摘要：以筛选的16个基本面特征变量作为甲醇期货价格预测变量，利用ARMA、随机森林和XGBoost构建预测模型。依据三种模型的预测精度来检验基本面特征对甲醇期货价格的预测能力。研究结果发现：经过归一化处理后的预测变量并没能让模型的预测精度提高，反而较大程度地增大了模型预测的不稳定性；特征变量甲醇毛利、甲醇现货价格、甲醇产量对预测模型的贡献度一直处于前列（对模型贡献度最高的），这表明这三个特征变量对甲醇期货价格具有预测意义；利用随机森林和XGBoost能够建立有效的甲醇期货价格预测模型。（考虑一下是不是要进行一下跟踪交易）（摘要中目前还缺少一个与时间序列分析之间进行的比较）

关键字：期货价格；价格预测；随机森林；XGBoost；（ARIMA）

引言

期货交易是市场经济发展到一定阶段的产物，成熟的期货交易市场在一定程度上可以认为是完全竞争的市场，是经济学中最理想的市场形式。我国第一个商品期货交易所的出现要追溯到上世纪90年代，而世界上最早出现期货交易场所的时间是19世纪50年代，相比于一些国外期货市场而言，我国的期货交易市场还不太成熟，期货理论对现实期货交易的支撑还不够。国内的甲醇期货于2011年10月28日在郑州商品交易所上市，到今年已有十年之久，相关企业和投资者对甲醇期货的关注度也在不断上升，对甲醇期货价格进行分析对国内甲醇期现货市场的健康发展以及相关企业和投资者进行甲醇期现货交易具有重要意义。

在现有期货市场的研究中，研究方法主要分为数理统计和机器学习方法。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。机器学习方法包括决策树、支持向量机、随机森林、人工神经网络等【4】。随着计算机技术的发展，机器学习也逐渐被应用到金融领域的研究中，并表现出很好的研究性能【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]。机器学习被认为是计算机通过自身的不断观察和学习来提升自己与真实世界交互能力的一门学科[8]。Breiman指出随机森林算法模型不需要考虑多元回归模型中的多重共线性问题[9]，Pérez-Rave等（2019）指出随机森林算法可以有效避免模型的过拟合，是一个能够做出精确预测的强鲁棒性算法[11]。Chen和Guestrin（2016）提出的XGBoost算法已被成功应用于多个领域。黄卿等（2018）将XGBoost应用到股指期货的变动方向的预测上，验证了该方法可以有效应用于我国金融领域的研究【黄卿，谢合亮. 机器学习方法在股指期货预测中的应用研究——— 基于 BP 神经网络、SVM 和 XGBoost 的比较分析 [J].】。

本文基于ARIMA、随机森林和XGBoost回归模型构建对照模型，以甲醇期货价格作为ARIMA的输入变量，以甲醇产业链上下游产品的基本面数据作为随机森林和XGBoost模型的输入，将其应用于甲醇期货价格序列的预测分析，分析产业链上下游产品的基本面特征变量对甲醇期货价格预测的有效性。

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业链上下游产品的基本面数据对金融资产价格预测具有重要意义【13】。

1.甲醇产业链结构分析

甲醇是有合成气生产的重要化学品之一，它既是一种燃料，也是一种化工原料，目前市场上交易的甲醇主要以工业甲醇为主，其生产所需的原材料主要为煤、焦油、天然气等，其在生产过程中不会产生任何副产品，并且消耗过程中没有污染物的排放。甲醇自工业化生产以来，生产工艺也随着技术的进步以及能源结构的变化发生了较大变化，由原先的以煤和焦炭为原料进行生产，转变为以天然气为主要原材料进行生产。由于我国独特的能源结构的分布呈现出“富煤贫油少气”的特征，国内甲醇的生产依旧以煤为主。目前国内以煤作为原料甲醇生产产能占总产能的65%，以天然气为原料的产能占总产能的19%，以焦炉气为原料的产能占总产能的16%。

甲醇的下游产业链也十分丰富，涉及到化工、建材、能源、医药、农药等众多行业，在我国经济中扮演着十分重要角色。它不仅是一种重要的化工原料，可以用来生产甲醛、二甲醚、醋酸等，还是一种重要的燃料，用于加入汽油掺烧或者代替汽油作为动力燃料，同时还是农药生产的重要原料等。

2.备选预测变量选取与预处理

经过上述甲醇产业链结构分析可知，甲醇产业链的上游产品主要为煤、天然气、焦炉气，下游产品主要有甲醛、二甲醚、醋酸等。理想情况下，根据生产理论，上游产品作为下游产品的生产原料，其销售价格将作为下游产品生产的直接费用，并直接反映到下游产品的销售价格；从采购成本角度出发，上游产品的库存量也将直接影响到下游产品的采购成本，并呈现出负向相关关系；从供需关系的角度出发，当产品的供给充足时，价格会出现下降的趋势，反之则会出现上升的趋势，从而呈现出负相关关系，而产品的产能将直接影响产品的库存量，并表现出正相关性；从商品期货现实交易过程来看，期货交易最终将以现货的方式进行交割，在众多期现货价格研究中也已证实两者之间的强相关性。

根据上述分析，结合我国甲醇产业链的实际情况，以及数据的可获得性，本文选取的样本时间区间为2016年1月2日至2021年5月31日，选择的上下游产品分别为煤和醋酸，备选的基本面特征为：煤现货价格、产量、进口量、库存量；甲醇现货价格、产量、进口量、出口量、库存量、毛利；醋酸现货价格、进口量、出口量、库存量、消费量、毛利，共16个特征变量。其中三种产品的现货价格来自生意社网站，甲醇期货价格取自大智慧甲醇主力合约收盘价，其余基本面特征变量样本来源于钢联客户端。

