**基于产业链甲醇期货价格预测重要变量筛选**

**作者：**

**摘要：**本文使用随机森林和XGBoost算法，从甲醇产业链角度出发寻找预测甲醇期货价格的特征变量。样本选取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经过归一化后的基本面预测变量构造对照模型，采用R2、MAPE来评估模型的表现。研究结果显示，无论是否进行归一化，模型的预测精度没有较大提高，但是经过归一化处理后所得的两种树模型的预测误差变得很大；无论是否对数据进行归一化处理，甲醇毛利、甲醇现货价格、甲醇产量三个变量对两种树模型的重要程度均处于最前列，对模型的预测具有重要意义。

关键字：期货价格；价格预测；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、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对金融资产价格的预测具有重要意义[13]。

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。

**2.预测变量的处理方法**

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预测变量与被预测变量相关性检验结果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 甲醇产量 | 甲醇出口量 | 甲醇进口量 | 甲醇库存量 | 甲醇现货价格 | 甲醇毛利 | 天然气产量 | 天然气进口量 |
| 期货价格 | Pearson 相关性 | -.054 | .270\*\* | -.506\*\* | -.457\*\* | .948\*\* | .959\*\* | -.122\*\* | -.009 |
| 显著性（双侧） | .059 | .000 | .000 | .000 | .000 | .000 | .000 | .754 |
| Kendall相关系数 | .010 | .143\*\* | -.344\*\* | -.322\*\* | .814\*\* | .849\*\* | -.122\*\* | -.010 |
| Sig.（双侧） | .596 | .000 | .000 | .000 | .000 | .000 | .000 | .618 |
| Spearman相关系数 | .005 | .195\*\* | -.506\*\* | -.458\*\* | .956\*\* | .968\*\* | -.173\*\* | -.039 |
| Sig.（双侧） | .855 | .000 | .000 | .000 | .000 | .000 | .000 | .170 |
|  | | 天然气库存量 | 天然气现货价格 | 冰醋酸出口量 | 冰醋酸进口量 | 冰醋酸库存量 | 冰醋酸消费量 | 冰醋酸现货价格 | 冰醋酸毛利 |
| 期货价格 | Pearson 相关性 | -.091\*\* | .425\*\* | .344\*\* | -.483\*\* | -.248\*\* | -.052 | .582\*\* | .379\*\* |
| 显著性（双侧） | .002 | .000 | .000 | .000 | .000 | .070 | .000 | .000 |
| Kendall相关系数 | -.090\*\* | .307\*\* | .250\*\* | -.301\*\* | -.175\*\* | .002 | .536\*\* | .324\*\* |
| Sig.（双侧） | .000 | .000 | .000 | .000 | .000 | .915 | .000 | .000 |
| Spearman相关系数 | -.123\*\* | .455\*\* | .347\*\* | -.430\*\* | -.252\*\* | .009 | .751\*\* | .480\*\* |
| Sig.（双侧） | .000 | .000 | .000 | .000 | .000 | .745 | .000 | .000 |
| \*\*. 在置信度（双测）为 0.01 时，相关性是显著的。 | | | | |  |  |  |  |  |

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。

**三、研究方法**

根据表1所示，各预测变量和被预测变量之间表现出强线性相关性的较少。根据各预测变量之间的相关性分析可以发现，大部分变量之间呈现出显著的相关性，限于篇幅没有将预测变量之间的相关性分析结果附上。方匡南等人于2011年指出，随机森林不需要考虑共线性问题，能够很好地处理高维度数据，对异常值和噪声有很好的容忍度，不易出现过拟合，且预测准确率高[14]。XGBoost作为梯度提升树模型的扩展，具有高稳定性和预测精度并且具备很强的拟合能力（fitting ability）[15]。根据预测变量和被预测变量所呈现出来的特点，本文采用随机森林和XGBoost回归模型来预测甲醇期货价格，并发现对预测甲醇期货价格贡献率较高的特征变量。

1. **随机森林回归模型**

随机森林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确的预测。随机森林在构造单棵决策树的过程中除了在选择样本时使用随机抽样，同样选择随机抽样的方式从原特征空间中选择部分特征作为输入特征，此做法有效降低了模型的不稳定性。在构建单个决策树的过程中，本文选择以分层随机抽样的方式来出现数据泄漏的问题。由于本文采用的变量是时间序列，如果不考虑时间先后顺序二随机选择训练集和验证集，容易出现错误地将验证模型的部分信息作为已知信息对模型进行训练，有可能导致最终模型的无效[16]。

