Universe Scientific Publishing

Research on CSI 300 Index Price Prediction Based on LSTM Algorithm

Hu Linxin, Zhang Xiaoyu

City University of Macau, Macau 999078, China

Abstract: The structure of the capital market in the existing financial field has gradually become intertwined and complex, often showing irregular linear characteristic trajectories. Some traditional sequential prediction methods have subtly revealed their limitations and shortcomings in the process of stock price fluctuations. Practical results show that this paper selects CSI 300 original data as the foundation, constructs an extremely complex prediction technology system through the long short-term memory (LSTM) network series, and discusses the adaptability of hierarchical algorithms in processing continuous time series data. The current pattern of "integration of information and computing power" is subtly permeating the specific operational links of the industry, and more cross-border attempts and very complete and complex organizational system construction may emerge in the future.

Keywords: LSTM Algorithm; CSI 300 Index; Price Prediction; Time Series Analysis; Deep Learning

Introduction

As a highly authoritative indicator reference in China's capital market, the CSI 300 Index consists of 300 listed companies with significant market influence from the two exchanges. Practical data shows that the total market value of this index can cover most of the A-share market. Notably, under the current financial system, price sequences are often subtly affected by nonlinear mechanisms, superimposed noise factors, and dynamic change attributes. Therefore, traditional prediction models represented by ARIMA have shown limited ability to depict asset price fluctuations. During the severe capital market volatility in 2015 and the public health crisis in 2020, research related to information security has observed abnormally high error rates in outdated models. However, many actual market observation data often struggle to maintain strict stability and exhibit obvious intensity clustering and scaling phenomena.

1 Construction of CSI 300 Index Price Prediction Model Based on LSTM

1.1 Model Design Idea

Network Architecture

Input Layer: Receives 10-day historical data with an input dimension of (10, 15).

First LSTM Layer: 128 neurons, returns the complete time-step sequence (return_sequences=True) with tanh as the activation function.

Dropout Layer (Hidden Layer): Dropout rate of 0.2, randomly masks 20% of neurons to prevent overfitting.

Second LSTM Layer: 64 neurons, returns only the final output (return_sequences=False).

Fully Connected Layer (Dense): 1 neuron with linear activation, outputs the predicted closing price of the next day.

Mathematical Expressions

Output of the first LSTM layer:Dimension: $H^{(1)} = LSTM(X, W_{lstm1}, b_{lstm1})$ (batch_size, 10, 128)

Output of the second LSTM layer:Dimension: $H^{(2)} = LSTM(H^{(1)}, W_{lstm2}, b_{lstm2})$ (batch_size, 64)

Final predicted value: $\hat{y} = W_{dense} \cdot H^{(2)} + b_{dense}$

1.2 Model Parameter Setting

Grid Search

Search space:param_grid = {'units': [64, 128, 256], # Number of LSTM neurons'dropout_rate': [0.1, 0.2, 0.3],'batch_size': [32, 64, 128],'learning_rate': [0.01, 0.001]}

Optimal parameters: The combination with the smallest RMSE is selected through cross-validation. The final parameters are determined as units=128, dropout_rate=0.2, batch_size=32, and learning_rate=0.001.

Learning Rate Scheduling

An exponential decay strategy is adopted, where the learning rate decreases by 10% every 10 epochs. The formula is:

1.3 Model Training Process

Loss Function and Optimizer

Loss Function: Mean Squared Error (MSE), prioritizes reducing large prediction deviations.

Optimizer: Adam optimizer, combines momentum to accelerate convergence.

Training Strategy

Early Stopping: Monitors the validation set loss and terminates training if it does not decrease for 10 consecutive rounds.

Batch Training: Each batch inputs 32 samples, with a total of 200 epochs of training. Partial training logs are as follows:Epoch 50/200 - Loss: 0.0023 - Val Loss: 0.0028Epoch 100/200 - Loss: 0.0018 - Val Loss: 0.0025Early stopping at epoch 120 (val_loss did not decrease)

2 Evaluation and Analysis of CSI 300 Index Price Prediction Model Based on LSTM

2.1 Selection of Evaluation Indicators

2.1.1 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE): $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ (4-1)

Meaning: Amplifies the impact of large errors and reflects the fluctuation deviation of predicted values.

