Statistics and Application 统计学与应用, 2023, 12(4), 968-973
Published Online August 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.124100

Hans汉斯

气象因素对香蕉批发价格影响的预测研究：以XGBoost模型为例

黄文娟，温标堂

捷佳润科技集团股份有限公司，广西 南宁

收稿日期：2023年7月11日；录用日期：2023年8月1日；发布日期：2023年8月15日

摘　要

本研究通过集成XGBoost模型，采用主要批发市场历史价格数据和相关气象因素，开展了一项系统性的香蕉批发价格预测研究。利用参数优化及基于树的特征选择方法，实现了对模型预测精度的显著提升。结果证实了气象因素对香蕉价格有显著影响，进一步印证了我们的假设。因此，模型为市场参与者提供了准确的参考信息，有助于进行更精准的价格预测。未来的研究方向包括更深入的特征工程优化，以及探索其他可能影响香蕉价格的因素，以期进一步提高预测模型的性能和应用范围。

关键词

XGBoost，香蕉价格，预测模型，气象因素，参数优化

Predictive Study of Meteorological Factors on Banana Wholesale Price: A Case of XGBoost Model

Wenjuan Huang, Biaotang Wen

JJR Science and Technology Group Co., Ltd., Nanning Guangxi

Received: Jul. 11th, 2023; accepted: Aug. 1st, 2023; published: Aug. 15th, 2023

Abstract

This study conducted a systematic prediction of banana wholesale prices by integrating the XGBoost model with historical price data from major wholesale markets and relevant meteorological factors. The predictive accuracy of the model was significantly improved through parameter optimi-

文章引用: 黄文娟, 温标堂. 气象因素对香蕉批发价格影响的预测研究: 以 XGBoost 模型为例[J]. 统计学与应用, 2023, 12(4): 968-973. DOI: 10.12677/sa.2023.124100

zation and tree-based feature selection methods. The results confirmed that meteorological factors have a significant impact on banana prices, further validating our hypothesis. Therefore, the model provides accurate reference information for market participants, helping to make more precise price predictions. Future research directions include more in-depth feature engineering optimization, and exploring other factors that may affect banana prices, with the aim of further improving the predictive model's performance and applicability.

Keywords

XGBoost, Banana Price, Predictive Model, Meteorological Factors, Parameter Optimization

1. 引言

1.1. 研究背景与目的

香蕉，作为全球消费量庞大的水果，其价格波动主要受季节性、气象变化和供应量等多重因素影响，其中气象因素的影响尤为重要。因此，构建一个基于气象数据的香蕉批发价格预测模型具有重大的实用价值，可以为市场决策提供可靠的参考依据。本研究将利用 XGBoost 机器学习算法，并结合北京新发地、广州江南市场的香蕉历史批发价格和相关气象数据，构建价格预测模型。同时，通过对模型参数和特征进行优化，我们期望进一步提高预测模型的精度。

1.2. 相关工作

随着信息技术的快速发展，机器学习技术在许多领域得到了广泛的应用，包括价格预测。特别是对于香蕉这种大宗农产品，在过去的研究中，通常使用时间序列分析、自回归模型等传统统计方法来预测农产品价格，但这些方法对于处理复杂的、非线性的、动态变化的数据可能存在困难。然而，这些方法通常假设数据是线性的和稳定的，这在很多情况下并不适用。香蕉价格的波动受到天气、市场需求、运输成本等多种因素的影响，这些因素的复杂交互使得香蕉价格呈现出非线性和非稳定的特性。相比于传统的统计方法，机器学习算法能够更好地处理非线性和非稳定的数据相比之下，XGBoost 机器学习算法、随机森林等机器学习算法，能更好地处理这种复杂数据，且预测结果往往更准确。

2. 数据处理

2.1. 数据来源

本研究所用数据源丰富、相关性强，这为模型的准确性和可靠性提供了坚实的基础。我们收集了自 2017 年 1 月 1 日至 2023 年 5 月 31 日新发地、江南市场的逐日香蕉批发价格数据作为研究的目标变量。北京新发地和广州江南果蔬市场作为中国最大的农产品批发市场，其香蕉价格数据可反映市场供需和价格趋势。同时，我们还考虑了自 2017 年 1 月 1 日至 2023 年 5 月 31 日新发地市场以及国产香蕉主要产区(如广西隆安、云南西双版纳、广东湛江和海南陵水等地)的气象数据，包括各主要种植区日累计降雨量、最低温度、最高温度、平均温度、平均风速等。这些气象因素可能会直接或间接影响香蕉的生长状况和产量，进而影响供应量和价格。尤其在农业生产中，天气因素对作物的生长有着重要影响，例如，降雨