3.备选预测变量预处理

由于本文所选取的备选特征变量中，有些数据是日频，而有些数据是周频和月频，需要对数据进行预处理。结合现实交易过程，样本数据基本是在交易日的收盘时刻才能出现，因而为了做出具有实用价值的模型，也需要对样本进行预处理。本文的处理方式是将备选预测变量数据向后偏移一个交易日，既当日的甲醇期货价格与前一交易日的备选特征变量值构成一个样本。对于较大口径特征变量，本文采用的方式是将两条相邻样本日期之间内所有交易日的数据均取值为离交易日较近的样本取值。将经过上述处理的各个特征变量与甲醇期货收盘价进行匹配以形成最终的样本。

3.备选预测变量相关性检验及预测变量筛选

对上述经过预处理后获得样本数据进行相关性检验。本文选用SPSS统计软件对样本数据进行Pearson、Spearman、Kendall相关性检验，检验结果如下所示。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 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 | -.164\*\* | .270\*\* | -.506\*\* | -.457\*\* | .948\*\* | .959\*\* | -.122\*\* | -.009 |
| 显著性（双侧） | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .754 |
| Kendall相关系数 | -.136\*\* | .143\*\* | -.344\*\* | -.322\*\* | .814\*\* | .849\*\* | -.122\*\* | -.010 |
| Sig.（双侧） | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .618 |
| Spearman相关系数 | -.105\*\* | .195\*\* | -.506\*\* | -.458\*\* | .956\*\* | .968\*\* | -.173\*\* | -.039 |
| Sig.（双侧） | .000 | .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。 | | | | |  |  |  |  |  |

根据上表可以看出，在0.01显著水平（双侧检验）下，除了煤进口量、醋酸消费量之外的所有特征变量均与甲醇期货价格显著相关。根据相关性检验结果，本文将煤进口量和醋酸消费量之外的共14个备选特征变量选为预测变量。

本文仅将被预测变量与各备选预测变量之间的相关性检验结果呈现于上，根据Pearson相关性检验结果，各备选预测变量两两之间，共44对的相关性检验结果中，有33对相关性检验结果是显著的，表明预测变量之间是存在共线性的。（是不是放在这里不是很合适）

（以上变量的筛选是没有问题的，接下来就是处理实验的部分）

研究方法

根据被预测变量和预测变量之间的相关性检验结果，与被预测变量呈现出强相关性的预测变量仅为甲醇现货价格和甲醇毛利，其余变量呈现出中、弱相关性。根据各预测变量相互之间的相关性检验结果，在44对预测变量之间相关性检验结果中，有33对检验结果是显著的，表明预测变量之间是存在多重共线性。方匡南等（2011）指出随机森林不需要考虑共线性问题，对异常值和噪声有很好的容忍度，且不易出现过拟合，能够很好地处理高维度数据，且预测准确率高[14]。XGBoost作为梯度提升树模型的扩展，能够在保持高稳定性和预测精度的情况下保持良好的拟合能力（这里需要找一下这个结论出自谁）。结合甲醇期货价格所呈现出的特点（是否要考虑一下和比特币所呈现出的动态性进行比较）（这里可能需要画一张图），本文选用适应能力较强的随机森林和XGBoost回归模型对甲醇期货价格进行预测并找出对模型贡献度较高的预测变量，本文选用MAPE和R2作为模型的评价指标。

随机森林模型

接下来的任务是描述一下随机森林，然后进行实验，实验之后进行回测检验模型的有效性，

同时要在检验模型的有效性以前进行和现有模型的比较

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。

XGBoost时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。

ARIMA模型

实证分析

变量样本数据说明

本文选取2016年1月2日至2021年5月30日区间内甲醇期货主力合约收盘价作为被预测变量。备选预测变量为煤现货价格、产量、进口量、库存量；甲醇现货价格、产量、进口量、出口量、库存量、毛利；醋酸现货价格、进口量、出口量、库存量、消费量、毛利，共16个特征变量。其中现货价格数据来源于生意社网站，期货价格数据来源于大智慧行情系统，其他特征变量来源于钢联客户端。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前一交易日对应的特征变量样本值则选用已有的离本交易日最近的样本值与之匹配。

预测变量筛选结果

使用SPSS统计软件进行备选预测变量与被预测变量之间的Pearson、Spearman、Kendall相关性检验，检验结果如表1所示。在显著性水平0.01情况下，除了天然气进口量和冰醋酸消费量以外的备选预测变量均与甲醇期货价格显著相关，因而将天然气进口量和冰醋酸消费量以外的备选预测变量选为预测变量。

预测模型构建

考虑到本文所选取的预测变量的量纲不同，可能会对模型的预测结果产生影响。因此对预测变量和被预测变量进行归一化处理。归一化方式为：

其中为原始甲醇期货价格样本值，和为甲醇期货价格样本值中的最小值和最大值，为归一化后的样本值；为选择的第k个预测变量所对应的第j个样本值，为第k个预测变量样本值中的最小值和最大值，为第k个预测变量第j个样本值经过归一化处理后所对应的样本值。

分别将筛选的预测变量和经过归一化后的筛选预测变量作为随机森林和XGBoost模型的输入变量构建对照模型，分析所选的预测变量对模型的贡献程度。此外，将上述模型结果与ARIMA预测结果进行对比，分析基于产业链筛选的变量是否有助于提高甲醇期货价格的预测精度。

随机森林模型

XGBoost模型

ARIMA模型

模型预测结果