

基于多特征量分析XGBoost-EMD-LSTM模型的股票价格预测

郑 洁, 颜七笙*

东华理工大学理学院, 江西 南昌

收稿日期: 2025年3月18日; 录用日期: 2025年4月11日; 发布日期: 2025年4月21日

摘 要

高效且准确地预测股票价格一直是研究者们不断追求的目标, 由于股票价格呈现出明显的非线性、非平稳性及时序性的特征, 提升股价预测的准确性成为了研究的重点。提出一种基于多特征量分析极端梯度提升(eXtreme gradient boosting, XGBoost)模型以及基于经验模态分解(Empirical Mode Decomposition, EMD)和长短期记忆(long short-term memory, LSTM)模型的组合预测方法对股票收盘价进行预测。首先, 建立XGBoost模型对股票收盘价进行预测, 并将预测结果作为新变量添加到原始数据的特征中, 形成新的多元时间序列的数据; 其次, 用EMD算法分解新数据得到有限个本征函数; 最后, 建立LSTM模型并建模, 得到最后的预测结果。通过仿真实验结果表明, 本文的模型相比于RNN模型、单特征LSTM模型、多特征LSTM模型、XGBoost模型、EMD-LSTM模型、EMD-RNN模型, 所提出模型的预测精度更高、效果更好。说明提出的方法能够有效提高预测准确性, 具有一定的参考价值, 对于投资者的决策具有积极的指导意义。

关键词

股票价格预测, XGBoost模型, 经验模态分解, 长短期记忆网络

Stock Price Prediction Based on Multi-Feature Analysis Using the XGBoost-EMD-LSTM

Jie Zheng, Qisheng Yan*

School of Science, East China University of Technology, Nanchang Jiangxi

Received: Mar. 18th, 2025; accepted: Apr. 11th, 2025; published: Apr. 21st, 2025

*通讯作者。

文章引用: 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. DOI: 10.12677/aam.2025.144186

Abstract

Efficient and accurate stock price prediction has been a continuous pursuit for researchers. Due to the significant nonlinearity, non-stationarity, and time-series characteristics of stock prices, improving prediction accuracy has become a key research focus. This paper proposes a combined forecasting method for stock closing prices based on multi-feature analysis using the eXtreme Gradient Boosting (XGBoost) model and a hybrid model incorporating Empirical Mode Decomposition (EMD) and Long Short-Term Memory (LSTM). First, an XGBoost model is constructed to predict stock closing prices, and the prediction results are added as new variables to the original dataset, forming a new multivariate time series. Next, the EMD algorithm is applied to decompose the new data into a finite number of intrinsic mode functions. Finally, an LSTM model is built and trained to obtain the final prediction results. Stimulation results demonstrate that the proposed model outperforms the RNN model, single-feature LSTM model, multi-feature LSTM model, XGBoost model, EMD-LSTM model, and EMD-RNN model in terms of prediction accuracy and overall effectiveness. This indicates that the proposed method can significantly enhance prediction accuracy, providing valuable reference insights and offering positive guidance for investors' decision-making.

Keywords

Stock Price Prediction, XGBoost Model, Empirical Mode Decomposition (EMD), Long Short-Term Memory (LSTM)

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

股市的价格波动受到多种经济和社会因素的影响,在经济市场飞速发展的背景下,人均收入和大众投资理财意识不断提高,近年来股票作为一种投资理财工具逐渐被人们所关注。股票作为金融市场重要的一部分,不仅是我国经济发展强大的推动力,也是反映经济发展状况不可或缺的指标。

现阶段股票价格预测方法可分为三大类:传统时间序列模型、机器学习算法与深度学习算法模型以及组合预测模型。

传统时间序列模型在理论上具有较强的解释性,股票价格作为金融时间序列的代表,常常采用多元回归模型、ARIMA 模型以及 GARCH 模型等[1]-[5]方法来预测,尽管这些传统时间序列方法的股票预测模型在结构上较为简单,且计算的复杂度也相对较低,但在处理非线性和波动性较强的股票数据时,其拟合效果并不理想。

由于股票数据更新迅速且频繁以及数据量过大,因此这对分析其股票数据的方法以及其计算能力的要求也会越高。于是,基于海量数据的机器学习算法,如决策树、支持向量机、集成学习、神经网络、随机森林等算法[6]-[11]得到了不断发展与更新,并且这些机器学习算法在股票预测领域应用的范围更加广泛。黄卿等[12]人通过应用 BP 神经网络、支持向量机(SVM)和 XGBoost 模型对沪深 300 指数的数据进行预测,结果表明, XGBoost 的预测效果最好。王燕等[13]则利用网格搜索算法对 XGBoost 模型进行寻优,发现通过优化后的 XGBoost 模型在面对短期股票数据时,其拟合性能有了显著的提升。

近年来,基于神经网络的深度学习算法以强大的计算能力以及卓越的准确性,在股票市场的预测中

应用广泛。其中, 长短期记忆神经网络 LSTM 被证明是表现最佳的神经网络之一。最初, Elman 提出了循环神经网络(RNN), RNN 为处理时间序列数据拓宽了新的视野。随后, Hochreiter 等人[14]提出了长短期记忆网络, 该模型是在 RNN 模型的基础上进行了改进, 并且针对 RNN 在训练过程中遇到的梯度爆炸和梯度消失的问题, LSTM 模型通过引入时序性和选择性记忆的机制成功解决这一问题。因此, LSTM 模型在股票预测等领域展现出明显的优势, 成为深度学习应用中的一个重要方向。神经网络方法, 如卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)、长短期记忆网络(Long Short-Term Memory, LSTM)等[15] [16], 不仅能够处理复杂的时间序列数据, 而且预测的准确性也比较高。韩莹等[17]通过提出一种宽度学习与深度学习相结合的一个新型预测模型, 通实验结果表明该模型在多个评价指标上都有显著地提升。尽管 LSTM 在处理长依赖问题上具有显著优势, 但在面对长序列数据时, 仍然存在着信息丢失的挑战。

