Stock Return Prediction in Machine Learning

Group 10
120090651 马可轩
120090489 李卓宸
120020128 刘鑫宇
120090717 罗单丹
120020312 周阅月
120090900 田海川
120090856 余松霖
120090452 薛颖
120090814 陈芝霖
120090626 王子潇

Content

01 Stock Selection & Preprocessing Strategy

02 Fama-French Three-Factor model & Ridge Regression Model

03 Deep Learning Model

04 Prediction Performance

05 Reason Analysis

Stock Selection Strategy

Step 1:

- Start with a dataset of the stocks in **CSI500**, that includes stock codes, dates, and total market values.
- For each date, rank the stocks by their **market value.**

Step 2:

- **Recency Weighting**: assign weights to these rankings based on the recency of the data, with more recent dates getting higher weights.

000012	南玻A
002250	联化科技
002487	大金重工
002518	科士达
300395	菲利华
600435	北方导航
600839	四川长虹
603026	胜华新材
603355	莱克电气
603688	石英股份

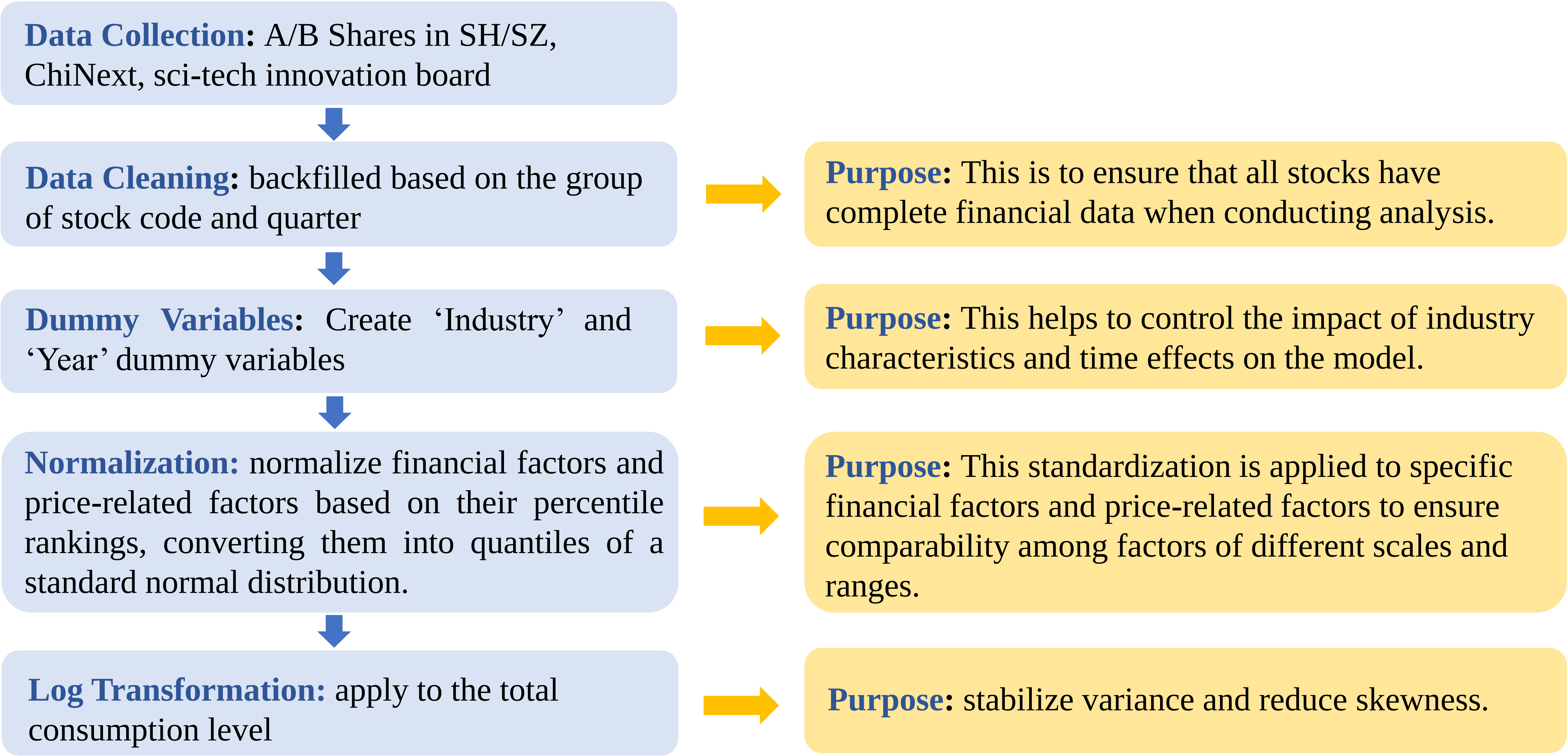
Step 3:

- **Applying the weights to the ranks**, so rankings from more recent dates have a greater impact on the final score. Then I sum these weighted ranks for each stock across all dates.

Step 4:

- **Select the top 10 stocks** with the **highest summed weighted ranks**, indicating that they have a high market value and have been more relevant in recent.

Preprocessing



Data Collection: A/B Shares in SH/SZ, ChiNext, sci-tech innovation board

Data Cleaning: backfilled based on the group of stock code and quarter

Purpose: This is to ensure that all stocks have complete financial data when conducting analysis.

Dummy Variables: Create ‘Industry’ and ‘Year’ dummy variables

Purpose: This helps to control the impact of industry characteristics and time effects on the model.

