广东省大学生创新实验项目

结题验收书

项 目 编 号：	S202210559116
项 目 名 称：	基于我国粮食经济发展战略背景下的小麦价 ARIMA-LSTM-XGBoost 组合预测模型研究
项 目 负 责 人：	苏晓钰
负 责 人 年 级 及 专 业：	2019 级经济统计学
联 系 电 话：	13690803668
指 导 教 师 姓 名：	林少萍
项 目 所 在 院 系：	经济学院
项 目 起 止 时 间：	起于：2022 年 3 月 止于：2023 年 3 月

广东省教育厅 制

2023 年 4 月

填表说明

1、本表前五项项由项目小组集体填写，除需亲笔签名外，其余部分均需采用打印稿，不够可加页；

2、本表第六项由指导教师核查填写，第七项由院系填写，第八项由学校验收专家组填写；

3、本表第三项以附录方式提交内容至少含：

附录一：公开发表的论文、获授权专利、论文发表或专利申请录用通知复印件或其它支撑材料

附录二：项目组成员科研总结，包括成功、失败的经验教训和心得体会

4、填写本表内容统一使用宋体，小四号字，单倍行距。

一、项目组成员（姓名栏本人签字）				
学号	姓名	学院	内招/外招	具体分工
2019052121	苏晓钰	经济学院	内招	负责学习 ARIMA、LSTM 等模型的原理，并对模型进行组合，构建出优化的预测模型，用 ARIMA、LSTM 等预测模型对小麦现货价格时间序列与各种变量之间的关系作出预测分析，检验比较小麦价格预测的准确率。之后，在完成前者的基础上总结几种预测模型对于小麦现货价格预测的优缺点。
2019051643	陈嘉妤	国际学院	内招	负责收集宏观粮食农业扶持改革政策、国内外政治、经济、社会影响因子的资料，确定课题研究方向，相对集中地对小麦价格变化进行定性分析。此外，对预测所得的价格走势提出未来产业布局和相关政策建议。并在论文中完成对这一部分的撰写。
2019052020	李敏怡	经济学院	内招	负责查找国内外小麦期现货价格等变量与社会环境变量，并学习掌握 R 统计分析软件的部分功能，将所构建的多个预测模型通过代码实现。最后，撰写论文的模型原理部分，并完成研究报告及相关材料的填写。
2020101603	阮炜霖	信息科学技术学院	内招	基于影响因素指标体系的未来值，构建 ARIMA、多变量的 XGBoost 模型与多变量的 LSTM 模型的组合模型，并将三组模型的预测值输入 SVR，采用贝

				叶斯优化算法进行调参，完成这部分代码的编写，并对各个模型的预测效果进行评价。
2020100730	郑昊天	信息科学技术与学院	内招	负责通过所构建的小麦期货价格的预测指标体系,基于传统的 ARIMA 模型、LSTM 模型建立经典的单变量时间序列模型。此外，基于影响因素指标体系,将小麦价格序列分解为线性趋势和非线性趋势，完成这部分代码的编写，并完成相关部分论文的撰写。

二、项目实施情况

1、完成任务情况	立项任务书的预期成果 （1）调研报告：通过对小麦市场公开数据建立 ARIMA-LSTM-XGBoost 预测模型分析，结合国家农产品扶持政策与小麦宏观经济环境的信息，完成一份关于小麦期货价格预测模型的报告《基于我国粮食经济发展战略背景的 ARIMA-LSTM-XGBoost 小麦价格组合预测模型》。 （2）学术论文：将调研报告整理归纳为一篇完整的学术论文，选择合适的期刊或会议，进行投稿发表。 （3）学术报告：向国家农业农村部递交报告，帮助相关部门通过本项目预测模型更好地预测小麦价格走势，降低市场风险。	实际获得的成果 （1）为课题研究准备基本的资料， 研究国内小麦生产状况、农业提振政策改革及国际粮食供给趋势，参考相关领域专家意见，完成对课题所处客观环境的分析。 （2）成功进行数据处理，编写相关计算程序，建立理论模型。 （3）完成对中国各省区小麦产出和价格的测度，并对外部环境因素进行定量分析， 在此基础上分析我国小麦现货价格波动,完成对未来小麦现货价格预测,设计理论最优计量模型。 （4）学术论文暂未完成投稿，调研报告、学术报告均已完成。

