

随机森林与XGBoost在股票价格预测中的应用

申 晨, 徐 硕

河北金融学院河北省金融科技应用重点实验室, 河北 保定

收稿日期: 2025年8月24日; 录用日期: 2025年9月22日; 发布日期: 2025年9月28日

摘 要

金融市场具有强随机性、非线性、非平稳性等特征, 对股票价格预测提出了挑战。本文引入支持向量机(SVM)、决策树、随机森林、XGBoost等四种机器学习算法, 通过数据收集、数据预处理、模型训练、结果分析等步骤, 对股票价格进行预测。实验结果表明, 四类模型均能在一定程度上捕捉股价变化趋势; 综合考虑预测效果和稳定性, 随机森林和XGBoost模型更具优势, 为预测股票价格提供了有益参考。

关键词

机器学习, 股票价格, 随机森林, XGBoost

Application of Random Forest and XGBoost in Stock Price Prediction

Chen Shen, Shuo Xu

Hebei Key Laboratory of Financial Technology Application, Hebei Finance University, Baoding Hebei

Received: August 24, 2025; accepted: September 22, 2025; published: September 28, 2025

Abstract

Financial markets are characterized by high randomness, nonlinearity, and non-stationarity, posing significant challenges to stock price prediction. This study introduces four machine learning algorithms—Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost—to predict stock prices through a process encompassing data collection, data preprocessing, model training, and result analysis. Experimental results demonstrate that all four models can, to some extent, capture the trends in stock price movements. When considering both predictive performance and stability, the Random Forest and XGBoost models exhibit superior advantages, offering valuable insights for stock price prediction.

Keywords

Machine Learning, Stock Price, Random Forest, XGBoost

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

股票市场作为经济运行的“晴雨表”，其价格波动受到多元因素交织的影响：从宏观层面看，GDP增速、货币政策调整、通货膨胀率等经济指标会直接对市场预期产生作用；从中观层面看，行业周期的更迭以及产业政策的导向会重塑板块的估值逻辑；从微观层面看，企业的财务表现、重大战略决策、管理层变动等信息会驱动个股价格出现分化。另外，市场参与者的情绪波动、新闻舆情的传播、突发地缘政治事件等非经济因素，会让股价走势更加复杂。因此，股票价格呈现出明显的强随机性、非线性、非平稳性等特征[1][2]，传统基于财务分析或技术指标的预测方法难以有效捕捉复杂的动态规律。

机器学习技术在处理高维数据、识别非线性关系、动态建模等方面具有明显优势，成为突破传统预测方法局限的关键工具。支持向量机[3][4]、决策树、随机森林[5]、XGBoost [6]等算法借助对历史数据的学习，可自动提取隐含特征、构建复杂预测模型，为解析股价波动的潜在模式提供了新途径。

本文首先收集股票历史交易数据，并通过数据清洗、特征工程等步骤提高数据质量，分别采用支持向量机、决策树、随机森林、XGBoost 等四种机器学习算法构建股票价格预测模型，对比分析不同算法在预测股票价格方面的表现差异，最终得到适合于股票价格预测的模型和参数。

2. 数据获取和预处理

2.1. 数据获取

本文采用的股票价格数据主要通过调用 Tushare 股价数据接口完成。以沪深 300 指数为例，通过 Tushare 接口和 pandas 库，可以方便地获取 2020 年 2 月 25 日至 2025 年 3 月 1 日沪深 300 指数的日线数据。获取的数据字段说明和部分数据实例分别如表 1、表 2 所示。

Table 1. Data field description

表 1. 数据字段说明

名称	类型	描述
ts_code	str	Ts 指数代码
trade_date	str	交易日
close	float	收盘点位
open	float	开盘点位
high	float	最高点位
low	float	最低点位
pre_close	float	昨日收盘点
vol	float	成交量

Table 2. Partial data examples**表 2.** 部分数据实例

ts_code	trade_date	close	open	high	low	Pre_close	Vol
000300.SH	20200302	4069.66	4066.30	4126.71	4020.91	4132.84	219,313,559
000300.SH	20200303	4091.36	4070.10	4132.78	4061.13	4123.85	219,369,777
000300.SH	20200304	4115.05	4084.08	4110.53	4067.44	4073.02	182,424,730
000300.SH	20200305	4206.72	3988.73	4031.15	3928.07	4084.88	231,823,115
000300.SH	20200306	4138.50	3968.84	4092.54	3968.84	3940.05	172,475,717
000300.SH	20200309	3997.13	4125.14	4147.65	4067.91	4069.67	208,102,035
000300.SH	20200310	4082.73	4078.50	4115.26	4060.73	4091.36	197,416,402
000300.SH	20200311	4028.42	4150.21	4215.85	4127.04	4115.05	161,656,111
000300.SH	20200312	3950.91	4158.20	4181.31	4135.39	4206.73	157,060,458
000300.SH	20200313	3895.31	4063.08	4063.08	3995.13	4138.51	188,296,722