Normalization: normalize financial factors and price-related factors based on their percentile rankings, converting them into quantiles of a standard normal distribution.

Purpose: This standardization is applied to specific financial factors and price-related factors to ensure comparability among factors of different scales and ranges.

Log Transformation: apply to the total consumption level

Purpose: stabilize variance and reduce skewness.

Selection of Indicators

Company Operational Indicators

These factors typically represent the fundamental financial health and performance metrics of a company.

- **PB** (Price-to-Book Ratio)
- **PE** (Price-to-Earnings Ratio): a measure of market expectations and growth prospects.
- **DivYield** (Dividend Yield): Provides insight into the income generated by an investment in stocks relative to its price.
- **PS** (Price-to-Sales Ratio): A valuation metric that compares a company's stock price to its revenues.
- **PCF** (Price-to-Cash Flow Ratio): Assesses the value of a company's stock price compared to its operating cash flow.

Basic Indicators (Trading)

These factors are typically related to the trading aspects of stocks and are used to understand the trading environment or market sentiment toward a company.

- **Turnover:** Reflects the trading volume or liquidity of the stock.
- **MA10, MA20** (Moving Averages): Used to smooth out price data to identify trends over 10 and 20 days.
- **VWAP** (Volume Weighted Average Price): Gives an average price a security has traded at throughout the day, based on both volume and price. It is important because it provides traders with insight into both the trend and value of a security.

Selection of Indicators

Technical Factors

Technical factors are derived from statistical analysis of market activity, such as past prices and volume.

- **RSI** (Relative Strength Index): Measures the magnitude of recent price changes to evaluate overbought or oversold conditions.
- **MACD** (Moving Average Convergence Divergence): A trend-following momentum indicator that shows the relationship between two moving averages of a security’s price.
- **BIAS:** A technical analysis indicator that compares the closing price to a moving average to identify trends.
- **For more indicators, please refer to the Appendix**

Economic and Market Indicators

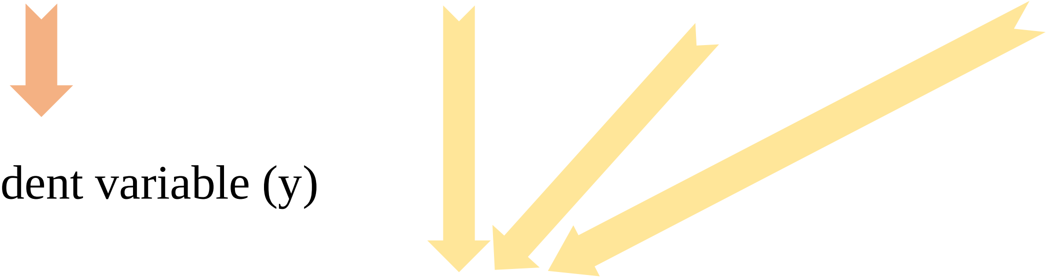
These factors take into account broader economic and market signals which can affect the financial markets.

- **GDPGrowth**: Reflects the growth rate of the economy, which can impact company earnings and stock performance.
- **InflationRate**: Inflation can influence the discount rates used to value stocks and affect a company's input costs and consumer demand.
- **TotalConsumptionLevel**: Represents consumer spending which drives a large part of economic activity and can therefore affect company revenues.

Baseline Model

Fama-French Three-Factor Model

$$E(R_{it}) - R_{ft} = \beta_i[\underline{E(R_{mt} - R_{ft})}] + s_i\underline{E(SMB_t)} + h_i\underline{E(HML_t)}$$



Dependent variable (y)

Independent variable (X)

Additional X-variables:

- **GDPGrowth**: GDP growth
- **totalConsumptionLevel**: total consumption level
- **Ind_1-Ind_5**: dummy variables from financial sector to industrial sector.

Data

Training Set: consisting of the first 1500 observations for each stock

Test Set: consisting of the remaining observations.

Clustered OLS Regression

Improve

- **Intra-group Correlation**: In many financial datasets, observations within the same group (such as the same year, industry, or market) often exhibit correlation with each other.
- **Heteroskedasticity**: Financial time series data often display heteroskedasticity, meaning that the variance of errors varies at different time points.

Regularized Linear Regression Model

Ridge Regression Model

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \underline{\lambda} \sum_{j=1}^{p} \beta_j^2 \right\}$$

The regularization parameter, controlling the degree of penalty applied to the coefficients.

Variables:

- ➢ **y**: ‘ri-rf’
- ➢ **X:** In addition to the existing independent variables in the baseline model, we also selected 28 variables from the previously selected indicators to add to the ridge regression model, such as ‘roe_Ttm’, ‘roa_Ttm’, and ‘current_Ratio’, etc.

Data

The same procedure as Q1.