因此, 学者们提出了组合预测模型来提高股票价格预测的准确性。朱菊香等[18]利用卷积神经网络(CNN)提取数据特征, 并将其输入到 LSTM 网络中, 同时使用 CEEMDAN 分解算法对 PM2.5 序列进行分解, 其结果表明, 该模型预测 PM2.5 的浓度的准确性较高。大量的实验表明, 合理选择并组合不同模型能够显著提升预测的精度和准确性[19]-[21]。张彪等[22]通过 ELM-EMD-LSTM 组合模型对船舶运动姿态进行预测, 并且将组合模型与 LSTM 模型、ELM-LSTM 模型进行对比实验, 结果表明组合模型的预测精度更高, 且平均绝对误差与均方根误差最小。Ren [23]将 XGBoost 模型与 CNN 模型组合在一起, 并应用组合后的模型对图像进行分类, 结果表明该模型在分类任务中表现最佳。此外, 郭长东[24]首先利用 SVM 对股票进行预测, 并将预测的结果输入到 XGBoost 模型中, 最终形成一个 SVM-XGBoost 的组合预测模型, 结果表明该混合模型的预测精度比单一模型的预测精度更优。

通过对以上文献的研究可以发现, 将不同的模型组合在一起的预测方法通常比单一模型的效果更好。这是因为通过不同模型的组合, 可以利用各个模型的优势进而提高整个模型的预测性能。因此, 组合模型对于提高模型预测的准确性和鲁棒性方面均有明显的效果。为了进一步提高股票价格预测精度, 本文提出一种 XGBoost-EMD-LSTM 混合模型。此模型结合 XGBoost、EMD 和 LSTM 各自的优点长处对股票价格进行预测。本文利用中国石油股票和中国石化股票历史数据进行预测实验, 通过仿真实验结果表明, 与其他模型相比, 本文建立的模型在股票预测方面有着较高的预测精度且评价指标最优, 验证了本文混合模型的有效性。

2. 研究方法

2.1. XGBoost 模型

XGBoost (Extreme Gradient Boosting)算法是 2016 年陈天奇博士[25]提出来的一种新的算法, 又被成为极端梯度提升算法。作为一种提升方法, XGBoost 的主要核心思想是通过把一组弱分类器结合起来, 然后建立一个强分类器。XGBoost 是通过多棵决策树的预测结果进行叠加的一种模型, 即由多个基模型组成的集成模型。在每一次的迭代中, 训练的树模型都会对当前的误差进行修正, 假设第 t 次迭代要训练的树模型为

$$\hat{y}_i^{(t)} = \sum_{k=1}^K f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (1)$$

(1)式中: $\hat{y}_i^{(t)}$ 为第 t 次迭代后样本 i 的预测值, $\hat{y}_i^{(t-1)}$ 为前 $t-1$ 棵树的预测值, x_i 表示第 i 个点的特征向量, K 表示决策树的数量。 $f_t(x_i)$ 为第 t 棵树的模型

XGBoost 模型的目标函数分为两部分构成, 一个是损失函数、另一个是模型的复杂度。具体的表达式可以用以下方式表达:

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (3)$$

(2)式和(3)式中: $L(\Theta)$ 和 $\Omega(\Theta)$ 分别表示模型的损失函数和模型复杂度; T 表示叶子节点数; ω_j 代表节点的权重; γ 和 λ 表示两者的惩罚系数。

2.2. EMD 经验模态分解

经验模态分解(Empirical Mode Decomposition, EMD) [26]方法是由黄锬与其他学者于 1998 年提出的一种里程碑式的信号处理方法。

EMD 具体分解步骤如下:

步骤 1: 设原始序列为 $x(t)$, 取出 $x(t)$ 的全部极大值和极小值;

步骤 2: 分别通过三次样条插值函数拟合 $x(t)$ 的上下包络线, 记为 $x_{\max}(t)$ 和 $x_{\min}(t)$;

步骤 3: 计算上下包络线的均值 $m_1(t)$, 即

$$m_1(t) = [x_{\max}(t) + x_{\min}(t)]/2 \quad (4)$$

步骤 4: 令 $h_1(t) = x(t) - m_1(t)$, 如 $h_1(t)$ 若满足本征模态函数的两个条件, 则得到一条本征模态 IMF, 否则, 令 $x(t) = h_1(t)$, 将以上步骤重复进行;

步骤 5: 根据以上步骤进行运算, 可以逐步分解出第 1, 2, ..., n 个本征模态量和一个残差 $r(t)$ 。

此时, 原始序列 $x(t)$ 的分解完毕, 可以表示为

$$x(t) = \sum_{i=1}^n IMF_i(t) + r(t) \quad (5)$$

对于同一模态函数中又包含了很多周期特征。

EMD 分解流程图如图 1 所示:

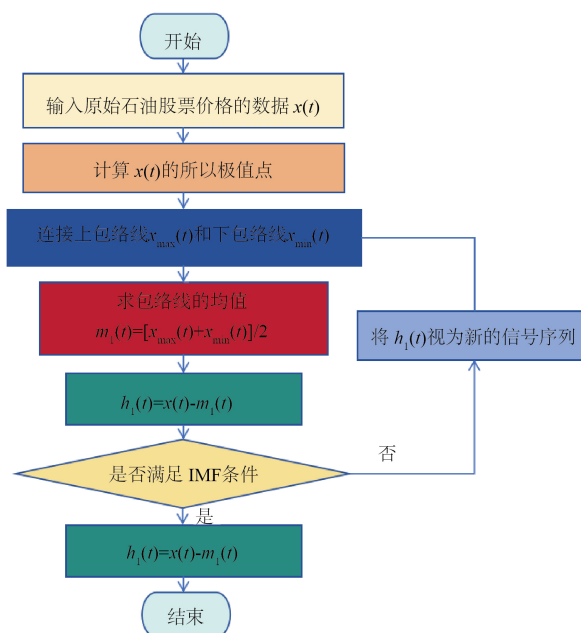


Figure 1. EMD decomposition flowchart
图 1. EMD 分解流程图

2.3. LSTM 长短期记忆神经网络模型

1997 年 Schmidhuber 等人[14]为了解决 RNN 递归神经网络模型当中存在的梯度消失、梯度爆炸问题以及长期依赖问题, 提出了更优的 LSTM 长短期记忆神经网络模型。LSTM 模型结构图如图 2 所示:

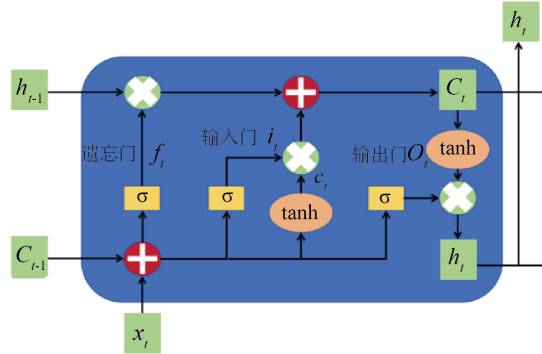


Figure 2. LSTM model structure diagram
图 2. LSTM 模型结构图

从图 2 中可以看出, LSTM 模型在细胞单元主要由三个不同的门构成: 遗忘门、输入门、输出门。其中, C_{t-1} 为 LSTM 上一时刻细胞层的细胞状态, h_{t-1} 为 LSTM 上一时刻细胞层的输出, 3 个 σ 是 sigmoid 激活函数, $\tanh$ 层是为了创建新的细胞状态, 它为双曲正切激活函数。

处理步骤:

遗忘门: 主要是把输入数据 x_t 和上一状态存留下来的 h_{t-1} 整合在一起形成 $[h_{t-1}, x_t]$ 作为遗忘门的输入, 之后经过 sigmoid 激活函数处理后得到新的权值向量 f_t , 其中 f_t 的表达式如下:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

式中: W_f 为遗忘门的权重, b_f 为遗忘门的偏差, σ 为激活函数。

输入门: 主要是把输入数据 x_t 和上一状态存留下来的 h_{t-1} 整合在一起形成 $[h_{t-1}, x_t]$ 作为输入门的输入, 经过 $\tanh$ 函数层产生了新的权值向量, 然后与 C_{t-1} 进行合并更新, i_t 为控制向量, 更新后的公式为:

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (7)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (9)$$

式中: $\tilde{C}_t$ 为通过 $\tanh$ 激活函数创建的新的候选向量, C_t 表示 t 时刻的细胞状态, W_i , W_c 为输入门的权重, b_i , b_c 为输入门的偏差。

输出门: 把输入数据 x_t 和上一状态存留下来的 h_{t-1} 整合在一起形成 $[h_{t-1}, x_t]$ 作为输出门的输入, 经过遗忘门、输入门将 C_{t-1} 更新为 C_t 后, 将 C_t 经过 $\tanh$ 函数层处理后将有效信息提取出来, 然后通过计算 $\tanh$ 层得到的结果与 sigmoid 激活函数层得到的结果就会得到最终的输出, 其表达式为:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t * \tanh(C_t) \quad (11)$$

式中: W_o 为输出门的权重, b_o 为输出门的偏差。

3. 实证分析

3.1. 数据分析

本文选取中国石油股票 2012 年 2 月 22 日至 2024 年 6 月 1 日的历史数据作为研究对象来进行分析预测，共有 2982 条数据，其中包括收盘价、开盘价、最高价、最低价、成交量、次日收盘价、涨跌幅等多个数据特征，部分数据如表 1 所示。

Table 1. Partial raw data
表 1. 部分原始数据

序列	收盘价	最高价	最低价	开盘价	成交量	次日收盘价
1	10.27	10.26	10.28	10.2	12.46 M	10.31
2	10.31	10.27	10.38	10.23	21.43 M	10.43
...	...	...	...	...	...	...
2981	10.2	10.36	10.53	10.16	173.35 M	10.14
2982	10.14	10.11	10.22	10.08	135.39 M	10.16

为了更好地评估模型的预测性能，现将本文的数据集划分为测试集和训练集，其中训练集是用来训练模型以及数据的拟合，而测试集是用来检验模型的泛化能力。具体将本文数据划分为 8:2 的比例，其中将 80% 的数据用于构建训练集，剩余 20% 的数据用于构建测试集。