2.2. 数据预处理

首先对数据进行缺失值处理, 先使用线性插值法填充, 若仍有缺失值则用前向填充。然后针对收盘价和成交量, 分别创建滞后 1 期、3 期和 5 期的新特征列[7], 以捕捉数据的历史依赖关系, 最后删除包含缺失值的行, 以保证数据集的完整性和一致性。最后对特征数据进行标准化处理, 并将数据集按 8:2 的比例划分为训练集和测试集。

3. 模型训练和参数选取

为了有效评估实验的结果, 本文选取在股价预测中常用的四个指标: MSE (均方误差)、RMSE (均方根误差)、MAE (平均绝对误差)、MAPE (平均绝对百分比误差) [8], 具体公式见(1)~(4):

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

3.1. SVM 股价预测模型

本文采用径向基函数作为统一核函数, 系统剖析惩罚函数参数 C 的取值对模型性能的影响。实验中设定 $C = \{0.1, 1.0, 10.0\}$ 作为对比组, 依靠标准化评估指标体系对模型预测效能进行量化分析, 实验结果表明, 当 $C = 1.0$ 时, 模型在测试集上的综合表现优于 $C = 0.1$ 的情况, 体现出对训练数据噪声的合理鲁棒性。虽然 $C = 10.0$ 时的评价指标数值优化, 显示出更强的训练数据拟合能力, 但考虑到惩罚函数取值过大会增加模型过拟合风险, 即可能过度捕捉训练数据中的局部噪声特征而降低泛化能力, 本文权衡模型拟合精度与泛化性能, 最终选定 $C = 1.0$ 作为最优惩罚函数参数, 该选择保证了模型对股价序列趋势特

征的有效学习，又凭借正则化约束避免了过拟合问题，为后续股价预测模型的稳定性提供了可靠支撑。不同惩罚函数参数预测结果如表 3 所示，预测可视化结果如图 1 所示。

Table 3. Comparison of Predictions for different penalty function parameters
表 3. 不同惩罚函数参数预测对比

惩罚函数参数	MSE	RMSE	MAE	MAPE
C = 0.1	319223.78	564.99	510.11	14.46%
C = 1.0	284817.18	533.68	490.84	13.87%
C = 10.0	114312.18	338.10	311.12	8.68%

