Machine Learning in Multifactor Stock Selection: Based on Japanese Pharma Sector

Yuetian Zhang

Contents

1 Introduction

For our final project in Equity Markets and Quantitative Trading (EN.553.441/641), we, as a group, focused on creating an innovative trading algorithm that incorporates machine learning and historical back-testing to practice some of the methods and ideologies studied in class. As our area of focus for the algorithm, we ventured into the Japanese biopharmaceutical industry; in early conversations with Professor Miller (an industry expert in the Japanese financial markets), we elected this area of focus due to its niche nature, and novel opportunities for further learning in foreign equity markets. In this report, we will begin by detailing the historical context of the Japanese equity markets, particularly the pharmaceutical sector. We will then explain our data collection and cleaning methods and subsequent selection of fundamental factors best suited for machine learning and back-testing. Following our factor analysis, we will provide insight into our machine learning and implementation processes, highlighting the results. This paper will conclude with some final analysis, and growth opportunities for our work going forward.

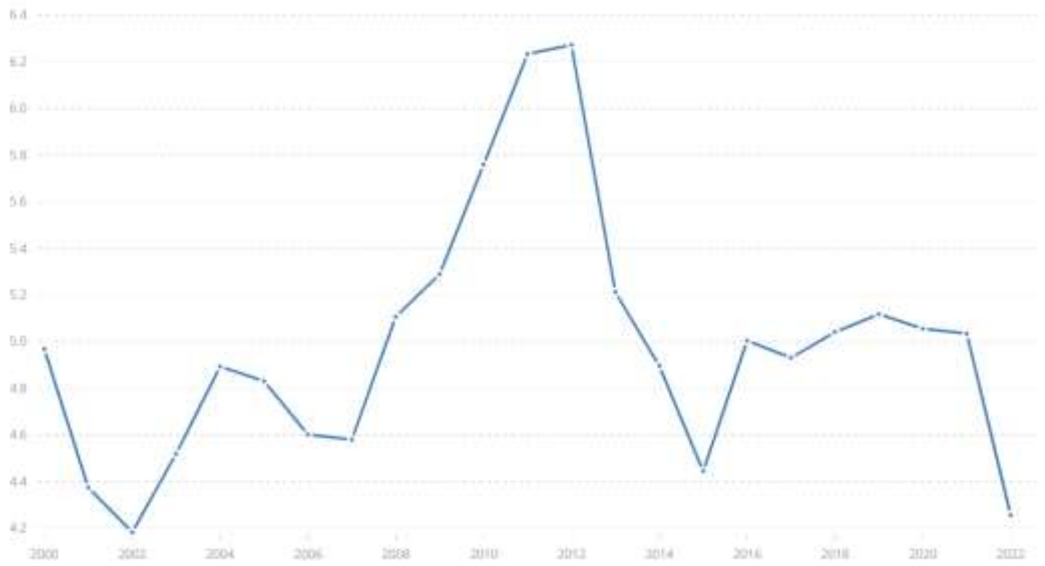
2 Japanese Market Analysis

2.1 Economic Overview

Japan's economic narrative has recently been marked by a contrast between growth and challenges. Although Japan's GDP grew by 1.9% in 2023, the yen has depreciated sharply due to the U.S. interest rate hikes, and the yen's exchange rate against the U.S. dollar has been severely affected. However, this growth masked a contraction in the second half of the year. Japan's economy also faces the challenge of passively imported inflation, which hit 3% in 2023, a 20-year high, signaling a departure from Japan's history of price stability.

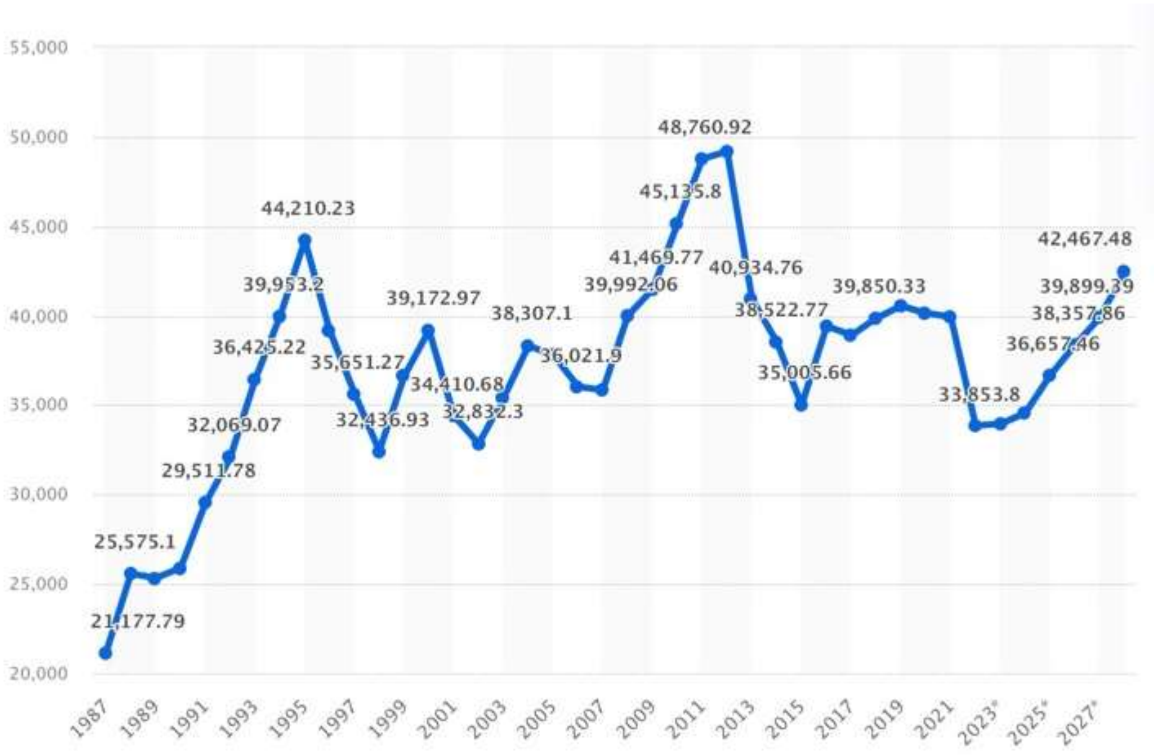
As the Bank of Japan considers abandoning its long-term negative interest rate policy by 2024, it has already engaged in forms of economic tightening through modifications to its yield curve control policy. The move is part of a broader shift as Japanese investors, faced with rising FX hedging costs due to monetary policy disagreements with other central banks, are turning to capital repatriation, as evidenced by large net sales of foreign bonds in 2022.

Kazuo Ueda is about to be appointed as the new governor of the Bank of Japan, adding to the uncertainty about the future of monetary policy. At the same time, the combined effects of the Bank of Japan's policies and global quantitative tightening are causing a withdrawal of global liquidity, with potential consequences for international markets. Overall market risks are like a riptide, and the current optimism may be masking the risk of liquidity withdrawals. Given these dynamics, investors should remain vigilant for potential liquidity traps in the market.



Japan: Nominal Gross domestic product (GDP)

From 2003 to 2023, Japan's nominal GDP experienced a series of fluctuations. The period from 2003 up until the global financial crisis in 2008 showed an initial decline followed by substantial growth, peaking just before the crisis hit. This peak was followed by a sharp decline in 2008-2009, reflecting the global financial turmoil. A recovery phase is observed post-2009, with GDP rising until around 2011, but the devastating Tohoku earthquake and tsunami, along with the Fukushima disaster, caused another significant economic downturn. The years following 2011 show a gradual recovery and stabilization, which can be associated with the implementation of "Abenomics," a series of economic policies introduced by Prime Minister Shinzo Abe. The trend from 2015 to the onset of the COVID-19 pandemic in 2020 displays growth with some volatility, indicative of an adapting economy facing various global and domestic challenges. The pandemic brought about a noticeable downturn, signaling the economic impact of the health crisis. While there was a partial recovery post-2020, the graph ends with a sharp decline in 2023, the reasons for which would require further data and context to fully understand.



