ML Methods for Stock Price Time Series Prediction

Zhaozhong Qi
College of Engineering
Boston University
zqi5@bu.edu

Zemeng Wang
College of Engineering
Boston University
zmwang@bu.edu

Shuyan Zhang
College of Engineering
Boston University
shuyanzh@bu.edu

1. Introduction

When people think of investing, the first thing that comes to mind is the use of secret strategies to investigate the common sense and rules behind massive trades in every single transaction. Then, they are used to generate wealth by creating algorithms that simulate those most genius strategies. Economics tries to tap into the vast data sets behind the stock market. They explore the relationship between data and society, humans, and the underlying rules behind the market, and give these data a specific name - candidate factors. Specifically, candidate factors are unstructured data that are closely related to finance and may be of service to finance. This candidate data can be better used to leverage the machine learning domain, such as building appropriate modules (features and labels).

In this project, we will use machine learning models to investigate the correlations and patterns behind stock closing prices and candidate data. Our focus is based on the historical dataset, which is mainly influenced by macroeconomic developments and investor sentiment, etc. We use the LASSO regression algorithm to extract the factors in the dataset that have a high correlation with the closing prices, and then use Support Vector Regression and Kernel ridge regression to fit and predict the closing prices of Amazon and China The closing prices of Amazon and Ping An Bank Co. The accuracy of the final prediction is over 85%.

Stock prices are a common reflection of a company's development, so models with appropriate factors and weights are able to measure a company's profits and performance, as well as predict future performance. However, a large number of empirical but irrelevant features may lead to overfitting of machine learning models to the training data set, which then side-effects the accuracy of the model. Therefore, the screening of factors has been a common challenge in the industry and academia. Then, machine learning and adaptive trading methods adapt to the rapid changes in stock markets, so they have the potential to achieve more accurate prediction results than humans.

1.1. Time Series

Time series is a sequence of data points that occur in successive order over some period of time, such as earnings, GDP and stock price. It includes the change and movement of a certain item in a specific period of time. In investing, time series is utilized to track the price of a security. In the financial context, security is a kind of certificate or other financial instrument with monetary value and can be traded in the market. Time series analysis can be used to track the price of securities in both short and long term, from hour scale to month scale.[4]

In the report, we use time series data with open price, close price, high price, low price and so on to predict the close price of one day, five days and twenty two days ahead, which correspond to next trading day, last trading day of a week and a month respectively.

1.2. Research Procedures

Stock Selection Criterion

Decision model could not perform perfectly when the dataset has more noise, therefore two well known companies around the world are used:

Amazon: Most valuable U.S.-based internet company by market capitalization.

PingAn Bank: The third market capitalization in ShenZhen stock exchange.

Feature Selection

Numbers of factors associated with the close price and the limited realistic computation, it is inevitable to reduce the redundant factors and simplify the data with several significant factors.

Prediction

Time series prediction uses information among historical values and associated with patterns to predict future activity. This has most often been applied to trend analysis, periodic fluctuation analysis.

Evaluation

Evaluation is to examine the accuracy and feasibility of prediction.

1.3. Data Source and Platform

Historical datasets of PingAn Bank. Fifteen years stocking price and factors in daily records, China. (API: JoinQuant, `https://www.joinquant.com/help/api/help#name:JQData`)

Historical dataset of Amazon, Ten years stocking price and factors in daily records, United States. (API: Kaggle, `https://www.kaggle.com/`)

Reasons of Selection

In purposely, we first choose the datasets that are satisfied who possess a long enough history since its date of release of Initial Public Offerings(IPO). Since, in order to carry over enough magnitude of the historical value to feed the inquiry of the SVR and KRR model, we need to ensure it actually stands a long historical and stable enough. The point of view about why it also needs to be relative stabilization. It is just because SVR's Decision model could not be performing well under the model that has too much-unchecked noise with either empirical judgment or not. In addition, with the nonlinearity model, too. In other words, the price of the stock should never depend on either a single tweet or breaking news that could lead to bankruptcy mainly. Just like instances of some of the mood fluctuation factors suggested, etc. turnover volatility, 60 days average turnover rate. PingAn Bank created its IPO and was also being released as the initial stock market (000001.SSE) at Shenzhen Stock Exchange in 1984. Amazon issued its initial public offering of stock on May 15, 1997, at $18 per share, trading under the NASDAQ stock exchange symbol AMZN. And they are both very well-known listed enterprises around the world. We pick up the firms that have the stock market around the world (the United States and China) which possess the top highest popularity and market values, less fluctuation on account of uncertainty factors. It is a relative stabilization, but still, a nonlinearity model based on nearly 16 years of historical data since creation. Factors choosing are more reasonable and easier to be found in patterns following the objective law of the statistics and following the universal law of financial markets.

Platform

Edit and run the code in Jupyter notebook and colab.

1.4. Report Structure

The second section is literature review, generalizing and summarizing previous research and paper. The third section is problem formulation and solution approaches, introducing prediction models and corresponding solutions to it. The fourth section is implementation of the models and solutions in the third section. The experimental results with analysis and conclusion are in the fifth and sixth section. Description of individuals' effort, references and appendix are at the end of the report.