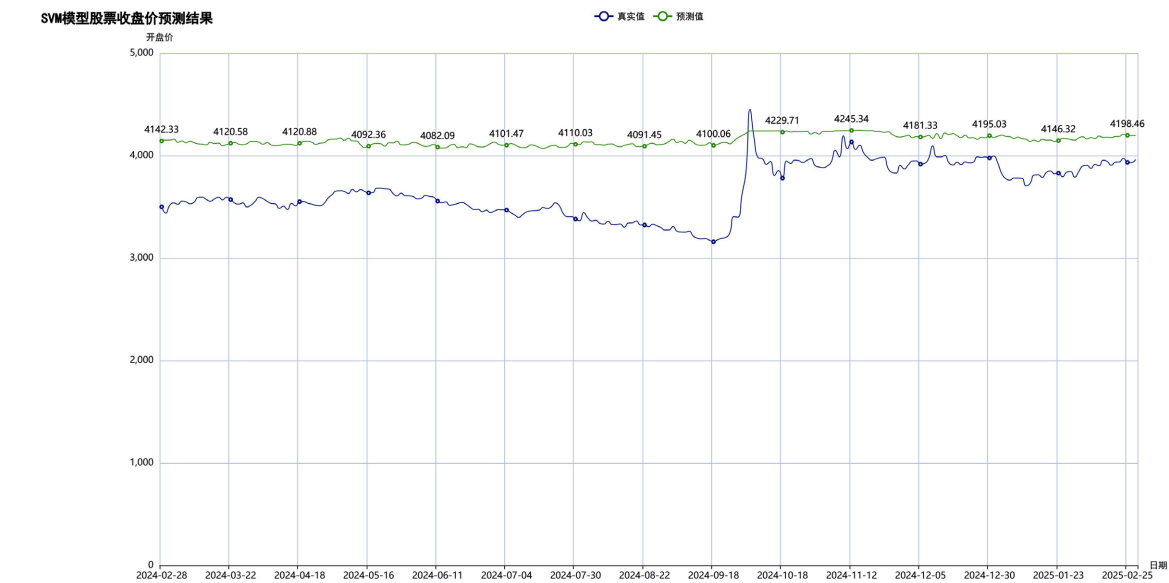


Figure 1. SVM model prediction results
图 1. SVM 模型预测结果

3.2. 决策树股价预测模型

Table 4. Comparison of prediction results with different tree depths
表 4. 不同树深的预测结果对比

树深	MSE	RMSE	MAE	MAPE
1	111761.12	334.30	274.35	7.86%
2	21143.94	145.40	112.54	3.16%
3	9797.65	98.98	80.22	2.23%
4	3255.95	57.06	42.46	1.16%
5	1686.25	41.06	24.33	0.66%
6	1632.80	40.40	21.42	0.57%
7	1576.64	39.70	20.18	0.54%
8	1843.16	42.93	23.21	0.62%
9	1677.96	40.96	19.54	0.52%
10	1643.69	40.54	19.05	0.50%

在决策树模型中，本文通过调整最大树深，分析决策树模型复杂度与泛化能力之间的平衡关系。实验结果显示，当树深从 4 层增加到 5 层时，模型在关键评估指标上呈现出明显优化，这反映出深度的提高让模型可捕捉到股票序列中更为复杂的非线性依赖关系，提升预测精度。当树深超过 5 层后，随着模型复杂度提升，测试集指标趋于稳定甚至出现微小退化。这说明过度增加树深可能致使模型学习到训练数据中的噪声或者局部异常模式，引发过拟合风险。综合考虑模型预测性能与复杂性，本文选取树深为 5 作为最优参数，保证模型可充分挖掘股价数据的潜在特征结构，又避免了过拟合问题。前十层树深的预测结果如表 4 所示，预测可视化结果如图 2 所示。

