Analysis and Forecast of the Chinese Stock Market

Jiyuan Liu,

Yuchen Wang, and

Junseok Yang

Department of Statistics, University of Illinois at Urbana-Champaign

STAT 430: Time Series ML

Prof. Hyoeun Lee

December 11, 2023

Abstract

Stock markets can be unpredictable, and investors are always trying to select stocks that maximize their profits and mitigate risks. This study aims to use traditional time-series analysis methods, clustering algorithms, and deep learning techniques to analyze multiple features for stock selection and forecast the return rate on 500 Chinese trading stocks. Our results show that traditional autoregressive integrated moving average models (ARIMA) can produce relatively accurate forecasts for up to 10 trading days. Clustering methods are ill-suited for analyzing stock characteristics, patterns, and trends. Gated Recurrent Units (GRU) is accurate at predicting stock returns and grouping stocks

Keywords: time series, stock market, machine learning

Introduction

In the stock market, there are many ways to choose high-quality stocks, but this project intends to analyze stock data in different industries and apply multi-factor stock selection. Some factors can act on a single stock, but some factors act on a certain industry, so the analysis of individual stocks and industries is crucial.

Traditional time-series analysis is used to forecast short and long-term industry-wide return rate trends. The return rate is an important attribute as this metric can be quantitative trading. An accurate forecasting model can give investors better profitability and a competitive edge and also be used in risk management as a protection for riskier investments.

Clustering analysis will be used to understand the underlying trend of return rate in the market and further bind with other features like industries for more useful insights. The goal is to understand the characteristics of the return rate change over time and evaluate different trend patterns over different industries.

Gated Recurrent Units (GRU) analysis uses different features to predict return rates for individual stocks to group stocks and choose the high-quality stocks we need. Using the existing features in GRU to predict the future return of each stock, we can then observe and analyze the prediction results. Group stocks were created based on predicted return value to observe its ability to select stocks, we chose specific indicators to observe the stability of stock forecasts.

About the Data

Our data contains trading and enterprise characteristics of 500 Chinese stocks that are collected from the first day of 2018 to the last day of 2021. The data has 19 variables and 1 label. The primary response variable is the stock's return rate at the end of each day. Some of the covariates include enterprise value, long short ratio, price-to-book ratio, turnover rate, and a few derivatives from these variables. Detailed descriptions of all the variables can be found in the appendix.

All of these feature variables are continuous numeric variables, however many of them are sparse. The data can be understood as 3-dimensional data. However, we have reduced the dimension to two by storing the trading and enterprise attributes for each stock for each day as a row.

The data was generously provided by Red Wall Taihe Fund Management Co., Ltd. through a previous internship to us for academic purposes.

Descriptive Statistics

Table 1.

Selected Key Variable Descriptive Statistics

	Average Return Rate (%)	**Average Enterprise Value (RMB)**	**Start Enterprise Value (2018-01-01)**	**End Enterprise Value (2021-12-31)**	**Growth Rate (%)**
Mean	0.000195	3.634559e+10	3.100029e+10	4.617061e+10	148.75
Standard Deviation	0.000730	1.951540e+11	1.550422e+11	2.414321e+11	200.12
Minimum	-0.003630	1.023521e+09	1.188578e+09	1.106626e+09	9.56
Maximum	0.002647	3.871058e+12	3.154605e+12	4.784026e+12	2423.28

The table above shows some basic summary statistics of the data. The average return rate is just over 0 with a standard deviation of 0.0007. The range of the values is not huge, suggesting that the return rate does not deviate much from 0. The enterprise value shows that the data includes a broad range of companies. We see that most companies have grown over the time that the data was collected, with the average growth rate being 1.5 times.

Statistical Methods

ARIMA Analysis

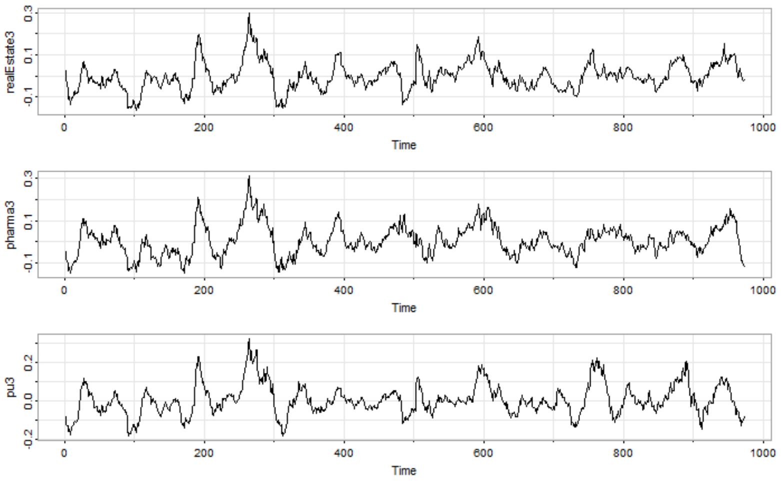
Since we have 500 stocks, this greatly hampers our ability to manually analyze each series. The stocks were grouped into their associated industries and the mean return rate was taken, and three industries were randomly selected, these industries are real estate, pharmaceuticals, and power and utilities. A number of autoregressive integrated moving average models will be fitted to forecast the immediate short-term and possibly longer-term return rates.