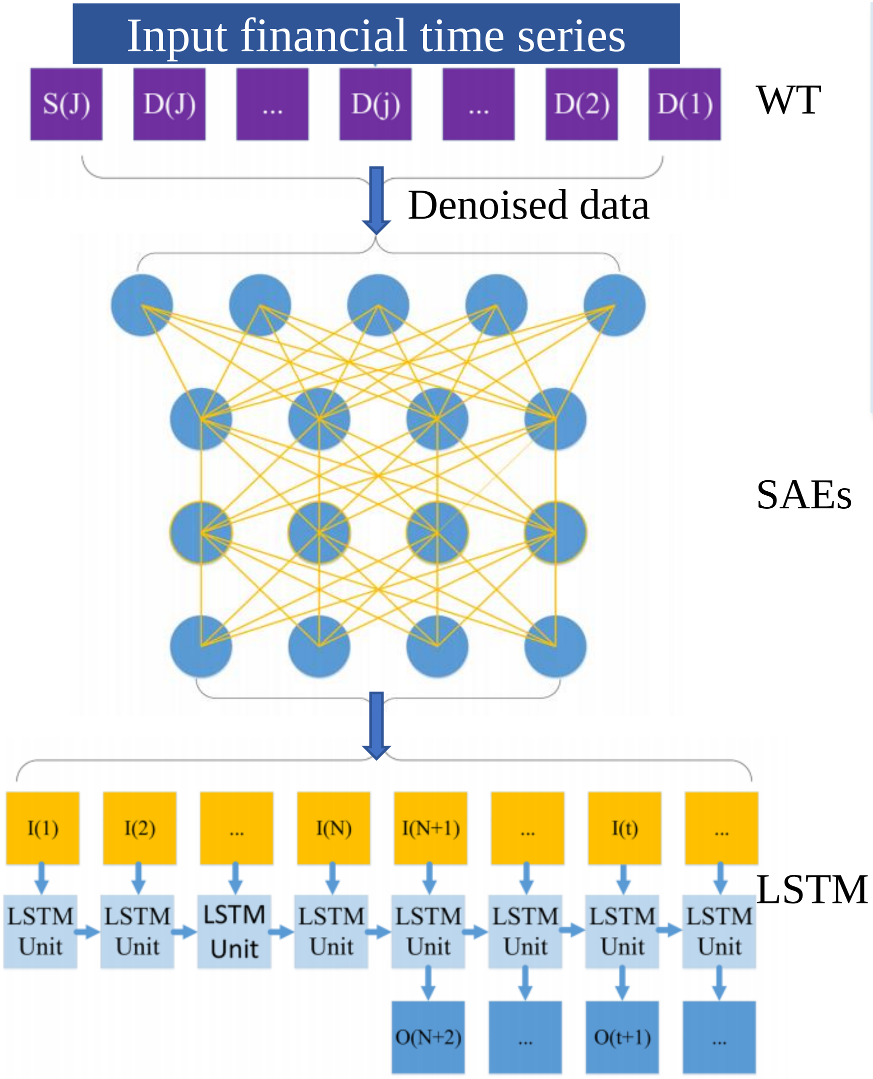
Advantages

- ➢ Ridge regression is a type of regularized linear regression method that **prevents overfitting and addresses multicollinearity** among predictors by adding an L2 penalty term (the sum of the squares of the coefficients) to the loss function.
- ➢ The L2 penalty helps **to shrink the coefficients that have less impact on the model**, making the model more stable and robust.

Hyperparameter Tuning

In Q2, we set λ as 0.5.

Overview of Deep Learning Models



Input financial time series

Denoised data

Model Framework

The stock market is difficult to predict with its noisy and volatile features. So a deep nonlinear topology should be applied to time series prediction. Our model includes 3 parts:**wavelet transforms** (WT), **stacked autoencoders** (SAEs) and **long-short term memory** (LSTM).

WT

WT is applied for data denoising since it can handle the unstable financial time series data.

+

SAEs

SAEs are used to learn useful features of the data by training the network to remove noise.