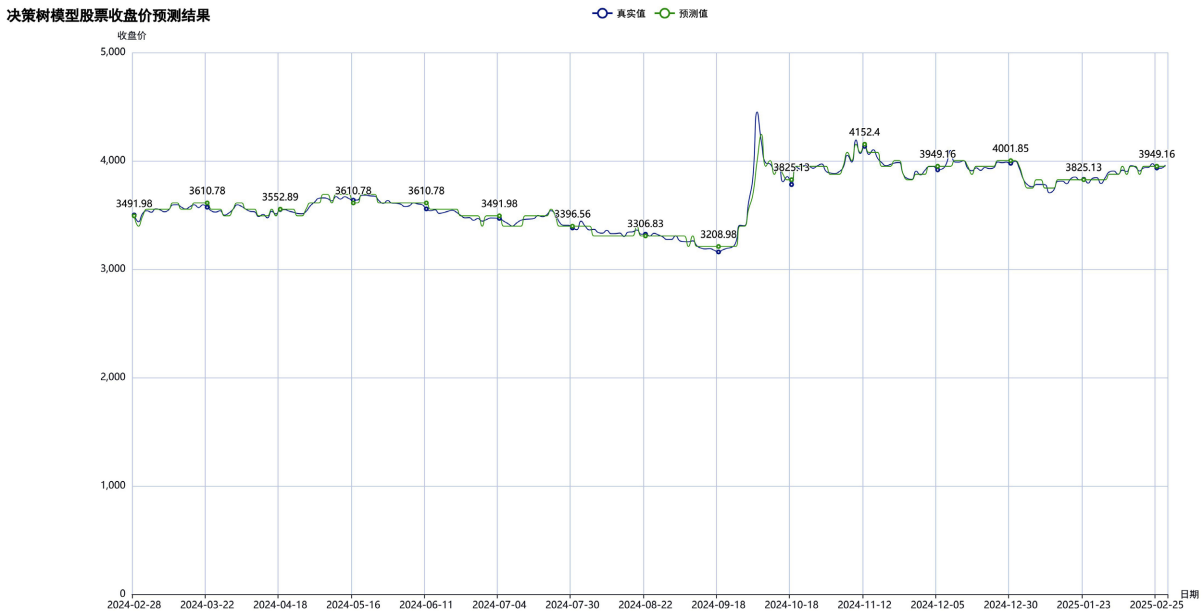


Figure 2. Decision tree model prediction results
图 2. 决策树模型预测结果

3.3. 随机森林股价预测模型

在随机森林模型参数选取过程中，同样选择固定树深为 5，选取树的数量分别为 50、100 和 200 进行实验评估。结果表明，当树的数量从 50 增加到 100 时，模型性能明显增强；进一步将树的数量提升至 200 时，尽管各项指标仍有细微优化，但提升幅度极为有限。考虑到计算资源消耗和模型复杂度，选择 100 作为树的数量参数进行后续预测，既能保证模型具备较高的预测精度，又能在性能提升与资源成本间达到较好的平衡，避免计算效率下降，减少过拟合风险。不同数量的树参数评估指标结果如表 5 所示，预测可视化结果如图 3 所示。

Table 5. Comparison of prediction results with different numbers of trees
表 5. 不同数量的树预测结果对比

树数量	MSE	RMSE	MAE	MAPE
50	1373.80	37.06	17.75	0.478%
100	1350.38	36.74	17.27	0.464%
200	1347.30	36.70	17.05	0.458%

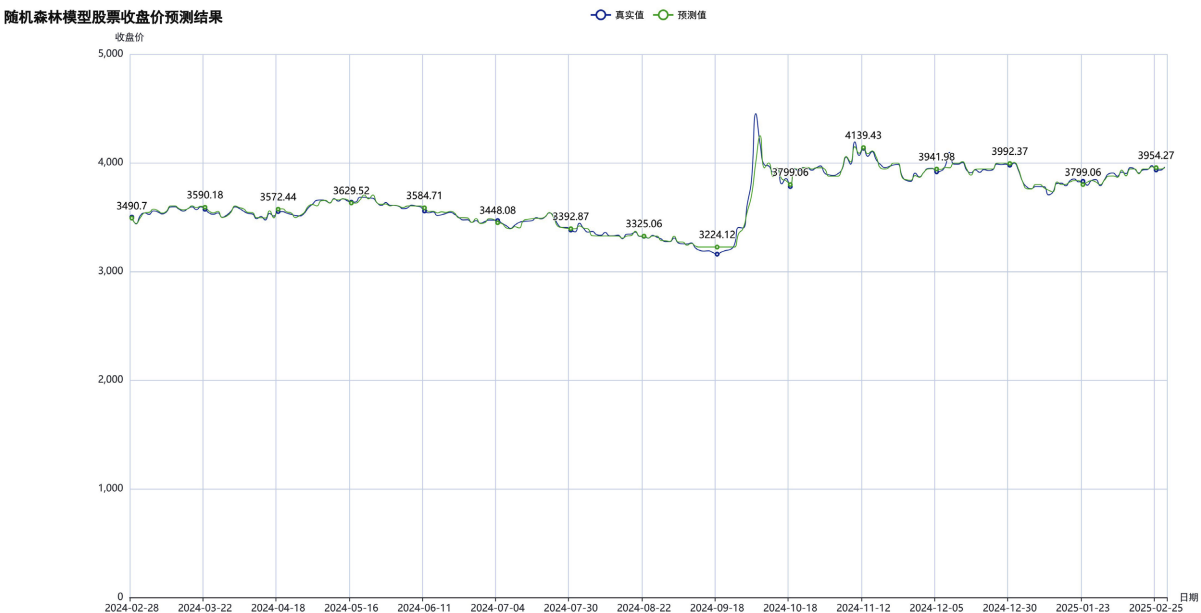


Figure 3. Random forest model prediction results
图 3. 随机森林模型预测结果

3.4. XGBoost 股价预测模型

在本次 XGBoost 模型参数优化过程中，针对最大深度(3, 5, 7)和学习率(0.01, 0.1, 0.2)进行交叉验证，当最大深度为 3 时，学习率 0.01 下模型欠拟合严重，0.1 和 0.2 时性能显著提升；最大深度为 5 时，学习率 0.1 对应的 MSE (1467.85)、RMSE (38.31)、MAE (19.59)和 MAPE (0.53%)均为该深度下最优；最大深度为 7 时，尽管学习率调整能改善性能，但整体指标不如最大深度为 5 的部分组合。综合考量，选择最大深度为 5、学习率为 0.1 作为后续实验参数，该组合在避免过拟合与欠拟合之间取得平衡，各项误差指标表现最佳，有助于提升股价预测的精准度与模型稳定性。结果如表 6 所示，预测可视化结果如图 4 所示。