Immediate Short-Term Forecast

To forecast the immediate short-term return rates, the data will be split into a training and testing set, with the testing set composed of rates of the last 10 trading days in 2021. The following figure contains the return rate time series plots for each industry.

Figure 1.

Time Series Plots of Return Rates for the Three Selected Industries

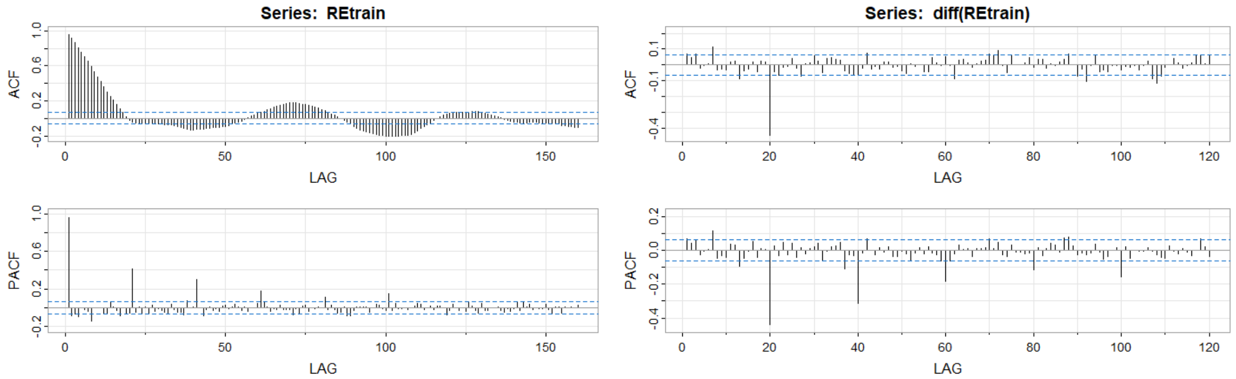


Note: realEstate3 is the return rate of real estate, pharam3 is of pharmaceuticals, pu3 is power and utility

Looking at the plots, we see that all three industries share very similar trends and patterns over time despite being very different, the return rates show a very mild upward trend; it is hard to spot any seasonal patterns. However, the volatility of the return rate changes over time.

Figure 2.

ACF and PACF plots for the Real Estate Return Rates and Differenced Return Rates



Looking at the ACF and PACF plots, we see that the series exhibits a seasonal pattern that seems to occur every 20 trading days, which is a "trading month". The pharmaceutical and power and utility series also exhibited near-identical patterns. KPSS test results indicate that all the series are not stationary. The series will be differenced, then seasonally differenced if needed. Figure 2 also shows the ACF and PACF plots of the differenced real estate return rates, again, both pharmaceuticals and power and utilities are very similar. The series does not exhibit non-seasonal AR or MA components, however, the one peak in the ACF at lag 20 and the tailing-off behavior seen in the PACF for the significant peaks for every 20 lags strongly suggested that there is an MA(1) component.

The differenced and seasonally differenced series ACF and PACF plots looked nearly identical to the only non-seasonally differenced ACF/PACF. This suggests that a

SARIMA(0,1,0)(0,0,1)[20] (Model 1) or SARIMA(0,1,0)(0,1,1)[20] (Model 2) model might be appropriate. In addition to these models, a stepwise search selected model for each of the series will also be fitted. The stepwise search will use the Akaike information criterion (AIC) to select its best model. The model it has chosen is ARIMA(1,1,1) for real estate, ARIMA(1,1,0) for pharmaceuticals, and ARIMA (1,1,2) (Model 3a, 3b, 3c) for power and utilities. To compare these models, both RMSE and visual inspection on the testing forecast will be used to determine the model performance.

Short-Term Forecast SARIMA Model Fitting

The following figures contain the diagnostics plots for the best-fitting model, SARIMA(0,1,0)(0,0,1)[20]. The model fits well, the SMA(1) term is significant at a 95% confidence level. However, we see that there are some assumption violations, the residuals seem to deviate from the normal Q-Q line at the tails, and we see some lags that fail the Ljung-Box statistics, suggesting that there is some autocorrelation in the residuals. The following table compares the testing RMSE values of all fitted models.