3.2. XGBoost-EMD-LSTM 模型预测

图 3 详细展示了 XGBoost-EMD-LSTM 模型的算法流程及其框架结构。具体步骤如下：

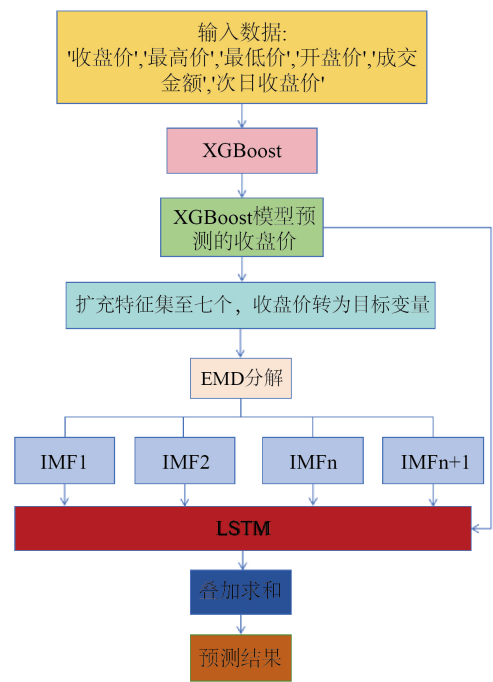


Figure 3. XGBoost-EMD-LSTM algorithm flowchart
图 3. XGBoost-EMD-LSTM 算法流程图

步骤 1: XGBoost 模型训练与预测。输入数据, 将数据分为训练集和测试集, 创建 XGBoost 模型来预测收盘价。

步骤 2: 数据预处理。在获得 XGBoost 模型对收盘价的预测值之后, 这些预测值被作为新的输入特征纳入模型。具体而言, 将 XGBoost 模型生成的收盘价预测值与原有的六个特征(收盘价、开盘价、最高价、最低价、成交量、次日收盘价)相结合, 扩充特征集至七个, 包括: 收盘价、开盘价、最高价、最低价、成交量、次日收盘价以及 XGBoost 预测收盘价, 形成新的多元时间序列的数据。

步骤 3: 数据分解。为了能够更好地提升预测精度, 采用 EMD 算法将输入的数据进行分解, 即分解为各个不同频率的样本模态分量。采用 min-max 归一法对输入特征集数据归一化处理, 将数值缩放在 0-1 之间, 计算公式如下所示:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (12)$$

其中: x' 为归一化后的数据; x 为原始数据; $x_{\max}$ 为极大值, $x_{\min}$ 为极小值。

步骤 4: LSTM 神经网络训练。针对各个子序列分量分别构建 LSTM 神经网络预测模型, 然后叠加在一起求和, 针对该模型使用 Adam 优化算法对 LSTM 网络进行训练, 调整模型参数为最小化损失函数;

步骤 5: 预测结果计算, 完成预测。

通过这一流程, XGBoost-EMD-LSTM 模型能够有效地整合传统机器学习算法与深度学习技术的优势, 实现对股票价格变动的高精度预测。

将 EMD 算法应用于收盘价数据中, 将其进行分解后得到的各个分量如图 4 所示。

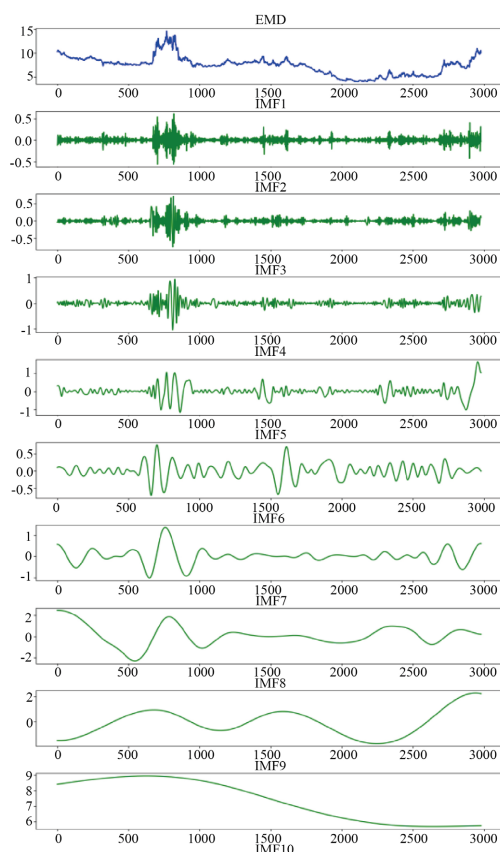


Figure 4. EMD decomposition diagram
图 4. EMD 分解图

对于其中的参数设置, XGBoost-EMD-LSTM 模型首先建立 XGBoost 的模型中, 将决策树的数量 (n_estimators) 设置为 200; 随机种子(random_state)设置为 420; 并将学习力(learning-rate)设置为 0.01。然后建立 LSTM 模型, 时间滑动窗口的宽度为 60, 用来预测未来一天的收盘价; 隐藏层的单元数量设置为 50 个; 采用 tanh 激活函数; 模型的优化器选择的是 Adma 算法、损失函数选择的是均方误差 MSE。根据设备的性能和模型的特征, 将训练次数(epochs)这一参数调整为 300, 批处理大小(batch-size)调整为 32, 其余的均为默认参数。根据这些参数设置来训练模型, 得到 XGBoost-EMD-LSTM 模型的可视化图如图 5 所示。

