Stock Price Prediction Based on Markov Chains

Siwei Wang *

School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding author: strange@bupt.edu.cn

Abstract. Short-term trend prediction in the stock market is of significant importance for effective market regulation by the government and optimizing resource allocation for investors. It has become a research hotspot in both academia and the industry in recent years. In addressing the long-term stock price prediction problem, a Markov Chain-based stock price prediction method is proposed. This method is based on the concept of state transitions in Markov Chains, where stock indicator data is transformed into state data. A transition probability matrix is generated, and predictions are made using matrix multiplication. Testing and verification are conducted using two datasets, BJ#430510 (Feng Guang Precision) and SZ#000009 (China Baoan). The results indicate that the stock price prediction model proposed in this paper exhibits high accuracy and stability.

Keywords: Stock Index, Long-Term Prediction, Markov Chain.

1. Introduction

The stock market serves important functions, such as financing, capital transformation, and price determination for stocks. On one hand, stock prices are influenced by various underlying factors. For example: national political and economic conditions, policies, and so on. Consequently, stock prices are often referred to as a "barometer" of the broader economy. On the other hand, stock prices have a significant impact on corporate decision-making and investor behavior, making them closely related to social and economic development and people's lives.

However, short-term stock price predictions may have limited value for long-term investment guidance, and the range of price fluctuations in the short term may not be substantial. Therefore, the pursuit of extreme short-term precision through model adjustments may not necessarily provide significant value in predicting long-term stock price trends. Many models struggle to achieve high accuracy in long-term stock price predictions. Consequently, forecasting long-term stock price changes has become a meaningful yet challenging research problem.

To address these challenges, this paper proposes a long-term stock trend prediction model based on Markov Chains. This model processes existing data to transform it into state data and builds a transition probability matrix. It predicts future stock price states using matrix multiplication. Finally, the model is tested and verified using data from two datasets provided by Tongda Xing (BJ#430510 - Fengguang Precision) and Shenzhen Stock Exchange (SZ#000009 - China Baoan). The results demonstrate that the stock price prediction model proposed in this paper offers higher accuracy and stability.

2. Related Work

Currently, algorithm-based stock predictions fall into two main categories:

Prediction methods based on time series statistical models [1]. These methods often involve the use of models like AEMA [2], ARIMA [3], and CARCH [4]. They establish relationships between past stock price data and future prices, primarily using time as the main variable. These methods typically require longer time spans and may not predict data at precise time points, and their price predictions may lack precision. However, they do offer some guidance for long-term investments.

Prediction methods based on machine learning. Within this category, prominent algorithms include multiple linear regression [5], decision trees [6], and k-nearest neighbors (KNN) [7]. These methods can achieve relatively accurate predictions of stock price changes in very short time frames by continuously incorporating new variables and adjusting parameters. However, these models are often complex, involve numerous variables and parameters, resulting in longer training and prediction times. Additionally, these models generally prioritize extreme short-term accuracy, which may lead to lower accuracy in long-term predictions.

3. Our Approach

3.1. The Relationship Between Markov Chains and Stock Prediction

The fundamental idea behind Markov chains is that all past information is already encapsulated within the current state, allowing us to predict the future based on the present. This concept significantly simplifies model complexity and reduces the time required for model training while still maintaining the ability to provide reasonably accurate predictions of general states.

In the context of stock price prediction, the characteristics of Markov chains can be aligned with the problem at hand. Specifically, for the prediction task in this paper, which aims to forecast the future price movements of stocks (rise, fall, or stability), these price changes can be treated as the changing states in a Markov chain. By calculating the transition probabilities between these states using historical data, a Markov chain's state transition matrix can be constructed. The following diagram illustrates the principle of using a Markov chain to predict stock prices (Figure 1):

