Introduction

In the stock market, "accumulation" refers to the strategic buying of stocks by large investors at low prices in preparation for future price increases. Identifying stocks nearing the end of this accumulation phase is of significant strategic importance to investors. However, traditional methods often struggle to accurately capture this complex process.

This study aims to develop a machine learning-based model to identify "strong stocks" nearing the end of their accumulation phase. Our approach combines traditional technical analysis indicators (such as Bollinger Bands and Average True Range, ATR) with advanced machine learning algorithms (such as Isolation Forest anomaly detection). This method not only extends the idea of applying machine learning to fundamental analysis as proposed by Zhang et al. (2020) [1], but also expands the framework of deep learning applications in finance suggested by Ozbayoglu et al. (2020) [2].

My model achieves its goal through three steps: first, identifying abnormal fluctuations in stock prices using technical indicators; second, applying machine learning algorithms to detect anomalous patterns within these fluctuations; and finally, analyzing the temporal distribution and characteristics of these anomalous patterns to identify stocks that may be nearing the end of their accumulation phase.

In addition, I have developed a comprehensive parameter optimization and backtesting system along with rich visualization tools. I believe that this entire process can provide a comprehensive, reliable, and easily understandable framework for analyzing abnormal market behavior, offering more valuable decision support to investors. Furthermore, this study draws on the computational basis of technical analysis as studied by Lo et al. (2000) [3].

Model Design

Dataset Description

The dataset used in this study contains information from the stock issuance date up to August 10, 2020. The data includes stock ID, trading date, opening price, highest price, lowest price, closing price, previous closing price, price change percentage, trading volume (in lots), and trading amount (in thousands of yuan). To facilitate better analysis and comparison of the output results, I selected a subset of stocks and categorized them into three sectors based on the companies they belong to: chemicals, healthcare, and estate.

Data Preprocessing

Steps for data preprocessing include:

Data Reading:

Use pandas to read the CSV files, ensuring the inclusion of necessary columns (highest price, lowest price, closing price, trading volume, trading date).

Date Handling:

Convert the trading date to datetime format and sort the data by date

Technical Indicators

Average True Range (ATR): Calculated using a 14-day period.

Bollinger Bands: Using 20-day moving average with a standard deviation multiplier of 2.

Relative Strength Index (RSI): Calculated using a 14-day period.

Volume Moving Average: Using a 20-day period.

MACD: Calculated using 12-day and 26-day Exponential Moving Averages (EMA).

Money Flow Index (MFI): Calculated using a 14-day period.

The selection of these technical indicators is based on the study by Tsantekidis et al. (2017), which demonstrated their effectiveness in predicting stock prices [4].

Feature Engineering

Additional Features: Include features for price change rate and volume change rate.

Handling Missing Values: Remove rows containing missing values.

Anomaly Detection Model

Inspired by the Isolation Forest algorithm proposed by Liu et al. (2008) [5], I used the Isolation Forest algorithm for anomaly detection. Key features include ATR, RSI, volume moving average, MACD, MFI, price change rate, and volume change rate (technical indicators).

Model Parameters:

- **contamination:** 0.1, assuming approximately 10% of the data points may be anomalous.
- **random_state:** 42, to ensure reproducibility of results.

Backtesting Process

The backtesting system follows these steps:

Parameter Grid Search: Optimize the following parameters

- **Minimum Duration:** [10, 20, 30] days
- **Maximum Duration:** [60, 90, 120] days
- **Price Change Threshold:** [0.05, 0.1, 0.15]
- **Volume Change Threshold:** [0.5, 1.0, 1.5]

For each set of parameters, filter anomalies using the given parameters, and then conduct simulated trading: buy on the next trading day after detecting an anomaly, hold for a fixed period (e.g., 30 days), and then sell.

Calculate the return for each trade and overall strategy metrics through the simulated trades, using the Sharpe Ratio as the primary evaluation metric.

The selection of the Sharpe Ratio as the evaluation metric is based on Sharpe's (1994) research [6]. The Sharpe Ratio is a crucial indicator for measuring portfolio performance, as it takes into account both returns and risk. The formula for calculating the Sharpe Ratio is as follows:

$$\text{Sharpe Ratio } = \frac{(Rp - Rf)}{\sigma p}$$

Where:

Rp is the expected return of the portfolio

Rf is the risk-free rate (in our model, we assume this to be 0)

σp is the standard deviation of the portfolio's returns

A higher Sharpe Ratio indicates a better risk-adjusted return, meaning the strategy yields higher excess returns per unit of risk taken. Based on the Sharpe Ratio, select the best-performing parameter combination.