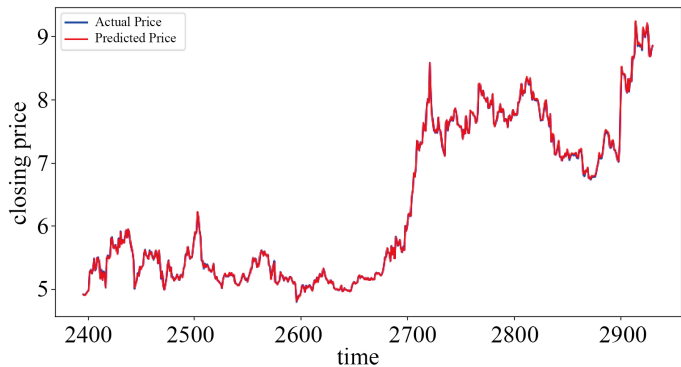


Figure 5. XGBoost-EMD-LSTM model stock price prediction curve
图 5. XGBoost-EMD-LSTM 模型股票价格预测曲线图

图 5 的分析结果揭示了本研究所提出的混合模型在预测性能上的显著优势。与单一的单特征 LSTM 模型、多特征 LSTM 模型以及 XGBoost 模型相比, 混合模型的预测值与实际值更为接近, 且能更准确地捕捉股价的波动趋势。特别是在原始数据出现较大波动时, 混合模型能够更精确地刻画数据的波动特性, 并始终保持与原始数据一致的趋势, 展现出最佳的跟随性和波动性。

为了全面验证 XGBoost-EMD-LSTM 模型在预测任务中的有效性, 本研究采用了损失曲线图作为分析工具。损失曲线图能够直观地反映模型在训练过程中的性能变化, 是评估模型泛化能力和调整超参数的重要依据。

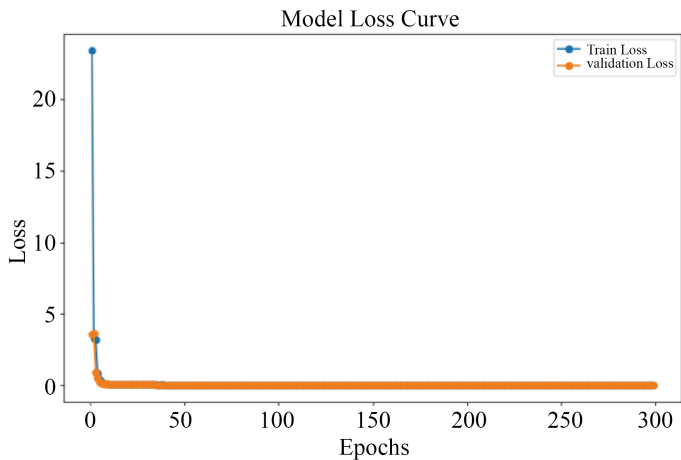


Figure 6. Loss curve
图 6. 损失曲线图

由图 6 所示, 整体来看, 模型的损失函数曲线表现出较为理想的趋势。损失值在训练过程中迅速下降, 并且训练损失和验证损失之间没有明显的差距, 显示了模型的良好泛化能力。模型的收敛速度较快且稳定, 训练过程没有过拟合, 最终损失值接近最优水平, 表明模型在所给任务上效果优秀。

3.3. 模型评价指标

本文选取的评价指标为平均绝对误差(MAE)、均方误差(MSE)、均方根误差(RMSE)、决定系数(R^2), 数学表达式如下,

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - y'_i| \tag{13}$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2 \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2} \tag{15}$$

$$R^2 = \frac{\left[\sum_{i=1}^N (y_i - \bar{y})(y'_i - \bar{y}') \right]^2}{\left[\sum_{i=1}^N (y_i - \bar{y})^2 \right] \left[\sum_{i=1}^N (y'_i - \bar{y}')^2 \right]} \tag{16}$$

式子中: N 为样本个数; y_i 为真实值; y'_i 为预测值; $\bar{y}$ 为平均真实值; $\bar{y}'$ 为平均预测值。 MAE 、 MSE 、 $RMSE$ 值越低, 表明模型的预测精度较高。 R^2 值越接近 1, 说明模型对数据的拟合性越高, 预测的准确性就越高。

为了进一步验证所提模型的有效性, 通过比较 RNN 模型、单特征 LSTM 模型、多特征 LSTM 模型、XGBoost 模型、LSTM-Transformer 模型、RNN-Transformer 模型、EMD-RNN 模型、EMD-LSTM 模型、XGBoost-EMD-RNN 模型以及本文研究的 XGBoost-EMD-LSTM 模型的 MSE 、 $RMSE$ 、 MAE 、 R^2 , 各模型的预测结果如表 2 所示。

Table 2. Evaluation metrics of different models
表 2. 各模型评价指标

模型	MSE	RMSE	MAE	R^2
RNN	0.12540	0.35412	0.28673	0.95022
单特征 LSTM	0.08011	0.28303	0.20437	0.96820
多特征 LSTM	0.03255	0.18042	0.16591	0.98707
XGBoost	0.06132	0.24763	0.21189	0.97566
LSTM-Transformer	0.00130	0.03606	0.02838	0.94181
RNN-Transformer	0.00164	0.04052	0.03419	0.92652
EMD-RNN	0.05313	0.23051	0.22377	0.97891
EMD-LSTM	0.02537	0.15927	0.15677	0.98992
XGBoost-EMD-RNN	0.01176	0.10847	0.07808	0.99235
XGBoost-EMD-LSTM	0.00009	0.00952	0.00720	0.99994