2、项目完成内容、关键技术及效果

本项目基于《中国农产品价格调查年鉴》中的 2009 年 1 月 4 日至 2021 年 12 月 17 日共 3255 个样本我国小麦期货价格序列的日数据、以及《中国统计年鉴》、国家粮油信息中心关于农业经济投入、人力投入、我国财政收支、小麦对外贸易、小麦国际进出口量等多个影响因素，建 立数学模型预测小麦期货价格。

第一部分：构成小麦期货价格的影响因素指标体系

计算 Pearson 相关系数度量变量之间的非线性相关程度通过随机森林特征重要性排序，最 终筛选出少数具有影响力的因素，构成预测小麦期货价格的指标体系，

第二部分：基于所构建的小麦期货价格的预测指标体系，建立模型进行预测

（1）通过传统的 ARIMA 模型建立经典的单变量时间序列模型；

（2）基于 XGBoost 模型和 LSTM 模型将无监督的时间序列数据进行转化，建立预测市场现 货价格的单变量 XGBoost 模型和单变量 LSTM 模型；

（3）为了提升模型预测效果，根据由随机森林提取的特征构建影响因素指标体系，分别构建多变量的 XGBoost 模型和多变量的 LSTM 模型并进行预测；

（4）参考 Stacking 集成学习算法在多领域中取得的丰硕成果，考虑构建 ARIMA、多变量的 XGBoost 模型和多变量的 LSTM 模型的组合模型。第一步，基于 ARIMA 模型对各个影响因素的未来值进行预测；接着，将指标体系中的预测值作为影响变量，构建出改进的多变量的 XGBoost 模型和多变量的 LSTM 模型，最终将传统 ARIMA 模型以及上述改进模型所得的三组预测值输入支持向量机回归 SVR 模型，并采用贝叶斯优化算法调参，最终得到优化的组合模型的小麦期货价格预测值。

第三部分：评价单一预测模型与组合模型的预测效果，并进行最优选择

通过计算得到测试集的误差、相对误差以及平均相对误差，并比较 ARIMA-LSTM-XGBoost 组合模型与各单一预测模型如 ARIMA、单变量的 XGBoost 模型、LSTM 模型等预测模型下的指标,最终选出最适合的预测模型，并应用于我国小麦市场发展趋势预测。

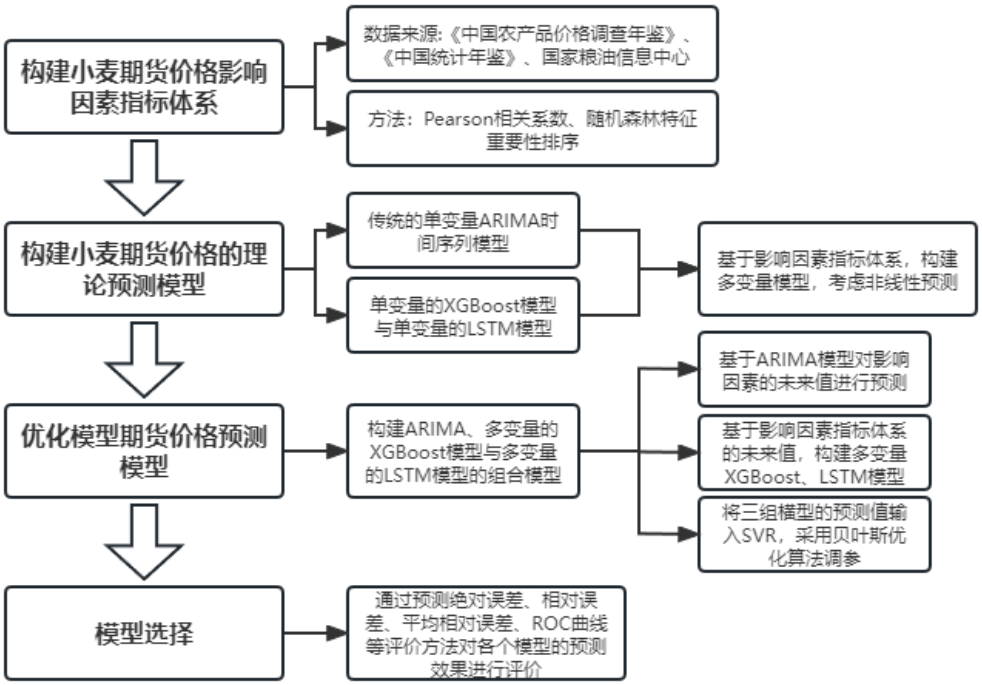
效果：

表 1 ARIMA 与组合模型在测试集预测效果

	真实值	ARIMA	预测残差	组合模型	误差	相对误差	平均相对误差
2021-05	2540.00	2538.918	-1.18	2540.10	0.10	0.00%	
2021-06	2522.22	2457.344	-31.90	2489.24	-32.98	-1.31%	
2021-07	2561.67	2503.054	-84.00	2587.05	25.38	0.99%	1.25%
2021-08	2595.56	2541.991	-11.70	2553.69	-41.87	-1.61%	
2021-09	2612.78	2539.56	-11.70	2551.26	-61.52	-2.35%	