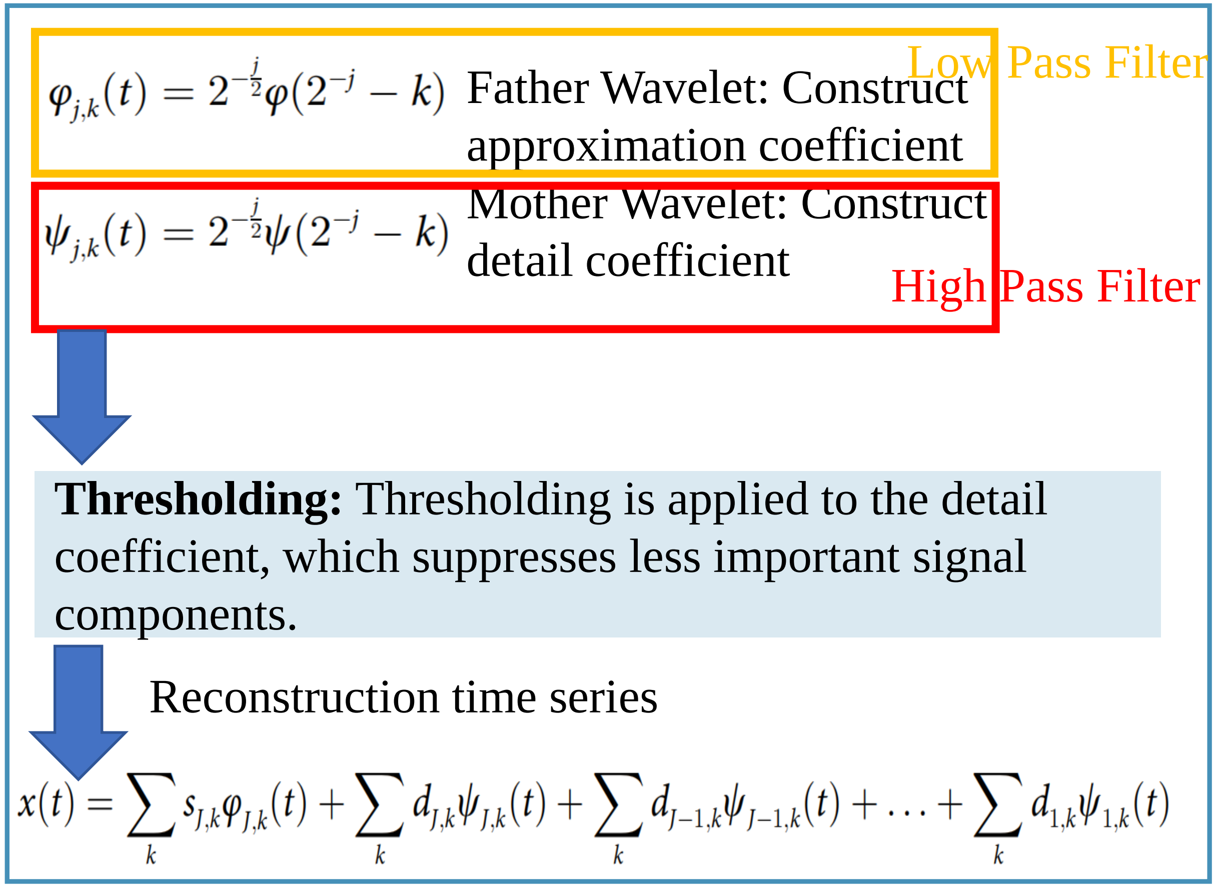
+

LSTM

LSTM is a type of recurrent neural network (RNN) that excels in learning from experiences to predict time series data.

Wavelet Transforms

Wavelet Transform (WT) is a method used in signal processing that provides a way to decompose and analyze signals at various scales or resolutions.



$\varphi_{j,k}(t) = 2^{-\frac{j}{2}}\varphi(2^{-j} - k)$ Father Wavelet: Construct approximation coefficient

Low Pass Filter

$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi(2^{-j} - k)$ Mother Wavelet: Construct detail coefficient

High Pass Filter

Thresholding: Thresholding is applied to the detail coefficient, which suppresses less important signal components.

Reconstruction time series

$$x(t) = \sum_k s_{J,k}\varphi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \ldots + \sum_k d_{1,k}\psi_{1,k}(t)$$

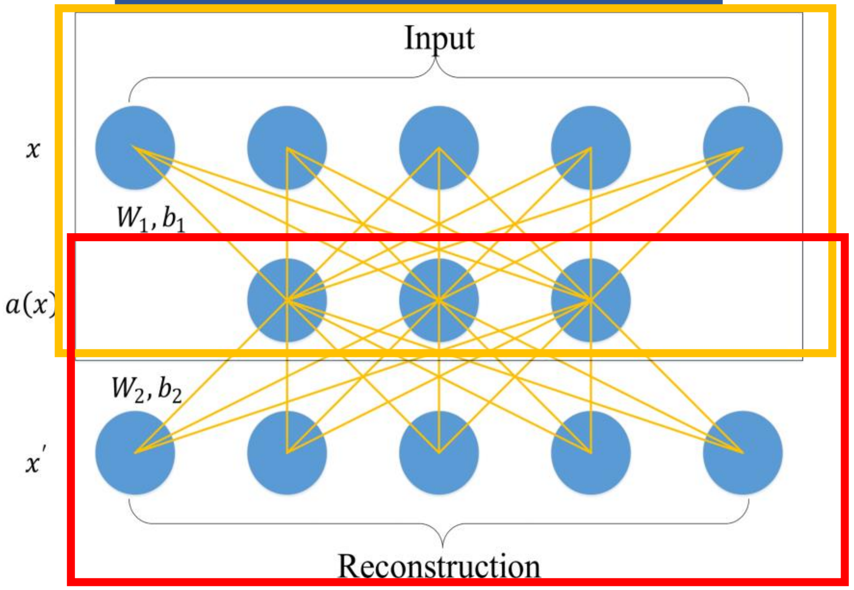
Key points

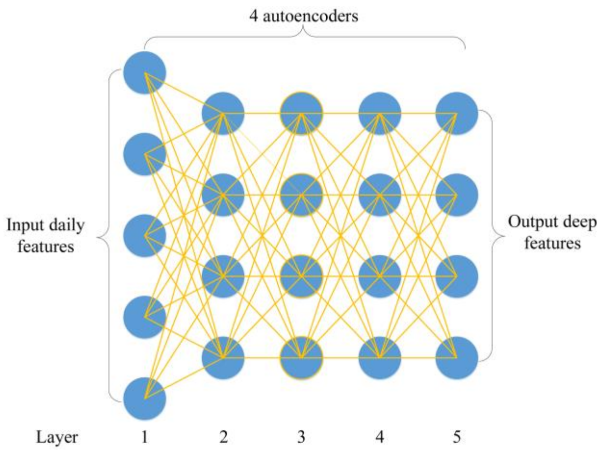
Wavelet Choice: The 'haar' wavelet is a common choice due to its simplicity and effectiveness.

Decomposition Level: 2 levels. The signal is decomposed into two sets of detail coefficients and one set of approximation coefficients.

Stacked Autoencoders

Single layer AE





Encoder

- Filter the input data to generate more abstract features, while reducing the dimension and removing some noise.

