Stock Return Prediction and Strategy Implementation Based on LightGBM

Authors: Zhu Zhixuan

Date: 2026-03-12T16:07:00+00:00

Abstract

Addressing the pain points of insufficient accuracy and weak practical applicability in traditional stock prediction models, this paper takes the Hong Kong stock Tencent Holdings (00700.HK) as the research object. We employ the LightGBM algorithm to construct a stock return prediction model and design a supporting quantitative investment strategy based on trend following and stop-loss mechanisms. The study utilizes daily data from 2018 to 2023 as the training sample, with the period from 2024 to 2025 serving as the out-of-sample backtesting interval. Empirical results demonstrate that the model achieves a 60% accuracy rate in predicting price direction, and the strategy realizes an annualized return of 36.65% with a maximum drawdown of only 12.21%. These results significantly outperform the buy-and-hold benchmark, providing a low-threshold Hong Kong stock quantitative investment solution for individual investors.

Full Text

Preamble

Stock Return Prediction and Strategy Implementation Based on LightGBM Zhu Zhixuan (School of Future Technology, South China University of Technology, Guangzhou, Guangdong 511442)

Abstract

To address the issues of insufficient accuracy and poor implementability in traditional stock prediction models, this study focuses on Tencent Holdings (00700.HK) in the Hong Kong stock market. We employ the LightGBM algorithm to construct a stock return prediction model and design a corresponding quantitative investment strategy based on trend following and stop-loss mechanisms. Using daily line data from 2018 to 2023 as the training sample and 2024 to 2025 as the out-of-sample backtesting period, empirical results demonstrate

that the model achieves a 60% accuracy rate in predicting price direction. The strategy yielded an annualized return of 36.65% with a maximum drawdown of only 12.21%, significantly outperforming the buy-and-hold benchmark. This research provides a low-barrier quantitative investment solution for individual investors in the Hong Kong stock market.

Keywords: LightGBM; Stock return prediction; Quantitative investment strategy; Hong Kong stock market; Machine learning applications

Stock prices are driven by multiple factors, including macroeconomics, industry cycles, market sentiment, and capital flows, exhibiting significant non-linear, non-stationary, and high-volatility characteristics. Accurate return prediction is a core prerequisite for investors to avoid risks and obtain stable returns, making it a vital research direction in financial engineering [**?**]. Traditional time-series statistical models are strictly limited by linear frameworks, leading to performance bottlenecks in stock prediction. While deep learning models possess strong non-linear fitting capabilities, they generally suffer from high computational requirements, complex parameter tuning, and poor interpretability, making them difficult for individual investors and students to use. Furthermore, although traditional GBDT algorithms are suitable for structured financial data, they struggle to balance training efficiency and prediction accuracy when dealing with high-dimensional price-volume features and large-sample time-series scenarios [**?**].

To address these challenges, this study utilizes the LightGBM algorithm to build a stock return prediction model for Tencent Holdings (00700.HK) and designs an implementable enhanced quantitative investment strategy. By optimizing the feature engineering system, time-series training process, and trading strategy design, we achieved effective return prediction and alpha generation within a standard computing environment. This provides a convenient and reproducible quantitative investment reference for individual investors [**?**].

Research into stock return prediction has developed into three main streams, each with its own strengths and limitations. Traditional statistical models, represented by ARIMA and GARCH, offer simplicity and strong interpretability but are limited by linear assumptions and fail to capture non-linear fluctuations, leading to insufficient accuracy in high-noise environments [**?**]. Machine learning models have become mainstream due to their powerful non-linear fitting capabilities. Gradient Boosting Decision Tree (GBDT) algorithms, such as LightGBM and XGBoost, offer a balance of accuracy, efficiency, and robustness for structured data while remaining accessible to individual investors [**?**]. Deep learning and hybrid models can capture long-range dependencies but require massive data, are prone to overfitting, and lack interpretability, making them less accessible to retail investors [**?**].

1.1 Core Algorithm: LightGBM

This study employs the LightGBM (Light Gradient Boosting Machine) algorithm, an efficient ensemble learning method developed by Microsoft based on GBDT. It is specifically designed for regression and classification tasks on structured data and is highly adaptable to financial time-series prediction [**?**]. Its core advantages include: 1. **Histogram Optimization:** Discretizes continuous features into fixed bins, significantly reducing computational complexity while preserving distribution information. 2. **Leaf-wise Growth Strategy:** Unlike traditional level-wise growth, LightGBM uses a depth-limited leaf-wise strategy that prioritizes splitting the leaf with the maximum information gain, leading to faster convergence and higher accuracy. 3. **Strong Regularization:** Utilizes L1 and L2 regularization and minimum gain thresholds to suppress overfitting, which is essential for noisy financial data. 4. **Sampling Enhancement:** Supports feature and data sampling to simulate ensemble effects and improve prediction stability. 5. **Time-Series Compatibility:** Integrates seamlessly with time-series cross-validation, ensuring that "past data predicts the future" without information leakage.

1.2 Research Framework and Data Partitioning

The research follows a core logic of **"Feature Engineering Construction - Time-Series Model Training -Quantitative Strategy Backtesting."** The target is daily market data for Tencent Holdings (00700.HK) sourced from the Wind terminal. Fields include opening, high, low, and closing prices, as well as trading volume and turnover. All prices are forward-adjusted, and non-trading days are excluded.

