标题

作者

林杰 季晨洋

摘要：本文使用三种机器学习算法，分别为含交叉项的二次多项式回归、随机森林、XGBoost模型，从甲醇产业链的角度出发寻找对甲醇期货价格预测模型较为重要的预测变量。样本选取自2016年1月至2021年5月期间交易的真实数据。以筛选的基本面预测变量和技术指标变量构造对照模型，采用R2、MAE、MSE、MAPE来评估模型的预测能力，研究结果显示，无论是否添加技术指标，甲醇毛利、甲醇现货价格以及冰醋酸进口量对模型的贡献程度均较高，表明这三项预测变量对甲醇期货价格预测具有重要作用。此外，在添加技术指标后，根据模型评估标准，二次多项式的预测准确度要高于其他两种树模型，对预测甲醇单交易日期货价格具有重要意义。

关键字：期货价格；价格预测；多项式回归；随机森林；XGBoost

一、 引言与文献综述

期货作为全球金融体系的重要组成部分，在调节经济方面具有重要的作用，对期货市场进行研究非常具有意义[1]。目前对期货的研究主要集中在能化、黑色系以及农业板块，能化产品主要集中在PTA、聚氯乙烯等，黑色系产品主要为螺纹钢、焦炭、铁矿石等，农业产品有小麦、玉米、大豆等[2]，对能化板块中的甲醇研究较少。

阅读国内外相关文献可以发现，对金融市场的研究方法主要分为数理统计模型和机器学习方法[3]。其中数理统计模型包括自回归模型、差分自回归移动平均模型、广义自回归条件异常差模型等。机器学习算法模型包括决策树（DT）、支持向量回归模型（SVR）、随机森林（RF）、人工神经网络（ANN）等[4]。随着计算力的提升，机器学习逐渐被应用到金融领域的研究中，并表现出良好的研究性能[5]。

Einav和Levin（2014）指出在经济学研究领域中，经济模型关注的重点主要放在能够代表工序变化的因素上面，研究人员需要对数据的异质性进行校正，并试图获得估计参数的异方差一致的标准误差和协方差。此外还需要通过不同模型来评估模型的鲁棒性[6]。与传统计量模型相比，机器学习算法对数据的前提假设较少，是以数据为驱动，将重点放在模型的输出结果上，进而揭示出特征对模型的解释能力[7]。机器学习是让计算机能够像人类一样行为，通过输入数据，不断进行观察和学习来提升自己与真实世界交互的能力的一门学科[8]。随机森林算法是集成多棵决策树的集成模型，Breiman指出不需要考虑多元回归模型中的多重共线性问题[9]，Mohd等2019，Mullainathan和Spiess，2017[10]，Pérez-Rave等分别与2019年，2017年，2019年[11]均随机森林算法指出可以有效地避免模型的过拟合，是一个能够做出精确预测的强鲁棒性算法。Alexandre Lucas等（2020）使用梯度提升、随机森林和XGBoost对欧洲地区居民电价进行预测得出XGBoost对能源的价格能够做出精确预测[12]。

二、 甲醇期货价格预测变量的筛选方法

期货价格预测效果取决于预测变量的选择和预测模型的构建。在预测变量的选择上，众多实证研究表明产品上下游产品的基本面数据[13]以及技术指标对金融资产价格的预测具有重要意义[14]。

1. 基本面指标的筛选

影响商品期货价格的因素有很多，从经济学原理的供需关系以及生产成本对供需关系的影响角度出发，期货商品的价格也应该在一定程度上遵循供需准则以及成本原理。从产业链的角度能够较为容易地找出影响商品期货价格的基本面数据。