Decoder

- Encoded data is restored to the original high-dimensional form, then calculate the reconstruction error of the model and pursue minimization.

$$\mathrm{argmin}_{W_1,b_1,W_2,b_2}[J] = \underset{W_1,b_1,W_2,b_2}{\mathrm{argmin}}\,[(1/2)\Sigma_{i=1}^{m}\|x_i - x_i'\| + J_{wd} + J_{sp}]$$

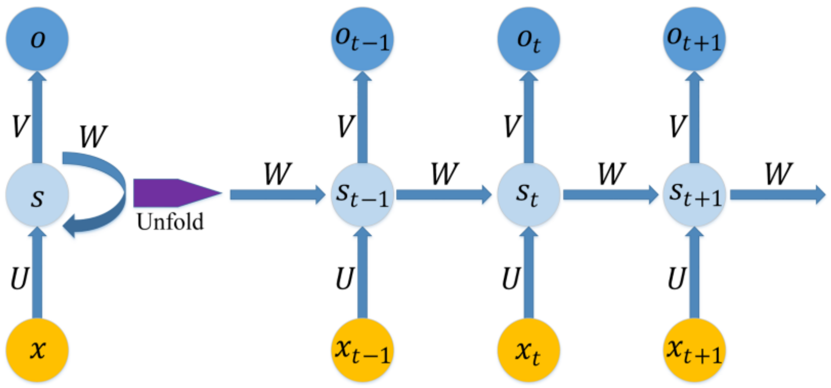
$$J_{wd} = (1/2)\lambda(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2)$$ Decay term

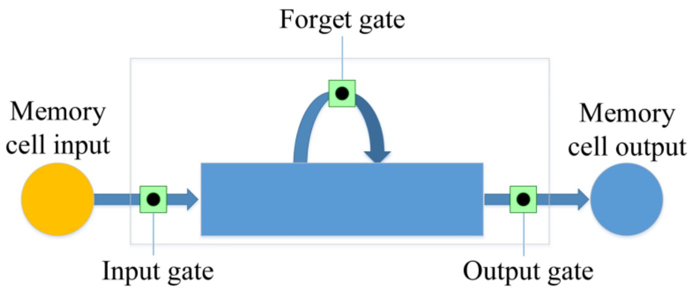
$$J_{sp} = \beta\Sigma_{t=1}^{m}KL(\rho \parallel \hat{\rho}_t)$$ Sparse penalty term

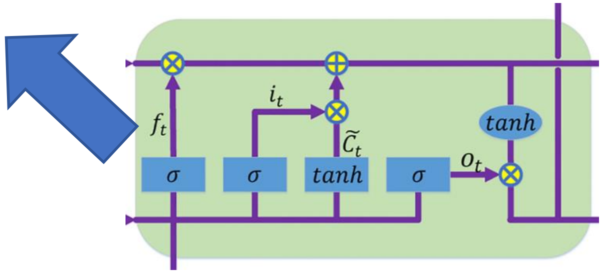
- **Initialization**: SAE initializes with unsupervised layer-wise training and fine-tunes the model through labeled supervised training.
- **Setting of hyperparameters:** Due to the data volume limitation, we cannot use very deep networks to avoid overfitting. The hidden layers are set to [20, 16, 8].

Long-short Term Memory

LSTM effectively captures long-term dependencies in time series for trend prediction.







- RNN vs LSTM which one is better?

RNN: has a depth structure in time dimension, but faces gradient disappearance and explosion problems when handling long sequences.

LSTM: can solve the gradient problem of RNN. Why?

Memory Cell:
Core component in LSTM, effectively addresses long-term dependency issues.

Gating Mechanism:

Input Gate controls how much new information can be added to the memory unit.

Forget Gate determines how much of the past information should be forgotten.

Output Gate determines which parts of the memory unit will be output to the next layer of the network.

Forecast Performance

Forecast Performance Measurements Metrics

$$MASE = \frac{\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|}{\frac{1}{N-1}\sum_{i=2}^{N}\left|y_i - y_{i-1}\right|}$$

$$R = \frac{\sum_{t=1}^{N}\left(y_t - \bar{y}_t\right)\left(y_t^* - \bar{y}_t^*\right)}{\sqrt{\sum_{t=1}^{N}\left(y_t - \bar{y}_t\right)^2\left(y_t^* - \bar{y}_t^*\right)^2}}$$

$$Theil-U = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - y_t^*\right)^2}}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t\right)^2} + \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t^*\right)^2}}$$

y_t is the actual value

y_t^* is the predicted value

N represents the prediction period

Benefit of the Selected Metrics

- MASE is calculated by MAE(Mean Absolute Error) divides MAD(Mean Absolute Difference) to scale the forecast error relative to the average absolute difference between consecutive observations.
- R is a measure of the linear correlation between two variables.
- Theil-U squares the deviations to give more weight to large errors and to exaggerate errors.