2. Literature Review

Time series analysis can be used to track and predict the trend of stock price.[4] It is proved that ARIMA (Autoregressive Integrated Moving Average) models can be used to build forecast models of stock prices, especially for short-term prediction.[1] Due to the self-adaptability and self-learning characteristic of CNN (Convolution Neural Network), there is a prediction model based on CNN, which performs high accuracy in identifying the trend of stock price.[10] HHM(Hidden Markov Model) has also been proved to be a good method to predict stock price, and the result is accurate to the trend of real stock price.[3] With an attention layer, LSTM(Long Short-Term Memory), a kind of neural network, can both overcome time lag limitations and predict stock price.[13] LDA(Linear Discriminant Analysis) combined with an online learning method is suitable for predicting stock movement.[12] By using the cluster information of some anomalies, a data mining approach is proposed to forecast the stock price in the market.[16] Random forest can also be used in stock movement prediction.[6] SVR (Support Vector Regression) is widely used in stock price prediction and has been suggested as a powerful predictive tool for stock price prediction in financial markets.[15] There is an optimal combination of PCA (Principal Components Analysis) with SVMs (Support Vector Machines) which has proved its effectiveness compared with traditional feature selection.[2] A combination of SVM and KNN (K Nearest Neighbor) approach has been proposed to forecast stock prices.[8] There is a SVR model based on Grey Relational Analysis which could improve the convergence speed and forecast accuracy. [5] A system has been developed for both regression and classification of stock price, which aims at predicting stock price and up and down trends respectively.[9] The directionality of the price movement is one of the most important attributes of a stock time series, but is seldom reflected in the prediction error of neural networks. [14]

3. Formulation and Solution Approaches

3.1. Feature Selection

Feature selection is to extract the several most effective factors among all factors according to the contribution of them to stock, then use these factors to demonstrate the change. In the model, we use LASSO (Least Absolute Shrinkage and Selection Operator) regression to select features. LASSO regression is a kind of linear regression (l_1-

penalty), and it is similar to Ridge Regression (l_2 penalty) based on the least squares difference mechanism. LASSO uses l_1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients.[11]

$$\sum_{n=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{1}$$

where y_j is the true result and $\sum_j x_{ij}\beta_j$ is the prediction results, β_j is coefficients. λ is a tuning parameter, which controls the degree of the penalty:
When λ equals zero, all parameters are preserved and the model is a linear regression.
With increment of λ, more coefficients are shrunk to zero and removed from the model and bias of the model increases as well. If λ is large enough, all coefficients will become zero.
With decrement of λ, more coefficients will be preserved and variance will increase.

3.2. Data Preprocess

Windowing
In the model, due to the time series' attributes, we access the historical data used to predict, utilize factor data of several days as the feature part and the future close price as the label. Window size can be treated as the sample number, which is the amount of days used in prediction. Horizon means predicting the closing price of the first few days of the future.

Normalization
Due to different units of different factors, it is necessary to normalize the raw data to avoid negative impact on the prediction results.

In the model, Min-Max Scalar is used to transform a feature, as it will preserve the shape of the dataset with no distortion and bias compared to Standard Scalar.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

where X is the raw value and X_{max}, X_{min} are maximum and minimum of the raw data respectively.

3.3. Prediction

Support Vector Regression (SVR)

SVR can address small capacity of data, non-linear recognition problems. Introduction of kernel function instead of inner product computation in higher dimensions converts nonlinear problem in original dimension to linear problem in higher dimension and avoids curse of dimensionality. It uses ε, insensitivity loss to add tolerance and reduce model complexity by minimizing $\| w \|^2$. Simultaneously, slack variables ξ_i, ξ_i^* are set to measure the deviation of training sample out of the ε sensitive tube shown in figure1:

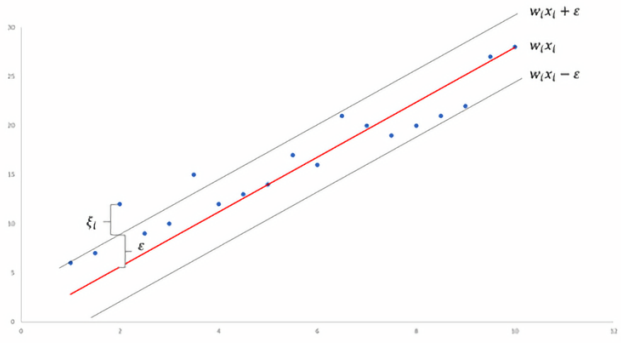

Figure 1. SVR with slack variables ξ and tolerance ε

$$\begin{aligned} \Phi(w,\xi) &= \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \varepsilon + \xi_i &\le y_i - f(x_i, w) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\ge o, i = 1, \ldots, n \end{aligned} \tag{3}$$

SVR performance depends on the configuration of kernel parameters C and ε.[7]

Kernel Ridge Regression (KRR)

KRR is to obtain parameter w, b through minimizing:

$$\frac{1}{n}\left\{\lambda \|w\|^2\right\} + \sum_{j=1}^{n}\left(y_j - w^T x_j - b\right)^2 \tag{4}$$

with training data and make decision using the same parameters to decide:

$$h(x) = w^T x + b \tag{5}$$

3.4. Evaluation

MAPE (Mean Absolute Percentage Error) is used to evaluate the performance of the model:

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{6}$$

where M is mean absolute percentage error; n is the number of times the summation iteration happens; A_t is the actual value and F_t is the forecast value.

The overall significant accuracy is measured by

$$(1 - MAPE) \times 100\% \tag{7}$$

MAPE is a scale-independent evaluation measurement technique, which is suitable for comparing accuracy across time series. The larger MAPE is, the more accurate the model fits.

