THE CNINESE UNIVERSITY OF HONG KONG, SHENZHEN

FIN3210 Fintech Theory and Practice

FINAL PROJECT

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

1. Stock Selection & Data Preprocessing Strategy

1.1 Stock Selection

Dataset Collection: Start with a dataset of the stocks in CSI500. For each date, rank the stocks by their market value.

Recency Weighting: Assign weights to these rankings based on the recency of the data, with more recent dates getting higher weights. Hence rankings from more recent dates have a greater impact on the final score. Sum these weighted ranks for each stock across all dates.

Top Stock Selection: Select the top 10 stocks with the highest summed weighted ranks, The 10 stocks selected are as follows:

code	short_name
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

1.2 Data Preprocessing Strategy

Data Collection: Initiating the process, the dataset encompasses A/B shares in SH/SZ, ChiNext, and the sci-tech innovation board.

Data Cleaning: Implementing a meticulous data cleaning approach involves backfilling based on the grouping of stock codes and quarters. This is to ensure that all stocks have complete financial data when conducting analysis.

Creation of Dummy Variables: ‘Industry’ and ‘Year’ dummy variables are created. This helps to control the impact of industry characteristics and time effects on the model.

Normalization: Financial factors and price-related factors are normalized based on their percentile rankings, converting them into quantiles of a standard normal distribution, which ensures comparability among factors of different scales and ranges.

Log Transformation: In the analysis of the variable denoted as "total consumption level," a log transformation has been implemented to stabilize variance and mitigate skewness within the dataset. This is because the variable is a macro-level variable, and the data frequency is relatively low.

1.3 Selection of Indicators

The factors are categorized into five groups: company operational indicators, basic indicators, technical factors, economic and market indicators, and additional firm characteristics. Specific factors are shown in the Appendix.

Company Operational Indicators: These factors represent the fundamental financial health and performance metrics of a company. They are derived from financial statements and are often used by investors to gauge a company's profitability, asset management efficiency, and market valuation.

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.

Technical Factors: These factors are derived from statistical analysis of market activity, such as past prices and volume. They are used to forecast financial or economic trends and to create various technical indicators.

Economic and Market Indicators: take into account broader economic and market

signals which can affect the financial markets.

Additional Firm Characteristics

2. Construction of the Fama-French Three-Factor Model (Baseline Model)

2.1 Model Overview

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

The baseline model is based on the classical Fama-French three-factor model, which is a well-known approach in financial economics for explaining stock returns. Beyond the classical three factors, the model incorporates additional variables like GDP growth, total consumption level, and various other indicators (Ind_1 to Ind_5) to potentially capture other systematic risks or patterns in the data.

2.2 Model Implementation

The selected factors include 'rm-rf', 'SMB', 'HML', 'GDP Growth', 'Total Consumption Level', 'Ind_1', 'Ind_2', 'Ind_3', 'Ind_4', 'Ind_5', forming the feature set denoted as X. The target variable Y represents the excess return (ri-rf).

The training set comprises the first 1500 observations for each stock, while the remaining observations constitute the test set. This division ensures a robust assessment of the model's performance on unseen data.

2.3 Use of Clustered OLS Regression

The Q1 model employs a clustered OLS regression, which is suitable for data that may have intra-group correlation. The clustering is done based on the 'year', which implies that observations within the same year may be more similar to each other.

3 Construction of Ridge Regression Model (Regularized Linear Regression Model)

3.1 Model Overview

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

The ridge regression model is used for a regularized linear regression model, which is based on a linear model that extends the classic Fama-French three-factor model by including additional factors and characteristics along with fixed effects.

The model further incorporates additional variables such as the book-to-market ratio (B/M), return on equity (roe_ttm), return on assets (roa_ttm), current ratio, and various other financial metrics and market indicators.

3.2 Model Implementation

The selected factors include 'rm-rf', 'SMB', 'HML', 'GDPGrowth', 'totalConsumptionLevel', 'Ind_1', 'Ind_2', 'Ind_3', 'Ind_4', 'Ind_5', 'B/M', 'roe_ttm', 'roa_ttm', 'current_ratio', 'PB', 'PE', 'PCF', 'DivYield', 'PS', 'NPM/CA', 'operatingMargin', 'liquidityRatio', 'cashRatio', 'LTDebt/E', 'turnover', 'BIAS', 'CCI', 'EMV', 'MA10', 'MA20', 'MACD', 'MTM6', 'MTM12', 'RSI', 'TRIX', 'VOSC', 'VRSI', 'VWAP', forming the feature set denoted as X. The target variable Y represents the excess return (ri-rf).