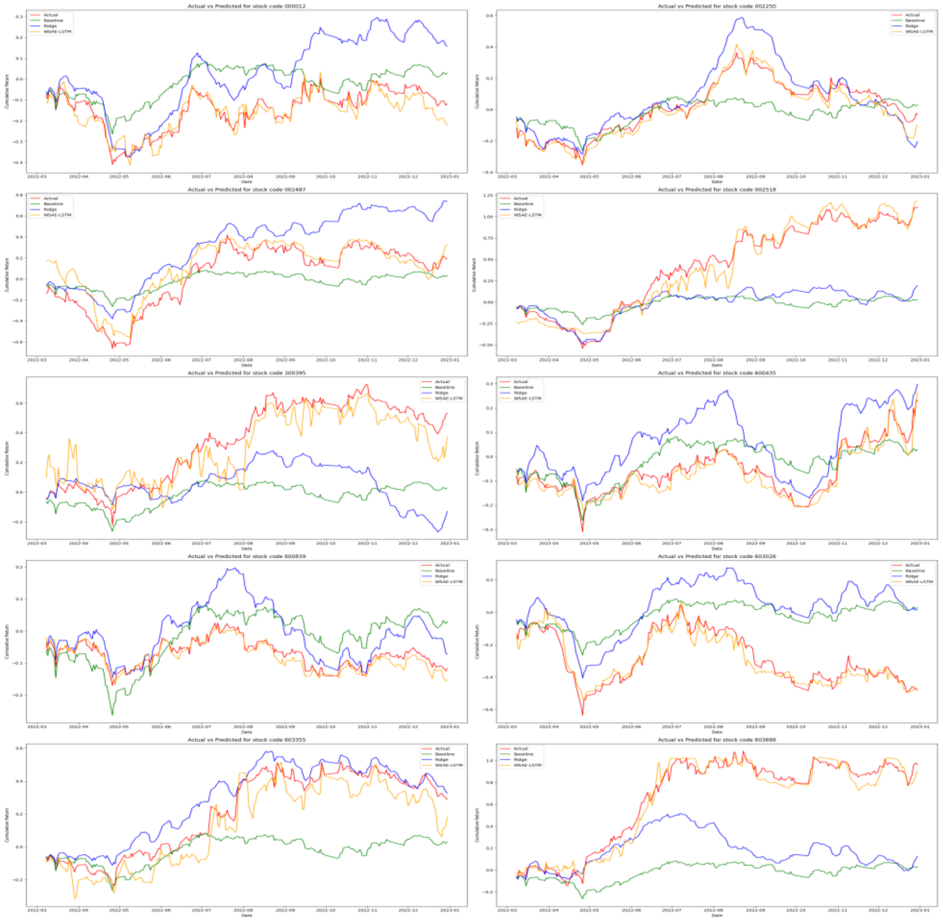
Forecast Performance

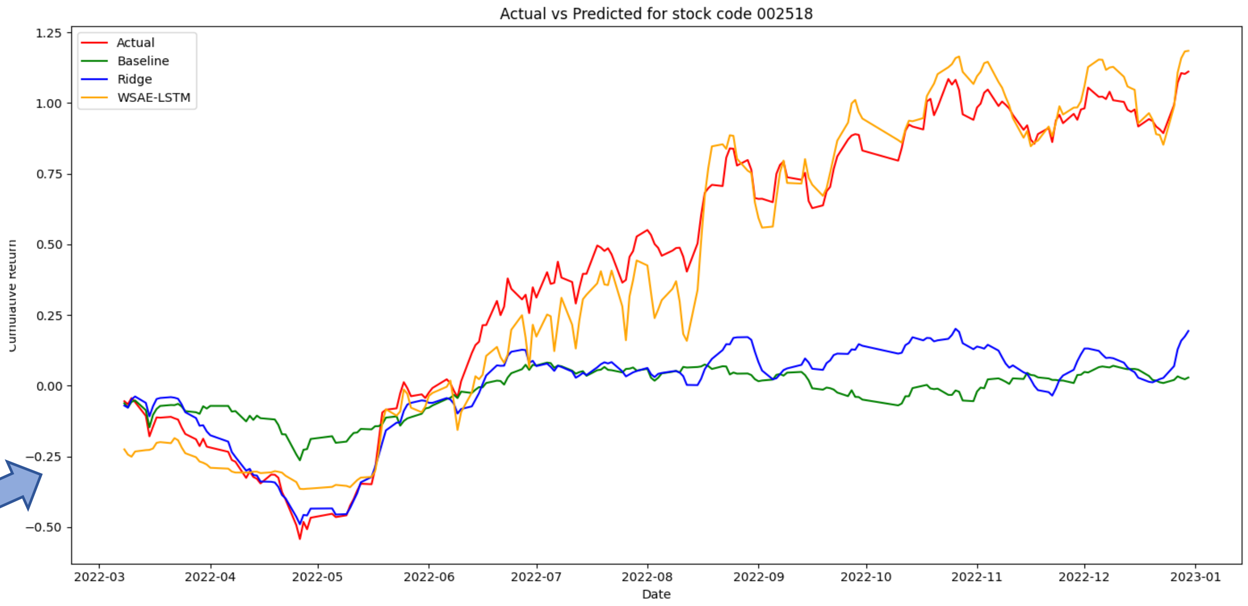
Metrics Results Showing

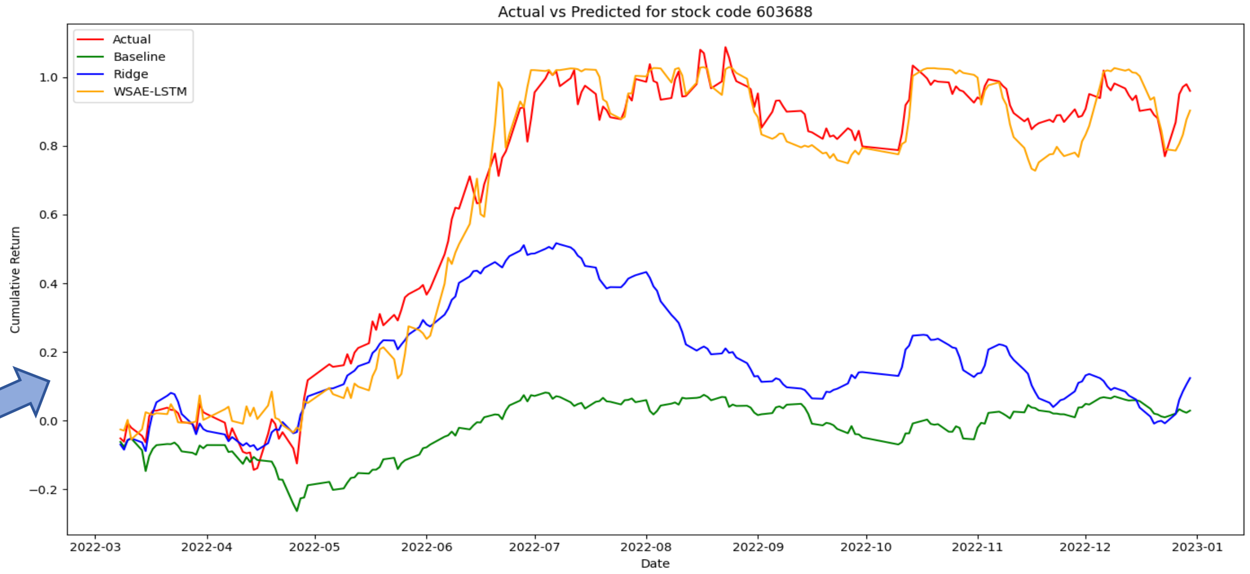
Stock ID	000012	002250	002487	002518	300395	600435	600839	603026	603355	603688
Panel A. MASE										
Baseline	6.285478647	5.612529512	6.285103923	15.25193198	14.38745866	3.933125979	6.826116251	10.80368289	11.44119166	21.61381062
Ridge	8.451251842	4.359207954	8.432936054	13.3523356	12.43632257	7.2517769	6.092826395	13.25305307	3.706676595	15.54865908
WSAE-LSTM	1.684395822	1.805343986	3.004938376	2.362637151	3.690210357	1.475738393	1.205477796	1.571444428	4.202634735	1.94305437
Panel B.Pearson Correlation										
Baseline	0.641223935	0.732723454	0.873270865	0.746405964	0.729339274	0.669000829	0.523361822	0.442546008	0.707933404	0.856454508
Ridge	0.915016437	0.925223163	0.927728666	0.859981482	0.312097239	0.910399899	0.891052455	0.634266368	0.975277725	0.510761327
WSAE-LSTM	0.914126702	0.984246965	0.942155065	0.983503707	0.912916344	0.972789495	0.941794524	0.956100205	0.969184919	0.9806499
Panel C.Theil-U										
Baseline	0.580981652	0.514401435	0.620414836	0.853751853	0.818403496	0.430288592	0.530270101	0.744053056	0.749043255	0.889876271
Ridge	0.554270299	0.26458038	0.40180002	0.694650477	0.672353351	0.521016352	0.517712846	0.729675142	0.136046921	0.581446604
WSAE-LSTM	0.118633286	0.117590036	0.222238639	0.075616049	0.145326307	0.126634872	0.096851905	0.073650545	0.170889032	0.050801452