Model	Window Size	Step Size	Horizon
1-day ahead	3	1	1
	8	1	1
	25	1	1
5-day ahead	3	1	5
	8	1	5
	25	1	5
22-day ahead	3	1	22
	8	1	22
	25	1	22

Table 1. Window Settings

4. Implementation

4.1. Feature Selection

In order to select features that are relatively more relevant to the value of closing price, a LASSO regression model is used on a scaled version of the original dataset. Only those features that have a coefficient different from 0 are considered. Steps describing how to use LASSO regression to extract several features that are highly correlated with the predicted values are given below:
Step1: Use 5-fold cross validation to select the most appropriate λ, which is the regularization parameter.
Step2: Get the weight of each feature in LASSO regression, which suggests the degree of correlation between feature and label.
Step3: Sort the weights according to the absolute value from large to small.
Step4: Increase the number of selected features according to the weight and restore the value of the original label according to these features and corresponding weights.
Step5: Calculate the mean-square-error with original label value in order to decide how many features to choose is the best choice using elbow method.

4.2. Data Preprocess

Before regression, data preprocessing is performed to obtain suitable training and testing data, as well as to normalize the data.

Normal Rectangular Windowing

The windowing operator transforms a given set of instances containing series data into a new set of instances containing single-valued instances. For this purpose, the window with the specified window size and step size moves through the series data and the attribute values lying horizon value at the end of the window is used as the label that should be predicted.

Table 1 shows the window setting that produces inputs for the regression analysis.

Normal Rectangular Windowing

By using the MinMaxScaler in the sklearn.preprocess module to normalize the data, they are scaled to the interval [0, 1], thus retaining the original data size ratio while using the same unit, facilitating subsequent calculations.

4.3. Regression

Training Phase
Step1: Read the dataset from local repository.
Step2: Split the dataset into training dataset and testing dataset and normalize them.
Step3: Select kernel types and special parameters of KRR and SVR(C, ε, γ etc).
Step4: Use gridsearch to accomplish a 5-fold cross validation and find the parameter of SVR and KRR based on the training set.
Step5: Exit from the training phase and apply trained models to the testing dataset.

Testing Phase
Step1: Apply the training model to the testing dataset for closing price prediction.
Step2: Calculate the mean absolute error and plot the graph of the predicted stock price.

5. Experimental Results

5.1. Feature Selection

LASSO regression: Selecting appropriate factors (feature) and eliminating the others by minimizing the sum of the squared residuals of l_1 penalty.

Partial-factors (Include open, high, low)

6 factors out of 30 factors in total (24 factors are eliminated). The weights of selected factors are shown in table 2.

All-factors (Not Include open, high, low)

39 factors out of 129 factors in total (90 factors eliminated). Using the elbow method, 9 features are selected from 39 non-zero weight factors. The weights of these factors are shown in table 3.

The elbow of the curve reflects the Mean-Square-Error drops apparently when the factors equal three out of thirty factors in total in figure2, and at ten out of one-hundred, thirty-nine factors in figure 3 in total respectively.

[1]Market_cap
[2]Circulating_Market_cap
[3]Net_Interest_expense
[4]Net_Invest_Cash_Flow_ttm
[5]Non_Operating_Net_Profit_ttm

Factor	High	Open	Low	Circulating_Market_cap	Market_cap	EMA5
Weight	0.4537	0.4576	0.0417	0.0236	0.0192	-0.0311

Table 2. Window Settings

Factor	M_cap[1]	CM_cap[2]	Price 1Y	ATR14	Variance120	VEMA26	NI_expense[3]	NICF_ttm[4]	NONP_ttm[5]
Weight	0.7232	0.1631	0.0571	0.0394	0.0321	-0.0300	0.0195	-0.0195	-0.0142

Table 3. Window Settings

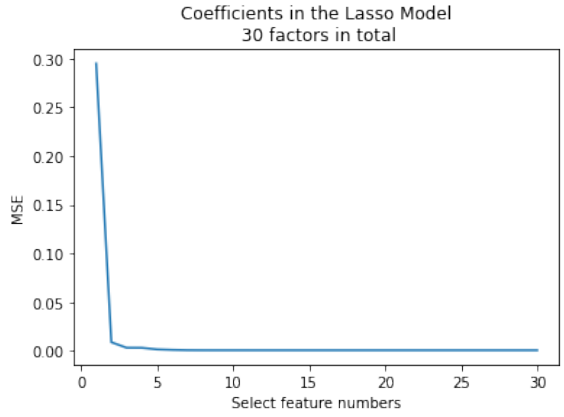

Figure 2. Results of LASSO implementation with partial factors

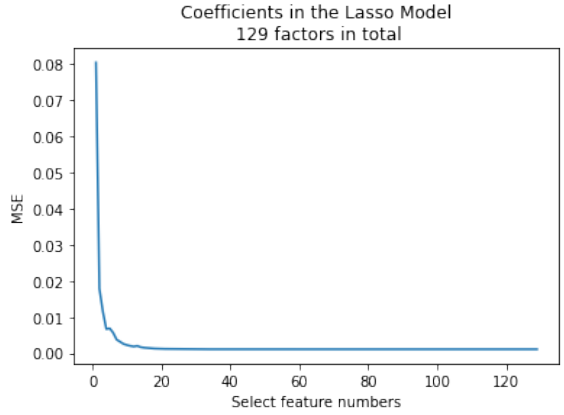

Figure 3. Results of LASSO implementation with all factors

5.2. Regression and Prediction

Partial factors (4 in total): In figure 4, the proportion of the dataset of training and testing are splitted by 10 years (83%) and 2 years(17%) since 2010. Windowing size is constant for each group that used to predict the next, fifth(next week), and 22th business days ahead(next month) of the current.

