

Hệ thống dự đoán giá chứng khoán

GVHD:

Trần Minh Quang

Lê Hồng Trang

Nhóm sinh viên thực hiện:

1913621 – Bùi Đắc Hưng

1911314 – Lương Thị Quỳnh Hương

2170535 – Lê Vũ Minh Huy

Outline

1. Tổng quan đề tài
2. Về dữ liệu
3. Mô hình ARIMA
4. Mô hình SVR
5. Mô hình Baseline
6. Mô hình LSTM
7. Tổng hợp kết quả

Tổng quan đề tài



GSG	14.5	15.5	13.5	14.2	3,243	14.25	6,376	14.3	1,295	14.3	10,986	0.29
KDC	35.95	38.45	33.45	35.75	1	35.8	1	35.85	1	35.95	1,387	0.8
MRR	31	33.15	28.85	30.2	1,172	30.25	191	30.3	1,264	30.3	13,429	0.79
MSW	96.2	102.9	89.5	92.7	170	92.8	364	92.9	197	92.9	7,668	0.39
MNG	106.8	114.2	99.4	104.1	184	104.5	220	104.6	198	104.6	983	0.87
NZL	31.7	33.9	29.9	31.9	1,087	31.9	32	31.95	50	31.7	1,116	0.8
RWL	51.5	57.2	49.8	51.6	101	51.9	31	52	65	52.6	8,947	0.39
PLX	70.6	76.5	65.7	68.7	8	68.8	319	69	671	69	2,957	1.39
REE	36.8	39.35	34.25	36	1,785	36.1	1,784	36.2	2,163	36.2	8,735	0.69
SOS	82.4	88.1	76.7	79.8	6,248	79.9	5,154	80	5,481	80	22,094	0.89
SAB	250	267.5	232.6	248.6	10	249	1,583	251.5	4	252	344	2.94
SUJ	17.6	18.8	16.4	17.1	21,248	17.9	1,079	17.2	16,600	17.2	22,198	0.49
SSE	36.5	37.95	33.05	34.36	1,200	34.8	740	34.46	1,000	34.9	29,293	1.97
SUW	13.1	14	12.2	12.95	12,326	13	28,188	13.05	6,327	13.05	3,933	0.059
VCB	58.7	62.8	54.6	56.8	3,183	56.9	155	57	23,332	57	24,077	1.79
VIC	132	141.2	122.8	129.5	26	129.8	54	129.8	1	130	18,977	2.69
VJE	107	216.7	183.3	193.6	2	184	411	194.3	100	194.3	3,729	2.29
WNM	176	188.3	163.7	168.8	50	169	2,161	189.1	179	189.1	13,179	0.99



GMD	40.15	47.95	37.35	41.2	1,571	41.25	60	41.3	20	41.3	6,542	1.154
HPG	40.85	44.1	33.8	47.5	15,580	47.65	100	47.6	1,781	47.7	35,598	0.354
HSG	24.6	26.2	22.8	24.65	6,696	24.6	6,306	24.65	698	24.65	7,128	0.154
KDC	13.4	14.3	12.5	13.5	9,560	13.65	8,144	13.6	19,595	13.65	14,122	0.254
KDC	33.5	40.25	36.75	41.35	340	41.4	413	41.5	199	42	3,965	2.54
MDB	25.4	27.5	23.85	26.4	5,406	26.45	5,120	26.5	12,332	26.55	62,343	1.354
MSW	75.7	82	71.4	79.2	209	79.3	100	79.5	160	80	3,321	3.74
MNG	131	138	121.9	131.7	195	131.8	90	132	2,154	132	3,873	1.24
NZL	11.6	16.8	14.3	13.8	0,365	13.9	710	13.95	675	14.2	2,981	0.74
NVI	65.1	68.6	60.8	65.3	1,221	65.4	1,262	65.5	148	65.5	3,825	0.44
PWD	23.35	24.99	21.75	22.9	11,680	22.95	9,020	23	44,266	23	55,867	0.359
REE	41.5	44.4	38.8	43	530	43.05	100	43.1	238	43.2	22,362	1.74
RGB	101.7	104.4	96	101.3	10,032	100.4	29,477	100.5	12,153	100.6	10,886	0.278
SAB	249.3	268.7	231.5	252	1,240	252.1	9	252.3	34	252.3	929	3.84
SAB	29.8	22.29	19.35	20.3	7,457	20.25	18,037	20.4	19,521	20.4	31,109	0.87
SSJ	28.8	30.8	26.8	29	3,827	29.05	38	29.1	76	29.1	18,978	0.34
STR	12.85	13.7	12	13.25	2,883	13.3	35,110	13.35	11,694	13.4	31,733	0.564
VCB	54.3	58.1	50.5	54.6	515	54.9	2,836	55	2,774	55	8,894	0.74
VIC	77.3	82.7	71.9	77.9	1,337	78.1	70	78.2	179	78.2	10,747	1.24
VNM	208.6	213	194	210.9	2	215	50	211.3	5	211.5	3,748	2.94

Về dữ liệu

Thu thập dữ liệu

Nguồn dữ liệu được thu thập bằng cách sử dụng API có sẵn của 1github <https://github.com/phamdinhkhanh/vnquant> được 276 stars. Nhóm sử dụng API để thu thập dữ liệu của 5 loại cổ phiếu từ năm 01012015 đến 01112022 từ sàn giao dịch VNDirect:

- **BVH**: Tập đoàn Bảo Việt
- **BID**: Ngân hàng thương mại cổ phần đầu tư và phát triển Việt Nam
- **CTG**: Ngân hàng thương mại cổ phần công thương Việt Nam
- **ACB**: Ngân hàng Thương Mại cổ phần Á Châu
- **FPT**: Công ty cổ phần FPT

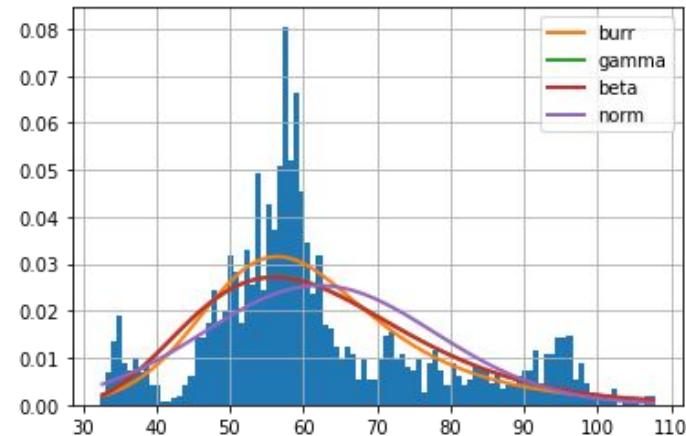
Dữ liệu đã được nhóm kiểm tra lại bằng cách so với giá trị trên <https://banggia.vndirect.com.vn/chung-khoan/danh-muc>

Mỗi bảng dữ liệu gồm 1955 dòng dữ liệu và 5 cột:

- **date**: ngày có thông tin về giá cổ phiếu
- **low**: giá trị thấp nhất của cổ phiếu trong ngày
- **high**: giá trị cao nhất của cổ phiếu trong ngày
- **close**: giá cổ phiếu lúc đóng cửa
- **open**: giá cổ phiếu lúc mở cửa

Trong bài tập lớn này, nhóm dự đoán giá trị cổ phiếu lúc đóng cửa, tức cột **close**.

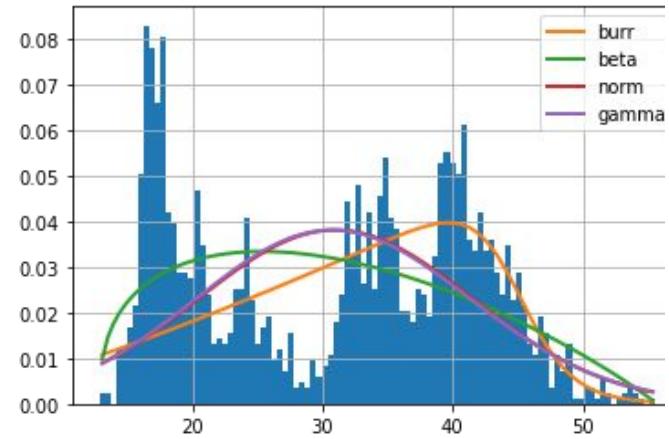
Phân tích dữ liệu



Biểu đồ giá cổ phiếu close theo thời gian và phân bố dữ liệu của BVH

Phân bố lệch về phía bên trái. Có thể đoán trong giai đoạn ở giữa đã xảy ra sự kiện gì đó với BVH như là ký kết hợp đồng lớn khiến cổ phiếu tăng vọt.

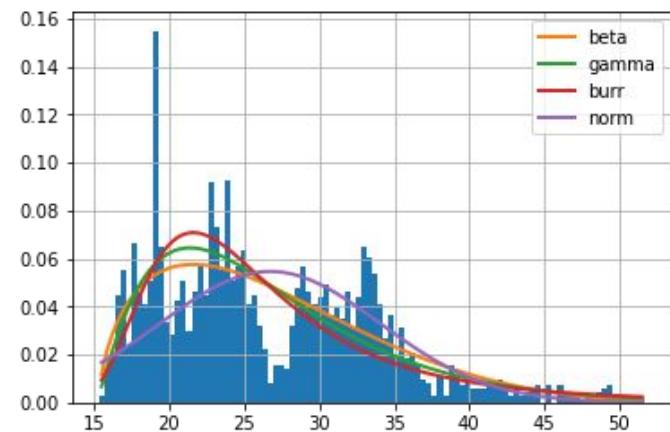
Phân tích dữ liệu



Biểu đồ giá cổ phiếu close theo thời gian và phân bố dữ liệu của BID

Phân bố hai đỉnh. Có thể thấy giá trị cổ phiếu lúc đóng cửa trước năm 2018 và sau năm 2018 tuân theo 2 phân bố khác nhau rõ rệt.

