E-Commerce Letters 电子商务评论, 2024, 13(2), 1770-1779
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.132216

Hans汉斯

基于深度学习对金融证券市场股票价格和风险问题的研究

杨　涛，马艳雨

浙江理工大学理学院，浙江　杭州

收稿日期：2024年3月4日；录用日期：2024年4月2日；发布日期：2024年5月20日

摘　要

中国金融证券市场已逐渐成熟，吸引广泛投资者参与，股票价格受内在价值和大盘指数波动影响，投资者需面对时机选择和风险价值分析的难题。本文针对中国金融证券市场的股票价格和风险预测问题，以晋城煤业作为研究对象，构建了基于深度学习的CNN-LSTM股票价格预测模型和基于信息熵与方差的风险度量模型。CNN-LSTM模型可以通过卷积神经网络提取局部空间特征，长短期记忆网络提取时间特征，并利用分位数法求局部性顶部或底部股票价格时间区间；而信息熵－方差模型综合考虑了信息熵和方差两个指标，度量股票收益的不确定性和波动性，构建了全面的风险度量模型。实证分析表明，股票价格预测模型能够较好地预测股票价格走势，并判断出局部极值的出现时间；而风险度量模型能够合理评估股票投资风险，风险值变化与股票实际波动性相符，从而为投资者投资行为提供有效支撑。

关键词

股票价格，CNN，LSTM，信息熵，风险度量

Research on Stock Price and Risk in Financial Securities Market Based on Deep Learning

Tao Yang, Yanyu Ma

School of Science, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Mar. 4th, 2024; accepted: Apr. 2nd, 2024; published: May 20th, 2024

Abstract

The Chinese financial securities market has gradually matured, attracting a wide range of inves-

文章引用: 杨涛, 马艳雨. 基于深度学习对金融证券市场股票价格和风险问题的研究[J]. 电子商务评论, 2024, 13(2): 1770-1779. DOI: 10.12677/ecl.2024.132216

tors to participate. Stock prices are influenced by intrinsic value and market index fluctuations, posing challenges for investors in terms of timing and risk-value analysis. This article addresses the issue of stock price and risk prediction in the Chinese financial securities market, focusing on Jincheng Coal Industry. We have developed a CNN-LSTM stock price prediction model based on deep learning and a risk measurement model based on information entropy and variance. The CNN-LSTM model utilizes convolutional neural networks to extract local spatial features, long short-term memory networks to capture temporal features, and employs quantile regression to identify local extreme points in stock prices over time intervals. On the other hand, the information entropy-variance model comprehensively considers both information entropy and variance, measuring the uncertainty and volatility of stock returns to construct a comprehensive risk measurement model. Empirical analysis indicates that the stock price prediction model performs well in forecasting stock price trends and identifying the occurrence time of local extremes. Meanwhile, the risk measurement model effectively assesses stock investment risks, with variations in risk values aligning with the actual volatility of stocks. This provides valuable support for investors in their investment decisions.

Keywords

Stock Price, CNN, LSTM, Information Entropy, Risk Measurement

1. 引言

中国金融证券市场经过多年的发展已经逐渐成熟，吸引了各行各业的广泛投资者参与[1]。在这个市场中，每支股票的价格不仅受到其内在投资价值的影响，还受到整体大盘指数的波动影响[2]。投资者在买卖决策时常常面临时机选择的难题，同时也面临着被套和踏空等风险，因此对股票的风险价值进行深入分析显得尤为重要。

自股票市场诞生以来，人们一直致力于研究股票价格预测的方法[3]，并涌现出许多预测模型。然而，传统方法在一定程度上无法满足对准确性的高要求，因此近年来人们纷纷将机器学习算法引入股票预测领域，希望能够获得更为准确的预测结果。机器学习算法的优势在于能够最大程度地模拟对象的具体特征，尤其在处理大规模复杂数据和进行预测等方面具备更大的优势[4] [5] [6]。