甲醇作为石化产业链中的一个重要的中间产品，其生产的原材料主要为天然气，其下游的产品主要有甲醛、MITBE、冰醋酸等。根据供需关系以及生产成本对其的作用，上游产品的出厂价格将作为下游产品的生产直接费用，并呈现出正相关关系；根据供需关系原理，上游产品的库存量将对下游产品的生产成本产生一定的影响，并呈现出负相关关系；而上游产品的库存量又与该商品的开工率以及生产利润具备一定的相关关系，并且呈现出正相关关系；产品的进出口量中的进口将作为国内该产品的库存量的重要组成部分之一，出口量将作为该商品需求量中的一部分，与该产品的库存量也具备一定的相关关系，分别表现出负相关性和正相关性。此外，期货交易最终将以现货的方式进行交割，换言之，现货的交易价格与该商品的期货价格紧密相关。因而选择的基本面数据主要为上下游产品的现货价格、期货价格、供给、需求、成本、利润几个方面。以上是从供需关系和生产成本的角度以及期货交易的方式出发所做出的假设，具体的相关程度还要通过数理统计方法的分析才能确定相互之间是否存在这样的关系。基于数据的可获性，本文选择2016年2月至2021年5月期间甲醇的产量、进口量、出口量、库存量、现货价格、毛利，天然气的现货价格、产量、进口量、库存量，冰醋酸的进口量、出口量、库存量、消费量、库存量、现货价格、毛利基本面数据作为备选预测变量，甲醇期货交易日收盘价作为被预测变量。数据的来源为大智慧、卓创资讯、钢联以及生意社。

在这些可获的数据中，数据有日频、周频、月频的，本文采用的处理方式是将周频、月频数据转换为日频数据，具体的做法是将特征数据中相邻两个日期中的所有交易日均取为两者之间后者的样本值。此外，从实用角度出发，将这些备选预测变量向前移动一个周期，即如果是日频数据，则将甲醇的期货价格与该备选预测变量的前一交易日数据进行匹配；如果是周频数据则将甲醇的期货价格与该预测变量前一周的数据进行匹配；如果是月频数据，则将甲醇的期货价格与该预测便量的前一个月数据进行匹配。

使用SPSS统计分析软件对上述这些备选预测变量与甲醇的期货价格进行Person、Kendall以及Spearman相关性检验，检验结果如表1所示。

表1（相关性检验表）



根据表1分析结果可以看出，除了甲醇产量、冰醋酸消费量以及天然气进口量与期货价格不显著相关以外，其余的备选预测变量在显著性水平0.01条件下均与甲醇期货价格显著相关。与我们此前的假设基本一致，并且符合经济学原理。因此，将冰醋酸消费量、甲醇产量以及天然气进口量以外的备选预测变量选作模型的预测变量。

1. 技术指标的筛选

技术指标种类很多，本文只选取了期货投资者经常使用的MACD、KDJ、RSI、WR、PSY、OBV共6技术指标作为备选预测变量，各备选技术指标的计算方法和备选参数如表\*所示。分别对

6 种技术指标，并为各指标选定备选参数，分别与甲醇期货收盘价进行相关性分析，按照相关系数较高者入选的标准对每个技术指标的备选参数进行选择。结果如表\*所示。

备选技术指标、备选参数

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | 计算公式 | | 备选参数 |
| MACD |  |  | |
| KDJ |  |  | |
| RSI | （其中为上涨天数涨幅平均值，为下跌天数跌幅平均值） |  | |
| WR |  |  | |
| PSY | （为日内上涨天数） |  | |
| OBV | （OBV的初始值取第1日成交量） |  | |

注:和 分别为第t日国债期货指数收盘价和成交量，和分别为第t日前日

内国债期货指数的最高价和最低价。

技术指标最优参数组合

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标 |  | Pearson相关性 | Spearman相关性 | Kendall相关性 |
| MACD-DEA | (10，10，10) | -.169\*\* | -.151\*\* | -.102\*\* |
| KDJ | (10，10，10) | .121\*\* | .117\*\* | .077\*\* |
| RSI | - | .152\*\* | .142\*\* | 095\*\* |
| WR | (5) | -.098\*\* | -.103\*\* | -.073\*\* |
| PSY | (5) | .085\*\* | .095\*\* | .068\*\* |
| OBV | - | .527\*\* | .597\*\* | .428\*\* |