1. **XGBoost回归模型**

XGBoost是机器学习中较新的算法，由Chen和Guestrin在2016年提出，可用于分类、回归和排序模型的创建，是基于梯度提升树的扩展。XGBoost回归模型是基于决策回归树。，并在目标函数中包含了正则化项来避免模型出现过拟合现象。XGBoost回归模型的目标函数为：

*其中*为梯度提升算法损失部分，为正则化项，为常数项。XGBoost模型的参数可以分为三个部分，分别为通用参数，booster参数，学习目标参数。当其中的booster参数被设置为‘gblinear’是模型是一个线性模型，如果被设置为‘gbtree’则模型是一个树模型，根据被预测变量和预测变量相关性分析的结果可以看出，呈现出强相关性的变量较少，仅仅为2个，因而本文选择树模型。通过对上述三类参数进行调优以选择相对最优的回归模型。在构建模型过程中，同样采用分层抽样的方式来降低不良模型出现的可能性。

**四、实证分析**

**（一）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。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个基本面特征变量。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，所以要对特征变量进行预处理，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条的数据。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。

**（三）模型预测结果**

**1. 原始数据模型回归结果**

随机森林和XGBoost回归模型的表现结果如下表所示。

表2原始数据回归模型结果

|  |  |  |
| --- | --- | --- |
|  | R2 | MAPE |
| 随机森林 | 0.990 | 0.012 |
| XGBoost | 0.991 | 0.012 |

选取两个回归模型中对模型贡献率最高的前5个变量，结果如下表所示。

表3原始模型贡献度最高特征变量

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 随机森林回归 | 甲醇毛利 | 甲醇现货价格 | 冰醋酸现货价格 | 甲醇产量 | 冰醋酸进口量 |
| XGBoost回归 | 甲醇毛利 | 甲醇现货价格 | 天然气进口量 | 甲醇产量 | 天然气产量 |

**2. 归一化数据模型回归结果**

表4归一化数据回归模型结果

|  |  |  |
| --- | --- | --- |
|  | R2 | MAPE |
| 随机森林 | 0.991 | 0.103 |
| XGBoost | 0.992 | 0.077 |

选取两个回归模型中对模型贡献率最高的前5个变量，结果如下表所示。

表5归一化模型贡献度最高特征变量

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 随机森林回归 | 甲醇毛利 | 甲醇现货价格 | 冰醋酸现货价格 | 甲醇产量 | 冰醋酸毛利 |
| XGBoost回归 | 甲醇毛利 | 甲醇现货价格 | 甲醇产量 | 天然气产量 | 天然气库存量 |

**（四）模型结果分析**

比较分别由原始数据以及经过归一化处理后的数据所构建的树模型的表现可以知道，在经过归一化处理之后，两种树模型的R2均提高了0.1%，模型的预测精度没有较大提高，但从MAPE表现结果来看，经过归一化处理后的两种树模型均表现地比原始数据构建的模型所表现出的结果要差，随机森林模型的误差增大了近10%，XGBoost模型误差增大了接近6%，因而从误差角度来看，将数据进行归一化处理后，模型的预测精确度反而变差了。比较表3和表5可以发现，无论是否将基本面数据归一化，甲醇毛利、甲醇现货价格以及甲醇产量这三个变量对两种树模型的重要程度都很高，可以认为这三个变量对预测甲醇期货价格具有重要作用。

根据两种树模型回归的结果可以看出，使用本文筛选基本面特征变量，可以将甲醇期货价格预测模型的R2保持在99%以上，能够很精准地预测甲醇期货价格，具有很高的实用价值。

**五 结论与启示**

首先，本文借助经济学原理中的生产成本以及供需关系准则角度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随机森林、XGBoost回归模型。将预测变量以及被预测变量归一化后，构建对照回归模型，对两种树模型的超参数进行调优，得到最终的两种回归树模型。

最后，根据模型得到的预测结果进行对比分析。结果发现：在将数据进行归一化处理之后，两种树模型的R2并没有明显的变化，两者仅提高了0.1%，但模型的预测误差却变得很大，因而归一化并没有改善模型的预测精度。甲醇毛利、甲醇现货价格、甲醇产量对模型的重要程度并没有因为对数据进行归一化处理而发生较大变化，可以认为这三个特征变量对甲醇期货价格预测具有重要作用。使用本文筛选后的基本面特征数据，模型的预测精度可以达到99%，能够很精准地预测甲醇期货价格，具有很高的实用价值。

**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]方匡南,吴见彬,朱建平,谢邦昌.随机森林方法研究综述[J].统计与信息论坛,2011,26(03):32-38.

[15]Q Chen. XGBoost: A scalable tree boosting system. The 22nd ACM SIGKDD International Conference ACM, 2016.

[16]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.