Table 2.

Test RMSE Table for All Fitted Models

	Real Estate			Pharmaceuticals			Power and Utility		
Model	1	2	3a	1	2	3b	1	2	3c
RMSE	0.039	0.041	0.047	0.074	0.118	0.121	0.050	0.057	0.059

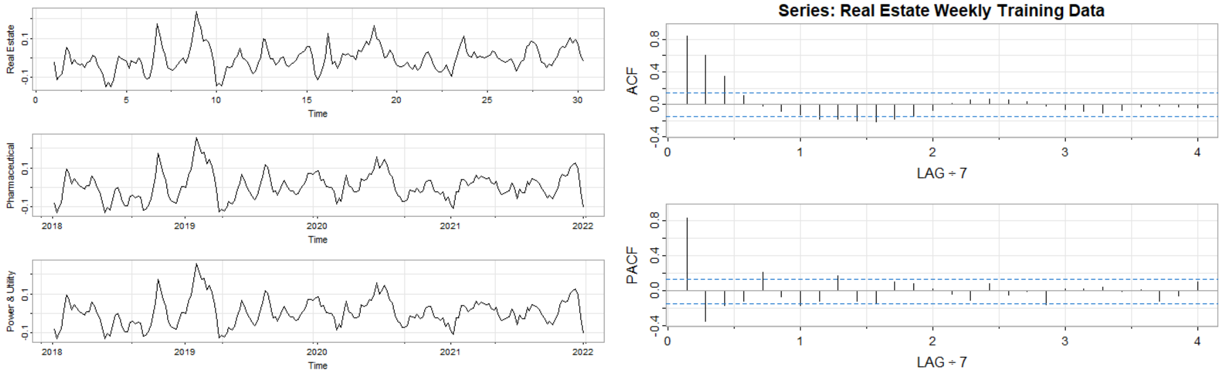
Model 1 fits the real estate data the best, but the other two industries also had low RMSE values. Therefore we will choose this as our final short-term forecast model. We will give the full forecasting result under the Result Section.

Aggregated Weekly Data Forecast

To perform the longer-term forecast, the data was aggregated into weekly data by computing the weekly average. As a brief summary, the SARIMA model chosen for the daily data performed poorly when it was tasked to forecast beyond 10 trading days. Therefore to give a better long-term forecast, the data was aggregated. Like our previous approach, the data will be split into training and testing sets, with the testing set consisting of the last 10 weeks of return rate data.

Figure 3.

Time Series Plots for Aggregated Weekly Data and ACF and PACF plots for the Weekly Real Estate Return Rates



Like the daily data, we see similar trends and patterns in the weekly data across three very different industries. We see that the data is quite volatile, which the series pattern is not consistent over time, as we see a huge rise and fall in early 2019 and a very consistent return rate in 2021. However, the ACF and PACF plots along with the KPSS suggest that the weekly data is actually stationary. The ACF and PACF plots suggest a possible ARMA(2,3) model as we see 3 significant peaks in the ACF and 2 in the PACF. There is a very mild weekly seasonal pattern, but it is not significant, which suggests that seasonal differences might not be needed. We still examined that possibility but the results of the seasonal differences suggest it was not appropriate to do so as it induced significant autocorrelation. The ARMA(2, 3) is fitted along with a stepwise search selected ARIMA model, which is SARIMA(2,0,1)(1,0,0) for real estate and power and utility and SARIMA (1,0,3)(1,0,0) for pharmaceuticals.

Longer-Term Forecast SARIMA Model Fitting

The two SARIMA models outperformed the manually selected ARMA(2, 3) in RMSE scores. The ARMA(2, 3) had no assumption violations and had significant AR(1), (2), MA(1), (2), (3) components. The stepwise search models also had no alarming assumption violations. However, only the SARIMA(2,0,1)(1,0,0) had significant coefficients when it was fitted to the power and utility series. These two models will be chosen as the final models and the results will be covered in the Results section.

Cluster Analysis

Clustering is a type of unsupervised learning whose data has no label or response variable Y, unlike any typical supervised learning such as regression or classification. Instead of predicting or classifying, the main interest is to find groups of observations that share similar features and understand any notable patterns or trends that are potentially underlying in both the clusters and data, which serves as descriptive analysis (James et al., 2023).