Japan: Gross domestic product (GDP) per capita in current prices

Plot from:https://www.statista.com/statistics/263578/gross-domestic-product-gdp-of-japan/

Starting from 2003, there is a gradual increase in GDP per capita, reaching a pre-financial crisis peak in 2007. This indicates an expanding economy during the early 2000s. The impact of the 2008 financial crisis is evident with a subsequent drop, reflecting the global economic downturn. However, a recovery trend is visible shortly after, with growth resuming until around 2011, when the Tohoku earthquake and the ensuing events may have contributed to a slight dip. The years following 2011 until around 2015 show recovery and a relatively stable increase in GDP per capita. The latter part of the decade appears to display a plateau, with minor ups and downs, indicating an economy experiencing slower growth or stabilization. Entering into the early 2020s, there's a notable decline which could be attributed to the economic repercussions of the COVID-19 pandemic. The latest figures in the graph show a modest recovery, although not to the levels seen in the late 2010s.

Comparing this GDP per capita graph to the nominal GDP trend, it is evident that while the nominal GDP graph displayed sharper peaks and troughs, the per capita graph shows a somewhat smoother progression. The nominal GDP graph’s sharp peak around 2007 followed by a steep decline illustrates the immediate impact of the financial crisis. However, the per capita graph, while still showing this trend, appears less volatile, which could be due to population changes mitigating the per capita figures. The dip in the nominal GDP graph around 2011 is also reflected in the per capita graph, likely due to the earthquake and tsunami. Both graphs display recovery post-2011, but the per capita graph shows a more consistent growth pattern, possibly due to gradual improvements in the economic standing of individuals. The effects of the COVID-19 pandemic are visible in both graphs with a clear downturn, but the per capita graph shows a less pronounced drop, suggesting that while the overall economy was hit, the impact per individual might have been cushioned by other economic factors or interventions. Both graphs end with a downturn around 2023, which could indicate current economic challenges or uncertainties facing Japan.

2.2 Japanese pharmaceutical market

The Japanese pharmaceutical market is currently facing major structural shifts and, like the broader pharmaceutical market, is challenged by rising healthcare spending. The Japanese pharmaceutical industry's response has included increased penetration of generic drugs, pricing pressure, restricted access for healthcare professionals, and a push for regionalization of healthcare. professionals, and a push for regionalization of health care. To be more detail:

Generic Drug Penetration: The use of generic drugs in the Japanese pharmacy market has increased dramatically, with its volume share increasing significantly from 48.8% in FY2013 to 69.9% in FY2017. The shift signals Japan's move away from its previous reliance on brand-name drug products that have traditionally enjoyed "long tail" profitability even after patent expiration.

Government-led drug price adjustments occur every two years, with cuts ranging from 5% to 7%. Annual price adjustments for drugs with higher pharmacy profit margins and other pricing reforms have added to pressure and signal that the pharmaceutical market is adopting tighter cost-containment measures.

Access and publicity campaigns

Japan’s Ministry of Health, Labor and Welfare’s monitoring of medical representatives has led to increasingly strict publicity guidelines due to cases of suspected violations. This development has limited the ability of pharmacies and pharmaceutical companies to engage directly with physicians, impacting traditional sales models.

Regionalized and Comprehensive Care

The establishment of community-based integrated care systems, designed to provide comprehensive health care services from hospitals to home services, reflects the structural reorganization within the healthcare services from hospitals to home services, reflects the structural reorganization within the health care sector. This approach requires the pharmaceutical market to move towards a more localized and region-specific strategy.

Japanese pharmacies and pharmaceutical companies are recalibrating their operating models in line with global standards while developing localization strategies to meet regional healthcare needs. This dual approach is critical to maintaining profitability and addressing the unique characteristics of the Japanese healthcare market.

Impact on the future

For the Japanese pharmaceutical market, GDP growth has had a positive impact on the market, while imported inflation, possible monetary policy adjustments, increased penetration of generic drugs, government adjustments to drug prices, and increased monitoring by medical representatives have all had a negative impact Influence. On the other hand, regionalization and comprehensive care of healthcare, adjustment of global standard operating models and localization strategies have a positive impact on the market.

3 Data Processing and Factor Construction

3.1 Data Cleaning Overview

3.1.1 Data Sources and Factor Construction

One of our biggest hurdles early in the project was collecting all of our data and organizing it in a way that would enable our machine-learning methods to be effective. In order to find this data, we first had to analyze the companies that comprise Japan's pharmaceutical industry. The TP.PHRM Index, short for Tokyo Stock Exchange TOPIX Pharmaceutical Index, covers the 37 largest companies in the pharmaceutical sector of the Japanese market and provided us with a foundation of companies, and performance history from which to build our model.

Following our discovery of the TP.PHRM we began our initial analysis of the data and the returns to individual factors. Utilizing Bloomberg, we were able to extract fundamental data for each company within the TP.PHRM index, recording current shares outstanding, free float percentage, average volume year-1, revenue per basic share, market capitalization, earnings per basic share, revenue year-1, price to earnings ratio, return on equity, beta to index, and net sales year-5. We recorded all of the key factors once yearly starting in 2005 until 2024, and compiled the performances into spreadsheets for factor analysis. For the most simple data, being daily price and volume, we were able to use Yahoo Finance's API to access these figures quite simply.

Subsequently, we use these data to construct factors, with the specific methods of factor construction detailed in the appendix.

3.1.2 Standardizing Factor Data and Handling Missing Values

Normalization: For the original factor data values, we perform standardization on the cross-section to achieve a mean of 0 and a standard deviation of 1.

Winsorization: Subsequently, we apply winsorization (with a cap of 0.5) to limit extreme values.

Filling missing data: For any missing data, we fill in the gaps using the mean of the cross-section.

3.2 Factors Derived from Volume and Price Data

For factors derived from daily frequency volume and price data, we varied the look-back period for each factor to calculate its returns and identify the most optimal factors. In this section, we analyzed the mean reversion factor, momentum factor, CPV (Correlation of Price and Volume) factor, and beta with index factor, with detailed calculation processes provided in the appendix. We evaluated the performance of each factor over different look-back periods, as illustrated in the accompanying diagram.

Factor Construction:

Factor Name	Data souce	Data Used	Data Processing Method
mr/ mom (mean_reversion and momtumn factor)	yahoo finance	Adj Close	Mean Reversion Factor: $mr = -\frac{1}{N}\sum_{i=0,1,\ldots,N} R_{t-i}$ $mom = -mr$ where: - N is the number of trading days in the calculation window (use one week, N = 5), - R_{t-i} is the daily return on day t-i
cpv (correlation between price and volume)	yahoo finance	Adj Close; volume	CPV: correlation between price change (return) and volume $CPV_t = corr(R, Vol\_Norm\,; window)$ where: - Vol_Norm is volume normalized by dividing the rolling average of the volume over one week. - R is the daily return - window use 1 week and 1 month

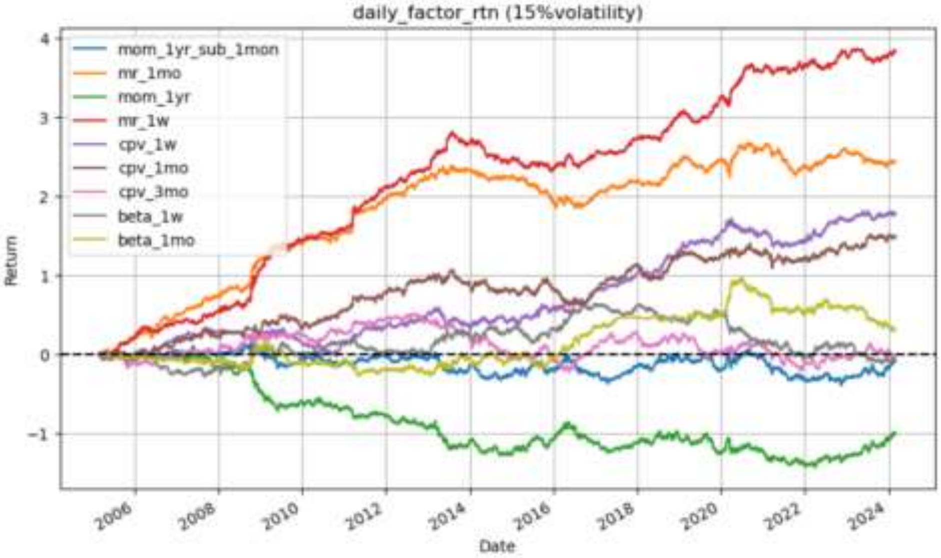
beta_tpphrm	yahoo finance	Adj Close	Computed as the slope coefficient in a time-series regression of excess stock return against index return.