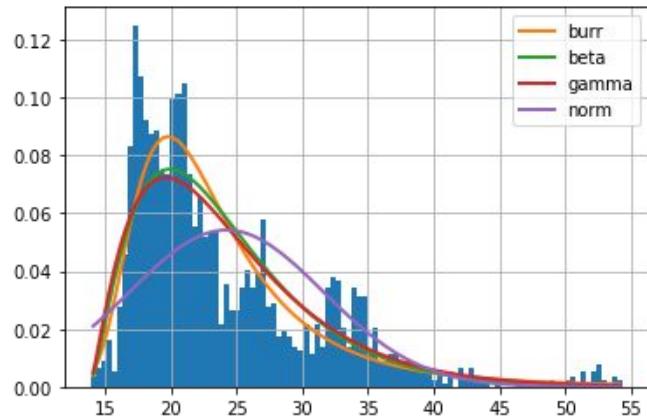
Phân tích dữ liệu



Biểu đồ giá cổ phiếu close theo thời gian và phân bố dữ liệu của ACB

Phân bố hai đỉnh. Tuy nhiên nhìn biểu đồ giá trị cổ phiếu ACB bằng mắt thường không phân biệt rõ rệt 2 giai đoạn giống như cổ phiếu BID.

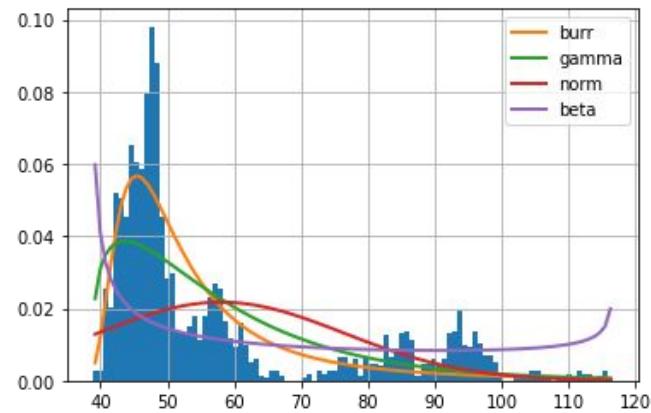
Phân tích dữ liệu



Biểu đồ giá cổ phiếu close theo thời gian và phân bố dữ liệu của CTG

Phân bố bị lệch về phía bên trái

Phân tích dữ liệu



Biểu đồ giá cổ phiếu close theo thời gian và phân bố dữ liệu của FPT

Phân bố 2 phía. Nhìn bằng mắt thường ta có thể thấy trước năm 2021 và sau năm 2021, cổ phiếu FPT có 2 phân bố phân biệt.

Phân tích dữ liệu

Nhận xét chung:

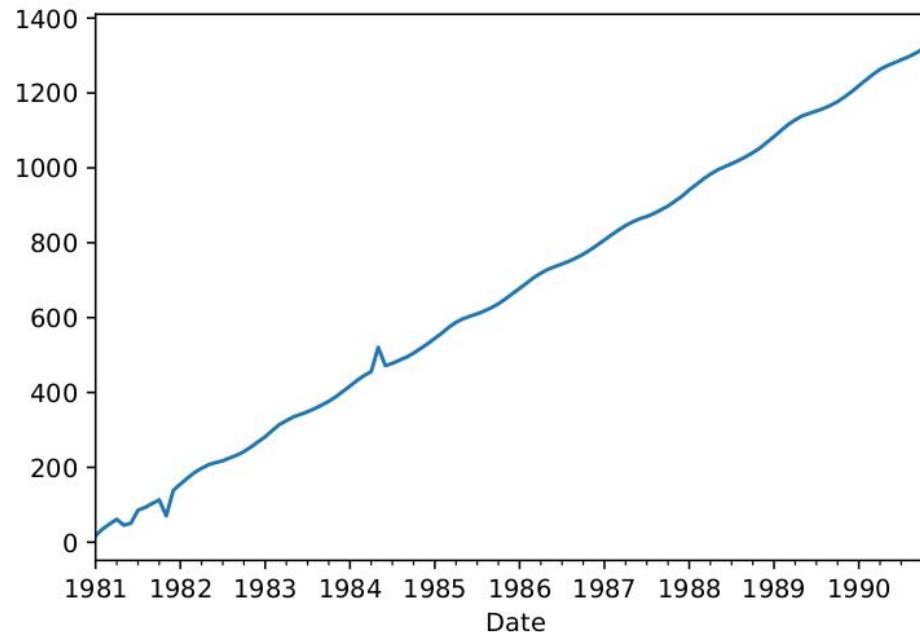
Trong 5 loại cổ phiếu đã xét, có 3 loại cổ phiếu thuộc lĩnh vực ngân hàng là BID, CTG, ACB có xu hướng giống nhau (đều đạt đỉnh vào ~ tháng 3-2018, và đầu năm 2021 đều tăng và ổn định) và có khoảng giá trị gần nhau. BVH và FPT thuộc 2 lĩnh vực khác nhau và có xu hướng khác với những loại cổ phiếu ngân hàng đã xét.

→ Có thể suy đoán các cổ phiếu cùng lĩnh vực có xu hướng giống nhau và ảnh hưởng nhau.

Mô hình ARIMA

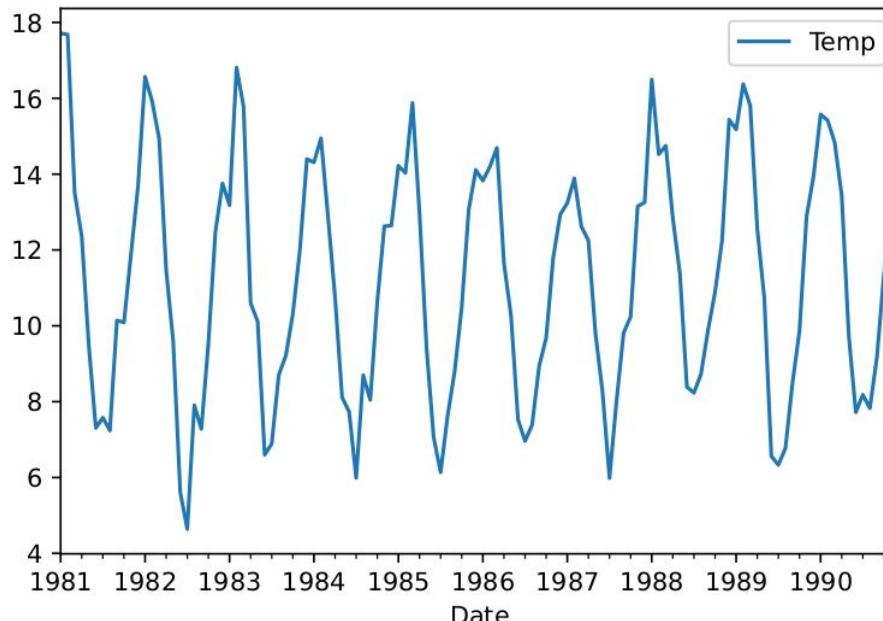
Một số đặc điểm của time series

Xu hướng (trend). Giá trị có khuynh hướng tăng hoặc giảm dần theo thời gian. Ví dụ



Một số đặc điểm của time series

Tính chu kỳ theo mùa (seasonal cycles). Dữ liệu có khuynh hướng lặp lại theo chu kỳ. Điều này là do dữ liệu bị ảnh hưởng bởi các yếu tố theo mùa (theo tháng hoặc theo năm, ...). Ví dụ:



Một số đặc điểm của time series

Tính ổn định (Stationary). Một giả thuyết thường xuyên xuất hiện trong các mô hình dự đoán như ARIMA là dữ liệu phải có tính ổn định. Một quá trình ổn định là một quá trình ngẫu nhiên trong đó tính chất thống kê không thay đổi theo thời gian. Nói cách khác, trung bình, phương sai, sự tương quan là hằng số. Vì thế, một time series ổn định sẽ không có xu hướng hay tính chu kỳ theo mùa.

Autoregressive (AR)

Mô hình autoregressive (AR) là một mô hình hồi quy trong đó giá trị quan sát tại một thời điểm là một tổ hợp tuyến tính các giá trị quan sát trong quá khứ. Tuy nhiên, mô hình AR không sử dụng hết tất cả giá trị quan sát trong quá khứ để dự đoán hiện tại, mà sẽ định nghĩa một tham số p là số giá trị quan sát trong quá khứ được sử dụng. Mô hình AR được định nghĩa như sau:

$$AR(p) = a_0 + a_1x_{t-1} + a_2x_{t-2} + \dots + a_px_{t-p}$$

trong đó, $a_i, 0 \leq i \leq p$ là hệ số trong mô hình hồi quy, $x_i, t-p \leq i \leq t-1$ là giá trị quan sát.

Moving Average (MA)

Mô hình MA đi phân tích sai số của các thời điểm trước để giảm nhiễu với mục đích dự đoán tốt hơn ở thời điểm hiện tại. Mô hình này cũng sử dụng một tham số q như là một cửa sổ trượt.

Mô hình MA được định nghĩa như sau:

$$MA(q) = \mu + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \dots + \theta_q\epsilon_{t-q}$$

trong đó, μ là trung bình của quá trình, $\theta_i, 1 \leq i \leq q$ là hệ số và $\epsilon_i, t - q \leq i \leq t - 1$ là sai số ở các thời điểm trước.

Integrated (I)

Yêu cầu sử dụng ARIMA là dữ liệu phải ổn định (stationary). Tuy nhiên, trong thực tế thì dữ liệu không phải lúc nào cũng ổn định (stationary). Một cách thường được dùng để chuyển đổi dữ liệu sao cho ổn định là lấy hiệu giữa 2 giá trị quan sát liên tiếp thay vì sử dụng dữ liệu ban đầu. Ta gọi tham số d là số lần dữ liệu được lấy hiệu, ví dụ:

$$\begin{aligned}d = 1 : \quad & \Delta x_t = x_t - x_{t-1}, \\d = 2 : \quad & \Delta^2 x_t = \Delta x_t - \Delta x_{t-1}.\end{aligned}$$

trong đó, $\Delta x_i, t - q \leq i \leq i$ là giá trị sau khi lấy hiệu d lần, $\epsilon_i, t - q \leq i \leq i + 1$ là white noise.

Một cách để chọn tham số d

a. Kiểm định Augmented Dickey Fuller Test (ADF Test) cho tham số d

Một cách trực quan, Augmented Dickey Fuller Test (ADF Test) dùng để kiểm tra dữ liệu time series có non-stationary hay không với giả thuyết null là dữ liệu non-stationary. Ta mong muốn bác bỏ giả thuyết này.