\*\*. 在置信度（双侧）为0.01时，相关性是显著的。

根据图2分析结果可以看出，这六个技术指标在置信度（双侧）为0.01的假设检验下均与甲醇的期货价格显著相关。因而将这6个技术指标同样作为甲醇期货价格的预测变量。

三 研究方法（研究方法这里有问题，没有说为什么选择这三个模型，应该要说一下选择模型的根据，然后分别对这几个模型进行描述，这才是正确逻辑顺序）

根据图1所示，在甲醇期货收盘价格与各预测变量之间的Pearson相关性检验结果中，两者表现为强线性相关的预测变量数量较少，此外根据图3甲醇期货收盘价曲线图所示，甲醇的期货收盘价呈现出了动态性，本文采用多项式回归、随机森林以及XGBoost方法，基于偏移的预测变量，对甲醇期货收盘价进行预测。（说了表现出动态性之后就立马说采用这几个模型是不是有点问题）

1. 构建多项式回归模型

多项式回归模型是简单线性回归模型的拓展，根据表1Pearson相关性分析的结果可以看出所选的基本面预测变量中与甲醇的期货价格呈现出强线性相关关系的变量占少数，仅为2个变量。根据图3所示，可以看出甲醇期货收盘价呈现出动态性，因而不宜采用线性回归的方式来对数据进行建模分析。根据相关矩阵可以看出，预测变量之间也呈现出不同程度的相关性（这里并没有添加相关系数矩阵，因而要换一种方式来说明，可以从经济学原理，变量之间的联动性出发），为了更好地拟合数据，本文选择有交叉项的二次多项式回归模型来对甲醇的期货价格进行预测。该二次多项式回归模型如下：

图3 甲醇期货收盘价呈现出动态性（前面也有提到动态性）

补图3

1. 构建随机森林模型

随机森林是一种有监督学习算法，可用作分类和回归模型。是决策树算法的集成算法，它将n棵决策树的预测结果结合起来以做出比单棵决策树更为精准的预测。除此之外，随机森林在构造单个决策树的过程中并没有将所有的特征作为该决策树的输入特征，而是采用随机抽样的方式从原特征空间中选取部分特征作为输入特征。此做法降低了预测方差较大模型的不稳定性，尤其是决策树模型。在本文中，随机森林预测模型的构建步骤是

第（1）步将处理过后的2016年2月2日至2021年5月30日共1298条交易样本按7:3比例进行拆分，占比70%的样本用作模型的训练，30%的样本作为对训练过后的模型进行检验。其中甲醇期货收盘价作为模型的输出变量，其余12个特征变量作为模型的输入变量。第（2）步使用交叉验证的方式在训练集上训练数据。第（3）步使用网格搜索的方式对模型的超参数进行调节以找出最佳的训练模型。

对随机森林的内容还是要做一定的补充（不一定要补充，看一下这个目的内容是什么，应该统一一下）

1. 构建XGBoost模型

XGBoost是机器学习算法中比较新的算法，由Chen和Guestrin于2016年提出，可用于分类、回归及排序模型的创建，它是基于梯度提升树算法原型进行扩展，显式增加了正则化项以控制模型复杂度，防止出现模型过拟合。通过对多个超参数进行调优来寻求最终的模型。训练时的目标函数由两部分构成，第一部分为梯度提升算法损失，第二部分为正则化项。损失函数定义为

，其中n为训练函数样本数，是对单个样本的损失，为模型对训练样本的预测值，为训练样本的真实值，正则化项定义了模型的复杂程度：

其中， 和 为超参数，为决策树所有叶子节点值形成的向量，T为叶子节点数。

以第t步为例，模型对第个样本的预测为：

其中是由第步的模型预测值，是需要新加入的预测值，结合泰勒二次展开式，上述目标函数可写为：

其中为损失函数的一阶导数，为损失函数的二阶导数。将定义为，定义为，则目标函数可写成：

再将叶子节点对应的权值写作：

从而目标函数可写作：

令，分别为数据集的左右结点，其中，则分裂后增益为：

其中，分别为左右树的分数，为分裂前分数。

四 实证分析

1. 变量样本数据说明

本文选取了2016年1月至2021年5月时间区间内甲醇期货主力合约交易日的收盘价作为被预测变量。预测变量分为基本面预测变量和技术指标变量。其中技术指标为。 其中基本面备选预测变量为甲醇的现货价格、产量、进口量、出口量、库存量、毛利，天然气的现货价格、产量、进口量、库存量，以及冰醋酸的现货价格、进口量、出口量、库存量、消费量、毛利共16个备选预测变量。其中现货价格数据来源生意社网站，产量、进口量、出口量、库存量、毛利数据来源于钢联客户端，期货数据来源于大智慧行情系统。由于基本面数据中包含日序列数据、周序列数据、月序列数据，本文采取的处理方式是交易日若无新周序列数据、月序列数据将沿用上一日的数据。此外，将预测变量相对甲醇期货价格数据向前偏移一个时间单位。