Sharpe Ratio for each factor:

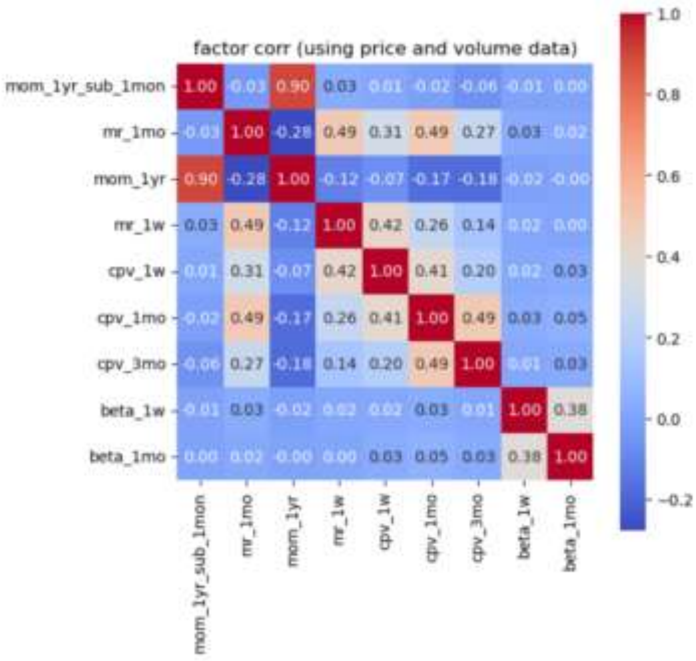
Factor Name	Sharpe Ratio
mr_1w	1.37
mr_1mo	0.87
mom_1yr_sub_1mon	-0.03
mom_1yr	-0.36
cpv_1w	0.63
cpv_1mo	0.54
cpv_3mo	-0.01
beta_1w	-0.03
beta_1mo	0.11

(mom: Momentum, mr: mean_reversion; cpv: Correlation of Price and Volume)

Factor Returns for each factor:



Correlation Heatmap:



Interestingly, the momentum factor appeared to exhibit negative predictive power, aligning with many previous studies on the Japanese market (e.g., AQR paper), leading us to utilize the negative of the momentum factor to represent mean reversion factors (mr_1w, mr_1yr). In selecting high Sharpe ratio factors, we also considered their correlation because we aimed to find factors representing different characteristics.

Looking at these graphs, we concluded that we would be most effective building out our model using factors that met the criteria of a Sharpe Ratio greater than 0.36 and with correlations less than 0.3. In this decision we tried to avoid overexposure to high correlation while also capitalizing on high expected returns. With these parameters set, we were able to reduce our area of focus to just five factors.

Here are the reasons for whether each factor was selected or not:

Factor Name	**Sharpe Ratio**	**Selected**	**Reason**
mr_1w	1.37	yes	Highest SR
mr_1mo	0.87		Although the Sharpe ratio is high, it has a high correlation with mr _1w and does not perform as well.
mom_1yr_sub_1mon	-0.03		Low SR
mom_1yr	-0.36	yes (as mr_1yr)	Highest absolue value of SR. Although the average return over one year usually represents momentum, it exhibits characteristics of m ean reversion here (the constructed momentum factor has inverse p redictive power). Considering that this will ultimately serve as inpu

			t data for machine learning, and given the peculiarities of the Japanese market (where momentum factors generally perform poorly), we have chosen to include this factor. We will use the negative of this factor and rename it to mr_1yr.
cpv_1w	0.63		Although the Sharpe ratio is high, it has a high correlation with mr_1w and does not perform as well.
cpv_1mo	0.54	yes	High SR
cpv_3mo	-0.01		Although the Sharpe ratio is high, it has a high correlation with cpv_1mo and does not perform as well.
beta_1w	-0.03		Low SR
beta_1mo	0.11		Low SR

Considering the performance of each factor (measured by the Sharpe ratio) and the correlation between factors (choosing the one with a higher Sharpe ratio among factors with high correlation), we selected cpv_1mo, mr_1w and mr_1yr in this section.

3.3 Factors Derived from Fundamental Data

For fundamental data, we obtain annual frequency data from Bloomberg and resample it to daily frequency(with specific processing methods detailed in the appendix) and then tested the returns attributable to these factors. The performance of each factor is presented below.

Factor Construction:

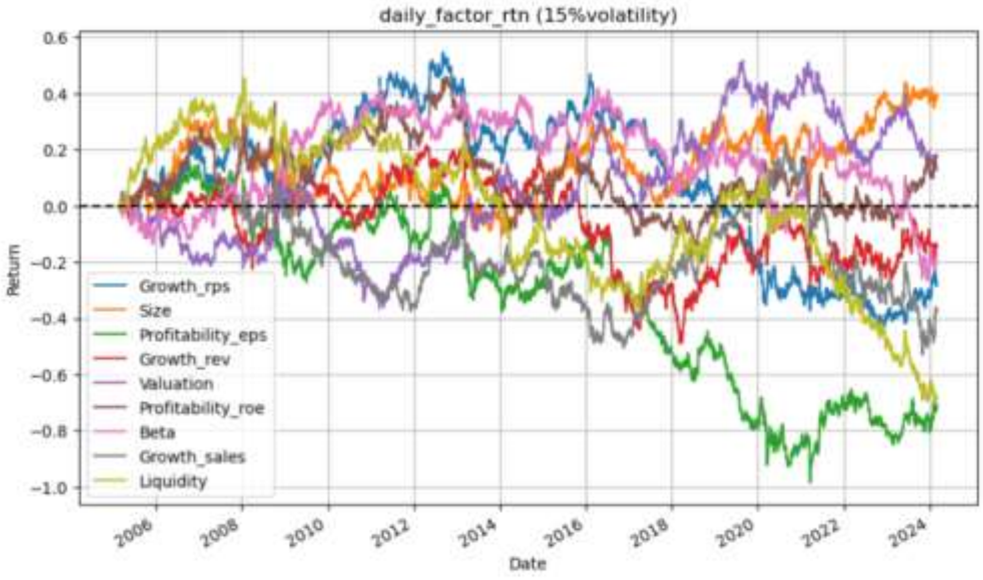
Factor Name	Data souce	Data Used	Data Processing Method
Beta	bloomberg	Beta:M-1	
Profitability_roe	bloomberg	ROE LF	
Growth_sales	bloomberg	Net Sales - 5 Yr Geo Gr LF	
Growth_rev	bloomberg	Rev - 1 Yr Gr:Q	
Growth_rps	bloomberg	Rev/Bas Sh T12M	
size	bloomberg	Market Cap	ln(Market Cap)
Profitability_eps	bloomberg	EPS T12M	
Valuation	bloomberg	P/E	

Liquidity	bloomberg	Average Volume:Y-1 & Curr Shares Out	Average Volume:Y-1/Curr Shares Out