Trong bài tập lớn này, ta tìm d sao cho p-value < 0.05 tức độ tin cậy là 95% và dùng d này làm để làm tham số cho mô hình ARIMA(p,d,q)

Một cách để chọn p, q

b. Vẽ PACF và ACF cho tham số p và q

ACF cho ta biết autocorrelation giữa y_t và y_{t-k} với k khác nhau. Nếu y_t và y_{t-1} tương quan, y_{t-1} và y_{t-2} tương quan thì y_t và y_{t-2} tương quan. Để giải quyết vấn đề này, người ta dùng PACF. PACF sẽ loại bỏ ảnh hưởng của các giá trị trung gian $y_{t-1}, y_{t-2}, \dots, y_{t-k-1}$ khi đo sự tương quan giữa y_t và y_{t-k} .

Một cách để chọn p, q

Dữ liệu tuân theo ARIMA($p,d,0$) nếu ACF và PACF thỏa :

- ACF giảm theo cấp số nhân hoặc có hình sin
- Có 1 significant spike tại lag p trong PACF, nhưng sau đó thì không.

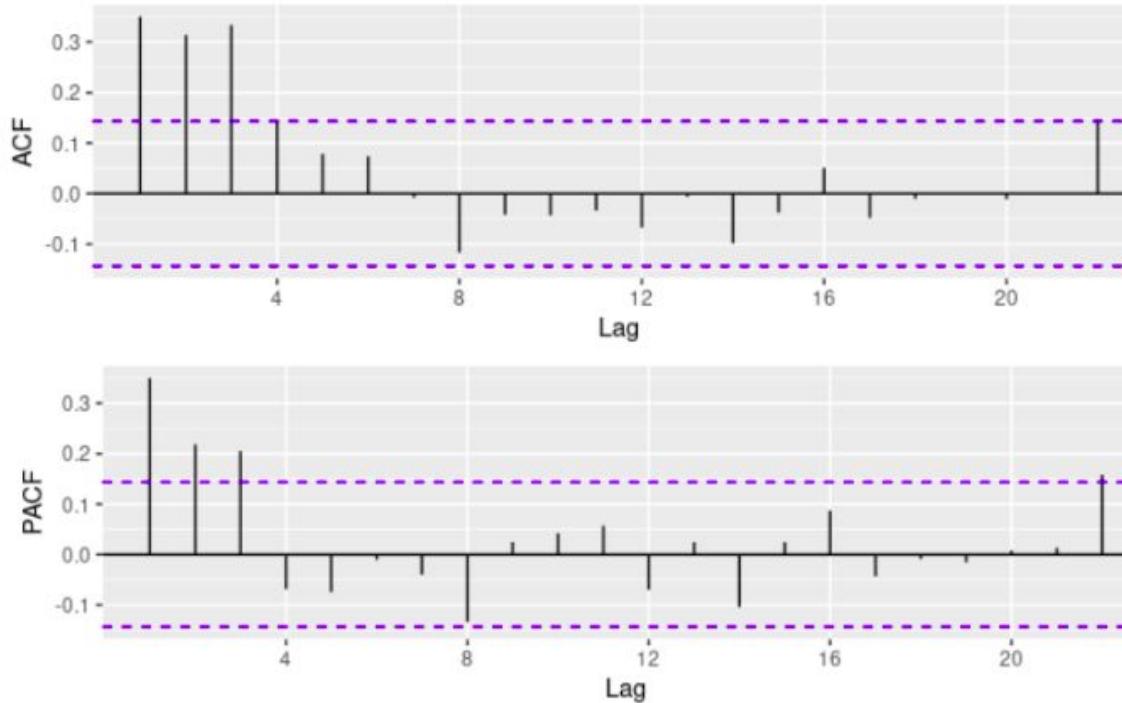
Dữ liệu tuân theo ARIMA($0,d,q$) nếu ACF và PACF thỏa :

- PACF giảm theo cấp số nhân hoặc có hình sin
- Có 1 significant spike tại lag q trong ACF, nhưng sau đó thì không

Ví dụ ở 2 hình trên, $p=3$ được coi là tối ưu vì ACF có thể được coi là giảm theo cấp số nhân và ở PACF có liên tiếp significant spike ở lag 1,2,3 nhưng sau đó thì không còn; $q=3$ được coi là tối ưu cũng theo lý luận như trên.

Một cách để chọn p, q

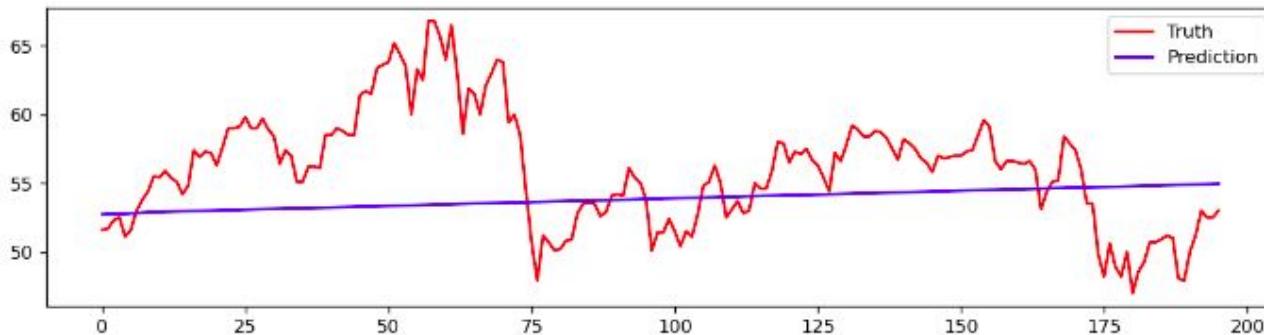
Ví dụ:



Ví dụ về ACF và PACF

Áp dụng ARIMA vào việc dự đoán

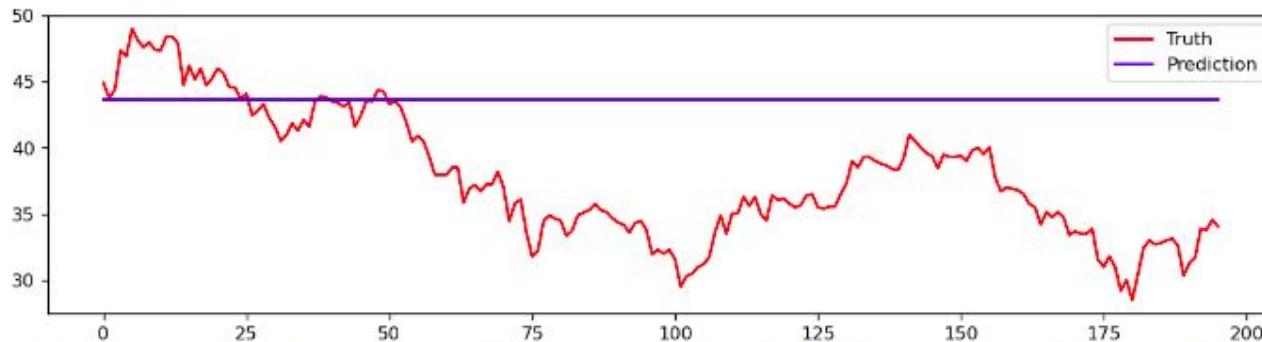
BVH



Hình 18: Trực quan hóa kết quả dự đoán của mô hình ARIMA (9,2,1) so với thực tế. Kết quả dự đoán của hình không được tốt vì nó chỉ cho ra 1 đường thẳng, không bắt được những lần lên xuống của dữ liệu.

Áp dụng ARIMA vào việc dự đoán

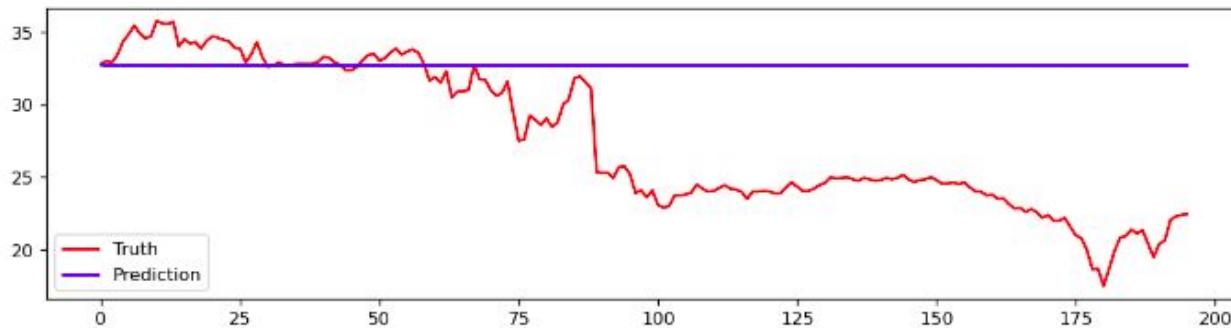
BID



Hình 21: Trực quan hóa kết quả dự đoán của mô hình ARIMA $(1,1,0)$ so với thực tế. Kết quả dự đoán của hình không được tốt vì nó chỉ cho ra 1 đường thẳng, không bắt được những lần lên xuống của dữ liệu.

Áp dụng ARIMA vào việc dự đoán

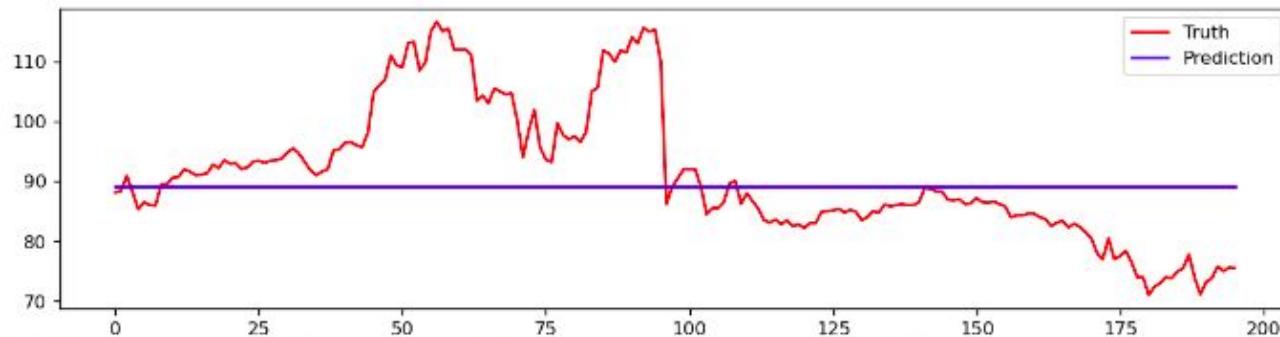
ACB



Hình 24: Trực quan hóa kết quả dự đoán của mô hình ARIMA (1,1,2) so với thực tế. Kết quả dự đoán của hình không được tốt vì nó chỉ cho ra 1 đường thẳng, không bắt được những lần lên xuống của dữ liệu.