量和温度可能影响香蕉的生长状况，而风速可能影响香蕉的运输。因此考虑这些因素对于模型的预测能力至关重要。因此，我们所用的数据集涵盖了价格、气象条件和产地等多个因素，为构建一个高效准确的预测模型提供了可能。

2.2. 数据预处理

在数据预处理阶段，我们对原始数据进行了清洗和修补，合并了价格数据、市场气象数据和产地气象数据，以建立一个综合的数据集。在数据预处理阶段，笔者以原始数据集为基础，执行了一系列的清理与修补步骤。这包括处理数据集中的缺失值和异常值的策略：

1) 删除包含缺失值的数据记录：这是一种直接而易于操作的策略。但是，如果缺失数据过多，这种方法可能会导致训练样本减少，进而对模型训练效果产生负面影响。

2) 应用数据插值技术：针对连续的时间序列数据，数据插值是填补缺失值的有效途径。举例来说，我们使用线性插值方法，该方法假设数据在时间上按线性规律变化，并根据周围的数据值计算出缺失数据点的值。

3) 使用预测模型补全缺失值：此策略借助其他模型预测缺失的数据值。例如，我们仅使用完整数据记录训练一个简单的模型，然后利用这个模型预测缺失的数据值。

3. 模型建立

3.1. XGBoost 模型

本研究选择了 XGBoost (eXtreme Gradient Boosting)模型作为香蕉价格预测的工具。XGBoost 是一种基于梯度提升决策树的集成学习算法，在工业界和学术界都有很高的应用价值，因为它不仅可以处理各种类型的特征，还可以有效地处理高维稀疏数据和大规模数据，具有强大的预测能力和优秀的泛化性能[1]。

XGBoost 模型在近年来的研究中得到了广泛的应用和认可。研究表明，XGBoost 模型在金融欺诈检测任务中具有更高的准确性[2]。同时，XGBoost 模型在处理大规模数据时具有较低的计算复杂度和较高的训练速度，尤其在医疗数据预测任务[3]，展现了其在大规模数据集上的高效性和可扩展性。除了高预测性能和可扩展性外，XGBoost 模型还具备自动特征选择的能力。通过自动筛选出最重要的特征，XGBoost 模型可以降低模型的复杂性并提高预测性能[4]。

但是，XGBoost 模型也存在一些局限性需要注意。首先，参数调整是确保 XGBoost 模型性能的关键。不正确的参数选择可能导致模型的过拟合或欠拟合。其次，XGBoost 模型对异常值和噪声比较敏感，可能导致模型的性能下降。

综上所述，XGBoost 模型在高预测性能、可扩展性和自动特征选择方面具有显著优势。然而，在实践中选择适当的参数配置、处理异常值以及噪声是确保模型性能和稳定性的关键。基于 XGBoost 模型在价格预测领域的成果以及其高预测性能、可扩展性和自动特征选择能力，本研究选择了 XGBoost 模型作为预测香蕉批发价格的工具[5]，本研究在实践中尝试了多种模型，并发现 XGBoost 模型在预测准确性和泛化能力方面表现出色。

3.2. 模型训练

1) 为了选择最具有预测能力的特征，我们采用了递归特征消除(RFE)算法。通过使用 XGBoost 回归模型作为评估器，我们选择了前 25 个最重要的特征，并得到了这些特征的名称。在选择重要特征的多次训练中，我们发现云南的气象特征在重要性排名中通常排在前 30%名，同时，经过我们把特征的数量从

20 个到 30 个之间多轮计算，最终发现 23 个或 25 个的特征值得到的训练结果相对更好，这些发现为我们下一步研究提供了一些思路。

2) 为了确保特征数据的准确性和可比性，我们进行了特征转换，采用了标准化(StandardScaler)方法来对选定的特征进行标准化处理。接下来，我们将数据集划分为训练集和测试集，其中 80%的数据用于训练模型，剩下的 20%用于评估模型的性能。

3) 在模型的参数优化过程中，我们采用了随机搜索(RandomizedSearchCV)方法。通过定义参数网格，包括 n_estimators (决策树数量)、max_depth (决策树最大深度)、learning_rate (学习率)、min_child_weight (子样本权重)、subsample (样本抽样比例)和 colsample_bytree (特征抽样比例)，我们搜索并选择了最佳参数组合；然后，基于最佳参数组合，我们创建了 XGBoost 回归模型，并使用训练集进行了模型训练。最后，在测试集上进行了预测，并计算了均方误差(MSE)和平均绝对误差(MAE)来评估模型的性能。

