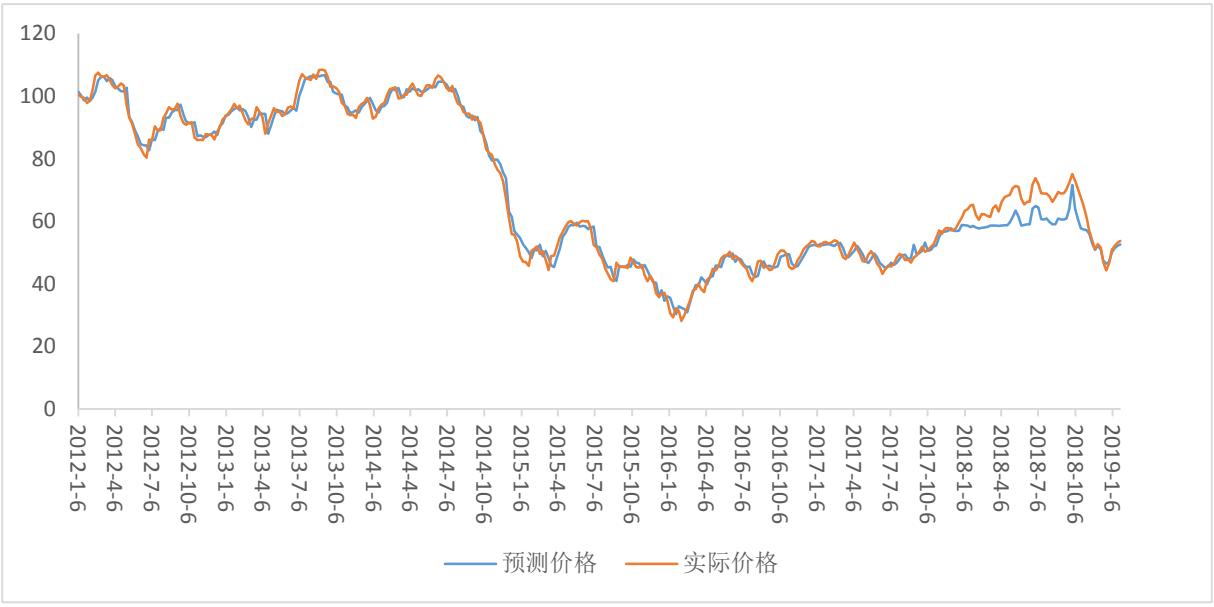
由此我们得到了最终的分类器。我们可以观察特征重要性的结果。因为不同的树纳入了不同的特征进行分类，那么就可以得知不同特征对于学习效果的影响。图片左边为均方误差标准下的特征重要性排序，也就是能使均方误差下降越大的特征越重要，图片右边为节点纯度标准下的特征重要性排序，越能提高节点纯度的特征越重要。可以看到，两种标准下重要性前六位的特征没有发生变化，且在均方误差标准下的特征重要性显著高于后四个特征。由于我们处理的是连续型变量的回归问题，因此主要还是参考均方误差标准下的结果。因此，除了滞后一期的期货价格以外，结果说明，CFTC 非商业持仓量、美国商业原油库存、美国原油出口量、WTI321 裂解价差、贝克休斯原油钻井数这几个指标对模型的贡献程度更高，对于 WTI 期货价格是更为重要的。

图 3：随机森林特征重要性



资料来源：兴证期货研发部

图 4：随机森林拟合结果



资料来源：兴证期货研发部

我们采用建立好的随机森林对样本内的 2012 到 2017 年以及未参与训练的 2018 年以来 WTI 原油价格进行拟合。从结果来看，随机森林对原油价格走势的拟合较好。我们分别测试了样本内外价格涨跌判断的正确率情况，分别为 57.88%和 79.31%。

4. 小结

随机森林是一种集成学习加决策树的分类模型，它可以利用集成的思想来提升单决策树的性能，使得结果更为稳定。随机森林引入了两个随机性：随机选择样本和随机选择特征。两个随机性的引入使得随机森林不容易陷入过拟合，并且具有良好的抗噪能力。

本文通过原油基本面框架寻找影响其价格的因子，基于随机森林对 WTI 周度平均收盘价进行拟合。实证结果表明，CFTC 非商业净持仓量、美国商业原油库存、美国原油出口量、WTI321 裂解价差、贝克休斯原油钻井数这几个指标是 WTI 期货价格的重要影响因素，随机森林对原油价格走势的拟合较好，价格涨跌判断的正确率较理想。

分析师承诺

本人以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。报告所采用的数据均来自公开资料，分析逻辑基于本人的职业理解，通过合理判断的得出结论，力求客观、公正，结论，不受任何第三方的授意影响。本人不曾因也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

免责声明

本报告的信息均来源于公开资料，我公司对这些信息的准确性和完整性不作任何保证，也不保证所包含的信息和建议不会发生任何变更。文中的观点、结论和建议仅供参考。兴证期货可发出其它与本报告所载资料不一致及有不同结论的报告。本报告及该等报告反映编写分析员的不同设想、见解及分析方法。报告所载资料、意见及推测仅反映分析员于发出此报告日期当日的独立判断。

客户不应视本报告为作出投资决策的惟一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的损失负任何责任。

本报告的观点可能与资管团队的观点不同或对立，对于基于本报告全面或部分做出的交易、结果，不论盈利或亏损，兴证期货研究发展部不承担责任。