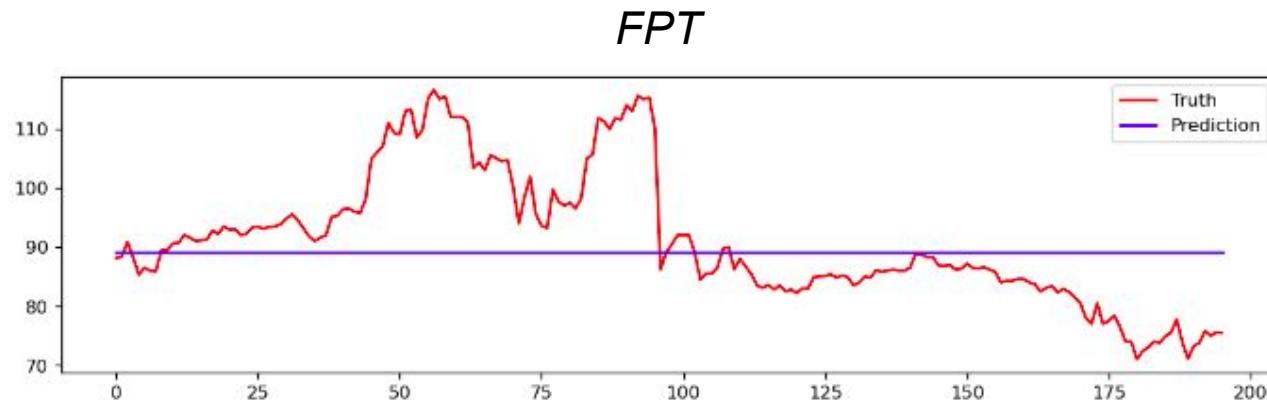
Áp dụng ARIMA vào việc dự đoán

CTG



Hình 27: Trực quan hóa kết quả dự đoán của mô hình ARIMA (1,1,0) so với thực tế. Kết quả dự đoán của hình không được tốt vì nó chỉ cho ra 1 đường thẳng, không bắt được những lần lên xuống của dữ liệu.

Áp dụng ARIMA vào việc dự đoán



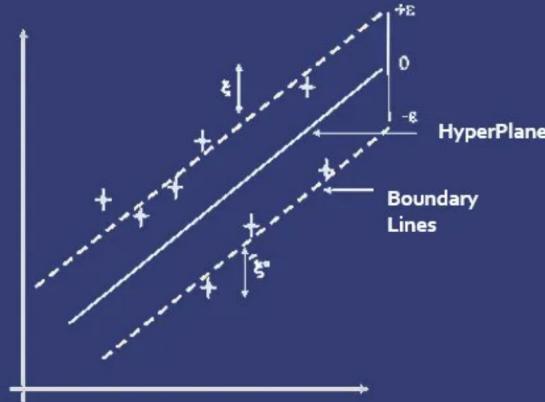
Hình 30: Trực quan hóa kết quả dự đoán của mô hình ARIMA $(0,1,0)$ so với thực tế. Kết quả dự đoán của hình không được tốt vì nó chỉ cho ra 1 đường thẳng, không bắt được những lần lên xuống của dữ liệu.

Mô hình SVR

Giới thiệu mô hình

Mô hình **Support Vector Regression - SVR** được tạo ra dựa trên ý tưởng từ mô hình Support Vector Machine - SVM. Mục tiêu của SVR là tìm được một siêu phẳng chứa được tối đa các training observation trong giới hạn margin ϵ (tolerance level)

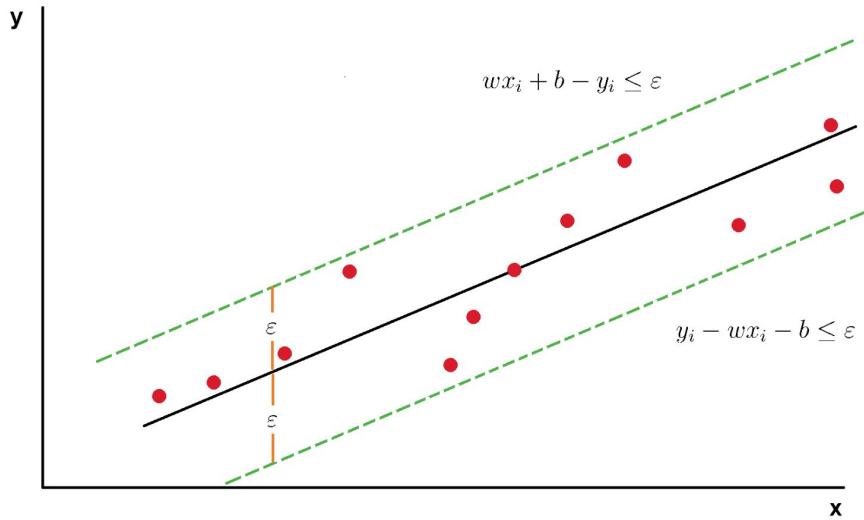
Support Vector Regression



Bài toán tối ưu trong SVR

Từ mục tiêu của mô hình SVR, dễ thấy rằng SVR là một bài toán tối đa hóa margin ε và sẽ tương đương với bài toán tối ưu như sau:

- Minimize: $\min \frac{1}{2} \|w\|^2$
- Subject to: $y_i - wx_i - b \leq \varepsilon$
 $wx_i + b - y_i \leq \varepsilon$

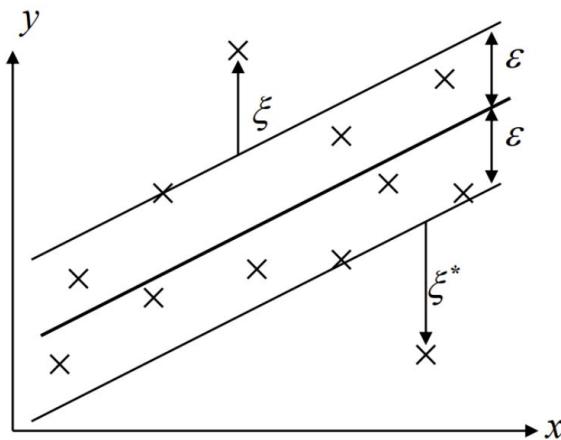


Support Vector Regression - SVR

Gọi y_i là giá trị thực thứ i và \hat{y}_i là giá trị dự đoán thứ i với \hat{y}_i có dạng $\hat{y}_i = wx_i + b$, từ các ràng buộc trong bài toán tối ưu trên, ta suy ra: $|y - \hat{y}| \leq \varepsilon$.

Soft margin

Từ các ràng buộc của bài toán tối ưu trên, dễ thấy rằng tất cả mọi quan sát không được nằm ngoài margin, và nó tương tự với trường hợp hard margin trong bài toán phân loại của SVM. Bài toán tối ưu này đôi khi sẽ không thể tìm ra được lời giải, vì vậy ý tưởng sẽ là sử dụng soft margin để thay thế.



Soft margin

Trong trường hợp này, bài toán tối ưu của SVR sẽ trở thành:

- Minimize: $\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$
- Subject to: $y_i - wx_i - b \leq \varepsilon + \xi_i$
 $wx_i + b - y_i \leq \varepsilon + \xi_i^*$
 $\xi_i, \xi_i^* \geq 0$

Các loại hàm kernel - Linear kernel

Đây là trường hợp đơn giản, với kernel là tích vô hướng của hai vector:

$$k(x, y) = x^T y$$

Trong thư viện sklearn, để dùng hàm nhân đa thức cho SVR, ta thiết lập kernel = 'linear'.
Hàm nhân tuyến tính nhanh về mặt tính toán và ít bị overfitting.

Các loại hàm kernel - Polynomial kernel

Công thức của hàm nhân đa thức như sau:

$$k(x, y) = (r + \gamma x^T y)^d$$

Trong đó, $d > 0$, $d \in \mathbb{R}$ chỉ bậc của đa thức. Ở đây, d không cần phải là một số tự nhiên vì mục đích chính của ta là cách tính kernel chứ không phải bậc đa thức. Trong thư viện sklearn, để dùng hàm nhân đa thức cho SVR, ta thiết lập `kernel = 'poly'` và các hệ số d (degree), γ (gamma), r (coef0).

Hàm nhân đa thức phù hợp để giải quyết các bài toán phi tuyến tính.

Các loại hàm kernel - Radial Basic Function kernel (RBF kernel)

Công thức của hàm nhân đa thức như sau:

$$k(x, y) = \exp(-\gamma \|x - y\|^2), \gamma > 0$$

Trong thư viện sklearn, để dùng hàm nhân RBF cho SVR, ta thiết lập kernel = 'rbf' và hệ số γ (gamma).