Multidimensional Analysis

I conducted the following multidimensional analyses:

Price Change Analysis: Calculate the price change amplitude during the anomaly periods.

Volume Change Analysis: Calculate the change in volume relative to the 20-day moving average during the anomaly periods.

Anomaly Duration Analysis: Calculate the duration of each anomaly in days.

Results Visualization

To better present the output results and facilitate subsequent analysis, I created three types of visualizations: (inspired by the techniques used by Bao et al. (2017) in financial time series analysis [7].)

Anomaly Distribution Scatter Plot: Shows the relationship between price change, volume change, and duration.

Stock Price Trend Chart: Displays the stock price trend 30 days before and after the anomaly.

Technical Indicator Time Series Chart: Shows the time series of stock price, RSI, anomaly markers, and peaks.

Experimental Results and Analysis

– Using Stock Number 600529 as an Example

Report Content

Upon completion of the analysis, the model automatically generates:

- **SQLite Database:** Stores the analysis results for each stock, including anomaly data, optimal parameters, and statistical information.
- **Text Report (analysis_report_[Stock_Name].txt):** Contains analysis time, details of the optimal parameters, statistical information, and anomaly statistics.
- **Graphical Report:** Includes anomaly distribution plots and time series plots of anomalies (analysis_report_[Stock_Name].png), as well as stock price trend charts (anomaly_price_[Stock_Name]_[Index].png).

Text Analysis

Optimal Parameters:

- Minimum Duration: 30 days
- Maximum Duration: 60 days
- Price Change Threshold: 15%
- Volume Change Threshold: 50%

Strategy Performance:

- Sharpe Ratio: 2.375 (high, indicating good risk-adjusted returns)
- Average Return: 20.40% (average return per trade)
- Win Rate: 100% (all trades were profitable)
- Profit-Loss Ratio: Infinite (since there were no losing trades)

Data Statistics:

- Total Data Points: 4,313
- Data Date Range: June 28, 2002, to August 10, 2020
- Average Price: ¥14.996921
- Average Volume: 40,606.609316 (shares)

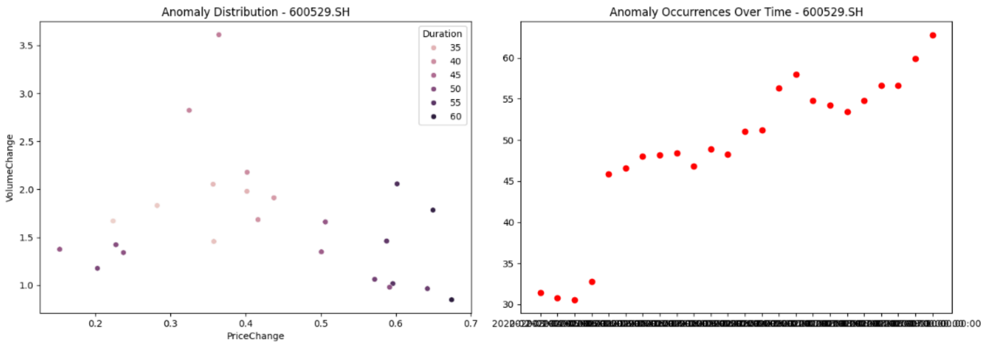
Anomaly Analysis:

- Total Anomalies: 24
- Average Price Change: 43% (average price increase during anomalies)
- Average Volume Change: 165% (average volume increase during anomalies)

Summary

This report shows that the analyzed stock exhibits significant characteristics, and the strategy has produced good risk-adjusted returns (high Sharpe ratio) and a relatively high win rate. Anomalous events are usually accompanied by significant price increases and volume surges, which may indicate potential "accumulation" behavior.

Graphical Analysis



analysis_report_600529.SH

This graphical report provides a time distribution chart of the stock's anomalous behavior, which is the first step in analyzing the anomaly patterns and potential "accumulation" behavior of stock 600529.SH.