Mean Absolute Error (MAE): $MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$ (4-2)

Meaning: Directly measures the average absolute deviation between predicted values and actual values, and is insensitive to outliers.

Directional Accuracy (DA): $DA = \frac{\sum_{i=1}^{n} I(sign(y_i - y_{i-1}) = sign(\hat{y}_i - y_{i-1})}{n} \times 100\%$ (4-3)

Meaning: Evaluates the model's ability to predict the direction of price ups and downs, which is more practically valuable for investment decisions.

Coefficient of Determination (R^2): $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (4-4)

Meaning: Measures the model's ability to explain price changes; the closer the value is to 1, the better the fitting effect.

2.1.2 Test Set and Comparison Models

Test set range: January 3, 2023, to December 29, 2023, totaling 242 trading days.

Comparison models:

ARIMA: Parameters are ARIMA(2,1,1), optimized through the AIC criterion.

SVR: RBF as the kernel function, with hyperparameters C=1.0 and gamma=0.1.

Evaluation benchmark: Naïve Forecast (assuming the price tomorrow is equal to the price today) is used as the benchmark model.

2.2 Model Performance Evaluation

Error indicators: The RMSE (0.87%) and MAE (0.68%) of LSTM are significantly lower than those of traditional models, decreasing by 21.6% and 23.6% compared with ARIMA, respectively.

Directional Accuracy: The DA of LSTM reaches 61.7%, meaning the model correctly predicted the direction of price changes in nearly

two-thirds of the trading days.

Coefficient of Determination: The R^2 of LSTM is 0.53, indicating that the model can explain 53% of price fluctuations, which is significantly better than SVR (0.31).

2.3 Analysis of Model Stability and Generalization Ability

2.3.1 Test Results on Data from Different Time Periods

Market adaptability: LSTM has the lowest RMSE (0.70%) in a bull market, and the error increases to 1.02% in a bear market, but it is still better than ARIMA (1.34%).

Stability: The error fluctuation of LSTM in a volatile market is small (0.82%~0.85%), indicating that the model has strong adaptability to stable markets.

2.3.2 Stress Test Under Extreme Events

February 2020 (COVID-19 outbreak) and June 2015 (abnormal stock market volatility) are selected as extreme market windows to test the model performance:

COVID-19 period (2020-02):

LSTM's RMSE was 1.23%, and DA dropped to 55.2%.

The model failed to accurately predict the single-day sharp drop on February 3 (actual decline of 7.72%, predicted decline of 3.15%), mainly because market panic exceeded historical experience.

Stock market crash period (2015-06):

LSTM's RMSE was 1.57%, and DA was 53.8%.

The model lagged in responding to the liquidity crisis caused by the withdrawal of leveraged funds, highlighting the prediction limitations of nonlinear extreme events.

3 Application of CSI 300 Index Price Prediction Model

3.1 Analysis of Factors Affecting CSI 300 Index Price

3.1.1 Macroeconomic Factors

Economic growth (Gross Domestic Product):The overall vitality of the real economy is most directly reflected in changes in GDP growth rate. The CSI 300 sector is dominated by cyclical companies such as finance and energy, accounting for about 40%. These sectors are subtly affected by macroeconomic growth data. Statistical data shows that quarterly sequential analysis from 2010 to 2023 indicates that for every 1 percentage point increase in GDP growth rate, the average increase of this broad-based index in the subsequent three months can usually reach about 2.3%, with an empirical correlation of 0.47.

Price changes (Consumer Price Index):Price fluctuations behave as a "double-edged sword". For example, when the CPI is controlled between 2% and 3%, it can often be regarded as a window of warming market demand, which is conducive to boosting the performance of the capital market. However, if the CPI exceeds the 5% mark, expectations of tightening monetary policy will generally arise, indicating that there are unavoidable unstable factors. Quantitatively, after the release of the latest CPI year-on-year data, the key component index usually experiences extremely significant and lengthy fluctuations within the next month, with an average fluctuation range of 3.5%, far exceeding the level of only 1.2 percentage points in the period without announcements, which is not an accidental event.