相较于使用单一模型 ARIMA 模型进行预测，各年份误差、相对误差均有所下降，整体的预测效果更优，这主要应该是由于 ARIMA 模型预测结果往往较真实值偏小，而 ARIMA-LSTM-XGBoost 模型对此进行了一定程度的修正

关键技术：


构建小麦期货价格影响因素指标体系
数据来源:《中国农产品价格调查年鉴》、《中国统计年鉴》、国家粮油信息中心
方法：Pearson相关系数、随机森林特征重要性排序
构建小麦期货价格的理论预测模型
传统的单变量ARIMA时间序列模型
单变量的XGBoost模型与单变量的LSTM模型
基于影响因素指标体系，构建多变量模型，考虑非线性预测
优化模型期货价格预测模型
构建ARIMA、多变量的XGBoost模型与多变量的LSTM模型的组合模型
基于ARIMA模型对影响因素的未来值进行预测
基于影响因素指标体系的未来值，构建多变量XGBoost、LSTM模型
将三组模型的预测值输入SVR，采用贝叶斯优化算法调参
模型选择
通过预测绝对误差、相对误差、平均相对误差、ROC曲线等评价方法对各个模型的预测效果进行评价

3、项目整体进度安排及实施情况（包括实验的工作量）

第一步：确定课题研究方向。了解当前国内小麦市场影响因收集宏观粮食农业扶持改革政 策、国内外政治、经济、社会影响因子的资料，了解小麦市场风险等情况、市场概况以及所面 临的问题，引入了本可以的研究对象，通过团队讨论以及专家帮助确定研究问题和研究方向。

第二步：对经济统计领域的预测模型进行学习与研究。通过查阅文献了解当前农产品价格 预测的研究现状，学习所使用的相关方法与原理，为后续分析小麦价格波动成因和价格预测奠 定了理论基础。

第三步：基于专业的统计知识构建多个预测模型并检验预测效果，其中包括 ARIMA 模型、 LSTM 模型、XGBoost 模型等。通过随机森林特征重要性排序挖掘出国内小麦价格的重要影响因 素，并将所构建的指标体系运用到预测模型中，提高预测率。针对单一模型的缺陷进行改进， 如，根据模型预测结果，对所使用的所有预测模型进行评价，并选出预测效果最好的模型。

第四步：整理结果，撰写报告。将本项目的成果以报告的方式呈现给中国农业农村部、中 国粮油信息中心等相关部门，为国家农产品扶持政策的施行进言献策。

3、项目完成情况的自评意见（包括实施过程中的成功与失败）

成功之处：

（1）本项目建立了有关小麦价格预测的指标体系，并利用相关系数与随机森林进行变量筛选，综合考虑了变量间的线性和非线性相关程度，筛选出对小麦现货价格变化贡献较大的特征，有利于提高模型的预测能力与拟合速度。

（2）本项目考虑到小麦现货价格同时具有线性与非线性特征，选择了适用于线性预测、短期预测精度较高的 ARIMA 模型拟合趋势，以及选择在处理非线性时间序列数据方面性能较好的其他模型，建立的组合预测模型相比于单一预测模型具有提高精度的优势。

（3）综合考虑了小麦价格时间序列本身的自相关图确定滞后阶数，发现小麦价格的影响因素（例如存在一个滞后 12 期，即小麦价格受到一年前同样月份的价格影响），从而帮助确定最优模型参数。

不足：

（1）我们采用了 2009 年 1 月至 2021 年 9 月的月度数据，其中包含了 18 年中美贸易战、19 年年末至今的全球新冠肺炎流行病等有可能使小麦价格产生大幅波动的关键时间节点，但由于统计的时效性，本次研究的许多指标数据由于缺乏 2022 年俄乌战争发生以来的数据，不能针对 2022 年 1 月至今的俄乌冲突小麦价格进行有效的预测分析，希望如果能进