Features and Significance:

Anomaly Distribution Plot (Left):

X-axis: Price Change (PriceChange)

Y-axis: Volume Change (VolumeChange)

Color: Represents the duration of the anomaly (Duration)

The left plot shows the distribution of anomalous events in terms of price change and volume change. For stock 600529, the anomalies are mainly distributed in the range of higher price changes and lower volume changes. Due to the relatively low volume change, a longer period is required to achieve a significant price change.

Anomaly Occurrence Time Series Plot (Right):

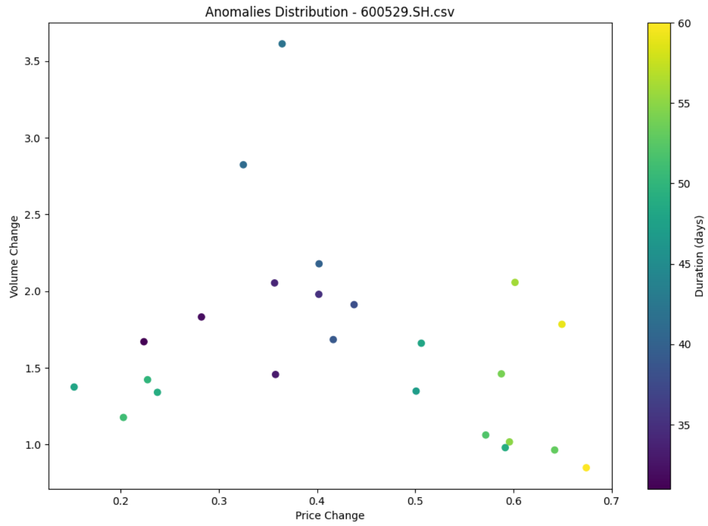
X-axis: Trading Date

Y-axis: Closing Price

Red Dots: Each red dot represents a detected anomalous event

The right plot shows the time distribution of anomalous events throughout the analysis period and their relationship with the stock price. For stock 600529, anomalies tend to occur during periods of higher closing prices. Since this stock exhibits an overall upward trend, it can be noted that anomalies mainly appear in the period from 2015 to 2020.

This visualization highlights that the stock's anomalous behaviors, which could indicate "accumulation" phases, are more frequent during significant price rises. The detailed distribution of price and volume changes, along with the duration of anomalies, helps to identify these potential accumulation behaviors effectively.



anomalies_distribution_600529.SH.csv

This chart provides an intuitive way to understand the anomalous behavior characteristics of the stock, helping to identify potential "accumulation" patterns and market anomalies.

Features and Significance:

X-axis (Price Change): Represents the price change during the anomaly period. According to the previous analysis report, the average price change is 43%.

Y-axis (Volume Change): Represents the volume change during the anomaly period. The average volume change is 165%.

Color (Duration): Represents the duration of the anomaly in days. The color bar shows the specific range of days, which can vary from 30 to 60 days (based on the optimal parameters).

Scatter Points: Each point in the chart represents a detected anomaly. There should be a total of 24 points, corresponding to the 24 anomalies mentioned in the report.

Distribution Characteristics:

The points are concentrated in the lower middle part of the chart, indicating that most anomalies have moderate price increases and significant volume increases.

The variation in color of the points indicates a wide range of anomaly durations.

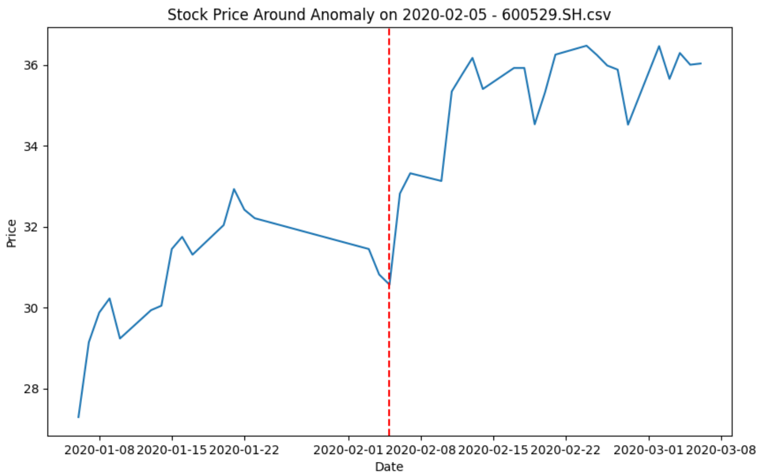
There is a notable outlier, indicating a particularly significant anomalous event.