With the original stock data, data manipulation was done to fit into diverse clustering algorithms. After the preprocessing step, each row represents a unique stock with its columns being a time range, which can be considered as data of multiple univariate time series. For instance, the first column is the first day of 2018 (i.e. January 1st, 2018), the second column is

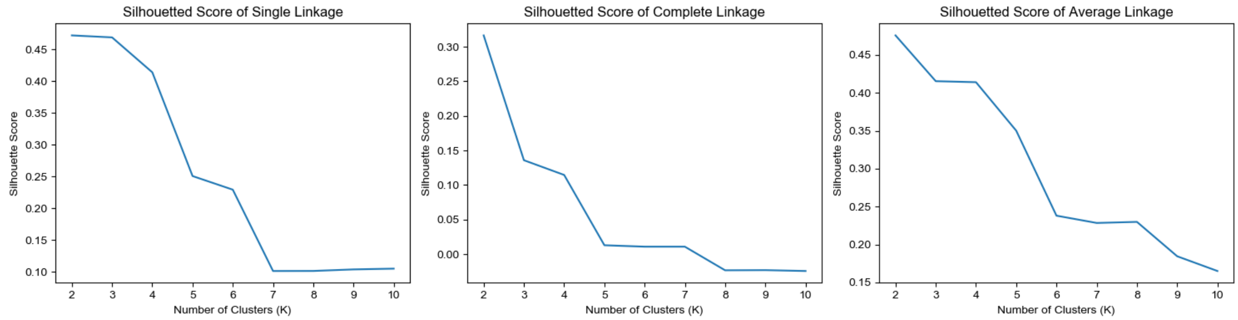
January 2nd, 2018, etc, and each cell represents the return rate of a stock of its corresponding date. Considering that there are a total of 973 days or columns in the data, reducing the dimension was inevitable for computation efficiency. As a result, combining the columns and averaging the return rate by month was able to narrow down the dimension to 48 (12 months per year, total 4 years from 2018 to 2021).

Two different clustering algorithms were tested: Hierarchical agglomerative clustering (HAC) and DBSCAN. It is known that HAC requires input data in a distance matrix form and DBSCAN can also accept distance matrix data as input, additional data transformation was processed for a more accurate comparison of the algorithms' performance. By calculating the distance between every pair of data, the dimension of data converts into an n-by-n matrix which n represents the number of observations. There are many methods of computing distance between data points, and although the original data has shrunk down by aggregating the average return rate by month, 48-dimension is still relatively large. A typical distance metric of Euclidean may suffer from a well-known problem of the curse of dimensionality, which is computationally complex and likely to hinder getting accurate results (Hastie, 2016). Thus, an alternative metric of cosine similarity was chosen for converting the data into a distance matrix form.

The clustering evaluation metric is another essential step to check the performance of a clustering algorithm. Silhouette score is one of the widely used metrics that shows how cohesive data points are within their assigned cluster and well separated from other clusters. By evaluating the cohesion and separation of the clusters, one can choose the optimal algorithm for clustering analysis (Ellison, 2021).

Hierarchical agglomerative clustering (HAC) was first tested to cluster the data. According to Tan et al. (2020), the algorithm assumes every data point is a cluster. By connecting the clusters based on proximity with different linkage methods, similar or close groups are combined into one cluster and this process continues until all data points or clusters form one big cluster, yielding a dendrogram which allows to see all the combining steps.

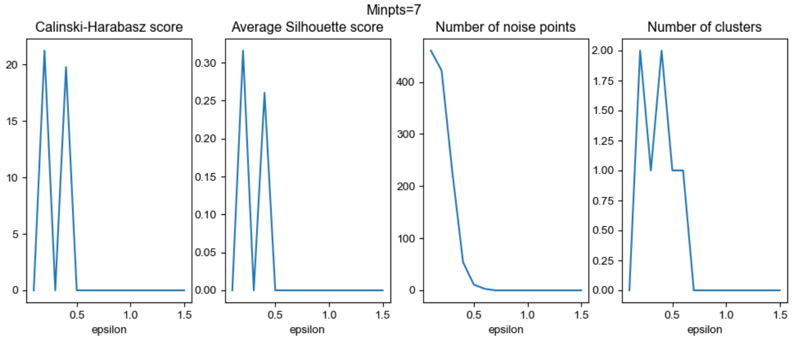
Figure 4.



From the figure above, it is clear that the overall silhouette scores are not high and tend to decrease as the number of clusters (K) increases. Although one can choose K=2 based on the score, it is not the only absolute standard on determining the cluster nor guaranteeing a meaningful outcome. One of the main research goals is to find several clusters representing both common and outlying patterns at the same time, K=8 with average linkage was chosen for hierarchical clustering.

DBSCAN was implemented as well for the analysis, which stands for Density-Based Spatial Clustering of Application with Noise. As it is in the name, its computations are based on density and therefore useful to identify and separate outliers from highly dense clusters. Referring to Tan, the two essential parameters to define before applying are ‘epsilon’ and ‘minpts’ which label every point as ‘core’, ‘border’, or ‘noise’. Epsilon (ε) is a radius from a point which then draws a circle boundary. Minpts is the minimum number of points that are required for a point to be defined as ‘core’. If there are more than ‘minpts’ within the boundary of ‘epsilon’ radius, the point is defined as ‘core’. If the first condition is not met but is still on the radar of core points, this becomes ‘border’. If none of the conditions are met, the point will be treated as ‘noise’.