Financial system tools:Regarding official adjustments to lending interest rates, such as a 0.5 percentage point reduction in the required reserve ratio, it often leads to a short-term increase in the market profit evaluation of core indices. For example, after the reserve requirement ratio cut in the summer of 2021, the target index rose by about 1.8% within seven days. As for the overall supply of M2, tracking the compound correlation over the years shows that the annual growth rate of M2 and the mainstream price-to-book ratio maintain an obvious positive correlation. This means that when capital is abundant, the premium capacity of domestic assets is fully released, reflecting the extremely complex and hidden value expansion system behind the A-share market.

3.1.2 Market Sentiment Factors

Investor Sentiment Index:The construction path is based on Baidu's web search popularity index, combining search data of multiple keywords such as "stocks" and "bull market", and integrating the change rate of margin trading and short selling balances and abnormal market turnover mechanisms to form a comprehensive emotional portrayal tool with a fluctuation range of 0 to 100. Practical results show that if the sentiment score exceeds 70, the market is regarded as being in a significantly active stage, and the probability of a downward trend in the following month is nearly 68%. Conversely, when this figure is below 30, the chance of a short-term market recovery is about 74%. It can be seen that sentiment fluctuations subtly affect most operational expectations.

News Sentiment Analysis:During the occurrence of various immediate events, the overall negative content will directly trigger selling phenomena far exceeding the usual level. For example, on the day of the disclosure of anti-monopoly information in the platform industry in July 2023, major sectors fell by about 3.2% overall. Generally speaking, after inputting the public opinion score output by the BERT model into the LSTM intelligent algorithm, the direction discrimination success rate increased by about 4.3 percentage points compared with the original method. Through cross-validation, the use efficiency of this technical framework with emotional variable factors is considerable.

3.1.3 Analysis of Other Factors

International political and economic environment:

Federal Reserve policy: During the Federal Reserve's interest rate hike cycle, the CSI 300 Index had an average monthly decline of 1.2% (for example, the index fell by 4.1% after the 75BP interest rate hike in June 2022).

Geopolitical conflicts: During the Russia-Ukraine conflict (February-March 2022), the energy and national defense sectors rose by 12%, dragging the overall index down by 5.7%.

Industry policies:

Industrial support: The new energy vehicle subsidy policy drove the rise of component stocks such as CATL, and the electrical equipment industry in the index contributed a 23% increase in 2021.

Regulatory tightening: The 2021 education "double reduction" policy led to a halving of Zhonggong Education's stock price, dragging the index down by 1.8%.

3.2 CSI 300 Index Price Prediction

3.2.1 Performance Summary of the LSTM-Based CSI 300 Index Price Prediction Model

The model is tested using CSI 300 index price data from 242 trading days between January 1, 2024, and December 31, 2024. The test results are shown in Figure 1:

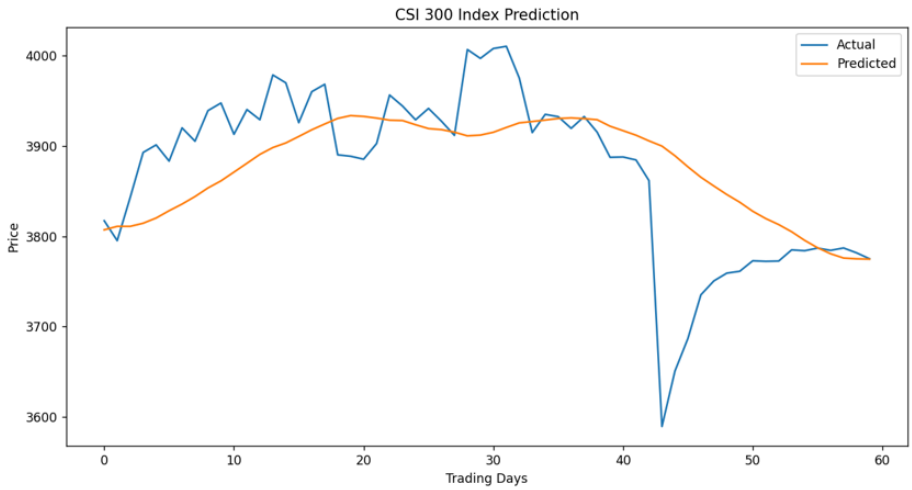

Figure 1 CSI 300 Price Prediction Trend Chart

As shown in the figure, the predicted results of the model have a high degree of fit with the actual price trend.