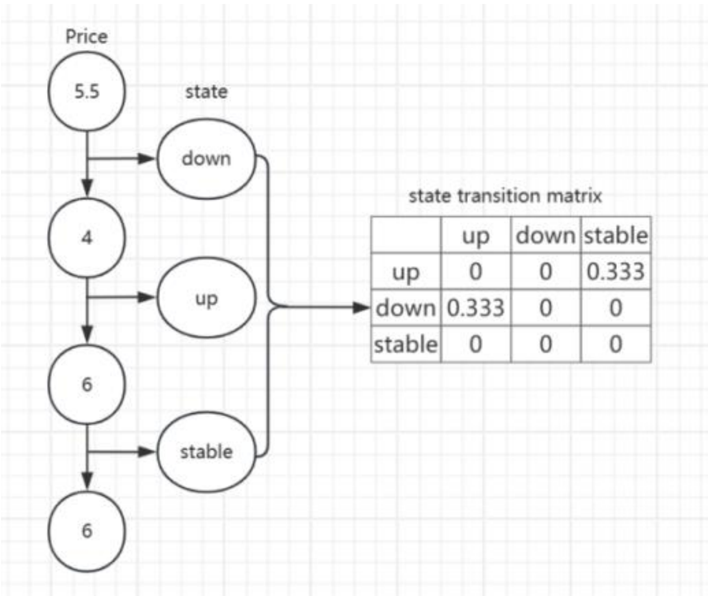

Figure 1. Markov Chain Stock Price Prediction Diagram

Through this approach, the model can avoid the dilemma of continuously attempting to increase the accuracy of price prediction by adding new decision variables. Furthermore, based on the stability of the state transition matrix, there is reason to believe that as the forecast time span increases, the results predicted by the Markov chain become relatively certain based on the collected historical data. This aligns with our expectations for stock prediction, ensuring the potential for returns from the stock in a future period.

3.2. The Main Work Done

3.2.1. Exploration of Existing Methods

First, the performance of various existing methods was studied. In this paper, several methods were primarily explored, including multiple linear regression, decision trees, and K-nearest neighbors (KNN). The decision variables used in these methods included the following: opening price, highest price, lowest price, closing price, and trading volume. Based on experiments, it was found that these methods exhibited the following characteristics in the test dataset used in this paper:

Increasing the number of decision variables can improve the accuracy of its predictions. Furthermore, the difference tends to increase as the prediction time horizon lengthens.

As the prediction time horizon lengthens, there is a noticeable decrease in the accuracy of several models.

Specific data are as follows (Table1-4):

Multivariate: Opening price, highest price, lowest price, closing price, trading volume.

Univariate: Closing price.

Table 1. Stock price prediction for one day ahead (multivariate).

model	accuracy	RMSE
DecisionTree	0.976446	0.731588638
LR	0.988138	0.637308294
KNN	0.986558	0.673529829

Table 2. Stock price prediction for one day ahead (univariate).

model	accuracy	RMSE
DecisionTree	0.976492	0.72267065
LR	0.987673	6.31E-01
KNN	0.979657	0.651254634

Table 3. Stock price prediction for thirty days ahead (multivariate).

model	accuracy	RMSE
DecisionTree	0.675002	3.239108711
LR	0.707762	2.899498085
KNN	0.767865	2.322349345

Table 4. Stock price prediction for thirty days ahead (univariate).

model	accuracy	RMSE
DecisionTree	0.606253	3.054305481
LR	0.71122	2.884052868
KNN	0.764868	2.46074329

3.2.2. Implementation of Markov Chain Method for Stock Price Prediction

Based on the concept of Markov Chains, in this section, the following three main tasks were carried out:

(1) Dividing the States of Stock Price Changes

The changes in stock prices were categorized into three states: rise, fall, and stability. Specifically, after recording all the data, a certain difference threshold (referred to as "diff" in this paper) was set to determine the next state following a given state. That is, if the current price is represented as P_0 and the price on the following day is P_1, the categorization is as follows: if $P_0 - P_1 \geq diff$, the current state is considered "down"; if $P_0 - P_1 \leq -5$, the current state is considered "up"; and if $|P_0 - P_1| \leq diff$, the current state is considered "stable." In this paper, various diff values were experimented with, including 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5.

(2) Obtaining the Transition Matrix

After converting all the data into state data, the next step is to record these state data in a transition matrix. First, it is necessary to record the probability of each state transitioning to another state after a specific number of days based on the forecasted time horizon. Then, this recorded data is filled into the matrix. For example, when diff = 5 and the time horizon is 30 days, the resulting transition matrix is as follows (Table 5):

Table 5. Transition matrix for thirty days ahead (diff = 5).

	up	stable	down
up	0.001	0.999	0
stable	0.001	0.999	0
down	0.001	0.999	0

(3) Obtaining the Prediction Results

After obtaining the transition matrix, obtaining prediction results simply requires performing matrix multiplication a corresponding number of times based on the expected forecast time horizon. By searching for the row with the highest probability in the matrix corresponding to the initial state, the model can predict the state after the specified number of days. Finally, the prediction results are compared with actual data to determine the prediction accuracy.