Figure 5.



The figure of comprehensive clustering evaluation scores connected with the number of clusters above suggests that the combination of epsilon=0.4 and minpts=7 is the optimal parameter value.

Deep learning (GRU) Analysis

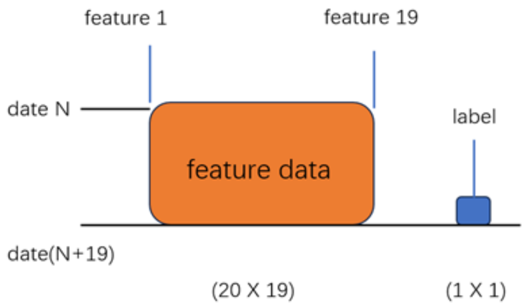
Gated Recurrent Units (GRU) is a gating mechanism in recurrent neural networks. Different from traditional machine learning, GRU uses the long-term feature data of a certain stock as a time series and takes the data in the past period as the input value. The biggest advantage of doing this is to maintain the persistence of information. Therefore, in the stock data and features, the data of individual features in the past time will affect data in the future. So, we select GRU to analyze the data.

In fact, RNN\LSTM\GRU are all relatively similar algorithms. The traditional RNN model is prone to the problem of gradient disappearance and is difficult to handle long sequences of data. The reason why the gradient disappears is essentially because the calculation method of the hidden layer state causes the gradient to be expressed as a continuous product. Therefore we are more inclined to choose LSTM and GRU. Since the amount of stock data is relatively large, there are about 500 stocks and about 1,000 trading days, and the total input data units are about 400,000. In order to facilitate parameter adjustment, we choose GRU, which has a relatively fast operation.

For the GRU model, we use the data from 2018-2020 as the training set, the data from 2021.01-2021.07 as the validation set, and the data from 2021.08-2022.12 as the test set. We select 20 as the time series length of a single input data. The 20-day return rate is used as labels, and the shape of the input single data is (1, 20, 19), where 19 is the number of features. After constructing the single data into the shape of (1, 20, 19), eliminate data containing missing values and stock data with a total trading market of less than 20 trading days. Then normalize the data on each day and compress the data size to 0-1.

Figure 6.

GRU Diagram



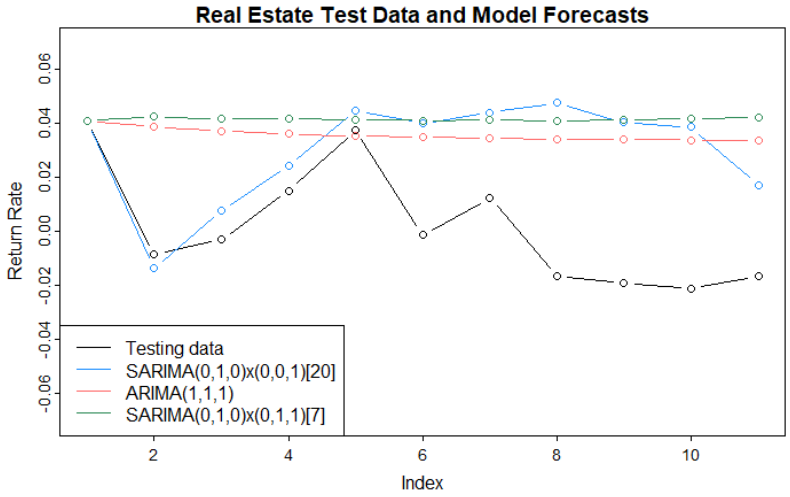
Results

ARIMA Analysis Outcomes

Let's first look at the forecast for the daily data, we see that the SARIMA(0,1,0)(0,0,1) was able to capture both the magnitude and the direction of the changes in the first 4-5 trading days, however, the magnitude is gradually lost after that. This suggests that the ARIMA model might not be appropriate for forecasting a longer term using daily data.

Figure 7.