Partial factors (30 in total): In figure 5, the proportion of dataset of training and test are splitted by 15 years (82%) and 3 years(18%) since 2005. Windowing size is constant based on each group that used to predict the next, fifth(next week), and 22th business days ahead(next month) of the current.

All factors (129 in total): Comparison between real stock price and the predictions of company - PingAn Bank of China. The proportion of dataset of training and test are splitted by 15 years (82%) and 3 years(18%) since 2005. Windowing size varied between 3 to 25 that used to forecast the next business day ahead of the current, shown in figure 67

All factors (129 in total): Comparison between real stock price and the predictions of company - PingAn Bank of China. The proportion of dataset of training and test are splitted by 15 years (82%) and 3 years (18%) since 2005. Windowing size varied between 3 to 25 that used to forecast the 22th business day(next month) ahead of the current, shown in figure 8 9

Regression result of prediction in Amazon among Partial Factors. The red line is the real stock price of the historical datasets. Green line indicates the prediction of a regression model based on only a single factor of the close price. Blue line is the prediction based on the selected factors after the screening of LASSO. Here, Partial Factors implied the algorithm of LASSO performed only under the maximum available choice of factors of thirty including the open, high and low price in figure 4. The label of the meaning of red, blue and green lines within the Regression result of prediction of PingAn Bank which are exactly the same in figure 5. The prediction differences of SVR and KRR in Amazon, and in PingAn's are, Amazon's model is only contributed and selected among factors of open, close, high, and lows; Nevertheless, PingAn's model is contributed and selected among thirty available factors.

The figure 6 7 which shows the summary of the results of prediction among two regression models by All Factors applied. The red, green and blue line which is labeled as the same as the plot of figure 4 5 indicates. It is the prediction based on a total of 129 factors before the screening of LASSO. After implementation, the result of prediction shown in figure 6 7 could be found as the following as shown. We tried to apply different sizes of windowing in order to make the model of prediction more comprehensive.

Window size could be observed as three, eight and twenty five, which means sampling the first three, eight and twenty five days series data respectively to forecast the price of the next business day ahead of the transaction of today's relative to the model.

The figure 8 9 which shows the summary of the results of prediction among two regression models by All Factors applied also. The red, green and blue line which is labeled as the same as the proceedings. Figure 8 9 could be found as the following as shown. Only difference between them and figures in part 4, figure 6 7, is that the plot applied varieties of different sizes of windowing to forecast the price of the 22th business day ahead of the transaction of today's relative to the model.

5.3. Performance

Performance of different combination of parameters is shown in the chart3 and chart4

6. Conclusion

In this study, we first applied LASSO with l_1 penalty regression by using 5-folds cross validation to select the most appropriate λ (regularization parameter) for the equation. Then use the lambda to calculate the degree of correlation between feature and label of each feature (factor of dataset) according to the calculated λ. Finally, we determined which feature (factor) is most relevant about the trend of the model (stock price).

The next step before applying the prediction model, we constructed the preprocessing of data in order to fit time series attributes in such a model of stock price prediction. We sample the datasets (amount of days used in prediction) into different window sizes, utilize factor data of several days as the feature part and the future close price as the label and normalize them.

Then we research plenty of algorithms, etc. Elastic net regression for the screen out feature algorithm, the ARIMA for predicting the model. We certified that not all of them could effectively fit our models. Actually, we found out SVR can beautifully address the small capacity of data of non-linear recognition problems. KRR can address nonlinear and stabilized models as well. Finally, we decided to use a combination of Support Vector Regression and Kernel Ridge Regression model to forecast the price of stock.

LASSO

When we apply the elbow method among the 30 factors, the algorithm determines the total of six relevant factors. The most relevant factor is the 'highest price' comes with the coefficient value at 0.4537 out of 1, which could be shown as table 2. When we apply the method among the 129 factors, the algorithm determines the total of 9 relevant factors, and the most relevant factor is Market_capitialization about 0.7232 out of 1.

If we check the prediction of the model (PingAn Bank) which has been shown in Figure 4, it is not hard to find our prediction curve is nearly compatible with the real price of the stock (closing price) involves fluctuation and trend about the ahead of one and five days group. Even if the MAPE increases as the time series increases, it still supports the relationship between the major relevant features and labels that has been successfully found, and approves the correction of the model.

In another word, the major factor of model PingAn Bank - Market capitalization among the 139 total factors model, which could be directly reflected by its stock properties. Based on the data resource section that we have researched, PingAn bank possesses the top highest popularity and market values in SZSE (Shenzhen Stock Exchange), which highly suggests its most noteworthy characteristic of stock might be the "Market Capitalization". And we use one of its domain factors to make the prediction which probably makes more sense in this case.

KRR

In the Amazon dataset, the highest significant accuracy is 93%, which is suggested by the fifth day ahead of current that shown at table4. And the least significant accuracy is 86%. In the PingAn comprehensive features dataset, the highest reached 98%, and lowest at 88%.

By comparison with the different time series model between the 1-day and 22-days ahead of the current figure6 and figure8, it predicts distinctly better while the time series is short (1-day) than the long (22-days). It also performs more accurately with enough features than less features (less factors here) in prediction. (Compare with table4 Amazon's and table5 PingAn's)

Compared with SVR, KRR performs inferiorly than SVR in the case of prediction based on fewer features as far as the conclusion in table4 suggested. However, it plays neck and neck or even slightly better sometimes, which is reflected by the comprehensive model among the multiple kinds of features screened by the l_1 ridge regression.