为解决股票价格和风险预测的问题，本研究采用了卷积神经网络(CNN)和长短期记忆网络(LSTM)构建了一个新颖的股票价格预测模型。同时，还采用了信息熵与方差相结合的方法，建立了一个全面的风险度量模型，为投资决策提供了重要的参考依据。通过利用 CNN 和 LSTM 等先进算法，可以更加准确地预测股票价格的走势，并通过综合考量信息熵和方差来量化风险，使得投资者能够更加全面地评估投资风险。这一综合性的研究框架为股票市场的理解和投资决策提供了新的思路和方法。

2. 研究现状

在当前的研究现状中，针对股票市场的价格趋势预测，研究者们广泛采用了上证 A 股 50 和沪深 300 指数的历史交易数据作为研究对象，并利用了 ARIMA 模型和 LSTM 模型等方法进行预测分析。这些研究旨在提高股票市场预测的准确性和可靠性，以指导投资决策。

文献[4]基于上证 A 股 50 的历史交易数据，采用 ARIMA 模型和 LSTM 模型进行股价趋势预测。实

证研究表明，基于 LSTM 模型的深度神经网络具有较好的预测精度，但未对过拟合问题进行充分解决。

文献[5]选取了沪深 300 指数的日交易数据、技术指标和估值指标作为样本数据，同样采用了 LSTM 模型进行未来一天收盘指数的预测建模，结果表明 LSTM 多特征输入模型相对较好，但在股价预测和涨跌预测中均存在过拟合问题。

文献[6]在 LSTM 模型基础上进行了结构改进和参数优化，使预测准确率提升 10%以上，且优于 SVM 和随机森林模型。然而，该研究仅使用了有限的特征数据进行预测，存在提高预测准确率的潜在空间。

综合而言，这些研究在股票市场预测方面取得了一定的进展，然而仍普遍存在一些共同的不足之处，其中包括模型过拟合问题以及对特征数据的不充分利用。解决过拟合问题、深入探索更加丰富的特征数据，以进一步提高股票市场预测的准确性和稳定性，仍然是未来研究亟需深入探讨的重要方向。

3. 研究设计

3.1. 研究假设

1) 假设以股票的每日收盘价来衡量股票的每日价格情况。
2) 假设股票的价格不受到其内在投资价格的影响。
3) 假设金融证券市场未来的行情由现在的行情决定。
4) 假设无人为操作股市走向，所有数据为随机数据。

3.2. 数据来源

本文研究所使用的数据，晋控煤业股票数据。选取晋控煤业作为研究对象，收集了其自 2006 年 1 月 1 日至 2022 年 5 月 17 日的完整股票数据。数据内容包括每个交易日的开盘价、收盘价、最高价、最低价、成交量、成交额等信息。

3.3. 数据预处理

使用 tushare 库可以轻松获取晋控煤业从 2006 年至今的交易数据。针对每个交易日，可以获取开盘价、最高价、最低价等九个指标。接下来，将这些指标按照每个连续的 365 个交易日构建一个 batchsize，其中该 batchsize 后的 30 个交易日的收盘价作为当前 batchsize 的目标值。因此，可以构建 3679 个这样的 batchsize。

3.4. 研究流程

本研究的流程如图 1 所示。首先按照 3.3 节的方法对数据进行预处理，然后使用处理后的数据构建 CNN-LSTM 模型。基于该模型得到 2021 年交易日的价格预测，使用八分位数法分析出股票价格出现局部性顶部和底部的时间区间。此外基于该模型预测出 2022 年 5 月 18 日起 3 未来 30 个交易日的股票价格，基于信息熵-方差的风险度量模型得到股票风险值。

4. 实证分析

4.1. 基于 CNN-LSTM 的股票价格预测模型

4.1.1. 模型准备

卷积神经网络(Convolutional Neural Networks, CNN)是一种前馈神经网络结构，通过卷积操作实现对输入数据的特征提取。CNN 具有表征学习能力，能够按照其深度结构对输入信息进行平移不变分类，因此也被称为“平移不变人工神经网络”。其局部感受特性使得 CNN 能够分别处理每个时间段的数据，并