4. Test

4.1. Test

The experiments in this paper primarily used data from two stocks, namely, BJ#430510 (Fengguang Precision) and SZ#000009 (China Baoan). The BJ#430510 dataset contains data for 712 days, while the SZ#000009 dataset covers 2337 days. Both datasets encompass seven variables: Date, Open Price, Highest Price, Lowest Price, Closing Price, Trading Volume, and Turnover. These datasets are sourced from Xinda Tong. The following are scatter plots for each dataset with the number of days on the X-axis and the closing price on the Y-axis.

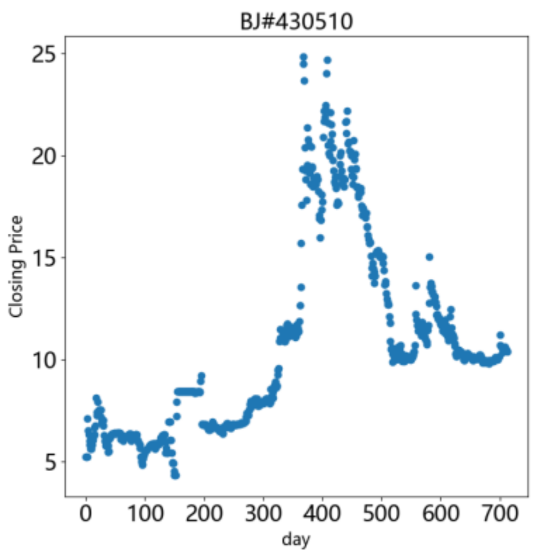

Figure 2. Scatter plot of Closing Price for BJ#430510

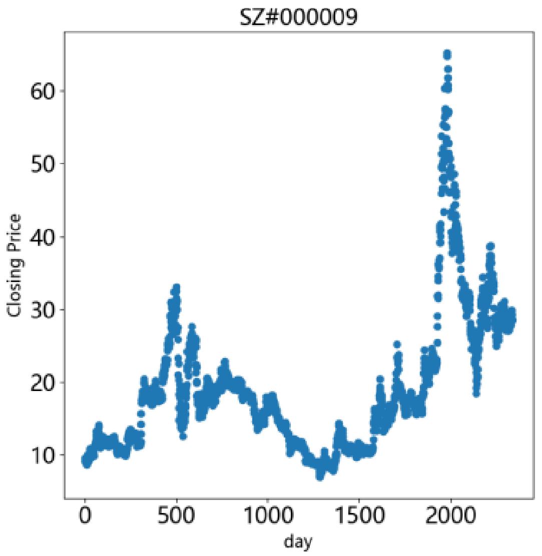

Figure 3. Scatter plot of Closing Price for SZ#000009

From the observation of the graph (Figure2-3), it's clear that both of these datasets exhibit significant fluctuations, which present a certain level of difficulty in achieving accurate predictions.

4.2. Presentation of Prediction Results

In this section, the accuracy change of various methods with increasing prediction time horizon under different difference thresholds will be presented mainly in the form of scatter plots. Different colors represent different difference thresholds, and the specific correspondences can be seen in the legend within the graph. Additionally, to illustrate the comparison of the predictive capabilities between the Markov method and other methods.

4.2.1. BJ#430510

1) Each method uses only the closing price as learning data.

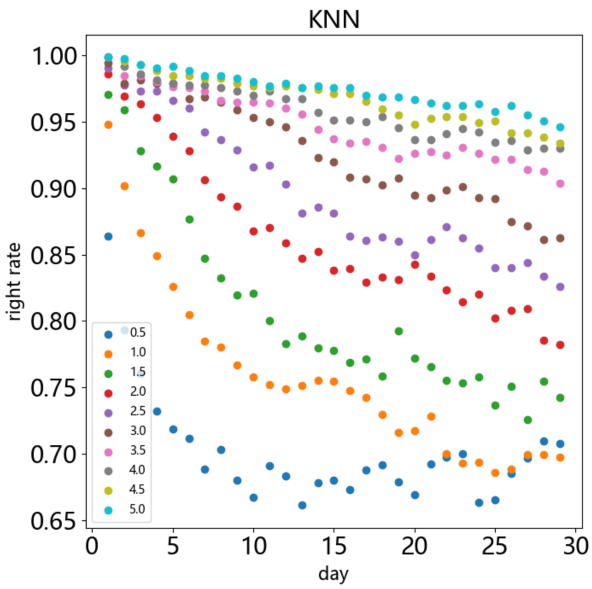

Figure 4. Scatter plot of Accuracy for KNN Algorithm (Univariate)