Table 6. XGBoost parameter cross-validation results
表 6. XGBoost 参数交叉评估结果

最大深度	学习率	MSE	RMSE	MAE	MAPE
3	0.01	78204.99	279.65	259.55	7.32%
3	0.1	1554.85	39.43	16.79	0.45%
3	0.2	1551.75	39.39	20.50	0.55%
5	0.01	76281.71	276.19	257.06	7.24%
5	0.1	1537.76	39.21	17.87	0.47%
5	0.2	1467.85	38.31	19.59	0.53%
7	0.01	76428.19	276.45	257.29	7.25%
7	0.1	1604.46	40.05	18.75	0.50%
7	0.2	1606.30	40.07	20.66	0.55%



Figure 4. XGBoost model prediction results
图 4. XGBoost 模型预测结果

4. 结果分析

本文采用的四种机器学习算法经过训练,生成各自的评估指标,从评估指标来看,SVM 模型的 MSE、RMSE、MAE 和 MAPE 均高于其他三种算法,预测效果较差,难以有效捕捉股价数据特征。决策树模型虽优于 SVM,但 MAE 和 MAPE 仍偏大,预测精度有待提高。随机森林模型的各项指标在四者中最优,借助集成策略有效降低方差,呈现出良好的预测稳定性与准确性。XGBoost 模型表现也较为出色,与随机森林接近。综合考虑预测效果与稳定性,随机森林和 XGBoost 在误差控制与拟合能力上更具优势[9][10],能为股价预测提供相对可靠的结果。

通过进一步分析,在股票价格预测场景中,SVM、决策树、随机森林与 XGBoost 的表现差异可能主要源于算法机制与金融数据特性的适配性。SVM 擅长高维数据处理,但对参数敏感且难以直接适应股票数据的非平稳时序特性,在市场剧烈波动时泛化能力受限;决策树善于捕捉非线性关系,但单树结构易受市场噪声干扰导致过拟合,且对时序数据波动敏感;随机森林通过集成多棵决策树显著提升了结果的稳定性;XGBoost 基于梯度提升框架,通过串行修正误差、正则化剪枝及并行计算,能高效处理高维金融数据(如多因子特征),同时抑制市场噪声,提升对复杂时序关系的拟合能力。

此外,在股票价格预测场景下,可以考虑采用集成方法进一步提升预测能力。集成方法可通过融合不同模型的互补信息进一步降低预测误差,同时也增加了计算复杂度、调参难度,在高噪声、非平稳的金融数据中的过拟合风险与可解释性损失亦不容忽视,需综合权衡利弊。

基金项目

2025 年度河北省金融科技应用重点实验室课题:融合时序特征与监管规则量化的证券市场异常交易监测研究(课题编号:2025010)。

参考文献

[1] 覃业梅,周帆,胡博颀,等. 基于样本卷积交互网络的股票价格预测及资产分配[J/OL]. 系统科学与数学: 1-21.

- <https://link.cnki.net/urlid/11.2019.o1.20250708.1517.060>, 2025-08-23.
- [2] 林娜娜, 秦江涛. 基于随机森林的 A 股股票涨跌预测研究[J]. 上海理工大学学报, 2018, 40(3): 267-273, 301.
- [3] 杨新斌, 黄晓娟. 基于支持向量机的股票价格预测研究[J]. 计算机仿真, 2010, 27(9): 302-305.
- [4] 丁宇昕. 基于 SVM 的股价预测模型研究[D]: [硕士学位论文]. 太原: 山西大学, 2021.
- [5] 邹婕, 李路. 基于随机森林的 SA-BiGRU 模型的股票价格预测研究[J]. 中国物价, 2023(11): 52-56.
- [6] 谷嘉炜, 韦慧. XGBoost-ESN 组合模型股价预测方法[J]. 牡丹江师范学院学报(自然科学版), 2022(1): 1-5.
- [7] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.
- [8] Hyndman, R.J. and Koehler, A.B. (2006) Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, **22**, 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [9] 郑洁, 颜七笙. 基于多特征量分析 XGBoost-EMD-LSTM 模型的股票价格预测[J]. 应用数学进展, 2025, 14(4): 558-571.
- [10] Basak, S., Kar, S., Saha, S., Khaidem, L. and Dey, S.R. (2019) Predicting the Direction of Stock Market Prices Using Tree-Based Classifiers. *The North American Journal of Economics and Finance*, **47**, 552-567. <https://doi.org/10.1016/j.najef.2018.06.013>