将长输入序列转换为由高级特征组成的更短序列，从而实现对序列空间特征的编码。

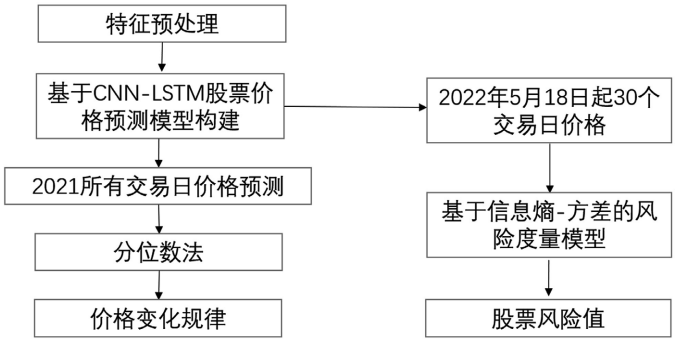

Figure 1. Research process diagram
图 1. 研究流程图

长短时记忆网络(Long Short Term Memory Network, LSTM)属于循环神经网络(Recurrent Neural Network, RNN)的一种。在处理较长序列数据时，传统 RNN 由于自身的记忆限制，无法充分结合之前的信息来赋予当前数据正确的权重，导致梯度消失问题。而 LSTM 网络通过引入门控机制，如遗忘门、输入门和输出门等，能够有效地保存和利用长序列中的重要信息，解决了 RNN 存在的长期依赖问题。

多层感知机(MLP, Multilayer Perceptron)是一种常见的人工神经网络结构，除了输入输出层外，中间可以包含多个隐层，通过多层的非线性变换实现对复杂数据的学习和表示。

基于 CNN-LSTM 的股票价格预测系统如图 2 所示，该系统主要包括五部分，分别是输入层、基于二维卷积神经网络的局部空间特征子模型、基于长短时记忆网络的时间特征提取子模型、用于融合全部特征的全连接层和输出层。

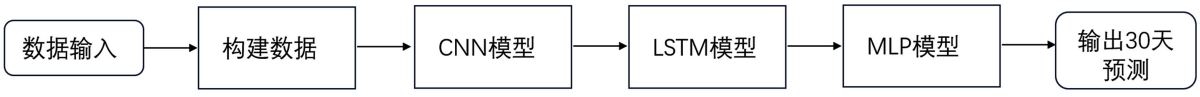

Figure 2. Flowchart of stock price prediction model based on CNN-LSTM
图 2. 基于 CNN-LSTM 的股票价格预测模型流程图

4.1.2. 模型建立

在局部特征提取阶段，对长度为 n 的股票数据序列做数据清洗，将其作为整个模型的输入，输入数据具体表示为 $X=\left[x_1,x_2,\cdots,x_n\right]^{\mathrm{T}}$，式中 x_i 代表股票的最高价、最低价、开盘价和收益量等。本研究采用两层 CNN 结构来提取历史数据中的空间特征，以实现局部感知并提高特征质量。在这两层 CNN 中，数据通过卷积核进行特征抽取。CNN 具有权值共享机制，有助于减少网络参数的复杂性，并提高整个模型的健壮性。此外，最大池化作为池化层的选择有利于特征降维，以避免网络训练过程中的过拟合现象。

随后，进行时间特征提取。为了提高特征提取的质量并降低整个网络的复杂度，本研究在时间特征提取子模型中引入了三层 LSTM 网络。LSTM 网络包括遗忘门、输入门、输出门和细胞结构，各自担负着关键的作用。遗忘门通过选择性过滤掉特征不明显的信息，有助于提高网络的存储记忆能力。具体的计算公式如下所示：

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

其中，f_t 是遗忘门的输出；h_{t-1} 和 x_t 代表遗忘门的输入；b_f 是偏置参数，W_f 是线性关系的系数。

输入门决定当前细胞的输入需要加入多少新信息，以确定细胞中哪些信息需要更新，哪些信息被用作备用更新内容。具体的表达方式如下：

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

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right)$$

上述公式通过 sigmoid 函数决定网络需要更新的数据，并通过 tanh 函数创建候选值 $\tilde{C}_t$，以去除网络中暂时不需要的信息，有效提高网络存储数据的能力。