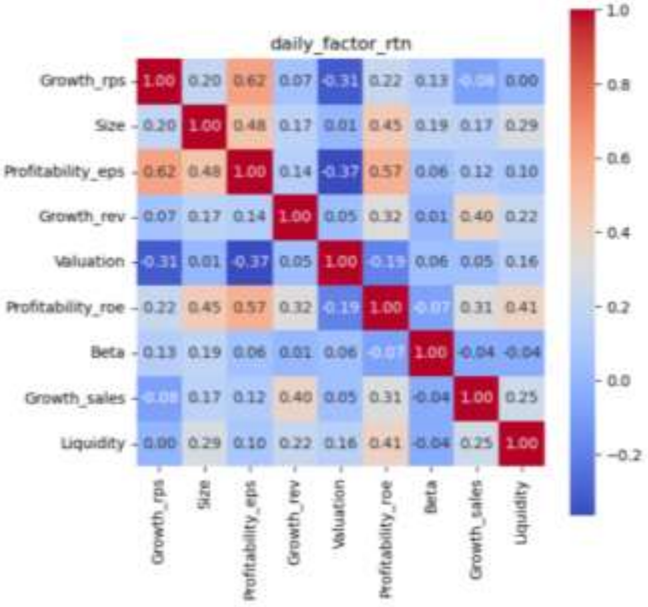
Sharpe Ratio for each factor:

Factor Name	**Sharpe Ratio**
Beta	-0.06
Profitability_roe	0.06
Growth_sales	-0.16
Growth_rev	-0.05
Growth_rps	-0.10
size	0.14
Profitability_eps	-0.26
Valuation	0.04
Liquidity	-0.24

Factor Returns for each factor:



Correlation Heatmap:



Here are the reasons for whether each factor was selected or not:

Factor Name	**Sharpe Ratio**	**Selected**	**Reason**
Beta	-0.06		low SR
Profitability_roe	0.06		low SR

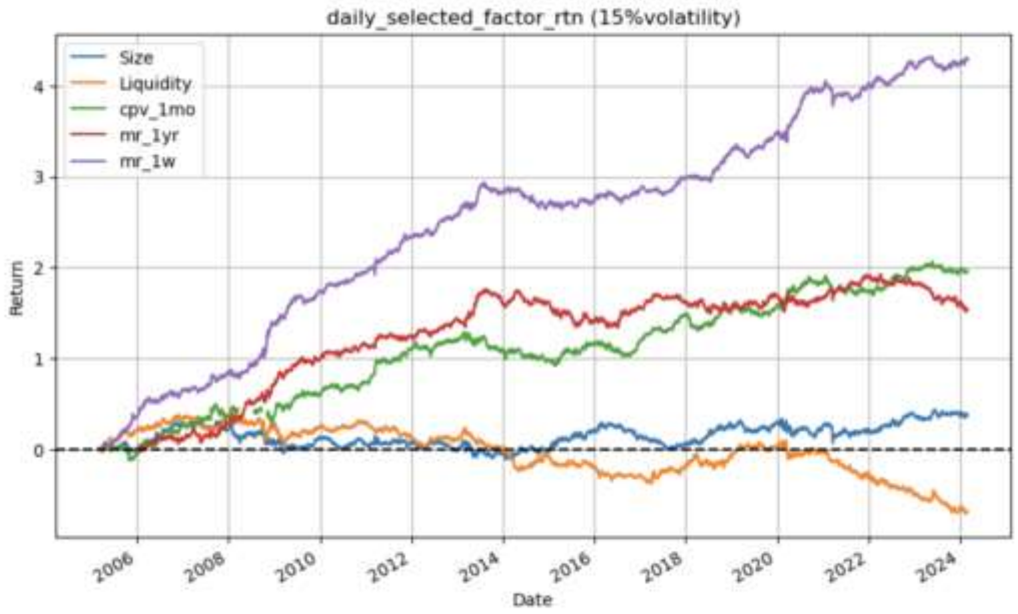
Growth_sales	-0.16		Although the Sharpe ratio is slightly higher, it contradicts financial intuition.
Growth_rev	-0.05		low SR
Growth_rps	-0.10		low SR
size	0.14	yes	slightly higher SR
Profitability_eps	-0.26		Although the Sharpe ratio is slightly higher, it contradicts financial intuition.
Valuation	0.04		low SR
Liquidity	-0.24	yes	The absolute value of the Sharpe ratio is very high, and indeed, there is sometimes a negative correlation between liquidity and returns.

3.4 Selected Factors

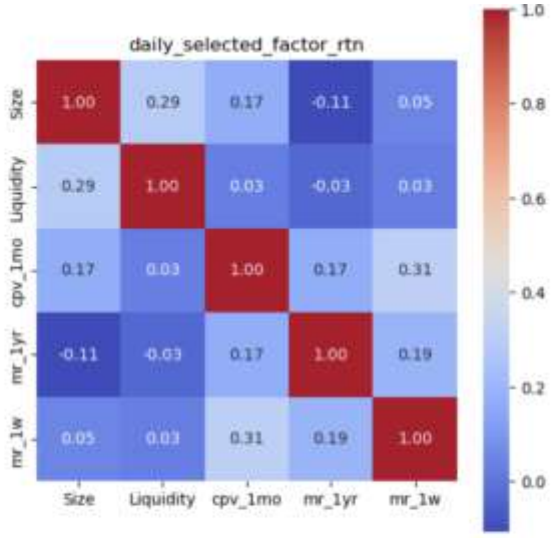
3.4.1 Performance of Selected Factors

The performance of the selected factors is as follows, showing that each factor possesses either positive or negative predictive power, and they have low correlation with each other, which is crucial for diversifying the predictive signals in our model. This multi-factor approach enhances the robustness of our investment strategy by combining insights from both market behaviors (volume and price data) and company fundamentals, leading to a more comprehensive and nuanced understanding of the market dynamics.

Factor Returns for each factor



Correlation Heatmap



These 5 factors as pictured above met the specifications and thus provided the basis for the training of our model. In the subsequent paragraphs we will thoroughly describe each of these factors.

3.4.2 Financial Explanation of Selected Factors

Size

In our fundamental data, size refers to the logarithm of market cap. Intuitively, and with the support of the above data, having a high market cap typically coincides with long-term profitability. According to the chart, size became increasingly more effective after 2014. As noted in our market analysis, during 2014 we saw a reduction in consumer spending as a result of monetary policy; this decrease typically bodes positively for pharmaceuticals both globally and in Japan.

Liquidity

Liquidity, as defined in our model refers to average volume over the past year, divided by current shares outstanding. During the 2008 global financial crisis, investors typically turned to safer, less volatile assets. After the 2011 East Japan earthquake, supply chain disruptions in the pharmaceutical industry may have affected production and sales. Large pharmaceutical companies with high liquidity stocks may face direct impacts, and panicked market reactions may lead to selling, affecting returns. Large pharmaceutical companies typically have a wider product line and may face more adjustments. High liquidity amplifies the impact on stocks and may lead to a decrease in returns. As we can see, since 2011, liquidity as a factor has had increasingly negative returns, and when we inversely trade, we are able to capitalize on the large absolute value of SR.

Correlation of Price and Volume – 1 month

Our most interesting factor incorporated into the model is the correlation coefficient between Price and Volume. Jieruo, one of our group members studied CPV as a factor for investment in a past internship and brought that knowledge with her into this project. According to Seeking Alpha, when a stock is behaving “correctly,” we notice that price and volume tend to move together, however we notice that in practice, when the correlation between price and volume is low, it is an indication of a good buy opportunity. When first analyzing factors, it became clear that the inverse of CPV performed well with the test set and had a low correlation to the other factors used for machine learning thus earning a role in the algorithm.