行后续的研究，能有效完善这一情况。

（2）神经网络模型需要大量的样本作为训练样本，对统计工作要求较高。由于我国目前统计发展仍落后于发达国家，许多经济数据都不能做到以月、日为步长公布，希望我们未来能更加有效提高地预测精度。

5、项目实施过程遇到的困难及解决方法

（1）我们在查找相关资料时，发现现有文献主要采用定性的方法，研究新冠肺炎疫情对中国粮食价格的影响、习惯将小麦价格与其它粮食关联，但对中国粮食价格的影响效应及作用机制没有得到充分关注，从而 忽略了小麦自身价格属性和变动情况。同时，在后疫情时代，粮食价格波动情况不同于以往，对于这一时期小麦现货价格的研究还处于未成熟阶段。我们认为对于粮食价格预测的模型可能受外部因素影响而使准确率发生变化，采用疫情前数据进行研究的模型需适时改进。于是，我们基于 Stacking 集成学习算法 改进现有的预测模型，考虑构建 ARIMA、多变量的 XGBoost 模型和多变量的 LSTM 模型的组合模 型：将三组模型的预测值作为支持向量机回归 SVR 模型的输入，并采用贝叶斯优化算法对 SVR 模型进行调参，解决影响因素的未来值的获取问题并显著提高预测精度。

（2）ARIMA、LSTM 模型能较准确地预测小麦现货价格的时间序列整体趋势，但对局部拟合效果不同，且部分拟合精度较低。该模型在外推时不方便加入外生变量，未考虑结构变动带来的影响，因此仅用 ARIMA、LSTM 模型可能导致未来的预测效果欠佳。

在实际应用中，通常可以结合各个模型的结构优势来构建组合模型，以达到更稳定高效的模型结构。此外，可以使用交叉验证等方法来评估组合模型的预测能力，并根据实际情况对模型进行调整。我们考虑将三种模型的优点结合，构建组合模型的思路如下：首先使用 ARIMA 模型来捕捉时间序列数据中的线性趋势和季节性模式。然后，使用 LSTM 模型来捕捉非线性关系和长期依赖性。最后，使用 XGBoost 模型来整合 ARIMA 和 LSTM 模型的预测结果，以提高预测精度。

相较于使用单一模型 ARIMA 模型进行预测，各年份误差、相对误差均有所下降，整体的预测效果更优，这主要应该是由于 ARIMA 模型预测结果往往较真实值偏小，而 ARIMA-LSTM-XGBoost 模型对此进行了一定程度的修正。

6、对指导老师、院系及学校的意见及建议

（1）希望学校可以多多举办相关创新创业比赛，如可以在校赛前先举行院赛，增加同院系之间学习的了解。

（2）希望院系可以开设创新创业、相关社团和公众号、微信小程序，团队之间招募、沟通更加方便。

（3）希望学校可以广泛开展创新创业方面的讲座，让同学们有更多这方面的知识储备。

三、项目成果

1、项目创新点

研究方法具有创新性：本课题基于 stacking 集成学习算法中分类器的特点，创新性地将传统的 arima 模型以及 LSTM、XGBoost 模型组合成多变量的预测模型。目前 XGBoost 算法和 LSTM 算法在多个领域的预测问题上取得了较好的成果，但二者在经济预测方面的研究较少，结合机器学习方法在经济序列预测的较好成绩，考虑引用影响因素、使用组合模型进行预测能有效地提高模型预测效果。

2、项目主要成果

本项目成功建立 ARIMA 模型、LSTM 模型、XGBoost 的组合模型进行预测精度的比较，并得出局部最优模型的结构，挖掘集成模型在预测小麦价格用途的可能性，丰富了有关市场价格预测的理论研究。

同时，本项目考虑到小麦现货价格同时具有线性与非线性特征，选择了适用于线性预测、短期预测精度较高的 ARIMA 模型拟合趋势，以及选择在处理非线性时间序列数据方面性能较好的其他模型，建立的组合预测模型相比于单一预测模型具有提高精度的优势。

我们考虑将三种模型的优点结合，构建组合模型的思路如下：首先使用 ARIMA 模型来捕捉时间序列数据中的线性趋势和季节性模式。然后，使用 LSTM 模型来捕捉非线性关系和长期依赖性。最后，使用 XGBoost 模型来整合 ARIMA 和 LSTM 模型的预测结果，以提高预测精度。