输出门决定模型的输出，首先是通过 sigmoid 层得到初始输出，然后通过 tanh 函数将 C_t 的值缩放到 [−1, 1]，再 sigmoid 得到的输出逐对相乘，从而得到模型的输出，表达公式如下：

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

$$h_t = o_t \cdot \tanh\left(C_t\right)$$

进一步，利用八分位数求局部性顶部或底部股票价格时间区间，本研究参考全年股票价格的整体趋势，运用改进后的八分位数法分析预测股票价格出现局部性顶部和底部的时间区间。

4.1.3. 模型结果分析

通过使用 Python 软件，绘制出基于 CNN-LSTM 的股票价格预测模型对 2021 年交易日的预测结果图，总计包含了 242 条数据。具体图示详见图 3。全部预测结果见附件一。部分预测结果见表 1。

在对 2021 年的股票价格进行全年的观察后，发现了一些显著的分布特征。前四个月的股票价格变化相对平缓，并于在 3 月股票价格跌至低谷。接下来的八个月，股票价格发生了较大波动，其中 9 月份股票价格上涨至全年的最高点。根据全年股票价格的变动特点，找到全年股票价格的中位数，依照此中位数将全年的价格分为两个时间段，在这两段时间内，将股票价格的最高点和最低点分别与股票价格的中位数相对应的四等分点进行对比，并在这些四等分点的股票价格处作水平参考线。图 4 仅展示部分参考线。

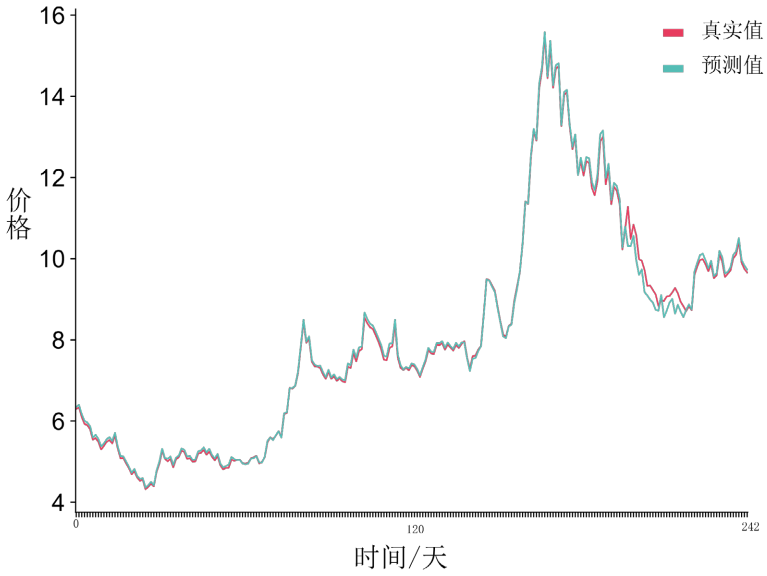

Figure 3. Prediction performance of stock prices on trading days in 2021
图 3. 2021 年交易日股票价格预测效果

Table 1. Translation of the real and predicted values of prices for the year 2021 trading days
表 1. 2021 年交易日价格的真实值与预测值

天数	日期	真实值	预测值
0	20210104	6.3	6.34761
1	20210105	6.34	6.399019
2	20210106	6.1	6.167511
3	20210107	5.93	6.003674
4	20210108	5.9	5.963681
5	20210111	5.8	5.873648
6	20210112	5.54	5.597471
7	20210113	5.58	5.66294
8	20210114	5.49	5.554958
9	20210115	5.3	5.37414
10	20210118	5.39	5.445477