对比表 2 的结果可以看出, XGBoost-EMD-LSTM 模型的 MSE 和 $RMSE$ 比 RNN 模型分别降低了 12.531%、34.46%; 比单特征 LSTM 模型分别降低了 8.002%、27.351%; 比多特征 LSTM 模型分别降低了 3.246%、17.09%; 比 XGBoost 模型分别降低了 6.123%、23.811%; 比 EMD-RNN 模型分别降低了 5.304%、22.099%; 比 LSTM-Transformer 模型和 RNN-Transformer 模型分别降低了 0.121%、2.645%和 0.155%、3.1%; 比 EMD-LSTM 模型分别降低了 2.528%、14.975%, 比 XGBoost-EMD-RNN 模型分别降低了 1.167%、9.895%; XGBoost-EMD-LSTM 模型的 MAE 比 RNN 模型、单特征 LSTM 模型、多特征 LSTM 模型、XGBoost 模型、LSTM-Transformer 模型、RNN-Transformer 模型、EMD-RNN 模型、EMD-LSTM、XGBoost-EMD-RNN 模型分别降低了 27.953%、19.717%、15.871%、21.657%、2.118%、2.699%、20.469%、14.957%、7.088%; XGBoost-EMD-LSTM 模型的 R^2 比 RNN 模型、单特征 LSTM 模型、多特征 LSTM 模型、XGBoost 模型、LSTM-Transformer 模型、RNN-Transformer 模型、EMD-RNN 模型、EMD-LSTM、XGBoost-EMD-RNN 模型分别增加了 4.972%、3.174%、1.287%、2.428%、5.813%、7.342%、2.103%、1.002%、0.759%。

根据误差值和 R^2 来看, 本文使用的模型误差最小, R^2 值最高, 模型的精度相较于其他的几种模型有很大提升。为了直观地展示不同模型之间的预测差异, 本文进行了对比分析, 将这十种模型的预测结果绘制成曲线图, 如图 7 所示。

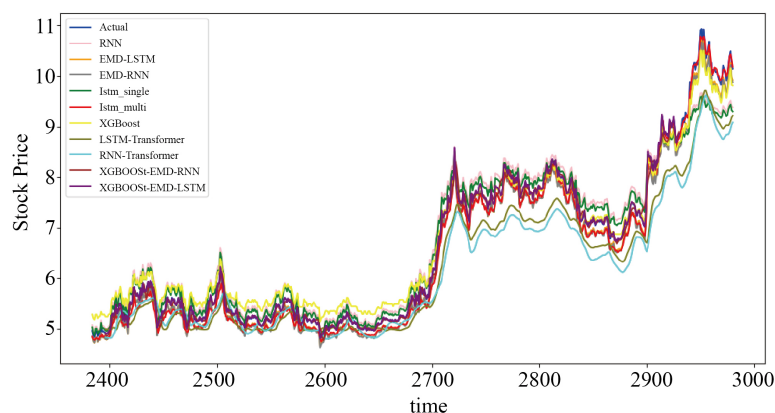


Figure 7. Comparison of different models
图 7. 各模型对比图

图 7 中展示了各模型的预测曲线与真实股价的对比。从图中可以看出, XGBoost-EMD-LSTM 模型(用紫色曲线代表)在整体趋势和价格波动的捕捉上, 表现出了优异的拟合效果。特别是在股票价格的剧烈波动期间, 该模型的预测曲线紧随真实价格波动, 而其他模型均有不同程度上的偏差。进一步的定量分析显示, 本文所研究的模型的 MSE 、 $RMSE$ 和 MAE 分别为 0.00009、0.00952 和 0.0072, 显著低于其他模型, 且 R^2 值达到了 0.99994, 表明它对数据的拟合程度远超其他模型。因此, 本研究中的模型不仅在预测效果上表现最佳, 而且在实际应用中具有较高的潜力和价值。

3.4. 中国石化股票价格和华建集团股票价格预测

为了分析本文模型在不同时间序列模式下的预测性能, 将模型应用于另两个数据集中, 即中国石化股票和华建集团股票(2014 年 9 月 16 日~2024 年 6 月 1 日)历史数据, 共包含 2360 条数据。循前述建模流程, 对中国石化股票的收盘价和华建集团股票收盘价进行了预测。图 8 和图 9 展示了预测效果, 而预测结果则详细列于下表 3 和表 4。通过这一扩展分析, 我们旨在评估模型在不同数据集上的泛化能力和