The full sample period spans from January 1, 2018, to April 24, 2025, divided into three non-overlapping subsets: - **Training Set:** January 1, 2018, to June 30, 2023 (for parameter fitting). - **Validation Set:** July 1, 2023, to December 31, 2023 (for hyperparameter tuning and early stopping). - **Test Set (Out-of-Sample):** January 1, 2024, to April 24, 2025 (for strategy backtesting and verification).

1.3 Feature Engineering System

Based on raw daily data, we constructed a 62-dimensional feature system across six categories: basic price features (6), technical indicators (21), volume-price relationship features (12), time-cycle features (8), and advanced derivative features (15). These cover price trends, volume changes, cyclical effects, and momentum reversals. Although some features exhibit linear correlation, tree-based models are insensitive to multicollinearity. Empirical validation showed that removing correlated features degraded performance; thus, all 62 features were retained.

1.4 Model Training Scheme

To address the non-stationarity of stock prices, we transformed absolute price prediction into a prediction of the log return for the next 6 trading days. This resolves unit root issues and aligns the prediction target with the strategy' s holding period. The label is constructed as:

$$target_t = \ln \frac{close_{t+6}}{close_t}$$

where $close_t$ is the closing price on day t.

To prevent look-ahead bias, we used a rolling origin validation scheme instead of random K-fold cross-validation [**?**]. The model is trained on the past 500 trading days to predict the next 60 days of returns, with the model updating every 60 days to capture the latest market patterns. Following grid search and validation tuning, the final parameters were set: regression task with MSE loss, `num_{leaves}=31`, `learning_{rate}=0.03`, `lambda_l1=0.1`, `lambda_l2=0.1`, `feature_{fraction}=0.8`, and `bagging_{fraction}=0.8`.

1.5 Quantitative Trading Strategy Design

We designed a trend-following strategy with a hard stop-loss, accounting for Hong Kong market transaction costs and T+1 rules: 1. **Buy Signal:** If the predicted 6-day log return $> 1.8\%$, execute a full-position buy at the next day' s opening price. 2. **Regular Sell Signal:** If the predicted 6-day log return $< -0.6\%$, execute a full-position liquidation at the next day' s opening price. 3. **Stop-Loss Signal:** If the opening price drops 5% below the cost basis, execute an immediate liquidation to control maximum loss.

2 Model Performance Evaluation

The model was evaluated on numerical accuracy and directional judgment: 1. **Numerical Accuracy:** The Mean Squared Error (MSE) was 0.003453 and Mean Absolute Error (MAE) was 0.042441. An average error of approximately 4.24% is considered reasonable for short-term stock return prediction. 2. **Directional Accuracy:** Reached 60.00%. This is significantly higher than the 50% of random guessing, indicating the model effectively captures key volatility features.

The prediction capability was further verified via a confusion matrix [Figure 1: see original paper]. Results show the highest precision for "hold" signals (79 samples), with clear differentiation for buy and sell signals. This reliability helps filter noise and avoid erroneous trades.

3.3 Backtesting Results and Analysis

With an initial capital of 100,000 CNY, the strategy completed 9 trades during the backtest period (January 1, 2024 –April 24, 2025). The net asset value (NAV)

trend is shown in [Figure 2: see original paper]. The strategy NAV grew steadily and remained above the benchmark for most of the period. During major market pullbacks in Q3 2024 and Q1 2025, the strategy exhibited significantly smaller drawdowns.

1. **Profitability:** The strategy achieved an annualized return of 36.65%, outperforming the benchmark (34.63%) by 2.02 percentage points. The total absolute return was 62.53%, providing an alpha of 3.72%. With a 55.56% win rate, the strategy demonstrates stable earning power.
2. **Risk Control:** As shown in [Figure 3: see original paper], the strategy' s maximum drawdown was only 12.21%, compared to the benchmark' s 23.49%. By exiting during extreme market conditions (e.g., April 2025) via the stop-loss mechanism, the strategy effectively preserved capital [**?**].
3. **Risk-Adjusted Returns:** The Sharpe ratio was 1.86 (vs. 1.0 for the benchmark) and the Calmar ratio was 3.00 (exceeding the target of 1.5). The profit-to-loss ratio was 2.31, meaning average gains were more than double average losses.
4. **Stability:** Of the 9 trades, 5 were profitable. The average profit per trade was 7,011.94 CNY. Losses were strictly limited by the stop-loss rules, verifying the strategy' s robustness.

3.4 Sensitivity Testing

Sensitivity tests on the buy/sell thresholds showed that within reasonable ranges, the strategy consistently outperformed the benchmark with drawdowns remaining under 15%. This suggests the strategy' s success is due to its core logic rather than over-optimization of parameters [**?**].

4 Conclusion and Outlook

This study demonstrates that LightGBM can effectively handle high-dimensional financial time-series data to provide high-precision stock return predictions on standard hardware. The resulting quantitative strategy achieved superior returns and balanced risk-reward profiles through effective stop-loss mechanisms.

Future research may focus on: 1) incorporating alternative data such as sentiment and macro indicators; 2) exploring dynamic position management and multi-asset diversification; and 3) developing adaptive parameter optimization to improve generalization across different market cycles.

Note: Figure translations are in progress. See original paper for figures.