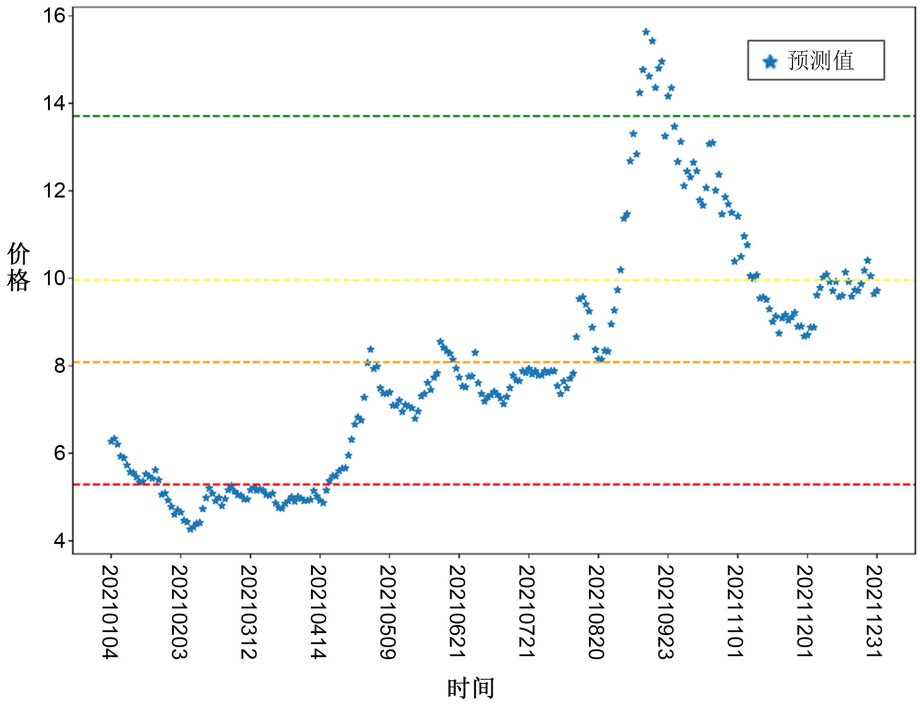

Figure 4. Scatter plot of stock price predictions on trading days in 2021
图 4. 2021 年交易日股票价格预测散点图

由图 4 可知，在 2021 年 1 月 4 日股票开盘后一段时间股票价格相对较低，预测在 2021 年 1 月 26 日至 2021 年 2 月 22 日股票价格出现局部性底部。之后 7 个月股票价格呈显著的上升趋势，预测在 2021 年 9 月 8 日至 2021 年 9 月 24 日内股票价格出现局部性顶部。2021 年最后两个月股票价格先下降后上升，预测局部性顶部出现在 2021 年 12 月 8 日到 2021 年 12 月 17 日之间。

4.2. 基于信息熵与方差的风险度量模型

4.2.1. 模型准备

由于证券市场受经济政策和资金流动等不确定因素的影响，投资者的实际收益往往与决策时期望的收益不一致，即投资存在风险. 在投资学中，投资风险可被刻画为未来收益的不确定性及其发生的概率。人们常常把风险度量[7]与方差联系在一起，因为方差是统计学中最常用的描述随机变量特性的指标，而风险中的损失和收益变化就是随机变量。

信息熵：设离散型随机变量 X，$P\left(X=x_i\right)=p_i$，$0\le p_i\le 1\left(i=1,2,\cdots,n\right)$，$\sum_{i=1}^{n}p_i=1$，则该离散概率事件 X 的信息熵[8]为 $H=-k\sum_{i=1}^{n}p_i\ln p_i$，其中，$k>0$，是一个取决于度量单位的常数，且规定 $0\ln 0=0$。设连续性随机变量 X，概率密度为 $f\left(x\right)$，则该连续概率事件 X 的信息熵为 $H=-\int f\left(x\right)\ln\left[f\left(x\right)\right]\mathrm{d}x$。

4.2.2. 模型建立

为了建立持有期股票价格预测模型，首先需要利用之前开发的股票价格预测模型，以便预测出未来 30 个交易日的每日股票价格 $R=\left[r_1,r_2,\cdots,r_{30}\right]^{\mathrm{T}}$。