经过评估，XGBoost 模型在测试集上的预测性能：均方误差(MSE)：0.4145887542028506，平均绝对误差(MAE)：0.46541164233216915，即真实价格之间的平均绝对差距约为 0.465 元/公斤。这表明我们的模型具有很好的预测能力，并且能够为市场参与者提供价格预测，

4. 结果与讨论

4.1. 比较分析

在模型优化的过程中，我们还比较了 XGBoost 模型与其他机器学习算法(如随机森林、线性回归和决策树)的性能。经过对比分析，我们发现 XGBoost 模型在验证集上的 MAE 值为 0.46，而随机森林回归模型 MAE 值为 0.64，线性回归模型的 MAE 值为 0.71，决策树模型的 MAE 值为 0.73，神经网络算法预测 MSE：0.76。这表明 XGBoost 模型在预测香蕉价格方面具有更高的准确性和泛化能力(见表 1、见附图)。

Table 1. Evaluation results (MAE) of different models on the test set
表 1. 各模型在测试集上的评估结果(MAE)

模型	XGBoost	随机森林	线性回归	决策树	神经网络
MAE	0.46	0.64	0.71	0.73	0.76

4.2. 模型适用性讨论

我们的研究显示，气象条件对香蕉价格的影响非常显著。特别是核心产地的最低温度、平均温度、风速等气象因素对香蕉价格有着重大的影响。这为理解和预测香蕉价格的波动提供了新的视角。

虽然该模型在预测新发地市场的香蕉价格方面有较好的表现，但这个模型主要是基于全国最大的两个批发市场具体数据训练出来的，因此它的适用性主要局限于新发地、广州江南市场的香蕉价格预测。这意味着如果想要将这个模型应用到其他市场或者预测其他水果的价格，可能需要重新收集相关市场或水果的数据，并重新训练模型。尽管如此，本研究为使用机器学习方法预测农产品价格提供了一种有效的框架和方法，只要有相应的数据，我们就可以根据这个框架和方法训练出适应特定市场或特定农产品的预测模型。

5. 展望

5.1. 多市场模型

在未来的研究中，我们设想将预测模型的视角拓宽到多市场的价格预测范畴，考虑引入更为复杂和

微妙的影响因素，如市场所在地的经纬度信息，进一步丰富模型的特征空间。这样的优化，可以使模型有能力预测多个市场的香蕉价格，甚至可能扩展至其他农产品的价格预测。通过引入这种地理经济学的视角，能进一步提升预测模型的性能。

5.2. 更多模型

即便现有模型已经在预测香蕉价格方面体现出了相当优异的表现，我们仍然发现存在一些可能性和机遇，进一步推动和提升模型预测能力的边界。其中之一便是利用更多的、可能更复杂的机器学习模型来进行价格预测。

5.3. 特征优化

特征工程无疑是机器学习建模过程中的核心环节，它涉及对原始数据的预处理、选择最具代表性的特征，以及构造能更好地反映数据模式的新特征。精心设计的特征不仅能够提升模型的预测性能，而且能够使模型更易于理解。因此，在未来的工作中，笔者计划在特征工程方面进行更深入的研究和优化，以期进一步提升模型的预测精度。

6. 结论

本研究成功地运用 XGBoost 模型，结合相关的气象数据，建立了一个预测新发地和广州江南果蔬市场香蕉价格的模型。这一模型的预测精度高，印证了气象条件对香蕉价格波动的显著影响，从而能为市场参与者提供准确的价格预测，有助于他们做出更优的决策。此外，这一模型提供了实质性的价值，因为其使用基于气象数据的预测方式，可以提供可靠的参考信息。未来的研究可以进一步探索其他可能影响香蕉价格的因素，以提高模型的预测能力。