Mean Reversion – 1 year & 1 week

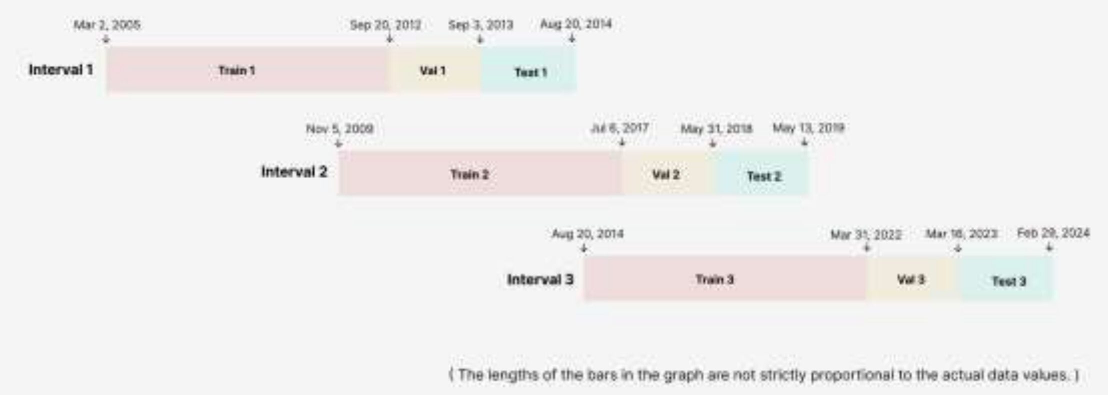
As noted earlier, in the market analysis, mean reversion has been historically effective in the Japanese equity markets and in pharma. 1 week and 1 year rolling windows were found to be most effective in our historical tests, and what these intuitively indicate is that the stocks tend to revert to means most effectively in 5-day, and one-year, windows; Reversion to mean refers to the tendency of individual stocks that overperform/underperform their sectors in the short run, to eventually correct themselves, and trade in a direction nearing the sector mean.

4 Model

4.1 Strategic Framework

4.1.1 Splitting the Training, Validation, and Test Sets

The data in this study spans from March 2, 2005, to February 29, 2024. Directly dividing the entire dataset into training, validation, and test sets might prevent the training set from learning the most recent data patterns. To fully utilize the data while considering the timeliness of the model, this study divides the data from March 2, 2005, to February 29, 2024, into three overlapping intervals for rolling backtesting. Each interval includes 50% of the total data (approximately 10 years), with the first 40% (about 8 years) as the training set, the middle 5% (about 1 year) as the validation set, and the last 5% (about 1 year) as the test set. The test set is considered as out-of-sample data for each interval. (See figure below)



(The lengths of the bars in the graph are not strictly proportional to the actual data values.)

4.1.2 Model Training and Selection

For each interval's data, we train with the training set, apply the model to the validation set to select parameters, decide on the depth of model training based on the performance on the validation set, and finally apply the model to the out-of-sample data (test set) of each round. First, we train models on the first two rounds of datasets and assess the out-of-sample performance to select the model that performs best on out-of-sample data. We believe that models with better out-of-sample performance are more suited for the Japanese pharmaceutical market. The selected model is then used for training and the final backtest in the third interval. Subsequently, a stock selection strategy based on the forecasted values is formulated, and backtesting is performed with the predicted values from the third interval's validation set to select the best-performing strategy. Finally, the previously selected best strategy is implemented for backtesting on the test set of the third interval.

This study opts to train using XGBoost and GRU (Gated Recurrent Unit) models, comparing the out-of-sample performance in the first two intervals to select the model that is both better performing and more stable out-of-sample. The selection criterion for this study is a higher cross-sectional IC (Information Coefficient) calculated out-of-sample. The selected model is then used for model training in the third interval.

4.2 Model Introduction

4.2.1 XGBoost (eXtreme Gradient Boosting)

XGBoost stands for eXtreme Gradient Boosting. It is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning technique that generates a stronger prediction by aggregating the predictions of multiple weak models.

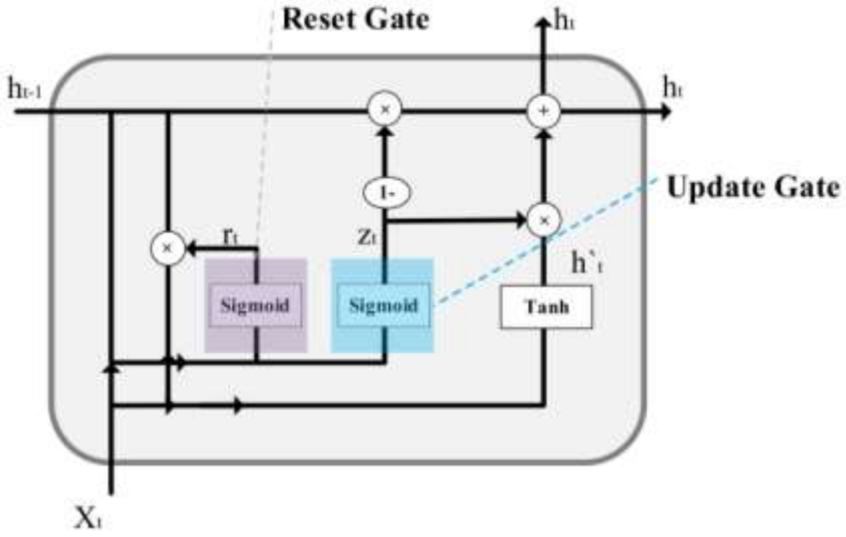
- Model: XGBoost’s model of choice is decision tree ensembles. All the independent variables are assigned weights, which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and then fed to the second decision tree.
- Tree Boosting: Just like other supervised learning models, XGBoost learns by defining an objective function and optimizing it. An additive approach is adopted to correct what the model has learned and add one new tree at a time, while taking into consideration the loss function.
- Regularization: XGBoost offers regularization, which allows for control overfitting by introducing L1/L2 penalties on the weights and biases of each tree. This feature is not available in many other implementations of gradient boosting. By defining the regularization term formally, we can evaluate the learning process and obtain models that perform well in the wild.
- Structure Score: At each level, the objective formula can be rewritten to obtain the best objective reduction once the tree model is reformulated, resulting in an equation that measures how good the tree structure is. The algorithm will then optimize one level of the tree at a time by computing a “gain” formula which includes 1) the score on the new left leaf 2) the score on the new right leaf 3) The score on the original leaf 4) regularization on the additional leaf. In the end, a left to right scan will calculate the structure score of all possible split solutions and outputs the best split.

4.2.2 GRU (Gated Recurrent Unit)

The Gated Recurrent Unit (GRU) model is a type of recurrent neural network (RNN) that is designed for processing sequential data, making it particularly useful for tasks involving time series analysis, natural language processing, and more. GRUs were introduced to solve the vanishing gradient problem that can occur with standard RNNs, making them more efficient and effective for long sequences.

The GRU simplifies the structure of the more complex Long Short-Term Memory (LSTM) network while maintaining a similar level of performance. It does this by using two gates:

- Update Gate: This gate decides the amount of past information to be passed along to the future. It helps the model to determine what information is relevant to keep from past steps, in order to make better predictions.
- Reset Gate: This gate is used to decide how much past information to forget. It allows the model to drop irrelevant information from the past, making it more flexible in handling various types of sequence data.



By utilizing these gates, GRUs can manage information flow within the network more efficiently than standard RNNs, making them highly effective for tasks involving sequential data, such as time series analysis, natural language processing, and more. Their simplified structure, compared to that of Long Short-Term Memory (LSTM) networks, allows for faster training times while still achieving comparable performance levels, which has contributed to their popularity in both academic and practical applications.