SVR

Significant accuracy applied by the same formula. In the Amazon dataset (Less Features), the highest significant accuracy is 99%, which is suggested by 1-day ahead prediction in table4. And the least significant accuracy is 96%. In the PingAn comprehensive features dataset, the highest reached 98%, and lowest at 87%.

In the time series nonlinearity model, the performance of SVR is similar to KRR. By comparison with the different time series model between the 1-day and 22-days ahead

of the current shown in figure6 and figure8, it predicts distinctly better while the time series is short (1-day) than the long (22-days). However, it shows a non-bad even good performance under the lack of features, like Amazon's table showed. Meanwhile, it performs very well on comprehensive windowing size and features, too, as the table5 suggested.

Since SVR can only address the small capacity of data of non-linear recognition problems. Compared with KRR, it performs in advantage compared with KRR in the case of prediction based on fewer features as far as the conclusion in table4 suggested. It plays as good as the prediction in the comprehensive model among the 129 factors screened by the l_1 penalty.

Prospect

In this study, We implemented LASSO Regression to eliminate irrelevant features, and used the Kernel Ridge Regression and Support Vector Regression to make the prediction. However, the model still has a lot of space in promotion and make-up.

The first is about data resources, we want to realize details and get more categorical of factors about the stock. In this case, the research about more accurate physical and psychological elements, and also the background of the company is necessary.

Second is reference about the uncompleted Data, such as the data factors missing or undeveloped in certain time series. Some of them are even fragmentary that could not be aligned to the same dimension of other factors. Then they even could not be utilized during the data-preprocessing. In this study, the way we deal with those issues is, we choose either to drop the factors data that might exist largely incomplete or using the "zero" to fill out the non-valid block within the time series dataset. But there are more intelligent ways, such as applied K-means, K-NN, Random forest method to complement the fragmentary blocks during the pre-processing.

And there exist more intelligent models. One of them is called: Keras' Long Short-Term Memory (LSTM). The principle of the model which has been introduced in the previous Literature part. The advantages of the model which involves it can store past important information and forget the information that is not. Meanwhile, LSTM can play a significant role in time series data, In time-series collection of datasets, especially the one data is proved dependent on another.

7. Description of Individual Effort

Data source selection and data acquisition: Zhaozhong Qi
LASSO: Zemeng Wang
Kernel Ridge Regression: Shuyan Zhang
Support vector regression: Zemeng Wang
Result evaluation and report writing: Zemeng Wang, Zhaozhong Qi, Shuyan Zhang

References

[1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 106–112. IEEE, 2014.

[2] C. Chun, Y. Liu, and J. Sun. Applied research of the algorithm combined of pca and svms on stock features. In *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, volume 2, pages V2–432. IEEE, 2010.

[3] A. Gupta and B. Dhingra. Stock market prediction using hidden markov models. In *2012 Students Conference on Engineering and Systems*, pages 1–4. IEEE, 2012.

[4] A. HAYES. What is a time series?, 2021.

[5] X. Hou, S. Zhu, L. Xia, and G. Wu. Stock price prediction based on grey relational analysis and support vector regression. In *2018 Chinese Control And Decision Conference (CCDC)*, pages 2509–2513. IEEE, 2018.

[6] K. Loke. Impact of financial ratios and technical analysis on stock price prediction using random forests. In *2017 International Conference on Computer and Drone Applications (IConDA)*, pages 38–42. IEEE, 2017.

[7] P. Meesad and R. I. Rasel. Predicting stock market price using support vector regression. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6. IEEE, 2013.

[8] D. Puspitasari and Z. Rustam. Application of svm-knn using svr as feature selection on stock analysis for indonesia stock exchange. In *AIP Conference Proceedings*, volume 2023, page 020207. AIP Publishing LLC, 2018.

[9] S. Ravikumar and P. Saraf. Prediction of stock prices using machine learning (regression, classification) algorithms. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2020.

[10] L. Sayavong, Z. Wu, and S. Chalita. Research on stock price prediction method based on convolutional neural network. In *2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, pages 173–176. IEEE, 2019.

[11] Stephanie. Lasso regression: Simple definition, 2015.

[12] T. Tantisripreecha and N. Soonthomphisaj. Stock market movement prediction using lda-online learning model. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 135–139. IEEE, 2018.

[13] D. Wei. Prediction of stock price based on lstm neural network. In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pages 544–547. IEEE, 2019.

[14] Y. Wei and V. Chaudhary. The directionality function defect of performance evaluation method in regression neural network for stock price prediction. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 769–770. IEEE, 2020.

[15] Y. Xia, Y. Liu, and Z. Chen. Support vector regression for prediction of stock trend. In *2013 6th international conference on information management, innovation management and industrial engineering*, volume 2, pages 123–126. IEEE, 2013.

[16] L. Zhao and L. Wang. Price trend prediction of stock market using outlier data mining algorithm. In *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, pages 93–98. IEEE, 2015.