在预测股票价格的基础上，需要计算出相应的区间数。根据我国证券管理规定，股票在单日内的涨跌幅度通常被限制在正负 10%之内。因此，本研究将股票实际收益区间划分为[−10%, 10%]，并将其均分为 q 个收益子区间。

$$\left[-10\%,10\%\right]=\bigcup_{i=1}^{q}\Delta_i,i=1,2,\cdots,q$$

为了研究子区间数对熵风险度量[9]值的影响，q 可取值如下 $q=10\cdot k$，其中 q 为子区间数，k 定义为子区间数目指数。

接下来，计算股票的收益率与子区间，设 $r_j\left(j=1,2,\cdots,d\right)$ 为股票的日收盘价，d 为数据采集天数，则日收益率可以表示为：

$$R_j=\frac{r_j-r_{j-1}}{r_{j-1}}$$

本研究将区间[−10%, 10%]均分为 $q=10\cdot k$ 个子区间，取步长为 $l=\frac{0.2}{q}$，子区间表示为：

$$\Delta_i=\begin{cases}\left[-10\%+\left(i-1\right)l,-10\%+il\right],i=1,2,\cdots,q-1\\ \left[-10\%+\left(i-1\right)l,-10\%+il\right],i=q\end{cases}$$

随后计算熵风险值，假设 R_j 落在第 i 个子区间内的次数为 n_i，记频数 $n_i=\frac{n_i}{d}$，则定义股票的熵风险值如下式：

$$H\left(k\right)=-\sum_{i=1}^{q}p_i\ln p_i$$

此外，还需要计算股票的方差。方差可通过下式计算：

$$D\left(R\right)=E\left[\left(R-E\left(R\right)\right)^2\right]=\sum_{i=1}^{30}p_i\left(r_i-E\left(R\right)\right)^2=\sum_{i=1}^{30}p_i\left[r_i^2-2r_iE\left(R\right)+E^2\left(R\right)\right]$$

其中，$E(R)$为随机变量 R 的均值。

最后计算风险值，风险值表示为：

$$L(R)=\lambda H_R(\varepsilon)+(1-\lambda)S_R(\varepsilon)$$

其中 $L(R)$ 表示事件 R 的风险，$H_R(\varepsilon)$ 为风险事件 R 在 ε 状态的信息熵[10]，$S_R(\varepsilon)$ 表示风险事件 R 的标准差，$0\le\lambda\le1$ 为常数，是决策者的风险偏好系数。