预测效果：

表 2 预测效果的模型结果比较

	真实值	ARIMA	LSTM	XGBoost	组合模型
2021-05	2540	2549	2494	2531	2540
2021-06	2522	2467	2502	2498	2489
2021-07	2562	2497	2507	2565	2587
2021-08	2596	2545	2551	2621	2554
2021-09	2613	2549	2584	2602	2551
平均相对误差		1.89%	1.51%	0.56%	1.25%

根据结果得出结论：在本项目中，LSTM 模型在训练集上的拟合性能不如 ARIMA 模型，但在测试集上的预测精度稍高。加入影响因素后，XGBoost 模型能够有效提升预测精度。ARIMA-LSTM-XGBoost 组合预测模型在小麦价格序列上的预测精度较高，稳定性更

好，但不及 XGBoost 网络模型。尽管如此，组合模型仍具有很高的应用前景，能够更好地捕捉时间序列数据中的非线性关系、噪声和异常值、长期依赖关系以及多元关系。

意义：当前全球各地疫情连续不断，阻碍各国进行有规律的农业生产，而近日俄乌局势动荡，乌克兰作为欧洲粮仓的地位被打破，导致以小麦为代表的国际与国内粮食价格被不断推高，直接增加了老百姓的生活成本与市场风险，也使国家难以对小麦价格进行调控。而对小麦价格的准确预测有利于国家积极应对市场波动并采取有效的政策策略，从而降低市场风险。基于结果提出相应政策建议，有助于及时调节市场供求关系，稳定粮食经济；并进一步推广到农产品价格预测领域。

四、项目科研日志或实验记录（可以以附件形式附在后面）

本项目基于《中国农产品价格调查年鉴》、《中国统计年鉴》与国家粮油信息中心的公开数据，从小麦国际和国内市场角度出发，探求影响小麦价格的重要因素，旨在对小麦发展趋势做出准确的预测。

第一部分：首先根据训练集数据，基于随机森林特征重要性排序构建了全面的小麦期货价格的影响因素指标体系，并构建出传统的时间序列分析模型与单变量的 LSTM 模型与单变量 XGBoost 模型。

第二部分：为了进一步构建更准确的小麦期货价格预测模型，我们基于所建立的影响因素指标体系，构建 LSTM 模型与 XGBoost 模型。

第三部分：最后考虑将构建的 ARIMA、LSTM 和 XGBoost 三种模型集成为一种多变量预测组合模型：将组合模型的预测值作为支持向量机回归 SVR 模型的输入，并采用贝叶斯优化算法对 SVR 模型进行调参，解决影响因素未来值获取问题并显著提高预测精度。

第四部分：综合组合模型与单一模型的预测结果，从绝对误差、相对误差、平均相对误差等指标比较模型，并选择出预测效果最好的模型。

五、经费使用情况

下拨经费： 10000 （元）

具体使用情况

用途	明细	单价（元）	数量	总金额（元）
图书资料费	查阅国内小麦生产状况、农业提振政策改革及国际粮食供给趋势资料	200/本	10 本	2000
数据采集费	向数据库付费获取数据	100/份	10 份	1000
咨询费	咨询高校教师和相关领域专家，解决技术难点	1000 元/人	4 人	4000
劳务费	酬谢指导老师与课题组成员的付出	375 元/人	4 人	1500
印刷费	课题组内部学习资料传阅	0.5/页	1000 页	500
交通费	咨询专家产生的往返路程费用	50/次	20 次	1000
合计				10000

六、指导教师意见

项目按期完成资料收集、文献综述、数据处理和编程、建立了 ARIMA 模型、LSTM 模型、XGBoost 的组合计量模型，利用最优模型对中国各省区小麦产出和价格做测度，并对外部环境因素进行定量分析。项目得出 LSTM 模型在训练集上的拟合性能不如 ARIMA 模型，但在测试集上的预测精度稍高， XGBoost 模型能够有效提升预测精度，ARIMA-LSTM-XGBoost 组合预测模型在小麦价格序列上的预测精度较高，稳定性更好等结论。结论有实际参考价值，达到预期目的。

项目已经完成调研报告和学术报告，经费使用符合规定。

同意结题。

亲笔签名：林少萍

2023 年 4 月 3 日

七、学院专家评审意见
专家签名： 年　　月　　日
八、学院意见
验收结果：优秀□　　合格□　　不合格□　　延期□ 签字盖章： 年　　月　　日
九、学校意见