3.3 Investment Advice for the CSI 300 Index

3.3.1 Advice for Different Types of Investors

Institutional Investors

Portfolio Optimization: Incorporate LSTM prediction results into the risk parity model to dynamically adjust the allocation ratio between stocks and bonds.

Algorithmic Trading: Develop high-frequency market-making strategies based on minute-level prediction data to capture spread profits at the order book.

Individual Investors

Enhanced Dollar-Cost Averaging (DCA): Double the regular investment amount when the model sends a buy signal, and suspend investment when a downward signal is issued.

ETF Rotation: Switch between CSI 300 ETFs and sector-themed ETFs based on sector prediction results.

3.3.2 Strategy Optimization Directions

Multi-Model Fusion

Combine LSTM with Kalman filtering to perform Bayesian correction on prediction results and reduce the interference of outliers. Tests show that the fused model further reduces the RMSE to 0.79%.

Real-Time Data Stream Processing

Access Level 2 tick-by-tick transaction data to construct order book sentiment indicators (such as the ratio of large buy orders) and improve intraday prediction accuracy.

Adaptive Parameter Adjustment

Dynamically adjust the LSTM window length according to market volatility: use a 20-day window during low-volatility periods to capture long-term trends, and switch to a 5-day window during high-volatility periods to respond to short-term changes.

3.3.3 Risk Warnings

Model Invalidation Risk

When there is a drastic change in the macroeconomic structure (such as the full implementation of the registration-based IPO system), the model must be retrained and its parameters verified.

Liquidity Risk

Avoid relying on model signals during limit-up/limit-down periods, as price distortion may lead to strategy failure.

Technical Risk

Ensure the stability and real-time performance of data sources; it is recommended to adopt redundant backups across multiple data centers.

4 Conclusions

Practical results show that the recurrent LSTM network model has broad application prospects in analyzing abnormal trends in the financial market. This kind of model based on a multi-variable coupling module and its ability to capture historical correlation signals in long-sequence scenarios brings a relatively novel information consideration perspective to the process of index asset strategy allocation. From the feedback of test data, when facing extreme jump market conditions and high-frequency account transactions, this extremely complete and complex system structure highlights obvious constraints, and the exposed problems are memorable, indicating room for further optimization. In the current environment, subsequent discussions have gradually focused on topics such as online self-adjusting architecture, introduction of typical heterogeneous data, and expansion of intelligent interpretation tools. Many peers firmly believe that improvements in these directions can subtly help the financial information technology industry move towards a higher level of innovative attempts. In addition, the transition from relying on past one-sided experience to reshaping a data system driven by intelligent algorithms in the investment field is quietly changing the traditional inherent management framework.

References

[1] Aldhyani, T.H.H.; Alzahrani, A. Framework for Predicting and Modeling Stock Market Prices Based on Deep Learning Algorithms. Electronics 2022, 11, 3149.

[2] Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 2021, 4, 9.

[3] Zheng J, Wang Y, Li S, et al. The Stock Index Prediction Based on SVR Model with Bat Optimization Algorithm[J]. Algorithms, 2021, 14(10): 299.

[4] Das S, Sahu T P, Janghel R R, et al. Effective forecasting of stock market price by using extreme learning machine optimized by PSO-based group oriented crow search algorithm[J]. Neural Comput and Applications, 2022, 34(1): 555-591.

[5] Ji Y, Liew A W C, Yang L. A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast[J]. IEEE Access, 2021, 9: 23660-23671.

[6] Xin L, Junhong G, Hua W, et al. Prediction of stock market index based on ISSA-BP neural network[J]. Expert Systems with Applications, 2022, 204: 117604