Daily Return Rate Forecasts

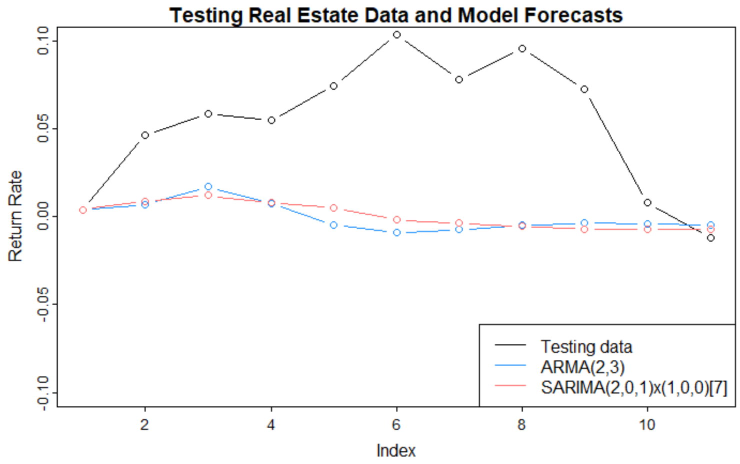


Note: The first value in the plot is the last known value/last value of the training set.

The weekly data, a visual inspection of the forecast on the test data revealed that neither model was fit to forecast the weekly data, as the model forecast is unable to capture the magnitude nor the direction of the true series. Considering that all the series were highly volatile, this suggests that ARIMA models might not be appropriate to forecast these return rates and instead the ARCH/GARCH models should be explored.

Figure 8.

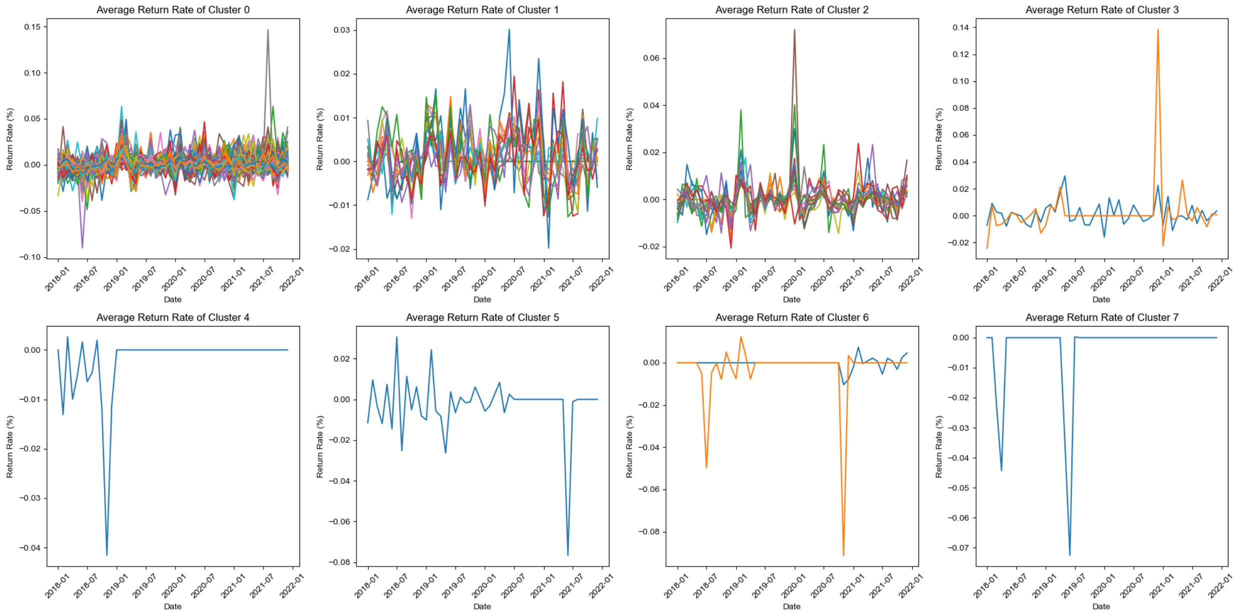
Weekly Return Rate Forecasts

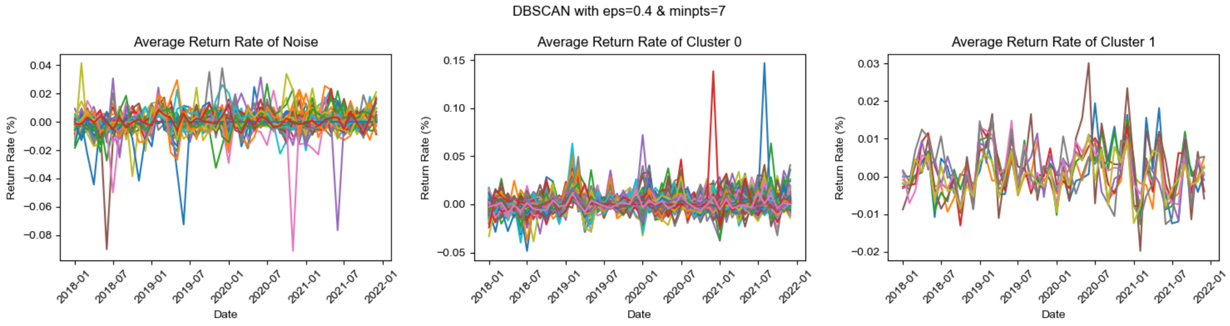


Cluster Analysis Outcome

The average return rate trends by cluster of HAC and DBSCAN can be found below.