Training and testing sets are the same as the baseline model.

3.3 Regularization of Model

Ridge regression, a type of regularized regression that includes an L2 penalty (squared magnitude of coefficients) in the loss function, is used to prevent overfitting and to handle multicollinearity among the predictors.

As for hyperparameter tuning, the regularization strength is controlled by the hyperparameter λ (lambda), which is chosen by the modeler. In this case, λ is set to 0.5.

4 Construction of Deep Learning Model

The stock market is difficult to predict with its noisy and volatile features. Hence, 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).

4.1 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. When applied to financial data, WT can be particularly useful for noise reduction, capturing important features of the data that might be overlooked by other methods.

The following are several key points regarding the model:

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

Decomposition Level: Level 2 is often used, which means the signal will be decomposed into two levels of detail coefficients and one level of approximation coefficients.

Denoising Process:

- Perform the wavelet decomposition on the time series to obtain the detail and approximation coefficients.
- Apply thresholding to the detail coefficients, which suppresses the less significant components of the signal, often assumed to be noise. Soft thresholding is typically used, which shrinks the coefficients towards zero.

- The standard deviation of the detail coefficients is often used to set the threshold value, with a common choice being half the standard deviation.

Reconstruction: Reconstruct the signal from the modified coefficients. After the detail coefficients have been thresholded, the inverse wavelet transform is used to synthesize a denoised time series.

Wavelet-based denoising is non-linear and adaptive, making it suitable for non-stationary financial time series, where the statistical properties change over time.

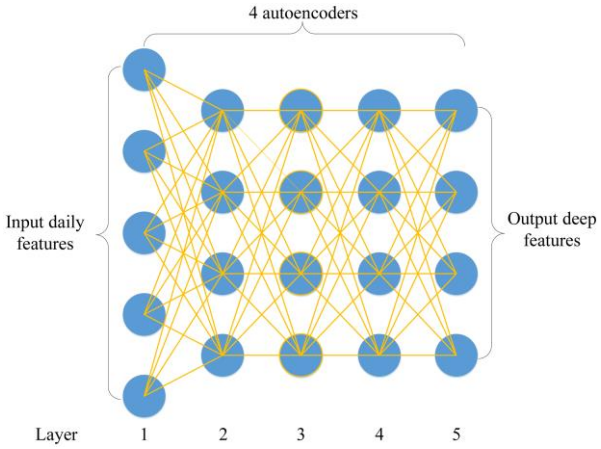
4.2 Stacked Autoencoders

Subsequently, the data undergoes processing through stacked autoencoders, emphasizing the extraction of essential features.

Encoder& Decoder: Initially, we scrutinize the architecture of a single-layer autoencoder. In the encoding step, the model generates more abstract features, effectively reducing data dimensionality and mitigating noise. Subsequently, encoded data is restored to the original high-dimensional form the decoding step calculates the reconstruction error, and the model undergoes optimization to minimize this error. The introduction of sparsity is a key aspect, enforcing a penalty term that maintains the inactivity of hidden layer neurons, contributing to computational efficiency and mitigating overfitting risks.

Initialization & Setting of hyperparameters: It's noteworthy that SAE initiates with unsupervised layer-wise training and subsequently fine-tunes the model through labeled supervised training. In addition, due to constraints related to data volume, our hidden layer sizes are configured to 20, 16, and 8. This strategic selection aims to strike a balance between model complexity and computational efficiency.

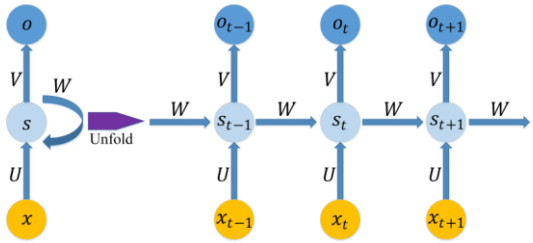
The schematic below illustrates an example of Stacked Autoencoders (SAE) with five layers, comprised of four single-layer autoencoders.

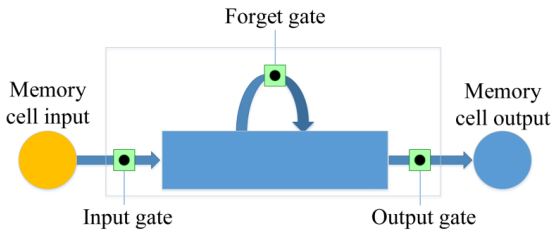


4.3 Long-short Term Memory

LSTM is employed as a key component of the model, primarily used for predicting stock prices based on high-level features generated by the SAEs. It's a type of recurrent neural network (RNN) that excels in learning from experiences to predict time series data, particularly adept at handling time steps of arbitrary sizes. The LSTM, in this model, contributes to improving the predictive accuracy by retaining and processing time-related information, addressing issues like vanishing gradients which are common in traditional RNNs. This inclusion of LSTM in the proposed model helps achieve better performance in terms of predictive accuracy and profitability, as demonstrated through tests on various individual stocks.