4.2.3. 模型结果

首先通过 Python 编程求解，预测出从 2022 年 5 月 18 日起股票未来 30 个交易日的价格，预测结果请参见图 5。

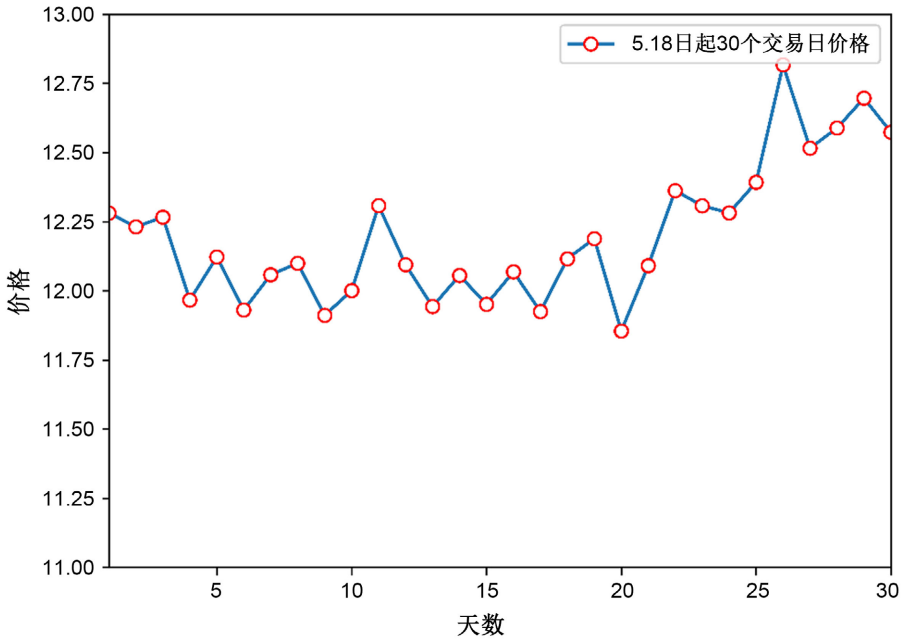

Figure 5. Stock price chart for 30 trading days of Jinkou coal industry starting from May 18, 2022
图 5. 2022 年 5 月 18 日起股票晋控煤业 30 个交易日的价格图

其次根据预测出的股票价格测算出 30 个交易日的日收益率，结果见表 2。

Table 2. Daily returns for 30 trading days starting from May 18, 2022
表 2. 2022 年 5 月 18 日起 30 个交易日日收益率表

时间	1 日	2 日	3 日	4 日	5 日
日收益率	−0.39%	−0.41%	0.29%	−2.44%	1.30%
时间	6 日	7 日	8 日	9 日	10 日
日收益率	−1.58%	1.07%	0.35%	−1.56%	0.75%
时间	11 日	12 日	13 日	14 日	15 日
日收益率	2.56%	−1.73%	−1.25%	0.94%	−0.87%
时间	16 日	17 日	18 日	19 日	20 日
日收益率	0.98%	−1.19%	1.60%	0.60%	−2.74%
时间	21 日	22 日	23 日	24 日	25 日
日收益率	1.99%	2.25%	−0.44%	−0.21%	0.90%
时间	26 日	27 日	28 日	29 日	30 日
日收益率	3.43%	−2.35%	0.58%	0.86%	−0.97%

接着根据日收益率和子区间划分数的不同，分别取区间数 5，10，15 计算股票持有 30 个交易日的信息熵风险值，结果见表 3。

Table 3. Entropy risk values for different intervals
表 3. 不同区间数的熵风险值表

区间数	5	10	15
熵风险值	2.238	3.151	3.489

然后计算 30 个交易日的标准差为 0.155，通过 $L(R)=\lambda H_R(\varepsilon)+(1-\lambda)S_R(\varepsilon)$，令 $\lambda=0.5$，得到最终的风险值，结果见表 4。