Forecast Performance

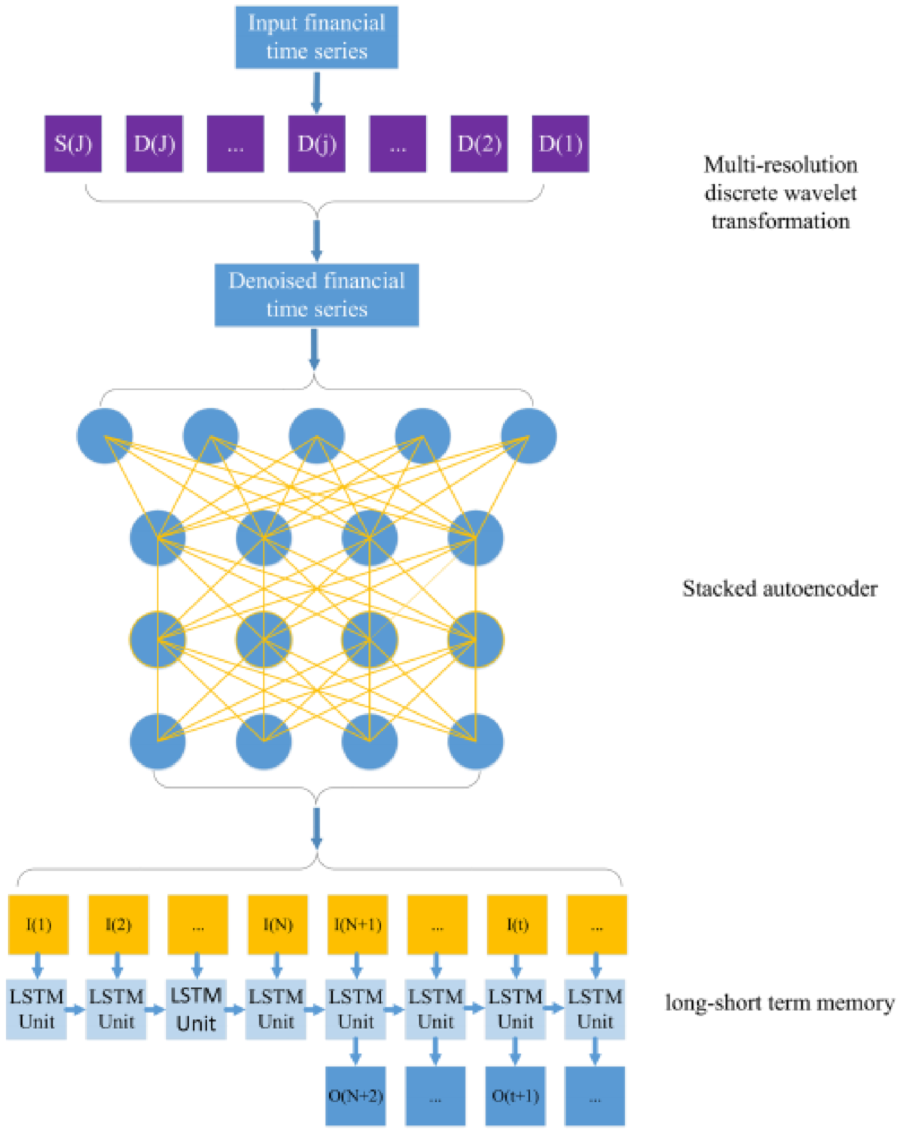
Predicted Data and Actual Data

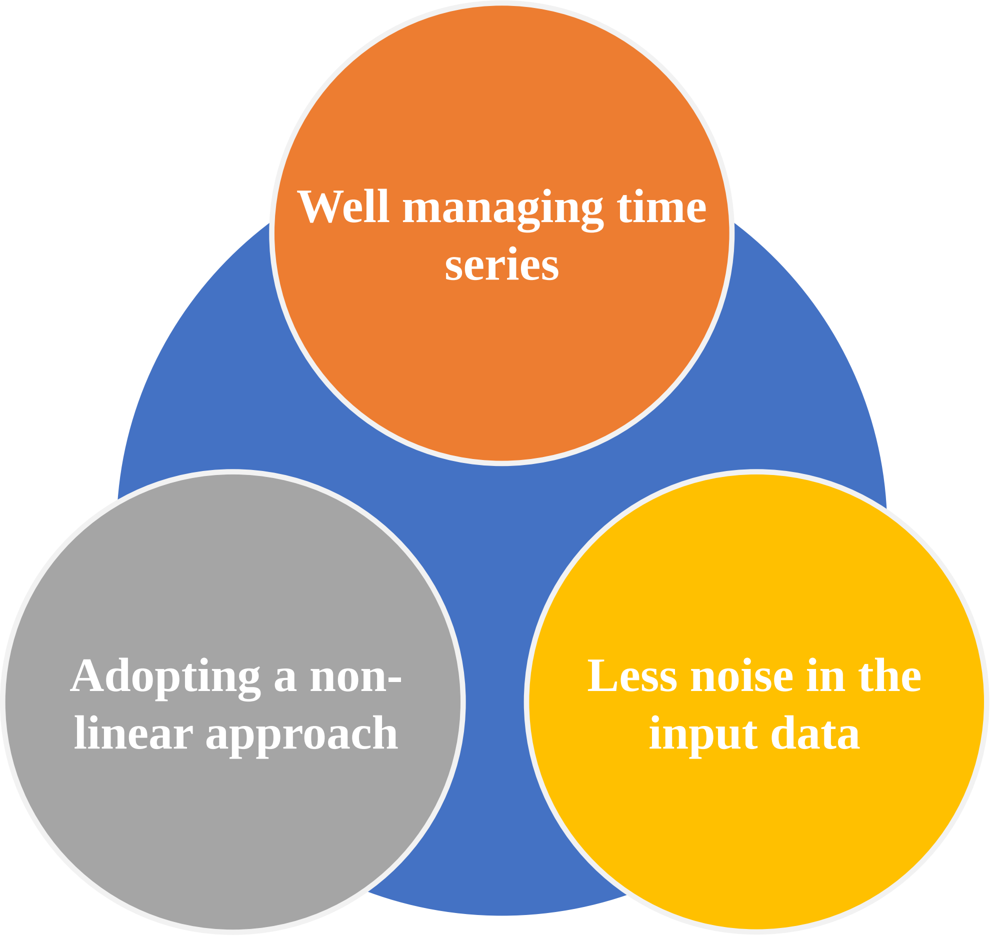






Reason Analysis





Well managing time series

Adopting a non-linear approach

Less noise in the input data

Appendix

Technical Indicators (Completed)

- RSI (Relative Strength Index): Measures the magnitude of recent price changes to evaluate overbought or oversold conditions.
- MACD (Moving Average Convergence Divergence): A trend-following momentum indicator that shows the relationship between two moving averages of a security’s price.
- BIAS: A technical analysis indicator that compares the closing price to a moving average to identify trends.
- CCI (Commodity Channel Index): An oscillator used to identify cyclical trends in a security.
- EMV (Ease of Movement): A volume-based oscillator that is designed to measure the ease of price movement.
- MTM6, MTM12 (Momentum): These are momentum indicators that measure the rate of rise or fall in stock prices.
- TRIX: Shows the percentage change in a triple exponentially smoothed moving average.
- VOSC (Volume Oscillator): Measures volume by comparing a short-period moving average with a longer one.

Additional Firm Characteristics

- ROE_TTM (Return on Equity, Trailing Twelve Months): Measures a corporation's profitability in relation to equity.
- ROA_TTM (Return on Assets, Trailing Twelve Months): Indicates how profitable a company is relative to its total assets.
- Current Ratio: A liquidity ratio that measures a company's ability to pay short-term obligations.
- LiquidityRatio, CashRatio: Indicators of a firm’s short-term liquidity and ability to use its cash to address immediate needs.
- LTDebt/E (Long-term Debt to Equity): Reflects the company's financial leverage and ability to meet long-term obligations.

Thanks for Listening
Q&A