In the financial sector, GRUs have found extensive applications due to their capability to model time series data effectively. Some of the key applications include:

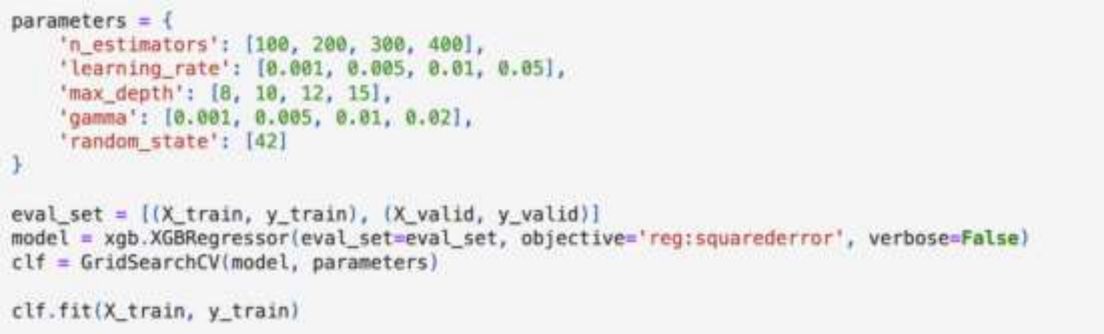
- Stock Price Prediction: GRUs can analyze historical stock data to predict future stock prices. By capturing the temporal dependencies in the stock market data, GRUs help in forecasting stock movements, which is invaluable for traders and investors.

- Fraud Detection: Financial institutions use GRUs to detect unusual patterns in transactions that may indicate fraudulent activity. The sequential analysis capability of GRUs allows for the identification of suspicious activities over time, enhancing security measures.
- Credit Scoring: GRUs are used in analyzing an individual's financial transactions and credit history over time to predict their creditworthiness. This helps in making more informed decisions on lending.
- Algorithmic Trading: GRUs can be employed to develop trading algorithms that predict market trends based on historical data. By analyzing patterns in market data, these algorithms can execute trades at optimal times, maximizing profits.
- Risk Management: By modeling financial time series data, GRUs assist in assessing and managing risk. They can predict potential market downturns or financial distress, allowing organizations to take preemptive measures.

4.3 Model Training

4.3.1 Training XGBoost

1. Model Construction:



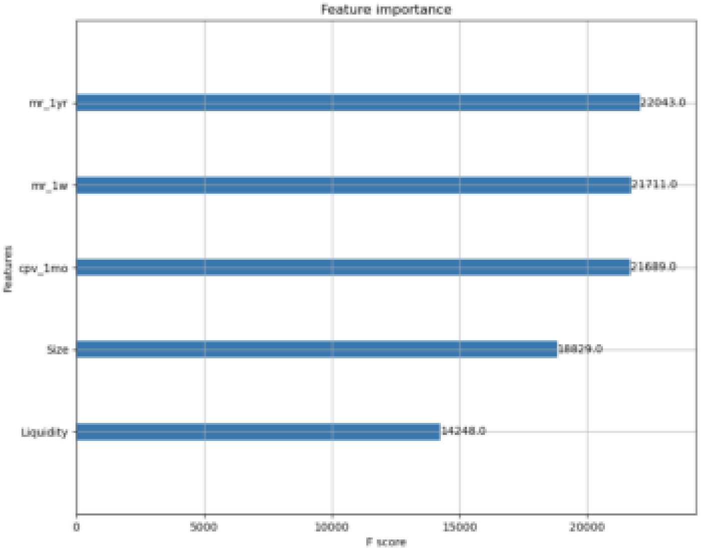
```
parameters = {
    'n_estimators': [100, 200, 300, 400],
    'learning_rate': [0.001, 0.005, 0.01, 0.05],
    'max_depth': [8, 10, 12, 15],
    'gamma': [0.001, 0.005, 0.01, 0.02],
    'random_state': [42]
}

eval_set = [(X_train, y_train), (X_valid, y_valid)]
model = xgb.XGBRegressor(eval_set=eval_set, objective='reg:squarederror', verbose=False)
clf = GridSearchCV(model, parameters)

clf.fit(X_train, y_train)
```

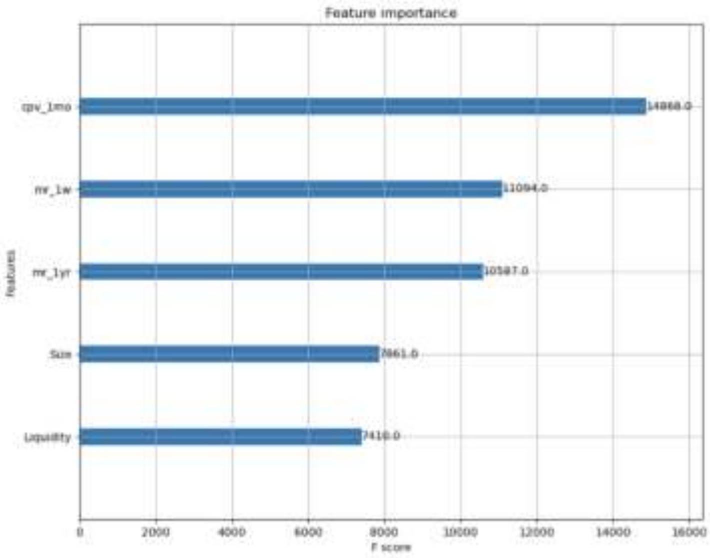
We considered a variety of parameters to train the model, most notably:

- learning_rate: This is the step size shrinkage used to prevent overfitting. After each boosting step, we can directly get the weights of new features, and learning_rate shrinks the feature weights to make the boosting process more conservative.
- gamma: Gamma is the minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.
- max_depth: Increasing max_depth will make the model more complex and more likely to overfit. objective: This determined the objective function in the learning process. We used reg:squarederror, regression with squared loss.

2. Feature Importance:

After training each interval of data, we outputted the best parameters undertaken by the model and plotted the f-score of each feature. f-score calculates the frequency of a feature being used to split the classification tree, can be used to identify the importance of each feature. The bar chart below shows the relative importance of features in the 1st and 2nd intervals. Note that the rankings are different—mr_1yr ranked the first and cpv_1mo the third for the first interval's data, while their ranking was flipped during the second interval.

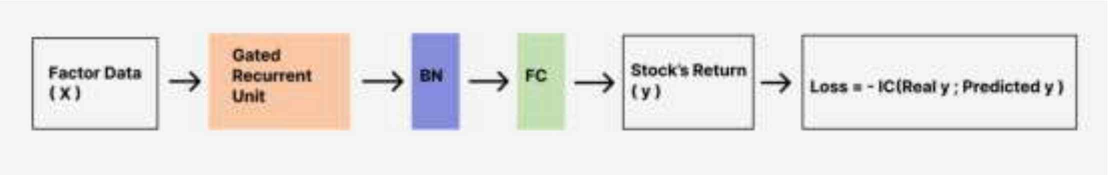




4.3.2 Training GRU

1. Model Construction:

   The model structure consists of a GRU layer, a Batch Normalization layer, and a Fully Connected layer, as shown in the diagram below.



- **Features X:** The factor values of individual stocks over the past 5 trading days (including today).
- **Label y:** The return rate of the individual stock for the next trading day.
- **Loss Function:** It is noteworthy that the loss function uses the negative of the Information Coefficient (IC) between predicted values and actual values. Predicting the actual return values of stocks is challenging; thus, we focus more on the performance of the prediction results in terms of cross-sectional IC during training. We aim for a higher IC, which is opposite to the direction of optimization of the loss function; hence, our loss function is the negative IC.
- **Optimizer:** Adam. Adam is a popular optimizer that adjusts the learning rate automatically, making it effective for a wide range of problems. Its main advantage is its efficiency in converging to the optimal solution faster than many other optimizers, especially in complex models and large datasets.

2. Techniques Applied:
   - **Learning Rate Decay:** Learning rate decay is a technique to adjust the learning rate over time. Initially, the learning rate is set to 0.01, and after every five epochs, the learning rate is multiplied by 0.1. This approach helps in making finer adjustments to the model weights as training progresses, potentially leading to better convergence.
   - **Early Stopping:** Early stopping is a method used to prevent overfitting by ending the training process early if there is no improvement. In this context, if the loss on the validation set increases for four consecutive epochs, training is stopped prematurely. This technique helps in saving computational resources and prevents the model from learning noise in the training data.
   - **Initialization with Previous Round's Training Results:** At the start of each training round, the model is initialized with the training results from the previous round. This approach ensures that the initial model has already learned patterns from the previous data and can continue learning based on updated data. This method is faster than starting training anew with fresh data.
   - **Backpropagation Through Time (BPTT):** GRU models are trained using a variant of backpropagation called backpropagation through time, which allows gradients to flow backward through the network and update weights. This process is essential for learning the dependencies between time steps in sequence data.

