

**Introduction to FinTech**

**Data Analysis Project 1**

**Predicting Stock Returns Using Machine Learning**

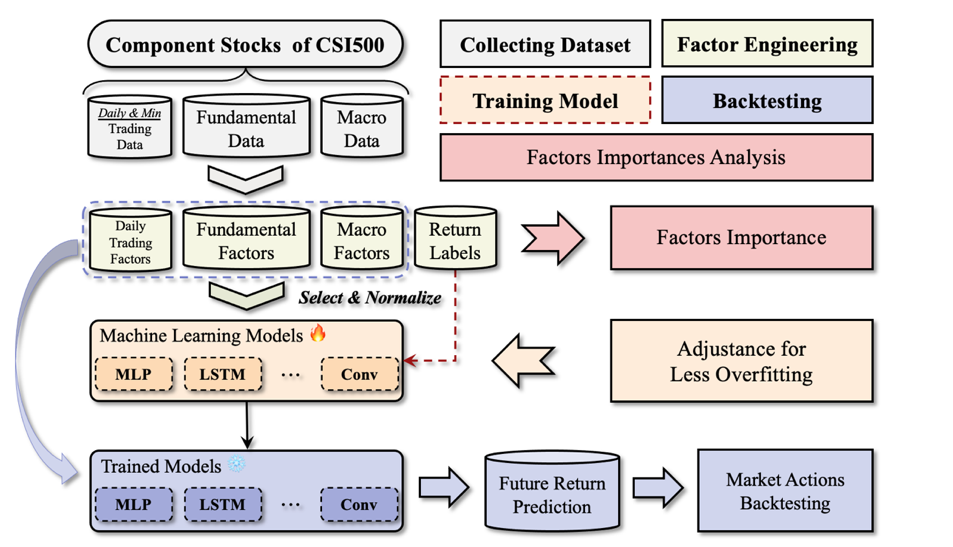
|  |  |
| --- | --- |
|  | |
| Kai Ren | 2401212437 |
| WeiJie Zhang | 2401212474 |
| Yang Lan | 2401212401 |
| YiMing Jiang |  |
| Rui Hu | 2401212392 |

**Peking University**

2024.11.4

# Abstract

For Data Analysis Project 1, we built a research framework for stock return forecasting using machine learning models to extract factor features end-to-end. The 500 constituent stocks in the CSI 500 (CSI 500) are modeled, and their 5.5-year trading data from January 1, 2019 to June 1, 2024 are collected, including daily and minute volume and price data, company fundamentals, and macroeconomic data. Based on these raw data, we study and construct 167 factors, including 26 daily trading factors, 64 high-frequency factors, 66 fundamental factors, and 11 macroeconomic factors. After eliminating the samples with serious missing factors, the factors are standardized using the Z-Score method on the cross-section. Based on this, we build the classical time-series neural network architectures such as MLP, LSTM, GRU, Transofrmer, and Conv to extract the factor features, which are used to predict the daily return of the opening price on this day. Under the premise that all data are used as training data, both for regression modeling and classification modeling, we obtain 100% of the theoretically derivable ACC and F1 within the sample, which validates the correctness of our modeling framework. Subsequently, we divided the training samples into three parts: training set (January 1, 2019 to June 1, 2023), validation set (June 1, 2023 to January 1, 2024), and test set (January 1, 2024 to June 1, 2024) according to time intervals to simulate real modeling scenarios in the industry. Based on this, we found obvious overfitting phenomenon and tried to adjust it in several aspects, such as adjusting the model architecture and introducing the lag period factor. Further, we constructed simple strategies and backtested them based on the predicted values over the test interval to obtain detailed returns. Finally, in order to better understand the financial implications of the factors in the modeling process, we used a tree model to provide a detailed analysis of factor importance.

