1. **Selection of 10 stocks**

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



1. **Data Preprocessing:**

Market data is filtered for specific market types (Markettype == 63, 沪深AB股和创业板和科创板).

Financial statement columns are backfilled based on the group of stock code and quarter, ensuring no missing values in these columns.

**Creating Dummy Variables:** Industry and year dummy variables are created to capture industry-specific and year-specific effects in the models.

**Normalization:**

A custom function *reverse\_normaliz*e is created to normalize financial factor values based on their percentile rank transformed to the quantiles of the standard normal distribution.

The normalization is applied to a set of specified financial factors and price-related factors.

Log transformation is applied to the total consumption level to stabilize variance and reduce skewness.

**SMB and HML Factor Calculation:**

A function is defined to calculate the Small Minus Big (SMB) and High Minus Low (HML) factors monthly. These factors are derived from stock returns categorized by size and book-to-market ratios.

Factors List:

'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'

**Company Operational Indicators:**

These factors typically 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.

* 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):**

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.

**Technical Factors:**

Technical 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.

* 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:**

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.

**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.

1. **Models:**

* **Q1 baseline model:**
  + X: ['rm-rf', 'SMB', 'HML', 'GDPGrowth', 'totalConsumptionLevel', 'Ind\_1', 'Ind\_2', 'Ind\_3','Ind\_4', 'Ind\_5''
  + Y: ri-rf
  + The data is divided into a training set and a test set, with the training set consisting of the first 1500 observations for each stock, and the test set comprising the remaining observations.

The Q1 model is based on the classical Fama-French three-factor model, which is a well-known approach in financial economics for explaining stock returns.

The model includes market risk premium (rm-rf), SMB (Small Minus Big) which represents the size premium, and HML (High Minus Low) which represents the value premium as its core factors.

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.

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.

* **Q2 Ridge model:**
  + X: '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'
  + Y: ri-rf
  + Training and testing sets are the same as Q1.

Q2 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.

Regularization: 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.

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.

1. WT

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.

**Some key points:**

**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.

The output is a DataFrame where each column represents the denoised time series of the input factors. This data can now be used for further analysis or modeling, with reduced noise potentially leading to more accurate insights and predictions.

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