4.4 Model Selection

In the evaluation of out-of-sample data performance across different intervals, we observed the following results for the XGBoost and GRU models:

Model	Interval 1 test set IC	Interval 2 test set IC
XGBoost	-0.0173	0.0075
GRU	0.0017	0.0279

It is evident that the XGBoost model's performance is not very stable, displaying variability in its predicted cross-sectional Information Coefficient (IC) with even negative values in out-of-sample data for the first interval. In contrast, the GRU model consistently exhibited positive IC values across both intervals. Therefore, the decision was made to proceed with the GRU model for training the data in the third interval, where it achieved an out-of-sample IC of 0.0306.

Model	Interval 1 test set IC	Interval 2 test set IC	Interval 3 test set IC
GRU	0.0017	0.0279	**0.0306**

Moving forward, different strategies will be formulated based on the predicted values and applied to the predicted values of the third interval's validation set to identify the most effective strategy. This selected strategy will then be applied in a backtest using the third interval's test set and the predicted data, aiming to validate the model's effectiveness and the strategy's applicability in real-world conditions.

5 Back Testing

5.1 Generating buy or sell signals

After confirming the model (GRU) and the main parameters we use, below are the steps and results of the back testing:

To generate buy and sell signals, we apply two algorithms.

a. Select based on cross-section rank:

On each trading day, ranking the predictive data of each stock on a cross-sectional basis, assigning buy signals (1) to the top n% of stocks and sell signals (-1) to the bottom n%; for the remaining stocks, remain no change (0).

b. Select based on threshold:

Calculate the statistical characteristics of the forecast data, using the mean plus or minus n times the standard deviation as the threshold for processing buy (1) or sell (-1) signals, with no signal (0) for the middle range. It is important to note that the test set data should not be leaked prematurely. When backtesting on the validation and test sets, the statistical data from the validation set should be used.

For each method, to reduce turnover rates, we also implement a rolling window. The predictive values are averaged over this rolling window before computing the signals. This approach smooths out the data, mitigating the impact of short-term fluctuations and potentially reducing the frequency of trades, thereby lowering transaction costs and improving the stability of the investment strategy. This technique allows for more consistent and reliable signal generation, aligning better with long-term investment goals.

5.2 Select the best strategy

These methods of signal processsing are tested on a validation set to determine which signal processing method performs best. The criteria for this determination include a combination of the Sharpe ratio, annualized returns from the backtesting results. The most effective method is then applied to the test set.

In addition to different signal processing methods, we also use various capital proportions during the backtesting phase to determine which approach yields the best results.

We will go through all parameter combinations, conduct back tests, and record performance metrics for each combination.

Fixed Parameter:

Parameter	Values
initial_capital	100000
trading_cost	0.05%

Parameter for cross-section rank method:

Parameter	Values
Top Quantile	0.2, 0.3
Bottom Quantile	0.2, 0.3
Rolling Window	1, 3, 5, 10
Capital Proportion	0.5, 1.0

Parameter for threshold-based method:

Parameter	Values
Buy Threshold	mean + 1*std, mean + 1.5*std
Sell Threshold	mean - 1*std, mean - 1.5*std
Rolling Window	1, 3, 5, 10
Capital Proportion	0.5, 1.0

Select combinations that perform well in both the top three ranked by Sharpe ratio and the top three ranked by annualized returns.The following are the superior combinations identified from testing various signal processing methods on the validation set.

Top three combinations ranked by Sharpe ratio：

Method	Buy Threshold/Top Quantile	Sell Threshold/Bottom Quantile	Rolling Window	Capital Proportion	Sharpe ratio	Annualized return
threshold	mean+ 1.5*std	mean - 1*std	5	1	2.09	4.41%
threshold	mean+ 1.5*std	mean - 1*std	5	0.5	2.09	2.20%
threshold	mean+ 1.5*std	mean - 1*std	10	0.5	1.83	1.38%

Top three combinations ranked by annualized returns：

Method	Buy Threshold/Top Quantile	Sell Threshold/Bottom Quantile	Rolling Window	Capital Proportion	Sharpe ratio	Annualized return
threshold	mean+ 1.5*std	mean - 1*std	5	1	2.49	4.41%
threshold	mean+ 1.5*std	mean - 1*std	3	1	1.30	2.87%
threshold	mean+ 1.5*std	mean - 1*std	10	1	1.83	2.76%

Based on performance, we have selected the following combination:

Method	Buy Threshold	Sell Threshold	Rolling Window	Capital Proportion
threshold	mean + 1.5*std	mean - 1*std	5	1

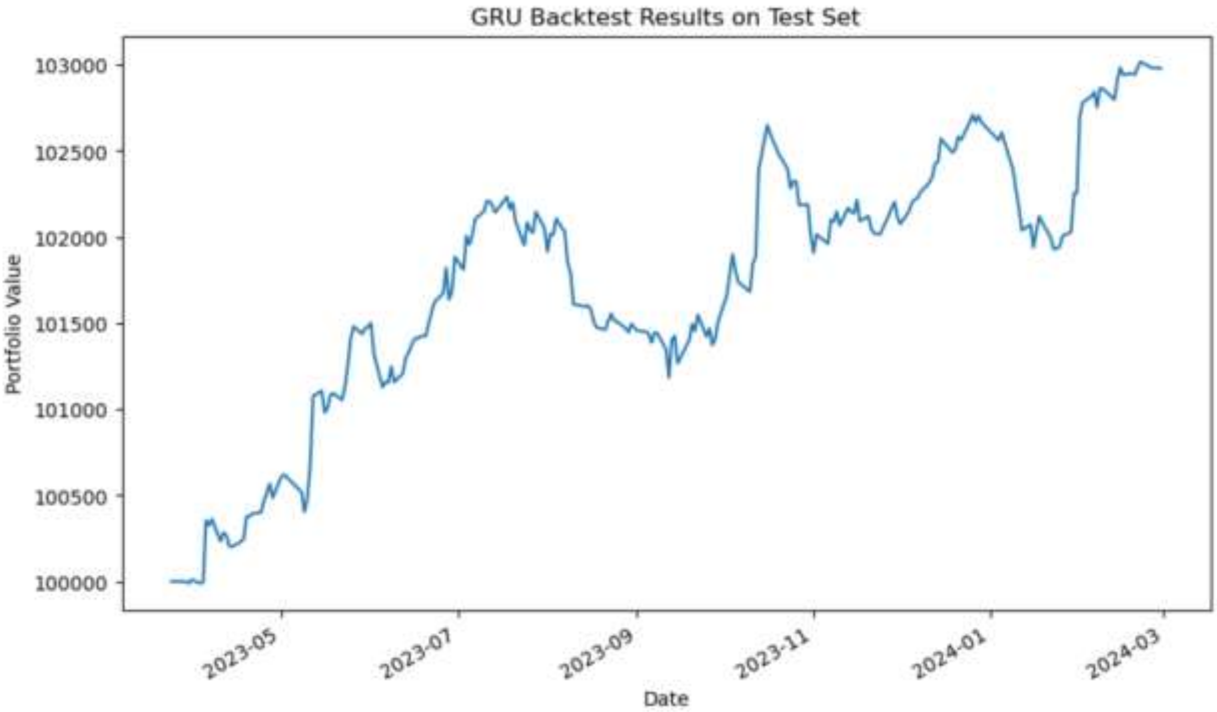
It's important to note that the calculation of the threshold consistently uses the statistical metrics from the validation set predictions

This combination features an aggressive buy threshold set at mean plus 1.5 times the standard deviation and a moderate sell threshold at mean minus the standard deviation. With a rolling window of 5 and a capital proportion of 1, this configuration aims to maximize returns while managing risk effectively.