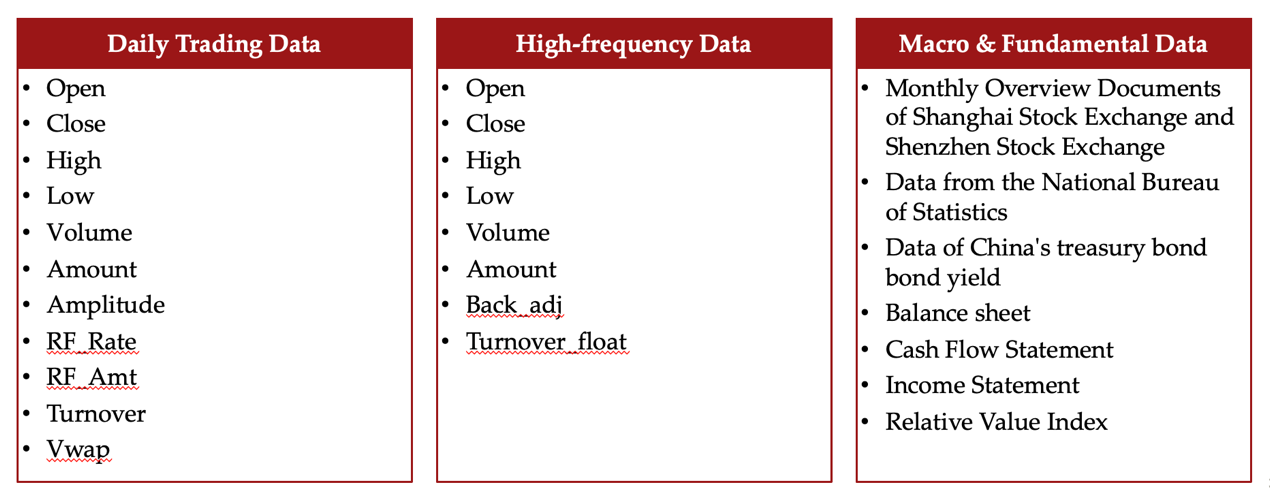


**Figure 1.** The Framework of Data Analysis Project 1

# Data Collection

For this project, we use data from January 1, 2019 to June 30, 2024 for the CSI 500 constituent stocks. The CSI 500 consists of 500 stocks from the Shanghai and Shenzhen markets with medium market capitalization and good liquidity. The fact that it covers a wide range of industries and is balanced and diverse enough to represent the overall performance of the Chinese market is the reason why we made such a selection.

In terms of time dimension, we choose the past five and a half years as the study interval, which is because it is difficult to mine sufficiently valid information with too short a time interval; while if the time span is too large, the accuracy of the results will be challenged by the fluctuation of the data over time, and at the same time too large a volume of data will lead to a larger computational cost. In addition, considering the subsequent modeling needs, i.e., through the rolling training and forecasting methodology, the factor model is trained using time series of 1 year or a few months, and then forecasting the return of 1 month or a few weeks in the future. Taking the above considerations into account, we believe that the time horizon of January 1, 2019 to June 1, 2024 is very reasonable and can meet the demand for studying current market changes, and the conclusions obtained are highly current and accurate. Finally, considering that the CSI 500 constituents will be adjusted periodically, we chose the 500 constituents of the CSI 500 index on September 30, 2024 as the modeling object.



**Figure 2.** Overall data included as source data.

For data collection, Figure 2. shows the overall data included as source data. We first collected the daily frequency trading data of all stocks in the above time interval based on the AkShare packet, which totaled 606,475 entries, smaller than the product of the number of stock samples and the number of trading days, 656,000 entries, due to the fact that some stocks were suspended from trading on some of the trading days, which resulted in missing data. Among them, China Baoan (000009.sz) has the highest number of trading days with 1,312 days, and Kodak Manufacturing (600499.sh) has the lowest number of trading days with only 277 days.

Second, we collect trading data at the minute frequency within the time interval for the subsequent calculation of the high-frequency factor. Finally, we collect macro data including SSE SZSE monthly profile files, NBS data, and China government bond yields in the CSMAR database during the time interval to calculate the macro factor, and collect annual financial data of companies such as income statements, cash flow statements, balance sheets, and relative value indicators for the subsequent calculation of the fundamentals factor.

For the label construction, we use the daily return under the next day's opening price as the forecasting target, i.e.

Which aims to predict future stock price movements using factor data up to the present.