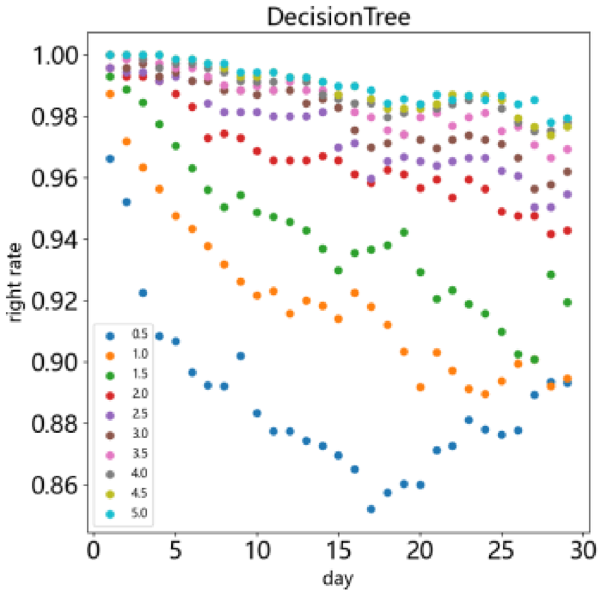

Figure 5. Scatter plot of Accuracy for Decision Tree Algorithm (Univariate)

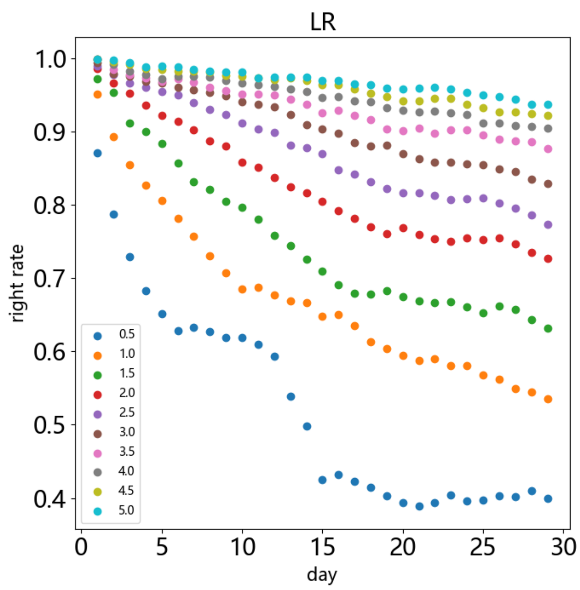

Figure 6. Scatter plot of Accuracy for Linear Regression Algorithm (Univariate)

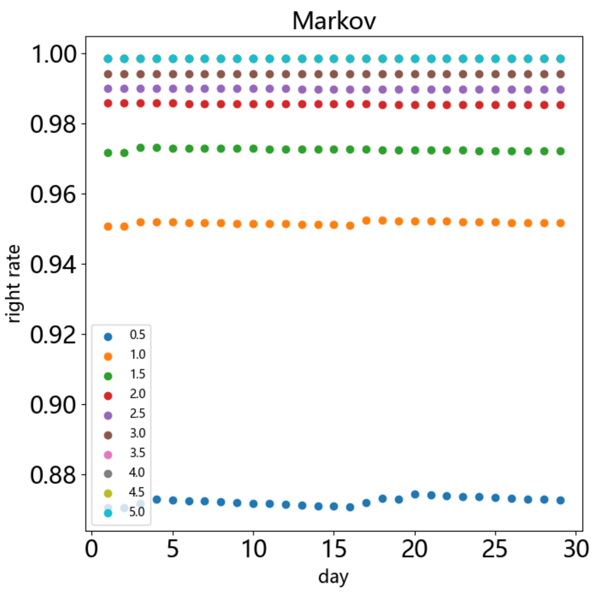

Figure 7. Scatter plot of Accuracy for Markov Chain Algorithm (Univariate)

4.2.2. SZ#000009

1) Each method uses only the closing price as learning data.