5.3 Backtesting on the test set

We finally came up with the optimal parameter combination for backtesting on test set. The back test results show:

Total return	2.98%
Annualized return	3.27%
Annualized volatility	1.57%
Sharpe ratio	2.09
Maximum drawdown	1.03%
Annualized turnover rate	1359.20%



Backtest Results Analysis:

In the backtesting on the test set, we achieved fairly decent positive returns after fees (Sharpe ratio of 2.09, annualized return of 3.27%). However, the annualized return of only 3.27% did not meet our initial expectations.

Reasons that lead the result lower than our expectation:

Model Fit: Although GRU and XGBoost models were employed, which are well-suited for capturing time series data and complex relationships in datasets respectively, they might not fully adapt to the intricate dynamics of the pharmaceutical market without considering upstream and downstream factors. These factors include the availability and cost of raw materials, the impact of regulatory changes, and the influence of market access strategies, which are crucial for accurately modeling the pharmaceutical sector.

Factor Selection and Industry Dynamics: The selected factors such as mr_1w (one-week mean reversion), mr_1yr (one-year mean reversion), and cpv_1mo (one-month correlation of price and volume) might not sufficiently account for the pharmaceutical industry's supply chain complexities. Including upstream factors like raw material cost variability, patent expirations, and R&D progress, alongside downstream elements such as distribution efficiency and market access barriers, could provide a more robust and realistic model. These aspects are vital as they directly impact production costs, revenue potential, and market penetration, influencing stock performance significantly.

Market Volatility: While the analysis acknowledges broad market volatilities such as financial crises and natural disasters, the specific volatilities related to the pharmaceutical supply chain, such as sudden changes in drug approval policies or healthcare reforms, should also be factored in. These elements can cause significant fluctuations in stock prices due to their direct impact on the pharmaceutical companies' operational capabilities and market expectations. By integrating upstream factors like drug patent pipelines and exclusivity periods, and downstream factors such as changes in healthcare provider preferences and patient access schemes, the model can capture a more accurate picture of the pharmaceutical landscape. This integration would allow for a better assessment of how external pressures and internal industry developments affect stock prices.

Data Handling and Model Training: The preprocessing of data and the model's training phase must also reflect the complexity added by these upstream and downstream factors. This could involve more sophisticated data engineering to capture the nuances of pharmaceutical market movements and ensure the model is trained on realistic scenarios reflecting the full spectrum of market dynamics.

By deepening the factor analysis to include these critical upstream and downstream components, the model not only becomes more aligned with the pharmaceutical industry's operational realities but also more sensitive to the specific economic forces that drive market behavior in this sector. This approach would likely improve the model's predictive accuracy and robustness, making it a more valuable tool for investors and analysts focusing on the pharmaceutical industry.

Reference

［1］Price and volume correlation. (2011, January 5). Seekingalpha. Retrieved April 17, 2024, from https://seekingalpha.com/article/244906-price-and-volume-correlation

［2］ASNESS, CLIFFORD S., et al. “Value and Momentum Everywhere.” The Journal of Finance, vol. 68, no. 3, 2013, pp. 929–85. JSTOR, http://www.jstor.org/stable/42002613. Accessed 17 Apr. 2024.

［3］GfG. (2023, February 6). XGBoost. GeeksforGeeks. https://www.geeksforgeeks.org/xgboost/

［4］Introduction to Boosted Trees — xgboost 2.0.3 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/tutorials/model.html

［5］Gao, Y., Wang, R., & Zhou, E. (2021). Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021, 1–8. https://doi.org/10.1155/2021/4055281

［6］Maio, P. F., & Santa-Clara, P. (2012b). Multifactor models and their consistency with the ICAPM. Journal of Financial Economics, 106(3), 586–613. https://doi.org/10.1016/j.jfineco.2012.07.001

［7］Japan Pharmaceutical Manufactures Association. (2024). DATA BOOK. https://www.jpma.or.jp/news_room/issue/databook/ja/rs40ob000000139v-att/DATABOOK2024_E_ALL.pdf

Appendix

Factor Name	Data source	Data Used	Data Processing Method	Explanation of Data Used	status
Beta	bloomberg	Beta:M-1		Estimate of a security's future beta. This is an adjusted beta derived from the past two years of weekly data, but modified by the assumption that a security's beta moves toward the market average over time. The formula used to adjust beta is : Adjusted Beta = (0.66666) * Raw Beta + (0.33333) * 1.0 Where : Raw Beta is (RK167, EQY_RAW_BETA) Equities: Values are calculated using Relative Index (PR240, REL_INDEX) for the security. Only the prices for the stock and its relative index are used in the calculation.	
Profitability_roe	bloomberg	ROE LF		Measure of a corporation's profitability by revealing how much profit a company generates with the money shareholders have invested, in percentage. Calculated as: (T12 Net Income Available for Common Shareholders / Average Total Common Equity) * 100 Where: T12 Net Income Available for Common Shareholders is T0089, TRAIL_12M_NET_INC_AVAI_COM_SHARE Average Total Common Equity is the average of the beginning balance and ending balance of RR010, TOT_COMMON_EQY If either the beginning or ending total common equity is negative, Return on Equity will not be calculated.	
Growth_sales	bloomberg	Net Sales - 5 Yr Geo Gr LF		Compound 5-year growth rate in sales. Calculated as: (Most Recent Revenue / Revenue Five Periods Earlier) ^ 0.2 - 1 * 100	
Growth_rev	bloomberg	Rev - 1 Yr Gr:Q		A percentage increase or decrease of sales revenue by comparing the current period with same period prior year. Calculated as: (Revenue from Current Period - Revenue from Same Period Prior Year) * 100 / Revenue from Same Period Prior Year	
Growth_rps	bloomberg	Rev/Bas Sh T12M		Ratio that computes the total revenue earned per share over the reporting period. Unit: Actual. Calculated as: Revenue / Weighted Average Shares Where: Revenue is IS010, SALES_REV_TURN Weighted Average Shares is IS060, IS_AVG_NUM_SH_FOR_EPS	
size	bloomberg	Market Cap	ln(Market Cap)	The monetary value of all outstanding shares stated in the pricing currency. Capitalization is a measure of corporate size.	selected
Profitability_eps	bloomberg	EPS T12M		Earnings Per Share: Earnings Per Share (EPS) is the portion of a company's profit allocated to each sharesholder. It is calculated based on Net Income Available for Common Sharesholders divided by the Basic Weighted Average Shares Outstanding. This field returns Bottom-line Earnings Per Share when FPDF Settings for 'Non-GAAP Adjustments.' Adjusted Override (DT094, FA_ADJUSTED) can be flagged 'Y' to return adjusted data for excluding abnormal items. Unit: Actual	
Valuation	bloomberg	P/E		Ratio of the price of a stock and the company's earnings per share.	

Liquidity	bloomberg	Average Volume:Y-1 & Curr Shares Out	Average Volume:Y-1/Curr Shares Out	Curr Shares Out: Total current number of shares outstanding. Average Volume:Y-1: prev 1 year average volume	selected
mr/ mom (mean_reversion and momtumn factor)	yahoo finance	Adj Close	Mean Reversion Factor: $mr = -\frac{1}{N}\sum_{i=0\sim N}^{\square} R_{t-i}$ $mom = -mr$ where: - N is the number of trading days in the calculation window (use one week, N = 5), - R_{t-i} is the daily return on day t-i		selected; using window of 1 week (mr_1w)
cpv (correlation between price and volume)	yahoo finance	Adj Close; volume	CPV: correlation between price change (return) and volume $CPV_t = corr(R, Vol\_Norm\ ; window)$ where: - Vol_Norm is volume normalized by dividing the rolling average of the volume over one week. - R is the daily return - window use 1 week and 1 month		selected; using window of 1 week (cpv_1w) and of 1 month (cpv_1mo)
beta_tpphrm	yahoo finance	Adj Close	Computed as the slope coefficient in a time-series regression of excess stock return against index return.		