预测准确性。

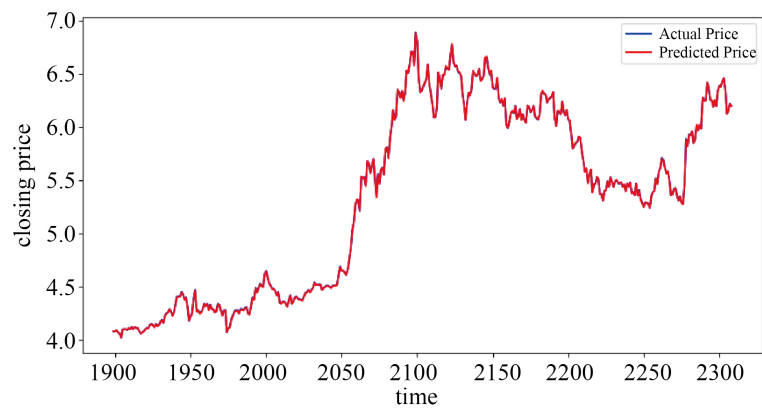


Figure 8. Sinopec stock hybrid model prediction curve
图 8. 中国石化股票混合模型预测曲线图

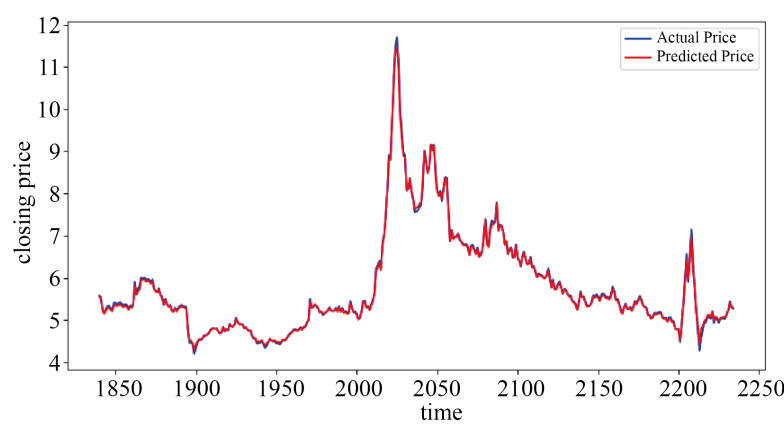


Figure 9. Huajian group stock hybrid model prediction curve
图 9. 华建集团股票混合模型预测曲线图

Table 3. Sinopec stock prediction results
表 3. 中国石化股票预测结果

序列	真实值	预测值	差值
1	4.08	4.0798040	1.96e-04
2	4.08	4.0800676	-6.76e-05
...	...	...	...
471	6.22	6.2224780	-2.478e-03
472	6.20	6.2005544	-5.544e-04

在对图 8 的分析中, 可以观察到本研究所提出的模型在应用于中国石化股票数据集时, 依然能够维持较高的预测准确性。模型的预测曲线与实际股价数据展现出高度的一致性, 这表明模型具有较高的拟合优度。通过表 3 和表 4 中的数据, 计算了实际值与预测值之间的差异(即实际值减去预测值), 结果显示这些差异相对较小, 进一步证实了模型预测结果与实际市场表现的接近程度, 从而反映出模型的预测效能较为出色。

Table 4. Huajian group stock prediction results
表 4. 华建集团股票预测结果

序列	真实值	预测值	差值
1	5.58	5.555662	2.4338e-02
2	5.53	5.4983454	3.16546e-02
...	...	...	...
471	5.30	5.31047	-1.047e-02
472	5.27	5.2858753	-1.58753e-02

为了量化中国石化股票价格的实际值与本模型预测误差之间的差异，本研究选取了平均绝对百分比误差(MAPE)作为评价指标。MAPE (Mean Absolute Percentage Error)衡量的是预测值与实际值之间绝对误差的平均值，表达式为

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - y'_i}{y_i} \right|$$

(17)

其中， y_i 是股票价格的真实值， y'_i 为股票价格的预测值。本文模型的 MAPE 评价指标与其他模型的预测性能如表 5 所示，可知，本文提出的预测模型在 MAPE 上均优于其他对比模型。

Table 5. MAPE values of different models
表 5. 不同模型 MAPE 值

模型	MAPE
RNN	4.03144
单特征 LSTM	3.35300
多特征 LSTM	26.1113
XGBoost	3.62826
LSTM-Transformer	0.22495
RNN-Transformer	0.24518
EMD-RNN	3.43104
EMD-LSTM	2.42332
XGBoost-EMD-RNN	0.31709
XGBoost-EMD-LSTM	0.21721

对于 XGBoost-EMD-LSTM 模型在华建集团的股票价格上预测的评价指标 MSE、RMSE、MAE、MAPE、 R^2 如表 6 所示。

Table 6. Evaluation metrics
表 6. 评价指标

模型	MSE	RMSE	MAE	MAPE	R^2
XGBoost-EMD-LSTM	0.00212	0.04610	0.03143	0.20481	0.99861

4. 结束语

股市因其具有高风险以及高收益的特性, 一直以来, 预测股票价格的有效性一直是投资者与研究者共同追求的目标。本文提出一种 XGBoost-EMD-LSTM 模型来对股票收盘价进行预测, 实证证明了其有效性和可行性。该混合模型的 MSE 、 $RMSE$ 、 MAE 、 R^2 分别为 0.00009、0.00952、0.00720、0.99994。通过与其他几种机器学习模型进行对比, 证实了 XGBoost-EMD-LSTM 模型的预测误差更小、拟合度更好。又通过把模型应用于另两个数据集中, 用模型的评价指标验证模型的有效性, 其结果表明: 本研究的模型在面对具有非平稳性和非线性特征的股票价格时, 展现出了卓越的预测能力。

但是, 本文还有一些局限性, 主要体现在对于模型的参数方面, 在模型训练过程中仅采用人为经验进行调参, 具有很大的不确定性和随机性。因此在后续的研究中将通过优化算法对模型进行调参, 可能会获得更好的预测效果。