Appendix

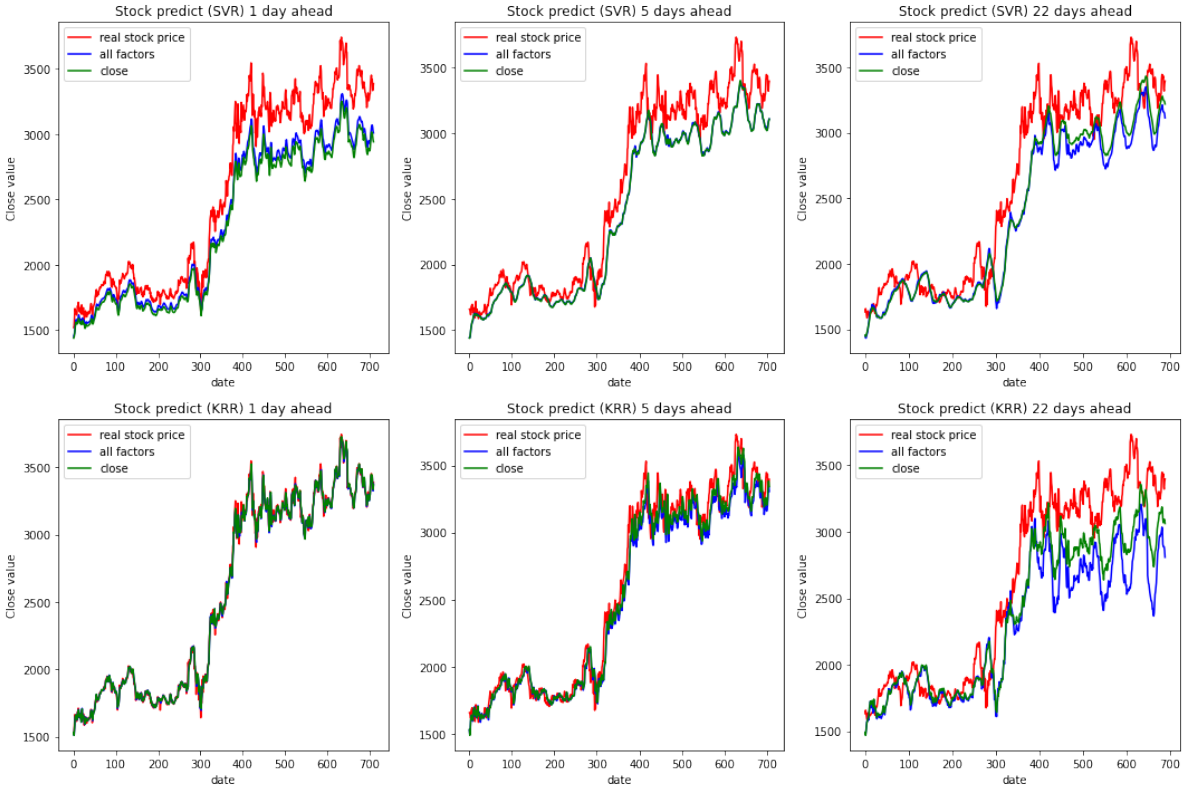

Figure 4. **Partial factors (30 in total):** Comparison between real stock price and Predictions of Amazon, United States.

Model	Factor	Window	horizon	Parameter				1-MAPE	
				C(SVR)	γ(SVR)	α(KRR)	γ(KRR)	SVR	KRR
1-day ahead	'Close'	3	1	10000	1e-5	1e-5	1e-4	0.99	0.89
	All factors	3	1	10000	1e-5	1e-5	1e-4	0.99	0.91
5-day ahead	'Close'	8	5	10000	1e-5	1e-5	1e-4	0.97	0.93
	All factors	8	5	10000	1e-5	1e-5	1e-4	0.96	0.93
22-day ahead	'Close'	25	22	10000	1e-5	1e-5	1e-4	0.99	0.91
	All factors	25	22	10000	1e-5	1e-5	1e-4	0.86	0.90

Table 4. Predict close price of **Amazon** by using constant window size

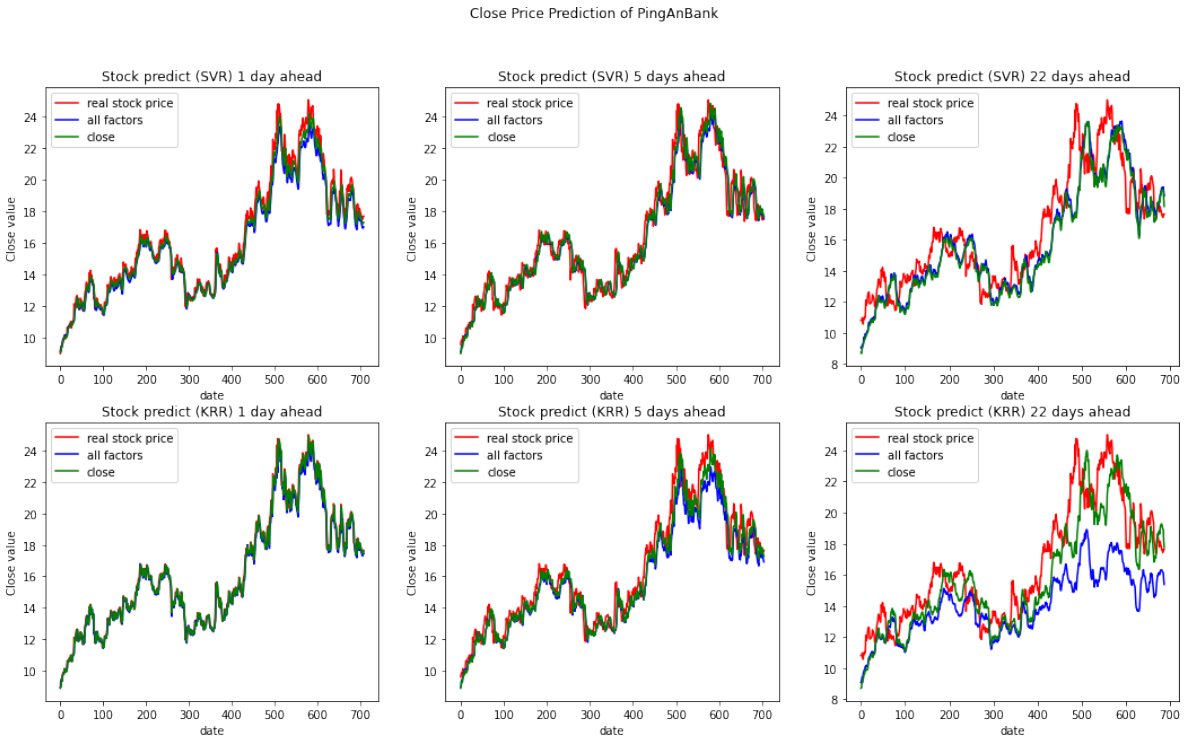

Figure 5. **Partial factors (30 in total):** Comparison between real stock price and the predictions of PingAn Bank of China.

