Single Factor Models Fitting, SMF

Single Factor Models Fitting Method:

1. **Regression Method**
2. **IC Method**
3. **Tiered Backtesting Method**

Regression Method

1. Data Collection

For example: 2021,01,01 - 2024,05,31; 1h

2. Factor calcualtion

For example: quote_asset_volume, number_of_trades, taker_buy_base_asset_volume, taker_buy_quote_asset_volume and some derivative factors

- Some specific factors
- Tsfresh
- Featuretools
- worldquant Alpha 101
- Guotaijunan Alpha 191

3. Data preprocessing

- Data cleaning: remove or correct missing, erroneous or abnormal data.
  - **Median-Based Outlier Removal**:
    - Let the value of a factor at time T be D_i.
    - Calculate the median D_M of D_i.
    - Calculate the median absolute deviation |D_i − D_M| and denote it as D_M1.
    - For all D_i, if D_i>D_M+5D_M1, then set D_i to D_M+5D_M1; if D_i<D_M-5D_M1, then set D_i to D_M-5D_M1.
  - **Missing Value Handling**:

        - Check if there are missing values in the new series of factor values D_i.
        - For missing values, fill them using the overall sample’ s mean or median. The choice between these methods depends on the distribution characteristics of the data (skewness or normality).
- Data transformation:
    - Z-score Normalization (Standardization)
    - Min-Max Scaling (Normalization)
    - Log Transformation, Box-Cox Transformation, Binarization, Binning (Bucketing)
    - Eg:
        - Standardization:
            - Calculate the mean and standard deviation of the factor values after handling missing values.
            - Subtract the mean from each factor value and divide by the standard deviation, so that the processed series of factor values approximates a standard normal distribution N(0,1).
        - Normalization:
            - Calculate the minimum ☒min*X*min and maximum ☒max*X*max values of the factor values after missing value treatment.
            - Normalize each factor value to scale between 0 and 1 using the formula: Xnorm=(X−Xmin)/(Xmax−Xmin)
- Data segmentation:
    - Train set: 2021,01,01 - 2023,12,31
    - Test set: 2024,01,01 - 2024,05,31

4. Model Assumptions

- Dependent variable(Y)
    - Forward return
        - Default: 1h, 4h, 8h, 1d, 3d, 7d, 14d
- Independent variable(X)
    - Default: 1h, 4h, 8h, 1d, 3d, 7d, 14d

5. Model selection

- **Linear model**:

    - Linear Regression
    - Lasso Regression
    - Ridge Regression
- **Non-linear model**:
    - Logistic Regression
    - Linear Discriminant Analysis, LDA
    - Quadratic Discriminative Analysis, QDA
    - Decision Trees Regression
    - Random Forest Regression
    - XGBoost
    - AdaBoost
    - Gradient Boosting Machines, GBM
    - Support Vector Machine Regression
    - Gaussian Processes
    - K-Nearest Neighbors, KNN
    - Neural Networks
    - RNN, LSTM, GRU

6. Model validation

- General Evaluation Criteria:
    - Accuracy
    - Mean Squared Error (MSE) or Root Mean Squared Error (RMSE)
    - R-squared (Coefficient of Determination)
    - Area Under ROC Curve (AUC-ROC)
    - F1-Score
    - Information Coefficient (IC)

7. Backtesting

- Initial Setup
    - Initial Capital: 0
- Signal Generation

    - Buy Signal (`1`): Predicted return exceeds a positive threshold.
    - Sell Signal (`-1`): Predicted return falls below a negative threshold.
    - No Action (`0`): Predicted returns are between these thresholds.
- Trading Costs
    - Assuming a transaction cost of 5 basis points (bps).
- Metrics Calculation
    - sharp ratio
    - return/CAGR
    - maximum drawdown
    - calmar ratio
    - a plot of performance curve

8. Outcome Explaination