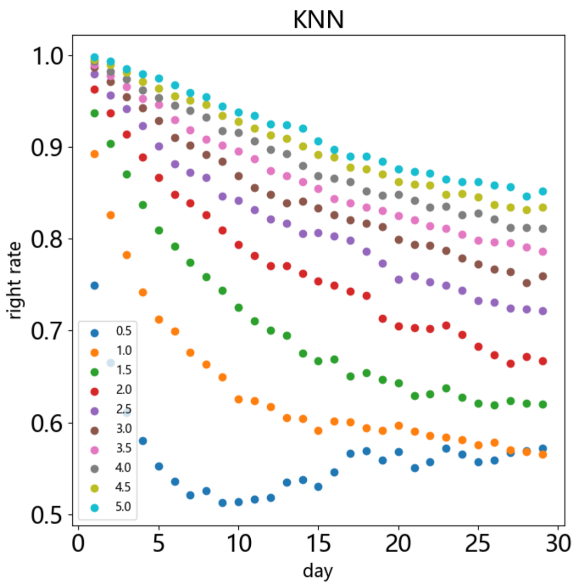

Figure 8. Scatter plot of Accuracy for KNN Algorithm (Univariate)

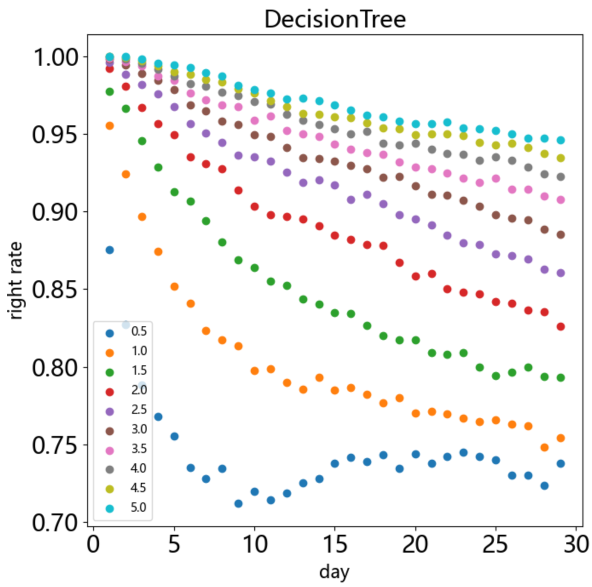

Figure 9. Scatter plot of Accuracy for Decision Tree Algorithm (Univariate)

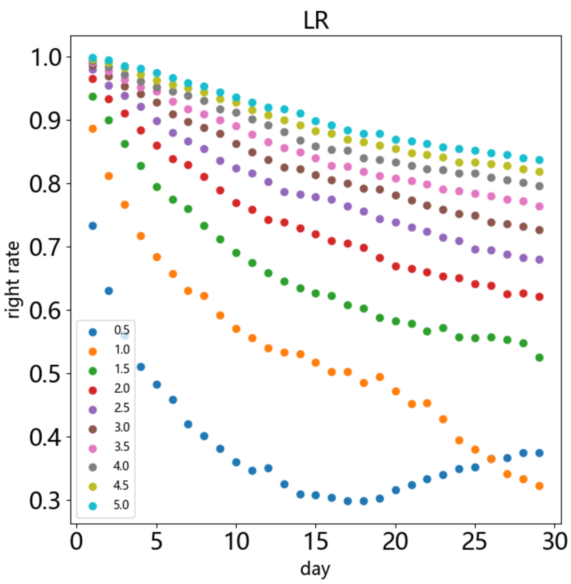

Figure 10. Scatter plot of Accuracy for Linear Regression Algorithm (Univariate)