Figure 9.





Note: HAC with average linkage of 8 clusters (above), DBSCAN with epsilon=0.4 & minpts=7 of 3 clusters (below)

It is clear to figure out that the hierarchical clustering algorithm was able to detect and identify some stocks with suspicious average return rate patterns from the main clusters with common trends whereas relatively inconsistent time-series lines were displayed from DBSCAN. While the main clusters showed some mild to notable consistent fluctuations, the outliers revealed several drastic drops with a long-term 0% return rate.

Figure 10.

Histograms Top Industries in Clusters

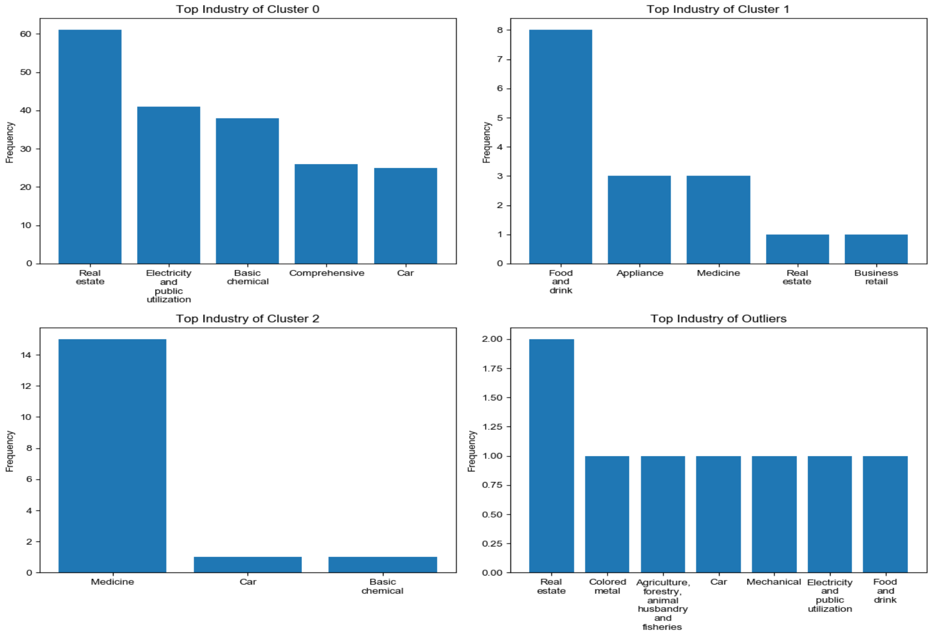


Table 3.

Key Descriptive Statistics by Clusters

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster Outlier**
Average Return Rate (%)	**0.000185**	**0.000976**	**0.000291**	**-0.001063**
Average Enterprise Value (RMB)	**3.419308e+10**	**1.321561e+11**	**1.407755e+10**	**1.170136e+10**
Average Growth Rate (%)	**146.20**	**266.11**	**114.52**	**177.92**

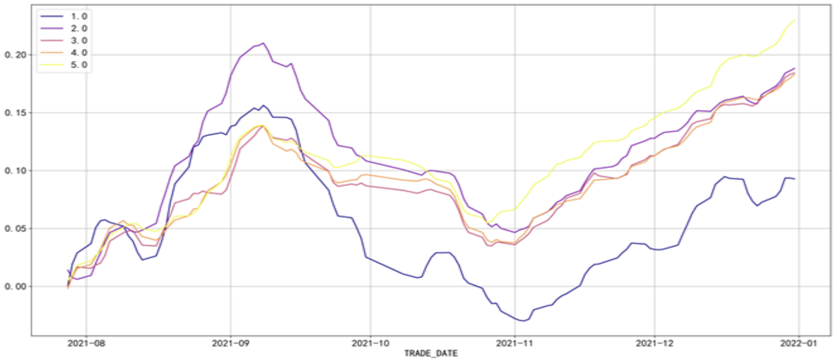
Note: Top industry by cluster, cluster 3 to 7 as outliers (above), Growth Rate = $(\frac{End\ Enterprise\ Value}{Start\ Enterprise\ Value})$ x 100 (below)

Further analysis was done by connecting other features such as industry and enterprise value, and more meaningful insights were detected. Every cluster had a different industry proportion; real estate for Cluster 0 and outliers, food and drink for Cluster 1, and medicine for Cluster 2. The first cluster showed mild to weak ups and downs from Figure 8 with a moderate average growth rate from the table above. On the other hand, cluster 1 had the highest average return rate with relatively extreme fluctuations than the first cluster with the average growth rate reflected this number from the table as well. The third cluster skyrocketed in 2020, and this can be connected to the fact that the pandemic started in the year. Plus, considering that most of the stocks in this cluster are related to medicine, this may be a factor that led them to grow fast in this specific time range. However, no significant increasing patterns were observed before and after the year 2020, which showed that the average return and growth rate were low. Lastly, the cluster with outliers had a negative return rate, meaning that the stocks were losing their value on average. Although the average growth rate is relatively high, this is due to the fact that there were some stocks with 0% return rate for a long time, resulting in no end enterprise value, and thus unable to calculate the growth rate properly.