Pattern Recognition:

Points near the Y-axis but far from the X-axis: These might indicate situations where volume increased significantly but price changes were minimal, potentially characteristic of early "accumulation" phases.

Points near the top-right corner: These might indicate the late stages of "accumulation," where both price and volume see significant increases.

This chart helps us better understand and identify the anomalous behavior characteristics of the stock by showing the relationship between price change, volume change, and duration, thereby identifying potential "accumulation" patterns and market anomalies.



anomaly_price_600529.SH

This chart displays the stock price trend before and after each detected anomaly. Such charts are generated for each detected anomaly, and by comparing these charts, one can discover common patterns or features, which are helpful for understanding and predicting the stock's "accumulation" behavior.

Features and Significance:

Time Range: Shows the stock price trend for 30 days before and after the anomaly date, totaling about 60 days of data.

X-axis: Represents the dates, ranging from 30 days before to 30 days after the anomaly.

Y-axis: Represents the closing price of the stock.

Red Dashed Line: Marks the specific date of the anomaly.

Blue Line: Represents the trend of the stock price over this period.

Potential "Accumulation" Characteristics:

If the stock price is relatively stable before the anomaly and starts to rise significantly after the anomaly, it may indicate the end of "accumulation" and the beginning of a price increase.

If the stock price remains relatively stable both before and after the anomaly, but there is a noticeable increase in trading volume (even though not directly shown in the chart), it may indicate ongoing "accumulation."

This chart provides a visual representation of the stock price movements around detected anomalies, helping to identify potential "accumulation" patterns and behaviors.

Analysis of "Stocks Approaching the End of Accumulation"

Based on the experimental results, the following aspects can be used to identify stocks that might be approaching the end of accumulation:

Anomaly Pattern Analysis:

Focus on Stocks with Anomalies Not Yet Showing Significant Price Increases: Look at stocks identified as anomalous but have not yet experienced substantial price rises.

Examine Anomaly Statistics in analysis_report_[Stock_Name].txt: Pay particular attention to average price change and average volume change.

Volume Analysis:

Look for Stocks with Gradually Increasing Volume but Little Price Change: Identify stocks where trading volume is rising, but price changes remain minimal.

Review the Anomaly Distribution Scatter Plot in analysis_report_[Stock_Name].png: Focus on points where volume change is high but price change is relatively small.

Price Consolidation:

Search for Stocks with Price Stabilization Within a Certain Range: Identify stocks where the price is consolidating in a narrow range.

Observe Stock Price Trends in anomaly_price_[Stock_Name]_[Index].png: Look for periods with minimal price fluctuations.

Anomaly Frequency:

Monitor Stocks with Increasing Frequency of Anomalies: Identify stocks where the frequency of detected anomalies is rising.

Check the Time Series Plot in analysis_report_[Stock_Name].png: Look at the density of anomaly points to assess increasing anomaly frequency.

Optimal Parameter Analysis:

Review Optimal Parameters in analysis_report_[Stock_Name].txt: If the optimal parameters indicate shorter durations and higher volume change thresholds, this may suggest that the stock is in the later stages of accumulation.

By focusing on these aspects, you can identify stocks that might be nearing the end of their accumulation phase and potentially poised for significant price movements.

References

[1]Zhang, Y., Chu, G., Li, D., Zhao, M., & Xie, J. (2020). Machine Learning for Stock Prediction Based on Fundamental Analysis. Expert Systems with Applications, 160, 113704. https://doi.org/10.1016/j.eswa.2020.113704

[2] Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications : A survey. Applied Soft Computing, 93, 106384. https://doi.org/10.1016/j.asoc.2020.106384

[3] Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The Journal of Finance, 55(4), 1705-1765.

[4] Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017, July). Forecasting stock prices from the limit order book using convolutional neural networks. In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 7-12). IEEE.

[5] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.

[6] Sharpe, W. F. (1994). The sharpe ratio. Journal of portfolio management, 21(1), 49-58.

[7] Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one, 12(7), e0180944.