（二）预测变量筛选结果

使用SPSS统计软件进行基本面变量和技术指标变量与甲醇期货收盘价格进行Pearson、Spearman、Kendall相关性分析。相关性结果如表1。在显著性水平0.01情况下，甲醇期货价格与冰醋酸消费量、甲醇产量以及天然气进口量之外的备选变量均显著相关，因而除了冰醋酸消费量、甲醇产量以及天然气进口量之外的变量均入选为预测变量。

1. 3种回归模型预测精度的比较

1.不含有技术指标的模型回归结果

在不将已经筛选出的技术指标预测变量添加到有交叉项多项式、随机森林、XGBoost回归模型时，各个回归模型预测结果的R2、MSE、MAPE如表\*所示。

表1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | MSE | MAPE | MAE |
| 多项式回归 | 0.993 | 1881.682 | 0.014 | 32.399 |
| 随机森林回归 | 0.990 | 1507.890 | 0.012 | 28.929 |
| XGBoost回归 | 0.991 | 1353.591 | 0.012 | 27.678 |

根据模型的结果选出对模型贡献率最高的前5个变量，结果如表\*所示。

表2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 多项式回归 | WR | 甲醇现货价格 | 冰醋酸进口量 | 甲醇毛利 | 冰醋酸库存量 |
| 随机森林回归 | 甲醇毛利 | 甲醇现货价格 | 冰醋酸现货价格 | 甲醇进口量 | 冰醋酸进口量 |
| XGBoost回归 | 甲醇毛利 | 甲醇现货价格 | 天然气产量 | 冰醋酸进口量 | 天然气进口量 |

2.含有技术指标的模型回归结果

将筛选出的技术技术指标添加到有交叉项多项式回归、随机森林以及XGBoost回归模型中，各模型预测的结果的R2、MSE、MAPE如表\*所示。

表3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | MSE | MAPE | MAE |
| 多项式回归 | 0.995 | 916.754 | 0.010 | 23.669 |
| 随机森林回归 | 0.993 | 966.167 | 0.010 | 23.159 |
| XGBoost回归 | 0.993 | 1052.282 | 0.010 | 27.524 |

根据模型的结果选出对模型贡献率最高的前5个变量，结果如表\*所示。

表4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 多项式回归 | WR | 甲醇现货价格 | 冰醋酸进口量2 | 冰醋酸出口量 | 甲醇毛利 |
| 随机森林回归 | 甲醇毛利 | 甲醇现货价格 | 冰醋酸现货价格 | OBV | 甲醇进口量 |
| XGBoost回归 | 甲醇毛利 | 甲醇现货价格 | 天然气产量 | 天然气进口量 | 天然气库存量 |

对比表1和表3模型评估标准，可以发现在添加技术指标到模型后，3种模型的R2分别从0.993、0.990、0.991提升到了0.995、0.993、0.993，并且3种模型的3种误差均得到了一定程度的缩小，这说明技术指标对预测甲醇期货收盘价具有明显的作用。根据表1可以发现，多项式回归模型除了在R2模型评估指标上表现的比其他两种树模型要好，在其他3种误差上表现的均比这两种树模型要差，在添加了技术指标之后除了在MAE这项评价指标上要比XGBoost回归模型要差之外，其余均优于这两种树模型。从MSE的值来看，在添加技术指标到模型后，随机森林的预测结果表现的比XGBoost更为平稳。除此之外，无论是添加技术指标预测变量之前还是之后，两种树模型的表现结果都较为相似。

比较表2和表4可以发现，无论是在添加技术指标预测变量前还是后，甲醇毛利、甲醇现货价格以及冰醋酸进口量对本文所选择的三种回归模型的贡献均较高，因而可以判断这三个预测变量与甲醇期货收盘价紧密相关，对预测甲醇期货收盘价具有重要作用。