GRU Outcome

After I get the predicted return rate, we use this data to classify stocks into five groups. The result shows that group 5 (the yellow line) with the largest predicted return rate has the highest return in the real market.

Figure 11.



There is another indicator “IC”. IC is the correlation value between the future return and the predicted return.

$$IC_t = corr([factor_{stock_1,t}, \dots factor_{stock_n,t}], [return_{stock_1,t+m}, \dots return_{stock_n,t+m}])$$

$$IC = \frac{\sum_{t=0}^{N} IC_t}{N}$$

When it is higher than 0.03, we can say the predicted return is performed stable in prediction. In the table, the 20-day mean value of IC is 0.0747, so it performs stable in the prediction.

Table 4.

IC Table

	0	5	10	20
Mean	**-0.0023**	**0.042**	**0.0544**	**0.0747**
Standard Deviation	**0.2178**	**0.1785**	**0.1627**	**0.1504**
T-statistics	**-0.1104**	**2.4278**	**3.4463**	**5.1135**

Discussion

The SARIMA model's performance was notable, suggesting that these types of models could be applied to individual stocks to give immediate future forecasts and mitigate risks for investors. The longer-term was less relevant as stocks were highly volatile due to other covariates that were not included in the model. While adopting SARIMA on aggregated weekly data to perform longer-term forecasts was suboptimal, ARCH/GARCH models should be explored more for these series considering high volatility.

For clustering analysis, the overall performance of clustering algorithms was relatively poor, and this could be an indication that the data was not clusterable initially. Some possible future work can be done by testing more algorithms such as with a wider range of parameter tuning might yield better performance.

The performance of GRU was positively notable in learning the time series impact of stock factors. Though the loss of some data on account of missing values is unavoidable when using GRU, it still performed well in the IC and stock grouping. It was able to select high-quality stocks when choosing the stock with the highest predicted value directly.

Lastly, different approaches to data preprocessing such as multivariate time-series by including more features could potentially allow the detection of sophisticated and useful insights relevant to the research focus.

References

Ellison, V. (2021). *Unit 3. Clustering Evaluation Metric* [pdf]. http://courses.las.illinois.edu/fall2023/stat207/syllabus.html

Hastie, T., Tibshirani, R., & Friedman, J. H. (2016). *The elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition)*. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning with applications in Python*. Springer International Publishing.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2020). *Introduction to data mining*. Pearson Education Limited.

Work Distribution/Contribution

Jiyuan Liu (jiyuanl2) - Deep Learning (GRU)

Yuchen Wang (yuchenw7) - ARIMA

Junseok Yang (jyang247) - Cluster Analysis

Appendix

Data Dictionary

Table 5.

Data Dictionary

Variable	**Description**
EV	Enterprise Value. Total value of a company, defined in terms of its financing
EV_EBITDA	EV is divided by EBITDA, which is the abbreviation for earnings before interest, taxes, depreciation, and amortization.
EV_SALES	Price-to-sales ratio of enterprise
long_short_change:	Do EWMA on long_short_ratio in different time windows and then find the difference between them
long_short_ratio:	The amount of security available for short selling versus the amount borrowed and sold
MOM_30	Momentum in 30-Day Stock Returns
MOM_60	Momentum in 60-Day Stock Returns
PB	Price-to-book ratio of enterprise
PCF_OT	Total market capitalization/net operating cash, Trailing 12 Months (TTM)
PCF_T:	Total market capitalization/net cash increase TTM
PE	Price–earnings ratio
PS_T	Price to Sales Ratio TTM
std_Nm_30	Volatility of stock rolling returns over 30-day window
std_Nm_60	Volatility of stock rolling returns over 60-day window
TURNOVER_RATE_30	Exchange ratio of stocks on the 30-days

TURNOVER_RATE_60	Exchange ratio of stocks on the 60-days
wgt_return_Nm_30	Multiply the stock return rate by the Exchange ratio of stocks and then get 30-day rolling average
wgt_return_Nm_60	Multiply the stock return rate by the Exchange ratio of stocks and then get 60-day rolling average
Closing_price	Stock’s return rate