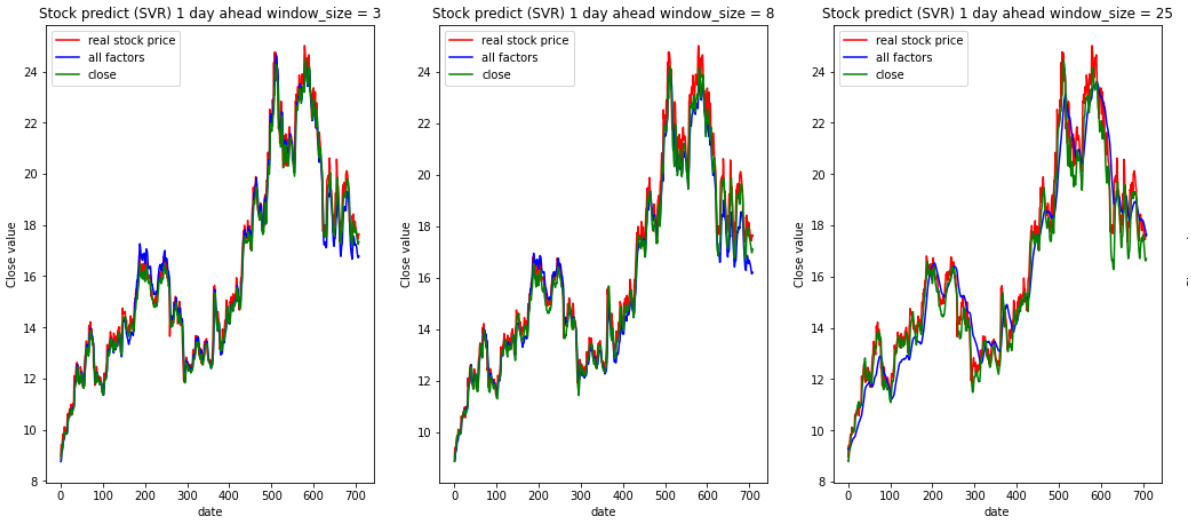

Figure 6. Predicted by Support Vector Regression ahead of a day.(SVR)

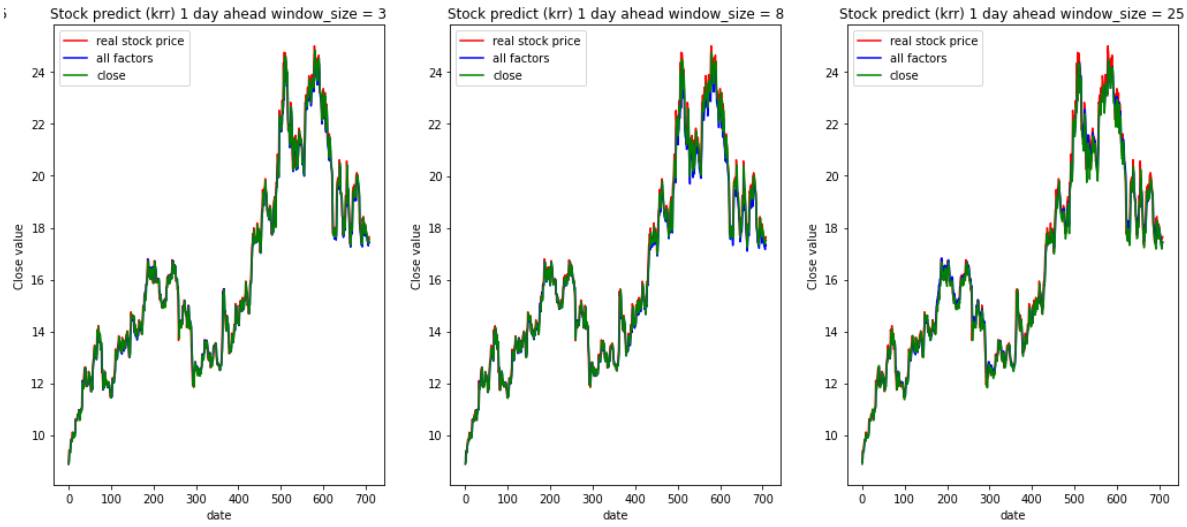

Figure 7. Predicted by Kernel Ridge Regression ahead of a day.(KRR)