五 结论与启示

首先，本文借助经济学原理中的重要理论，并根据现实市场交易中市场参与者所采用的常见的技术指标来初步选取模型的备选预测变量，对数据预处理之后使用相关性分析最终选取了模型预测变量。

其次，根据被预测变量甲醇期货价格，以及选取预测变量的特点选择含有交叉项的二次多项式、随机森林、XGBoost回归模型。构建对照回归模型之后，对两种树模型的超参数进行调优，得到最终的三个回归模型。

最后，根据模型得到的预测结果进行对比分析。结果发现：含有交叉项的二次多项式模型在未加入技术指标预测变量之前除了在R2这一模型评估标准上要优于两个树模型以外，在其余误差评估标准上均比这两个树模型表现的差，因而从预测结果的准确度以及稳定性角度来说，可以认为含有交叉项的二次多项式回归要表现的比两种树模型差。两个树模型表现的结果较为相似，但XGBoost模型整体表现要优于随机森林回归模型。在模型中添加技术指标后，三个模型在预测结果上均得到了一定提高。含有交叉项的二次多项式回归在R2、MSE、MAPE三个模型评估指标上要优于两个树模型，在MAE这一指标上要略低于随机森林模型所得出的结果，但高于XGBoost所表现的结果，因而可以认为在添加技术指标预测变量之后，二次多项式回归模型预测能力得到了明显提高，在准确性和稳定性方面要优于其他两个模型。此外，比较添加技术指标到模型前后的实验结果，发现甲醇毛利、甲醇现货价格以及冰醋酸进口量对模型的贡献度都比较高。

从本文实证分析的结论可以得到如下几点启示：一是从甲醇上下游产业链角度来选取甲醇期货收盘价的预测变量具有切实可行的意义；二是技术指标对提高模型预测准确度具有重要作用；三是在添加技术指标到模型中后，二次多项式的预测能力要比其他树模型更为准确。

参考文献

[1]周亚.基于XGBoost与弹性网络回归的集成模型对上证指数日极差的预测分析[J].中国物价,2021(05):68-71.

[2]李燕,何如海,廖宜静,聂雷.聚氯乙烯期货价格与现货价格传导关系实证研究[J].宿州学院学报,2013,28(01):14-18.

[3]范俊明,刘洪久,胡彦蓉.基于LSTM深度学习的大豆期货价格预测[J].价格月刊,2021(02):7-15.

[4]陈标金,王锋.宏观经济指标、技术指标与国债期货价格预测——基于随机森林机器学习的实证检验[J].统计与信息论坛,2019,34(06):29-35.

[5]范俊明,刘洪久,胡彦蓉.基于LSTM深度学习的大豆期货价格预测[J].价格月刊,2021(02):7-15.

[6]Liran Einav,Jonathan Levin. Economics in the age of big data[J]. Science,2014,346(6210).

[7]Alessandro Vespignani. Predicting the Behavior of Techno-Social Systems[J]. Science,2009,325(5939).

[8]李航. 统计学习方法[M].清华大学出版社:, 201905.403.

[9]Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

[10]Sendhil Mullainathan,Jann Spiess. Machine Learning: An Applied Econometric Approach[J]. Journal of Economic Perspectives,2017,31(2).

[11]Pérez-Rave,Correa-Morales,González-Echavarría. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes[J]. Journal of Property Research,2019,36(1).

[12]Lucas Alexandre,Pegios Konstantinos,Kotsakis Evangelos,Clarke Dan. Price Forecasting for the Balancing Energy Market Using Machine-Learning Regression[J]. Energies,2020,13(20).

[13]马郑玮,琚心然,姚瑜,冯异立.基于产业链视角对石油期货与PTA期货价格关联性的分析[J/OL].现代化工:1-16[2021-06-29].http://kns.cnki.net/kcms/detail/11.2172.tq.20200910.0845.005.html.

[14]尚玉皇,郑挺国,夏凯.宏观因子与利率期限结构:基于混频Nelson-Siegel模型[J].金融研究,2015(06):14-29.