Table 4. Risk values for different intervals
表 4. 不同区间数的风险值表

区间数	5	10	15
风险值	1.197	1.653	1.822

再对比晋控煤业(研究对象)与陕西煤业根据 2022 年 3 月 29 日~5 月 17 日 30 个交易日价格计算出其对应风险值，结果见表 5。

Table 5. Risk values for different interval divisions
表 5. 不同区间数划分的风险值表

区间数	5	10	15
晋控煤业风险值	1.154	1.592	1.84
陕西煤业风险值	1.049	1.441	1.553

由表可知，若投资人于 3 月 29 日同时购买晋控煤业与陕西煤业并持有一个月，则持有晋控煤业的风险要高于陕西煤业。

下面给出晋控煤业与陕西煤业 3 月 29 日至 5 月 17 日 30 个交易日的对应股票价格图，见图 6。

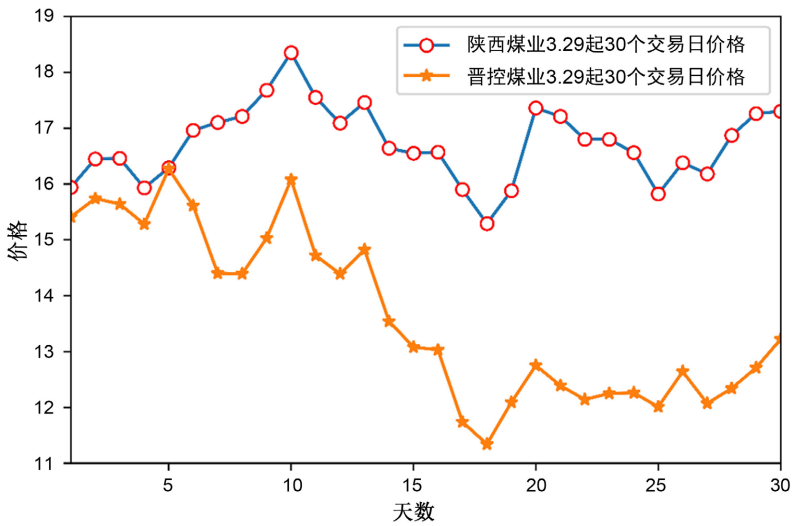

Figure 6. Comparative chart of the prices of Jinkou coal industry and Shaanxi coal industry over 30 trading days
图 6. 晋控煤业与陕西煤业 30 个交易日价格对比图

由图 6 可知，晋控煤业的波动性要比陕西煤业高，且晋控煤业的股票价格变化趋势向下，陕西煤业的股票价格变化趋势向上，无论在哪种区间划分下，晋控煤业的风险值都高于陕西煤业，因此可以证明，模型关于风险值测算的正确性。

5. 结论与建议

本文针对金融证券市场的股票价格和风险预测问题，综合运用深度学习、风险度量等方法，构建了 CNN-LSTM 股票价格预测模型和信息熵－方差风险度量模型，并通过实证分析验证了模型的有效性。研究结果表明，CNN-LSTM 模型能够较好地预测股票价格走势，并判断出局部极值的出现时间；信息熵－方差模型能够合理地度量股票投资风险，为投资决策提供支持。

基于上述研究结果，本文提出以下建议：

首先，投资者在购买股票时，可以参考本文构建的股票价格预测模型，预测未来一段时间内股票价格的走势，把握买卖时机，避免在局部高点买入或局部低点卖出。

其次，在评估股票投资风险时，投资者可以综合考虑本文提出的信息熵和方差两个指标，全面衡量股票收益的不确定性和波动性，选择风险值较低的股票，控制投资风险。

最后，通过案例分析可以看出，在风险值较高的情况下，股票价格出现下跌趋势的可能性更大。因此，投资者在风险值较高时应谨慎投资，或者及时止损，规避风险。

参考文献

[1] 丁晨. 倾听历史回音金融证券展示文化传承力量[J]. 中国收藏, 2023(12): 134-136.

[2] 岳飞冲, 张轩铭, 孔钰姝. 互联网金融资产证券化的发展情况分析——基于京东白条和阿里花呗[J]. 中国商论, 2023(4): 109-111. https://doi.org/10.19699/j.cnki.issn2096-0298.2023.04.109

[3] 杨晨烨, 袁泉. 一种股票价格预测方法及系统[P]. 中国专利, CN113935772A. 2022-01-14.

[4] 丁文绢. 基于股票预测的 ARIMA 模型、LSTM 模型比较[J]. 工业控制计算机, 2021, 34(7): 109-112+116.

[5] 张杰. 基于 LSTM 的股票预测实证分析[D]: [硕士学位论文]. 济南: 山东大学, 2020.

[6] 邓飞燕, 岑少琪, 钟凤琪, 潘家辉. 基于 LSTM 神经网络的短期价格趋势预测[J]. 计算机系统应用, 2021, 30(4): 187-192. https://doi.org/10.15888/j.cnki.csa.007855

[7] 杨鑫. 基于 KMV 模型的上市科技金融公司信用风险度量研究[J]. 中国物价, 2022(4): 78-80.

[8] Samunderu, E. and Murahwa, Y.T. (2021) Return Based Risk Measures for Non-Normally Distributed Returns: An Alternative Modelling Approach. *Journal of Risk and Financial Management*, **14**, 540. https://doi.org/10.3390/jrfm14110540

[9] 张雨. 我国证券市场系统性风险度量研究[D]: [硕士学位论文]. 贵阳: 贵州财经大学, 2021.

[10] 刘丹. 基于信息熵原理的辽东超采区地下水水位时空变化演变特征[J]. 水利规划与设计, 2021(5): 48-52.