参考文献

- [1] Panigrahi, S., Pattanayak, R.M., Sethy, P.K. and Behera, S.K. (2021) Forecasting of Sunspot Time Series Using a Hybridization of ARIMA, ETS and SVM Methods. *Solar Physics*, **296**, Article No. 6. <https://doi.org/10.1007/s11207-020-01757-2>
- [2] Xing, D., Li, H., Li, J. and Long, C. (2021) Forecasting Price of Financial Market Crash via a New Nonlinear Potential GARCH Model. *Physica A: Statistical Mechanics and Its Applications*, **566**, Article 125649. <https://doi.org/10.1016/j.physa.2020.125649>
- [3] Urom, C., Chevallier, J. and Zhu, B. (2020) A Dynamic Conditional Regime-Switching GARCH CAPM for Energy and Financial Markets. *Energy Economics*, **85**, Article 104577. <https://doi.org/10.1016/j.eneco.2019.104577>
- [4] Kim, C.B. (2018) Leverage Effect of HRCI Volatility and the Volatility Impact on Korean Export Container Volume before and after the Global Financial Crisis: Application of ARIMA-EGARCH and GIRF. *The Asian Journal of Shipping and Logistics*, **34**, 227-233. <https://doi.org/10.1016/j.ajsl.2018.09.006>
- [5] Rounaghi, M.M. and Nassir Zadeh, F. (2016) Investigation of Market Efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and Yearly Forecasting of Time Series Stock Returns Using ARMA Model. *Physica A: Statistical Mechanics and Its Applications*, **456**, 10-21. <https://doi.org/10.1016/j.physa.2016.03.006>
- [6] 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.
- [7] 李馨蕊. 基于 RF-SVM 算法的多因子量化选股研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2022.
- [8] 闫政旭, 秦超, 宋刚. 基于 Pearson 特征选择的随机森林模型股票价格预测[J]. 计算机工程与应用, 2021, 57(15): 286-296.
- [9] 白军成, 孙秉珍, 郭誉齐, 等. 融合三支聚类与分解集成学习的股票价格预测模型[J]. 运筹与管理, 2024, 33(8): 213-218.
- [10] 谢琳. 基于 LSTM-XGBoost 组合模型的股价预测研究[D]: [硕士学位论文]. 北京: 中央民族大学, 2022.
- [11] 邓建军. 基于 Attention-LSTM 与 XGBoost 集成机制的中国商品期货投资策略研究[D]: [硕士学位论文]. 成都: 四川大学, 2022.
- [12] 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]. 数学的实践与认识, 2018, 48(8): 297-307.
- [13] 王燕, 郭元凯. 改进的 XGBoost 模型在股票预测中的应用[J]. 计算机工程与应用, 2019, 55(20): 202-207.
- [14] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [15] Wang, J., Liao, L., Zhong, K., Deveci, M., du Jardin, P., Tan, J., et al. (2025) MRRFGNN: Multi-Relation Reconstruction and Fusion Graph Neural Network for Stock Crash Prediction. *Information Sciences*, **689**, Article 121507. <https://doi.org/10.1016/j.ins.2024.121507>
- [16] 曹超凡, 罗泽南, 谢佳鑫, 等. MDT-CNN-LSTM 模型的股价预测研究[J]. 计算机工程与应用, 2022, 58(5): 280-286.
- [17] 韩莹, 张栋, 孙凯强, 等. 结合长短时记忆网络和宽度学习的股票预测新模型研究[J]. 运筹与管理, 2023, 32(8): 187-192.

-
- [18] 朱菊香, 任明煜, 谷卫, 等. 基于 CEEMDAN-IGWO-CNN-LSTM 空气质量预测建模[J]. 计算机仿真, 2025, 42(1): 529-537.
- [19] 孙存浩, 胡兵, 邹雨轩. 指数趋势预测的 BP-LSTM 模型[J]. 四川大学学报(自然科学版), 2020, 57(1): 27-31.
- [20] Wang, Q. and Zhang, Y. (2022) Research on PM2.5 Pollution Prediction Method in Hefei City Based on CNN-LSTM Hybrid Model. *Journal of Physics: Conference Series*, **2400**, Article 012006.
<https://doi.org/10.1088/1742-6596/2400/1/012006>
- [21] 尤睿凡. 基于时间序列模型与机器学习的组合模型的股票价格指数预测研究[D]: [硕士学位论文]. 济南: 山东大学, 2021.
- [22] 张彪, 彭秀艳, 高杰. 基于 ELM-EMD-LSTM 组合模型的船舶运动姿态预测[J]. 船舶力学, 2020, 24(11): 1413-1421.
- [23] Ren, X., Guo, H., Li, S., Wang, S. and Li, J. (2017) A Novel Image Classification Method with CNN-XGBoost Model. *Digital Forensics and Watermarking*, Magdeburg, 23-25 August 2017, 378-390.
https://doi.org/10.1007/978-3-319-64185-0_28
- [24] 郭长东. 基于 XGBoost 模型的股票预测研究[D]: [硕士学位论文]. 延吉: 延边大学, 2021.
- [25] 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*, San Francisco, 13-17 August 2016, 785-794.
<https://doi.org/10.1145/2939672.2939785>
- [26] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., *et al.* (1998) The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **454**, 903-995.
<https://doi.org/10.1098/rspa.1998.0193>