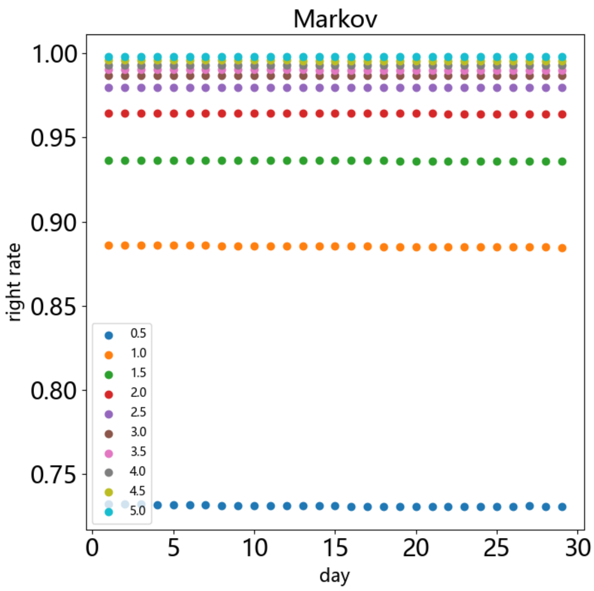

Figure 11. Scatter plot of Accuracy for Markov Chain Algorithm (Univariate)

Based on the scatter plots above, the following conclusions can be drawn (Figure 4-11):

(1) The prediction results of the Markov Chain exhibit strong stability. Regardless of the difference threshold used, the impact of the time horizon on the prediction results of the Markov Chain is relatively small. In contrast, other methods show a noticeable decline in prediction accuracy as the time horizon increases. This might imply that describing the market through the states of rise, fall, and stability may be the most intuitive and informative approach.

(2) The size of the difference threshold has a significant impact on the accuracy of the Markov Chain method within a certain range. It is observed that when the difference threshold is set to 0.5 for both datasets, the prediction accuracy of the Markov Chain is relatively low. However, when the threshold is increased to 1.0, there is a significant improvement in accuracy. When the difference threshold is further increased to 2.5, the change in prediction accuracy becomes less pronounced. This demonstrates that diverse investment risks faced by individuals with different investment styles.

When a wider price fluctuation range is acceptable, the probabilities of most states (stability) increase, leading to improved prediction accuracy.

Compared to other methods, the Markov Chain method exhibits the following three advantages:

(1) Stability over larger time horizons, making it valuable for longer-term predictions.

(2) Minimal parameter tuning and optimization are required, as the primary prediction method involves handling existing data to derive probabilities and matrix multiplication. This eliminates the need for extensive parameter tuning.

(3) Requires less data, in contrast to other methods that may require more data and additional variables to enhance the prediction model, the Markov Chain method can achieve relatively accurate predictions with minimal data requirements.

5. Conclusion

Short-term stock market trend forecasting is of significant importance for effective market regulation by the government and optimizing resource allocation for investors. Many existing prediction methods mainly focus on data processing and overlook the market information embedded in trend changes. Therefore, in this paper, the focus is on the change in stock price states. Using the Markov Chain method, numerical data is transformed into state data, leading to the generation of a transition probability matrix. The prediction results are obtained through matrix multiplication. Testing and verification are conducted using data from BJ#430510 (Fengguang Precision) and SZ#000009 (China Baoan) datasets. In conclusion, the proposed stock price prediction model in this paper exhibits higher accuracy and greater stability.

References

[1] Zhang, N. (2016). Research on Stock Trend Prediction Based on Time Series and Its Application in R Language. Modern Business, (23), 112 - 113.

[2] Lin, L., Chen, X., & Zhang, D. (2018). Application of ARMA Model in Stocks. Economic Research Reference, (26), 53, 146 - 148.

[3] Wu, Y., & Wen, X. (2016). Short-Term Stock Price Forecast Based on ARIMA Model. Statistics and Decision, (23), 83 - 86.

[4] Ke, X. (2016). Measurement of Stock Market Risk Based on GARCH Model. Contemporary Economics, (32), 12 - 14.

[5] Li, X. (2019). Application of Multiple Linear Regression and Time Series Models in Stock Forecasting. Technology Entrepreneurship Monthly, 32 (2), 153 - 155.

[6] Wei, X. (2007). Application of Decision Tree Algorithm in Stock Analysis and Forecasting. Computer Knowledge and Technology: Academic Edition, 5, 764 - 765.

[7] Wang, B., & Liu, Y. (2022). Stock Price Prediction Research Based on the KPCA-SVM-KNN Algorithm with Signal-to-Noise Ratio. Computer and Digital Engineering, 50 (4), 685 - 690.