The following is the basic structure of LSTM:





5. Prediction Performance of Three Models

5.1 Forecast Performance Measurement Metrics

In this paper, we choose three classical indicators (i.e., MAPE, R, and Theil U) to measure the predictive accuracy of each model. The definitions of these indicators are as follows:

$$MASE = \frac{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\frac{1}{N-1}\sum_{i=2}^{N}|y_i - y_{i-1}|} \quad R = \frac{\sum_{t=1}^{N}(y_t - \bar{y}_t)(y_t^* - \bar{y}_t^*)}{\sqrt{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2 (y_t^* - \bar{y}_t^*)^2}} \quad Theil - U = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - y_t^*)^2}}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t)^2} + \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t^*)^2}}$$

MASE is 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 is a relative measure of the difference between two variables. It squares the deviations to give more weight to large errors and to exaggerate errors. If R is bigger, it means that the predicting value is similar to the actual value, while if MASE and Theil U are smaller, this also indicates that the predicted value is close to the actual value.

5.2 Forecast Performance

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

The Metrics Results above show the specific indicators used to measure the accuracy performance of three models for each selected stock. According to the metrics, we can find that there is an obvious increase in the prediction accuracy with our WSAE-LSTM model:

MASE: With a substantially lower mean MASE (2.14), the WSAE-LSTM model demonstrates superior average forecasting accuracy, indicating its effectiveness in capturing underlying patterns in the time series data. Additionally, the lower standard deviation (0.87) and a narrower range between the 25th and 75th percentiles (1.28-2.36) for the WSAE-LSTM model suggest a more stable and consistent performance compared to the other models.

Pearson correlation: the Pearson correlation analysis highlights distinct performance trends among the three models. The WSAE-LSTM model stands out with the highest mean correlation, indicating a consistently stronger linear relationship between its predictions and the actual values. Furthermore, the WSAE-LSTM model exhibits both a higher minimum and maximum correlation, reinforcing its superior ability to capture varied patterns and consistently provide accurate predictions across a diverse set of stocks.

Theil-U: With the lowest minimum (0.05) and maximum (0.22) Theil-U values, the WSAE-LSTM model demonstrates resilience in handling both challenging and straightforward forecasting scenarios. Moreover, the lower 25th (0.09) and 75th (0.14) percentiles for the WSAE-LSTM model underscore its consistent superiority in capturing underlying patterns across the entire set of stocks compared to the Baseline and Ridge model.

In addition, the graph above compares the predicted data from the three models with the corresponding actual data for the time interval from March 2022 to January 2023. According to the figures, we can find that the Baseline and Ridge models have larger variations and distances to the actual data than the WSAE-LSTM model, which is consistent with the quantitative analyses conducted on Mean MASE, Pearson Correlation, and Theil-U values. This pattern suggests that the WSAE-LSTM model provides more reliable and accurate predictions across the given set of stocks when compared to the Baseline and Ridge models.

6. The Reasons Why the WSAE-LSTM Model Outperforms the Baseline and Ridge Model

6.1 Less Noise in the Input Data

Haar wavelet's ability to capture abrupt changes makes it well-suited for denoising, enhancing signal clarity by effectively removing noise and irrelevant fluctuations.

The unsupervised nature of SAEs allows the model to autonomously discover hierarchies of features, capturing intricate temporal patterns that may exist at different scales.

6.2 Consideration of Time Series

SAE: By considering the temporal aspect, the model gains a more nuanced understanding of how features evolve, contributing to improved forecasting accuracy.

LSTM: They are specifically designed for time-series forecasting, allowing them to inherently manage sequential data more effectively.

6.3 Forecast Adopting a Non-linear Approach

The non-linear capabilities of LSTM contribute to effective pattern recognition, allowing the model to capture intricate relationships that linear models might overlook.

Reference

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for industrial and applied mathematics.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*.

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

Appendix

Company Operational Indicators

- *PB (Price-to-Book Ratio)*: Indicates the market's valuation of a company compared to its book value.
- *PE (Price-to-Earnings Ratio)*: Shows how much investors are willing to pay per dollar of earnings, 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)

- *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.

Technical Factors

- *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.

Economic and Market Indicators

- *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.

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.