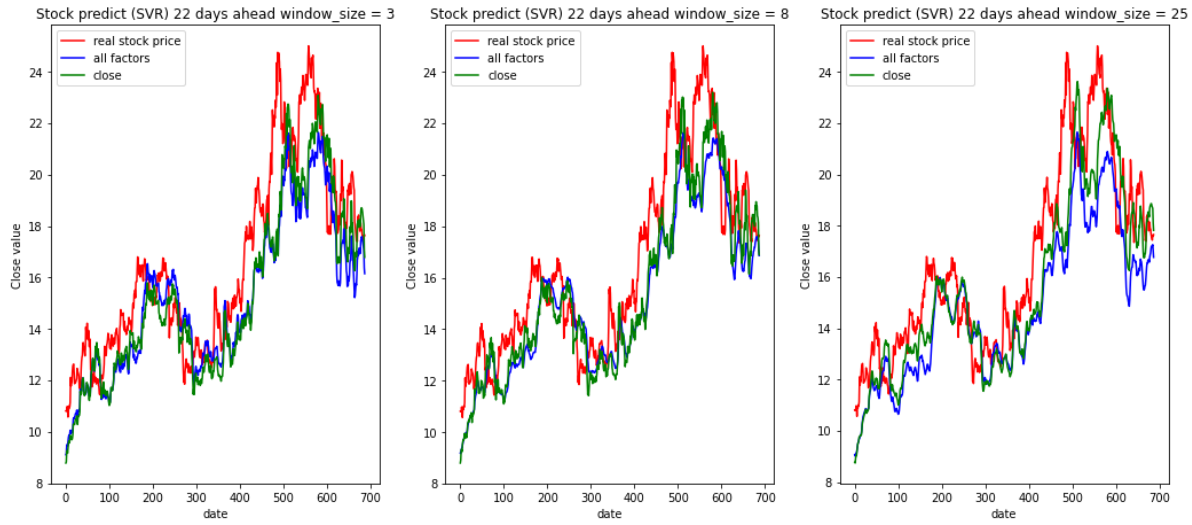

Figure 8. Predicted by Support Vector Regression ahead of 22 days.(SVR)

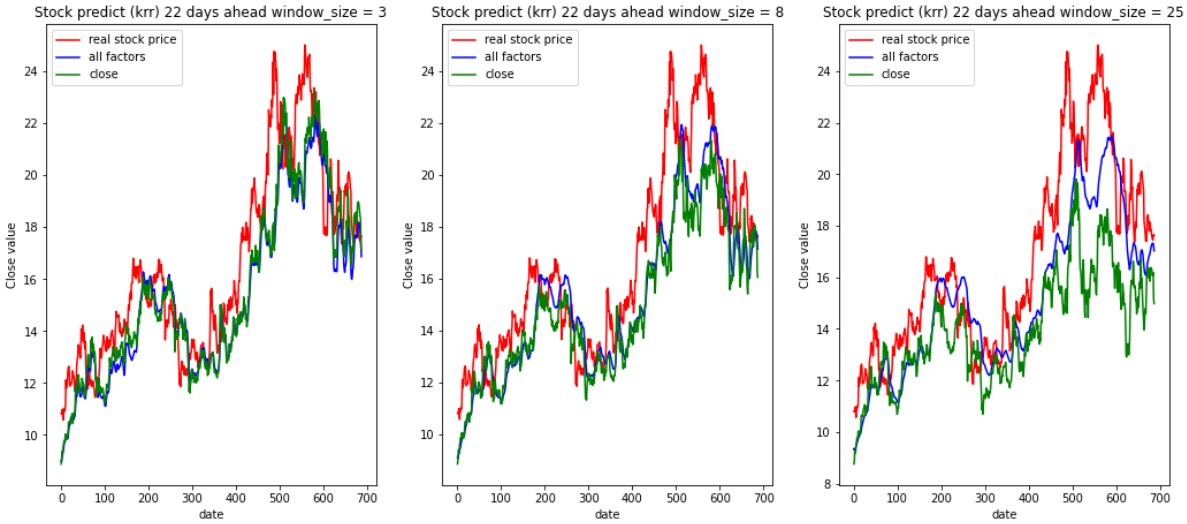

Figure 9. Predicted by Kernel Ridge Regression ahead of 22 days.(KRR)

Model	Factor	Window	horizon	Parameter				1-MAPE	
				C(SVR)	γ(SVR)	α(KRR)	γ(KRR)	SVR	KRR
1-day ahead	'Close'	3	1	100000	1e-5	1e-5	1e-4	0.98	0.98
		8	1	100000	1e-5	1e-5	1e-4	0.97	0.98
		25	1	100000	1e-5	1e-5	1e-4	0.97	0.98
	All factors	3	1	100000	1e-5	1e-5	1e-4	0.97	0.98
		8	1	100000	1e-5	1e-5	1e-4	0.97	0.98
		25	1	100000	1e-8	1e-5	1e-4	0.95	0.98
5-day ahead	'Close'	3	5	100000	1e-6	1e-5	1e-4	0.95	0.96
		8	5	100000	1e-6	1e-5	1e-4	0.96	0.95
		25	5	100000	1e-5	1e-5	1e-4	0.95	0.95
	All factors	3	5	100000	1e-5	1e-5	1e-5	0.96	0.96
		8	5	100000	1e-7	1e-5	1e-5	0.95	0.96
		25	5	100000	1e-7	1e-5	1e-5	0.95	0.95
22-day ahead	'Close'	3	22	100000	1e-5	1e-5	1e-4	0.89	0.90
		8	22	100000	1e-5	1e-5	1e-4	0.90	0.88
		25	22	100000	1e-6	1e-5	1e-4	0.9	0.82
	All factors	3	22	100000	1e-6	5e-4	1e-5	0.89	0.89
		8	22	100000	1e-5	5e-4	1e-5	0.89	0.89
		25	22	100000	1e-7	1e-5	1e-7	0.87	0.88

Table 5. Predict close price of **PingAn Bank** using different window size