参考文献

[1] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

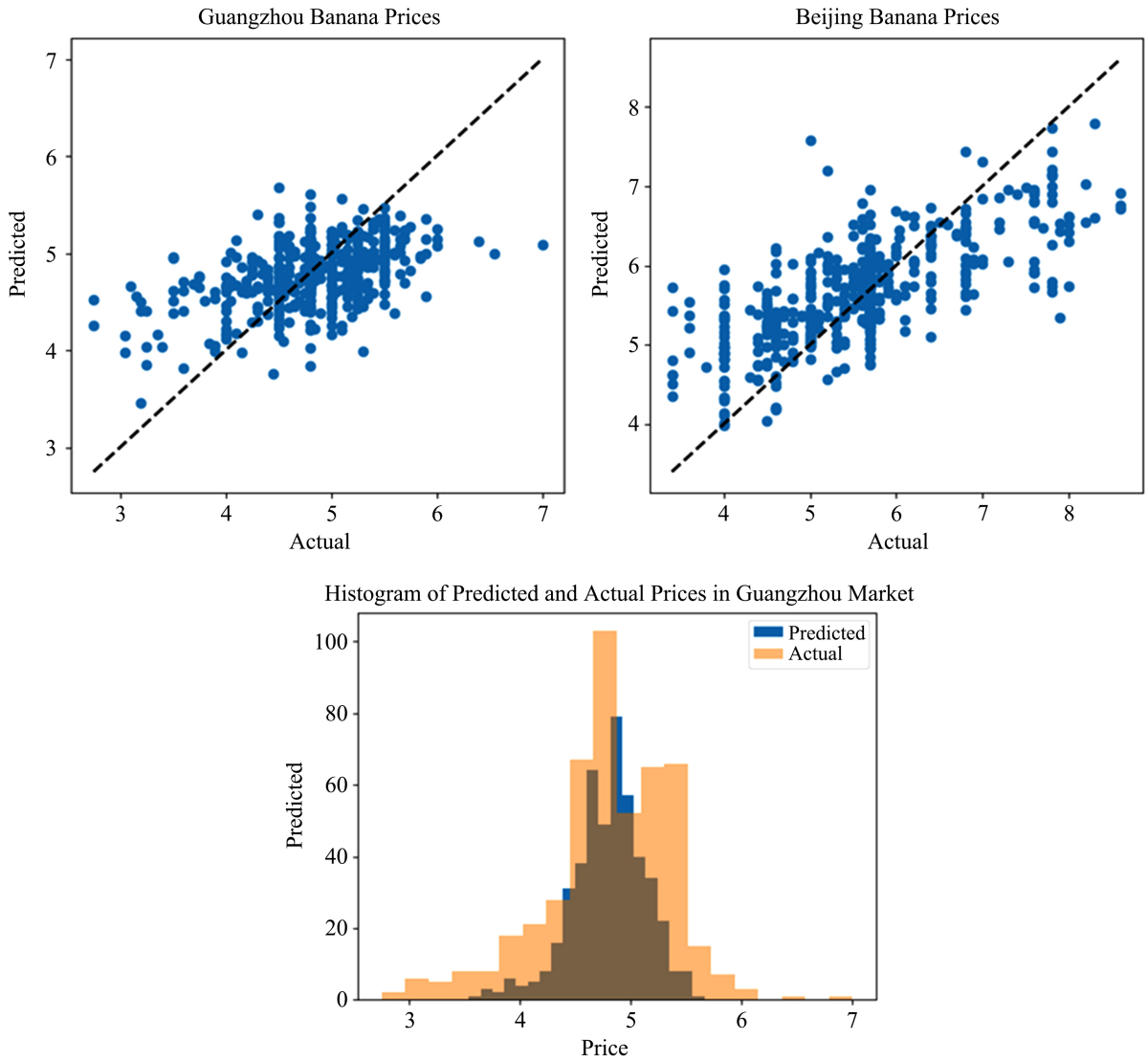
[2] 陈荣荣. 基于 GAN-XGBoost 的信用卡交易欺诈检测模型研究[D]: [硕士学位论文]. 杭州: 杭州师范大学, 2019.

[3] 王铭, 程振豪, 胡苗, 唐铭成, 徐福民, 王莉, 粘永健, 刘凯军. 基于 XGBoost 的 COVID-19 患者重症风险早期预测模型的建立与评价[J]. 陆军军医大学学报, 2022, 44(3): 195-202. https://doi.org/10.16016/j.2097-0927.202107161

[4] 郑列, 穆新宇. 改进的 XGBoost 模型在短租房价格预测中的应用[J]. 湖北工业大学学报, 2021, 36(2): 104-109.

[5] 孙志华, 刘浩. 后疫情时期中国生猪生产预测与展望——基于自回归 XGBoost 时序模型的实证研究[J]. 畜牧与兽医, 2021, 53(12): 140-146.

附　图



Plot. Scatter plot and histogram of predicted and actual values
附图. 预测值与实际值散点图、方直图