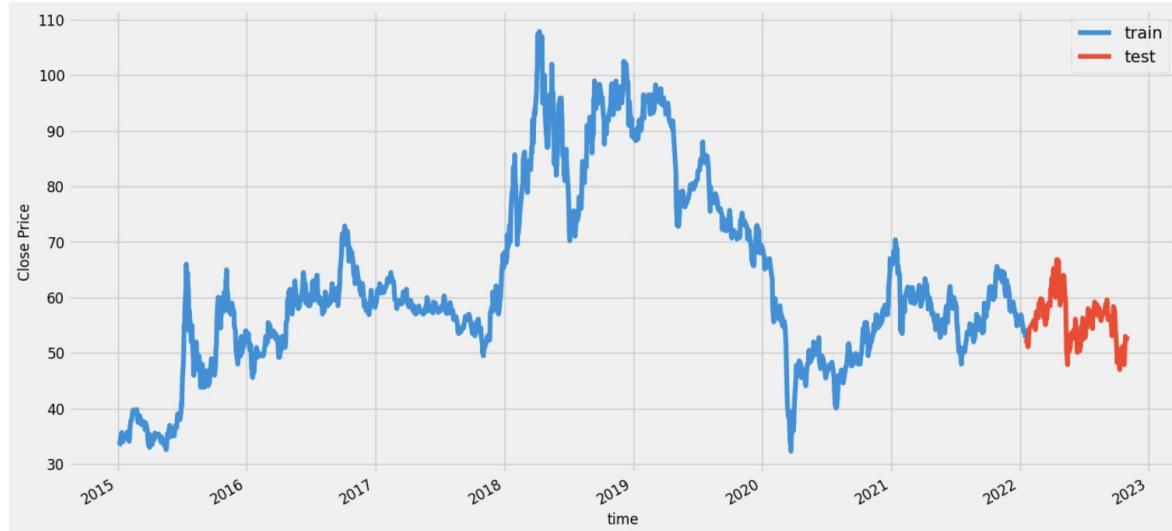
Hàm nhân RBF không chỉ hiệu quả về mặt tính toán mà còn phù hợp để giải quyết các bài toán phi tuyến tính. Trong thực tế, RBF kernel được dùng nhiều nhất.

Áp dụng mô hình SVR vào dự đoán giá chứng khoán

Để áp dụng mô hình SVR vào dự đoán giá chứng khoán, nhóm có một số thiết lập ban đầu như sau:

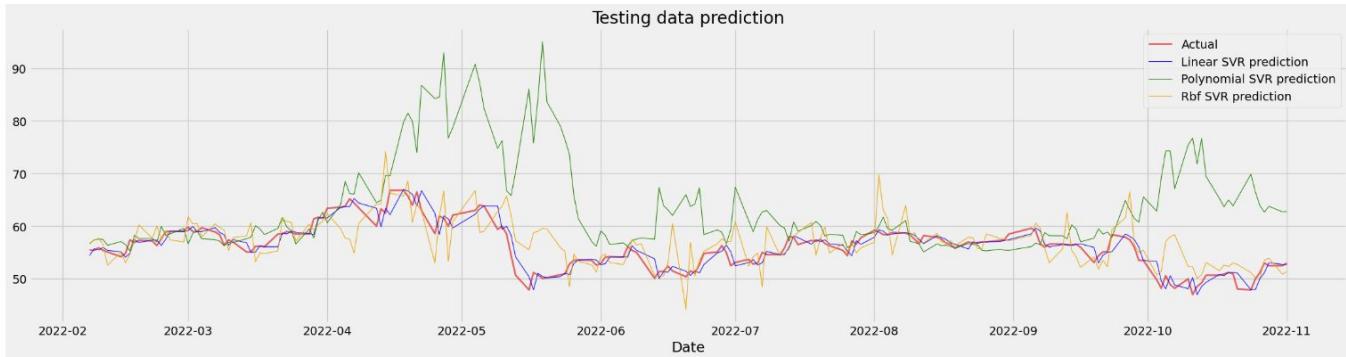
- Dùng 90% tập dữ liệu của từng công ty để train và 10% còn lại để test.
- Chuẩn hóa dữ liệu bằng StandardScaler
- Dự đoán dựa trên thuộc tính close của tập dữ liệu
- Dự đoán ngày hiện tại dựa trên dữ liệu của 9 ngày trước đó.
- Sử dụng mô hình SRV có hằng số $C = 1000$, với 3 loại hàm nhân: linear kernel; polynomial kernel với $d = 2, r = 0$; RBF kernel với $\gamma = 0.5$.

SVR - Dự đoán cho BVH



Trực quan hóa dữ liệu tập train và tập test của công ty BVH

SVR - Dự đoán cho BVH

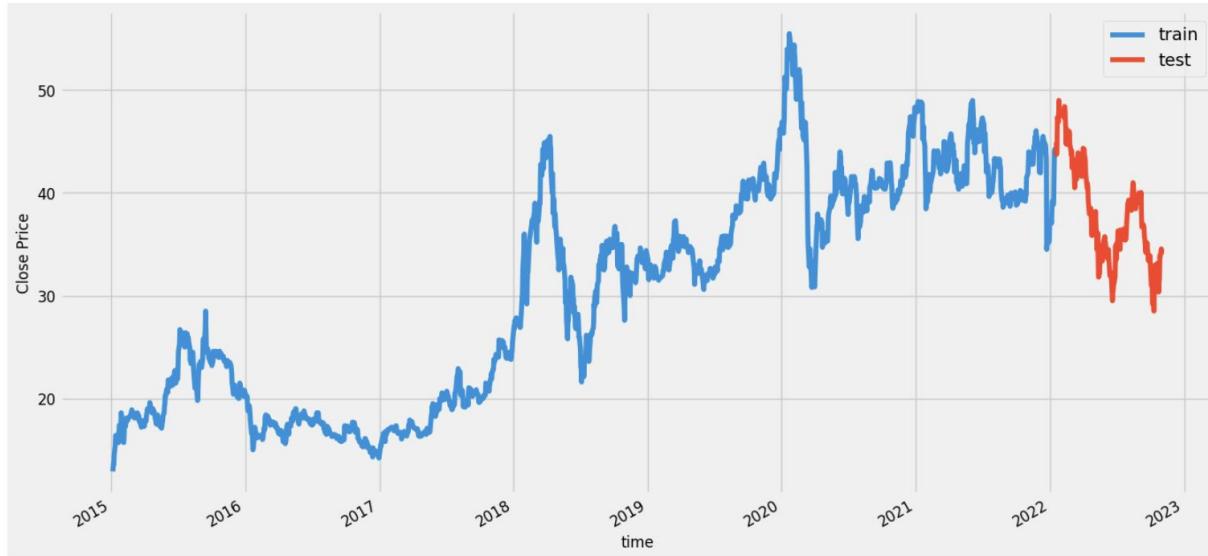


Trực quan hóa kết quả dự đoán so với giá trị thực trên tập test của công ty BVH

Bảng sau là kết quả một số độ đo error của mô hình trên tập test:

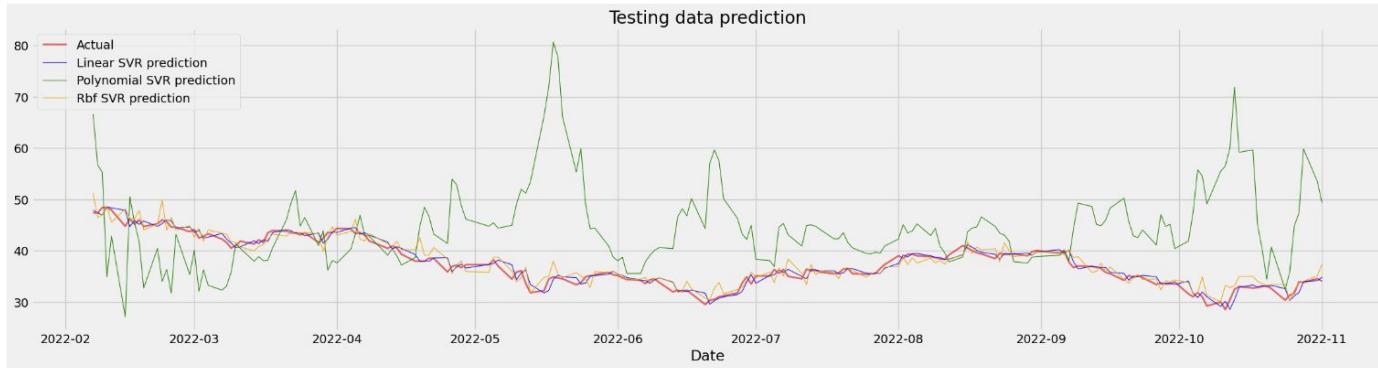
Kernel	MSE	MAPE	MAE
Linear	2.985	0.0189	1.0569
Polynomial	138.749	0.1429	7.6794
RBF	13.3278	0.0483	2.6793

SVR - Dự đoán cho BID



Trực quan hóa dữ liệu tập train và tập test của công ty BID

SVR - Dự đoán cho BID

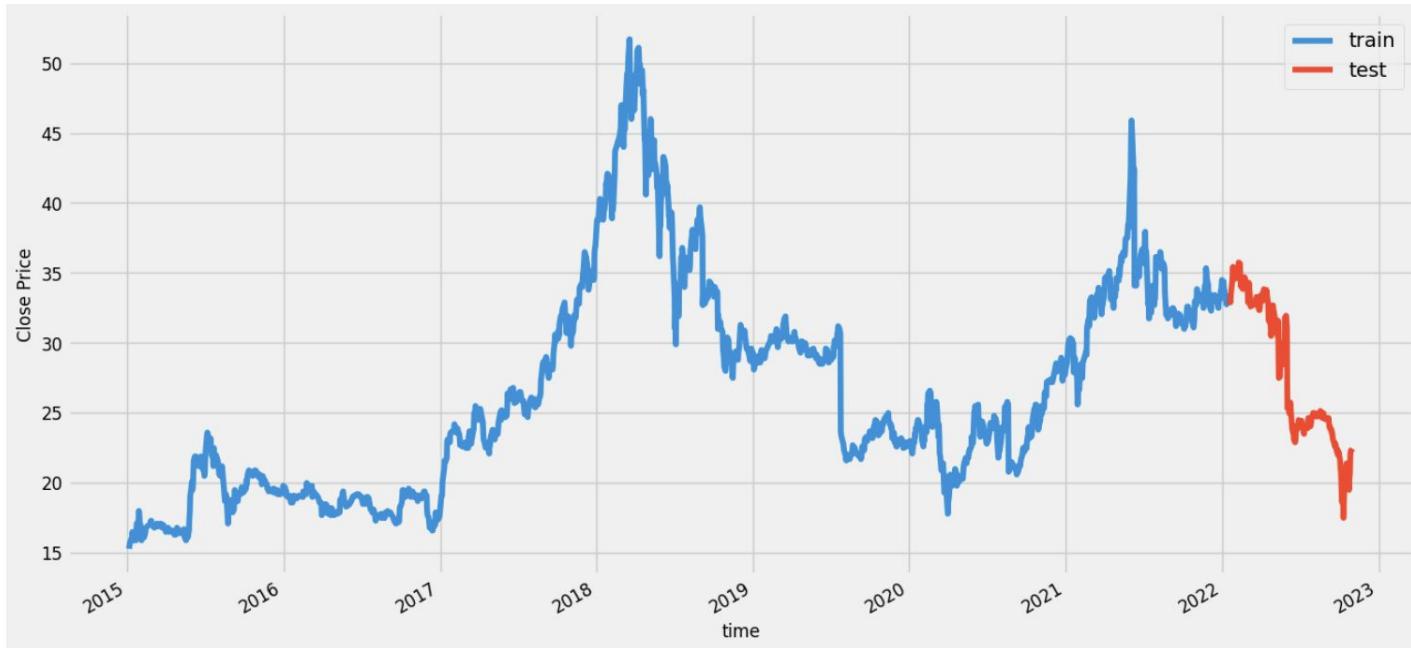


Trực quan hóa kết quả dự đoán so với giá trị thực trên tập test của công ty BID

Bảng sau là kết quả một số độ đo error của mô hình trên tập test:

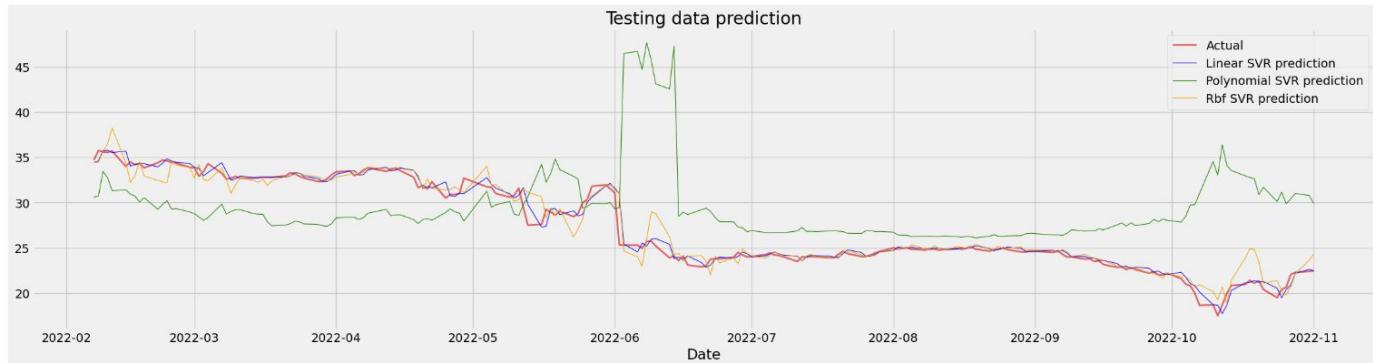
Kernel	MSE	MAPE	MAE
Linear	0.969	0.0206	0.7497
Polynomial	165.21	0.2717	9.5154
RBF	2.2836	0.0318	1.1728

SVR - Dự đoán cho ACB



Trực quan hóa dữ liệu tập train và tập test của công ty ACB

SVR - Dự đoán cho ACB

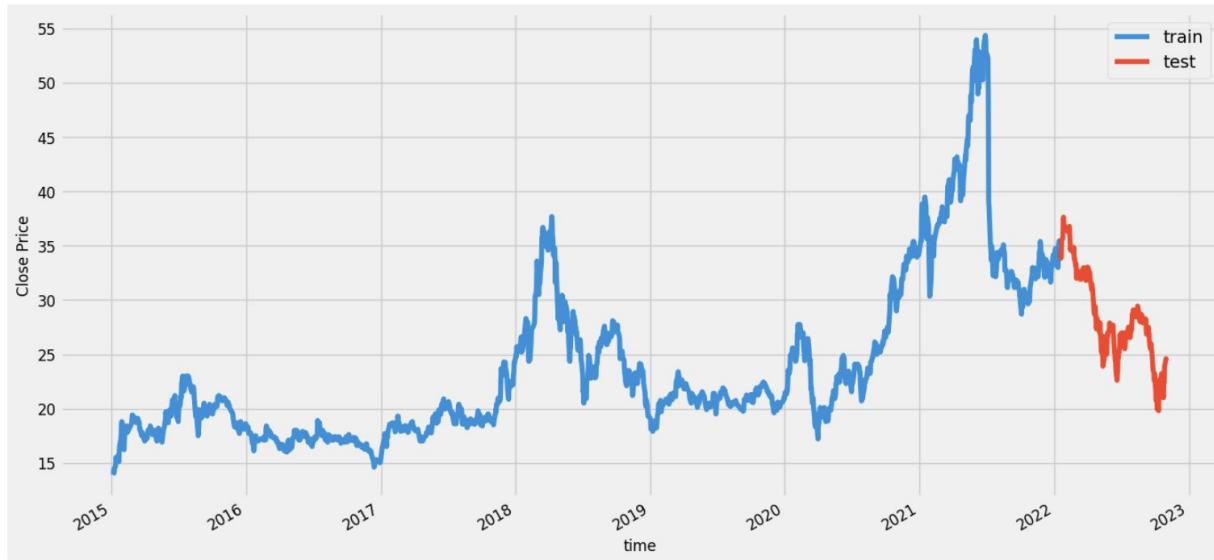


Trực quan hóa kết quả dự đoán so với giá trị thực trên tập test của công ty ACB

Bảng sau là kết quả một số độ đo error của mô hình trên tập test:

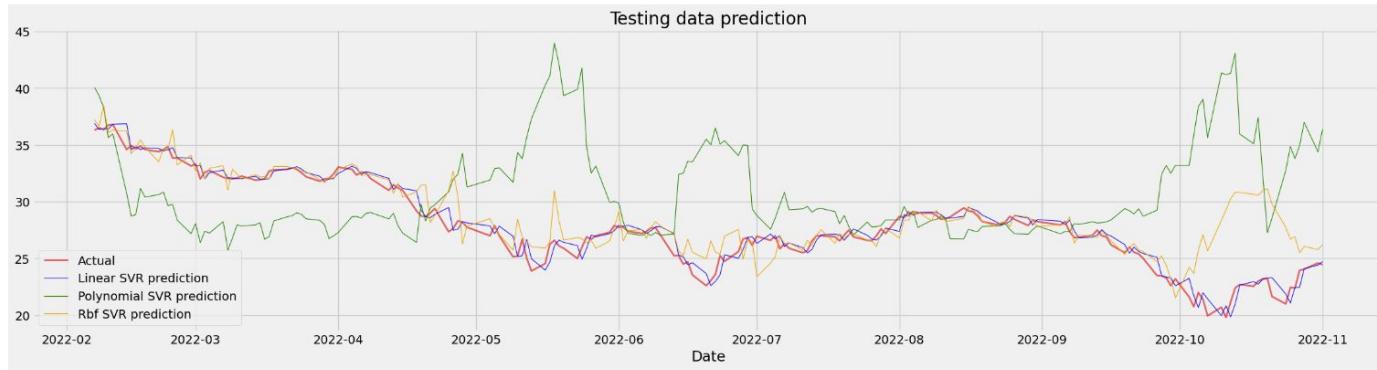
Kernel	MSE	MAPE	MAE
Linear	0.5135	0.0165	0.4347
Polynomial	45.0151	0.2014	5.0468
RBF	1.3568	0.0285	0.7462

SVR - Dự đoán cho CTG



Trực quan hóa dữ liệu tập train và tập test của công ty CTG

SVR - Dự đoán cho CTG

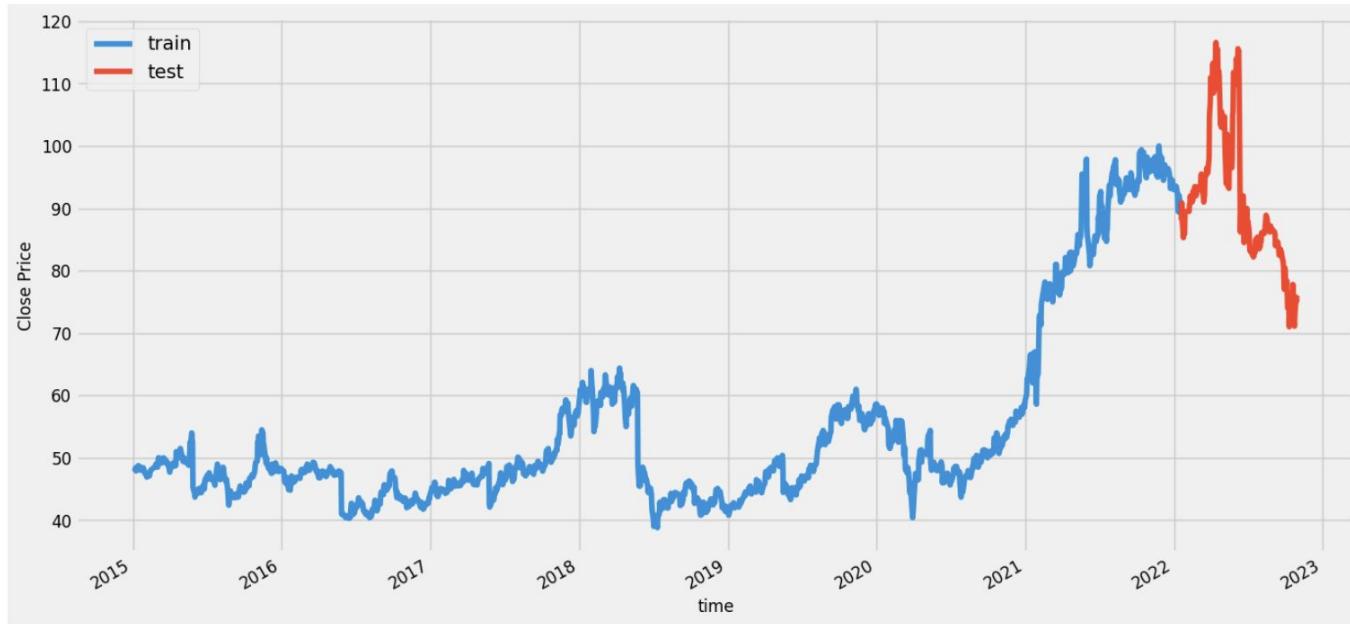


Trực quan hóa kết quả dự đoán so với giá trị thực trên tập test của công ty CTG

Bảng sau là kết quả một số độ đo error của mô hình trên tập test:

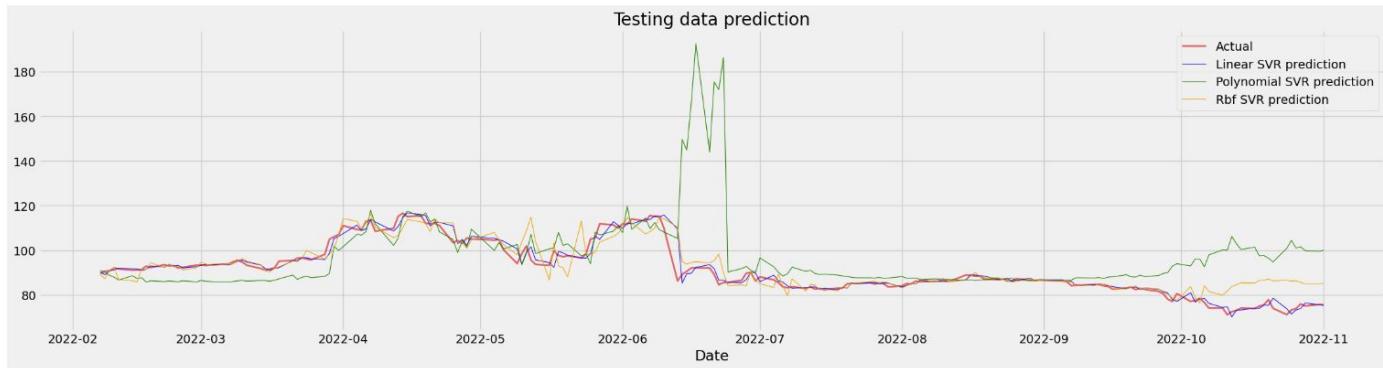
Kernel	MSE	MAPE	MAE
Linear	0.4982	0.01996	0.5247
Polynomial	52.4081	0.20798	5.2892
RBF	5.9562	0.0573	1.4004

SVR - Dự đoán cho FPT



Trực quan hóa dữ liệu tập train và tập test của công ty FPT

SVR - Dự đoán cho FPT



Trực quan hóa kết quả dự đoán so với giá trị thực trên tập test của công ty FPT

Bảng sau là kết quả một số độ đo error của mô hình trên tập test:

Kernel	MSE	MAPE	MAE
Linear	7.1543	0.0163	1.5021
Polynomial	369.4572	0.1186	10.134
RBF	27.9378	0.0386	3.3757

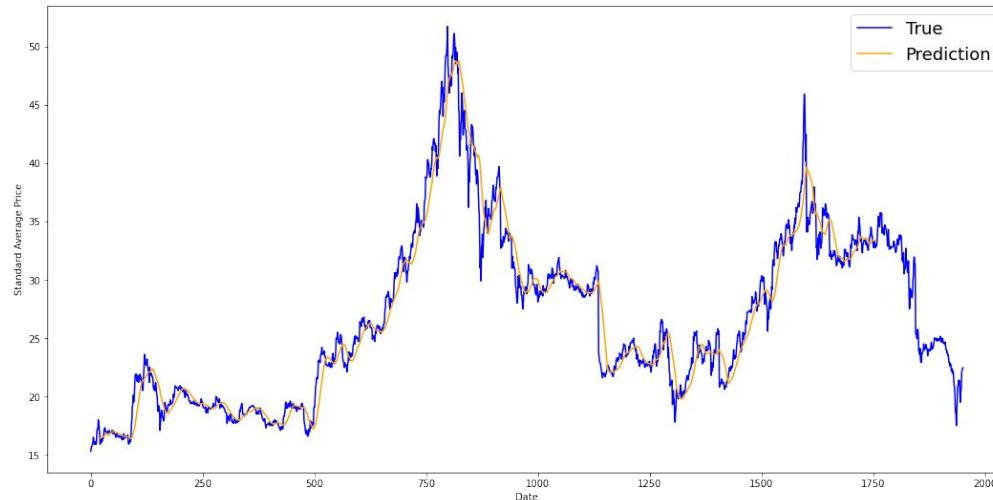
Mô hình Baseline

Mô hình One-Step Ahead thông qua việc trung bình giá

- Trung bình giá cơ bản (Standard averaging)
- Trung bình giá lũy thừa (Exponential moving average)

Mô hình trung bình giá cơ bản - Standard Averaging

Giá cổ phiếu ở ngày thứ n trong tương lai sẽ bằng trung bình cộng giá của K ngày trong quá khứ



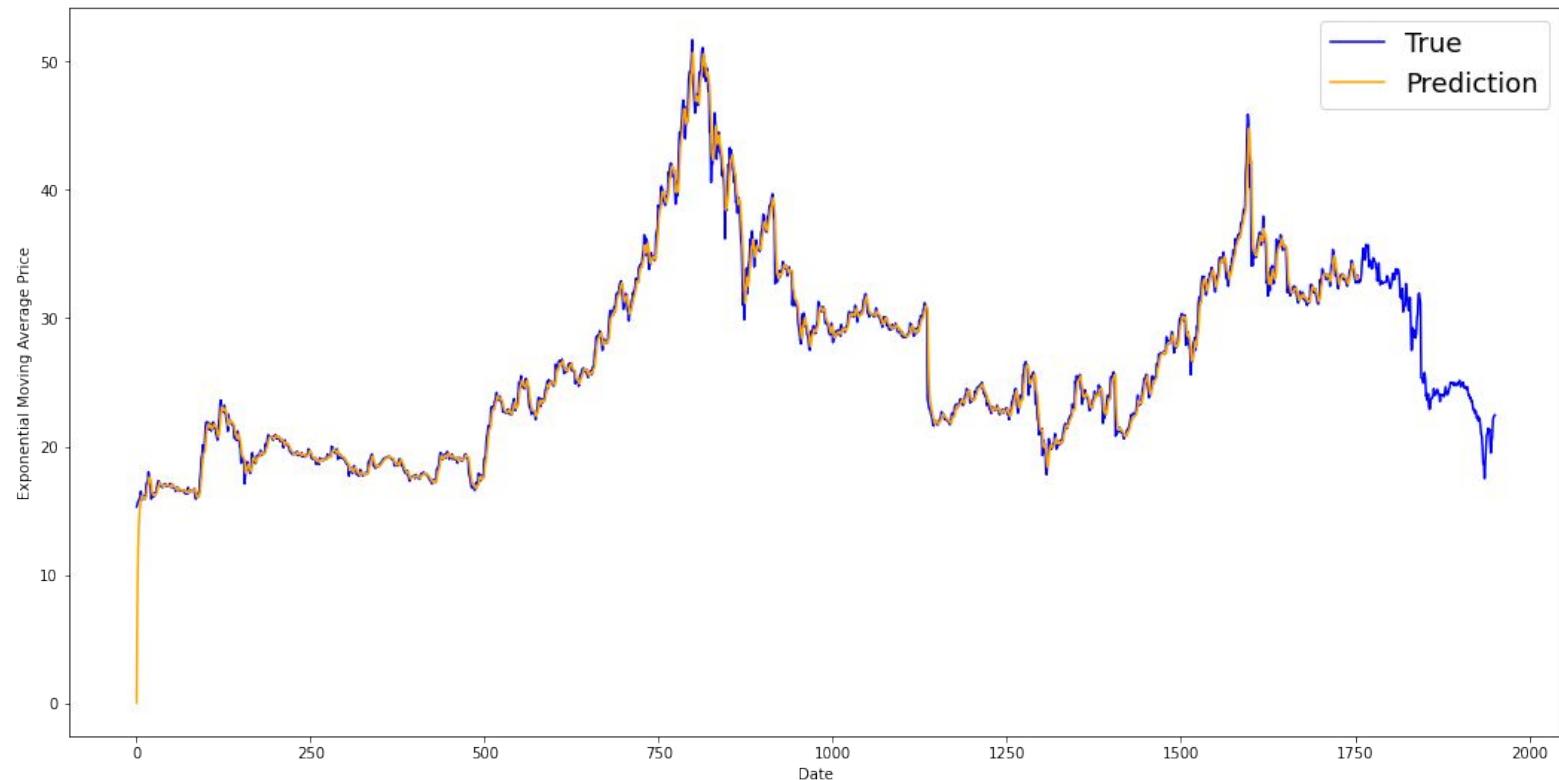
Mô hình trung bình giá lũy thừa - Exponential Moving Average

Giá cổ phiếu ngày thứ n trong tương lai được tính như sau:

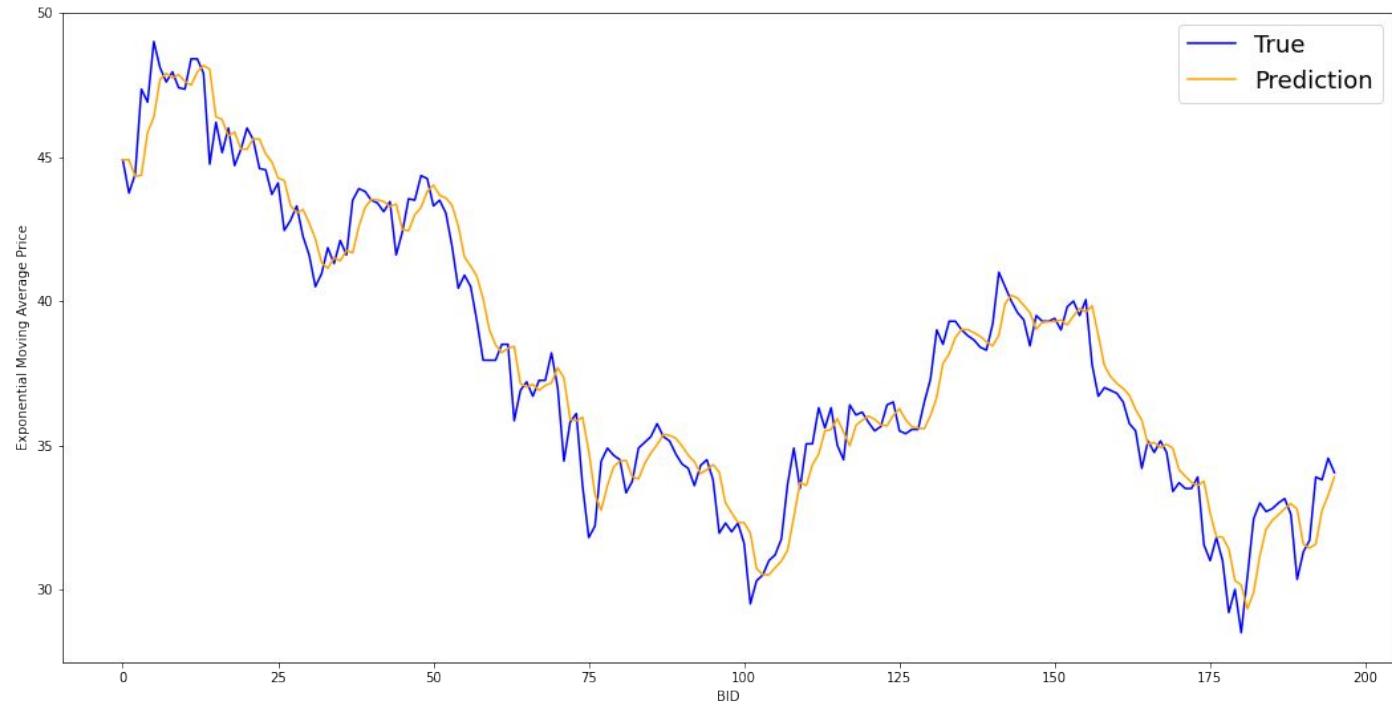
$$x_n = \text{EMA}_{n-1} = \gamma * \text{EMA}_{n-2} + (1 - \gamma)x_{n-1}$$

$$\text{EMA}_0 = 0$$

Mô hình trung bình giá lũy thừa - Exponential Moving Average



Mô hình trung bình giá lũy thừa - Exponential Moving Average



Mô hình trung bình giá lũy thừa - Exponential Moving Average

Vậy nếu nó đã tốt đến như thế

vì sao chúng ta lại cần những phương pháp phức tạp khác?

Mô hình trung bình giá lũy thừa - Exponential Moving Average

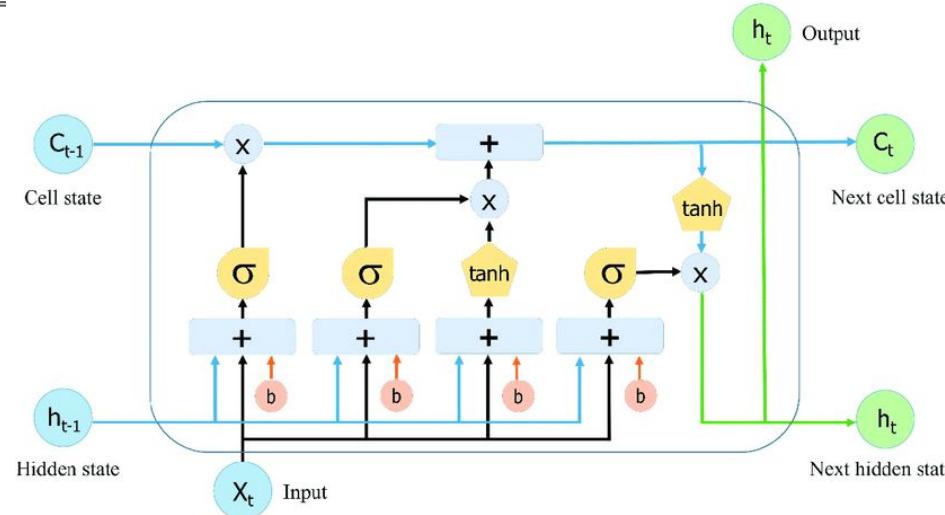
<i>Baseline</i>			
Stocks	MSE	MAPE	MAE
ACB	0.6615	0.0191	0.4985
BID	1.1850	0.0219	0.8046
BVH	2.7125	0.0211	1.1681
CTG	0.6313	0.0225	0.5996
FPT	9.3772	0.0186	1.7228
	2.9135	0.0206	0.9587

Mô hình LSTM

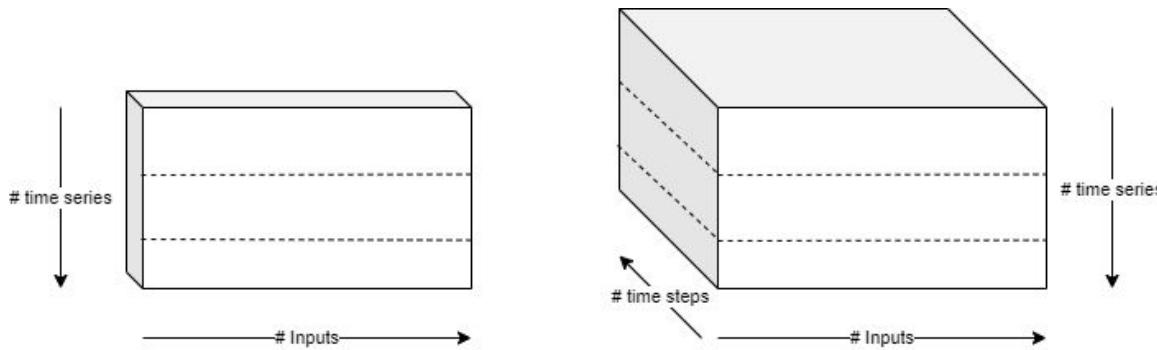
Mô hình Long short-term memory - LSTM

- Dữ liệu liên tục theo thời gian (time series),
giá trị trong tương lai phụ thuộc vào những biến động trong quá khứ.
- Mô hình LSTM cực kỳ mạnh trong việc xử lý các dữ liệu chuỗi thời gian,
đồng thời nó cũng có thể tìm ra những dữ liệu ẩn (hidden patterns) trong dữ
liệu ở quá khứ để đưa ra dự đoán xa hơn vào tương lai

Mô hình Long short-term memory - LSTM

**Inputs:** X_t Current input C_{t-1} Memory from last LSTM unit h_{t-1} Output of last LSTM unit**Outputs:** C_t New updated memory h_t Current output**Nonlinearities:** σ Sigmoid layer \tanh Tanh layer b Bias**Vector operations:** \times Scaling of information $+$ Adding information

Xây dựng mô hình LSTM



	Date	Close
0	19/08/2004	49.98265457
1	20/08/2004	53.95277023
2	23/08/2004	54.49573517
3	24/08/2004	52.23919296
4	25/08/2004	52.80208588
5	26/08/2004	53.75351715

Xây dựng mô hình LSTM

Layer (type)	Output Shape	Param #
<hr/>		
lstm_4 (LSTM)	(None, 3, 100)	40800
dropout_4 (Dropout)	(None, 3, 100)	0
lstm_5 (LSTM)	(None, 3, 50)	30200
dropout_5 (Dropout)	(None, 3, 50)	0
lstm_6 (LSTM)	(None, 3, 50)	20200
dropout_6 (Dropout)	(None, 3, 50)	0
lstm_7 (LSTM)	(None, 50)	20200
dropout_7 (Dropout)	(None, 50)	0
dense_1 (Dense)	(None, 1)	51
<hr/>		
Total params: 111,451		
Trainable params: 111,451		
Non-trainable params: 0		

Kết quả mô hình LSTM

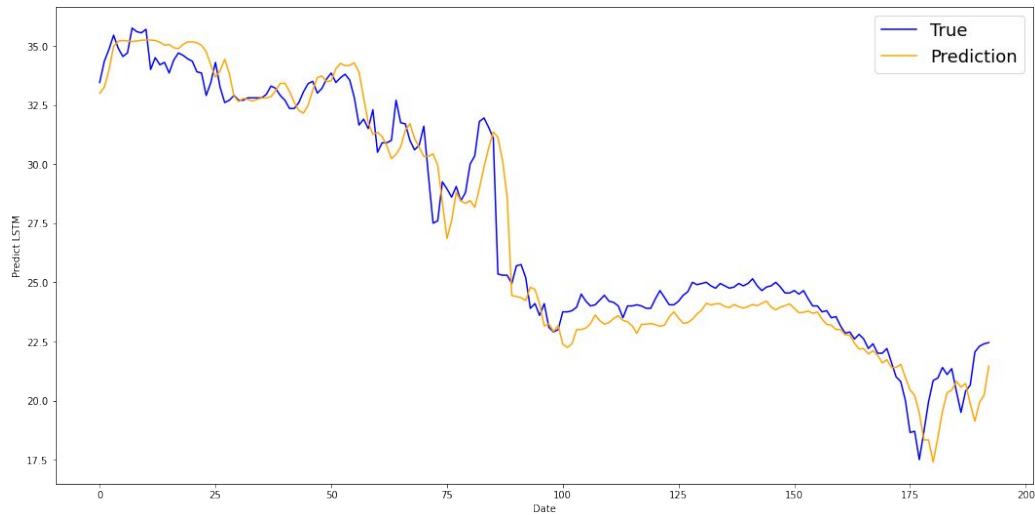
===== ACB =====

MSE: 1.420404601069918

MAPE: 0.034315722893661575

MAE: 0.8928547043874473

Epoch 27: early stopping



Kết quả mô hình LSTM

<i>LSTM</i>			
Stocks	MSE	MAPE	MAE
ACB	1.4204	0.0343	0.8929
BID	3.0530	0.0375	1.3989
BVH	19.6027	0.0643	3.4929
CTG	2.0949	0.0435	1.2302
FPT	1573.2482	0.4081	38.0363
	319.8838	0.1175	9.0102

Đánh giá mô hình LSTM

- Đồ thị dự đoán rất tốt trên những cổ phiếu có giá thấp
- Đồ thị dự đoán rất tệ trên những cổ phiếu có giá cao, điều này khả năng cao là do việc chưa chuẩn hóa dữ liệu.

Vì vậy việc chuẩn hóa dữ liệu rất quan trọng trong data mining, nó có thể làm một mô hình tốt cho ra kết quả cực kỳ sai lệch

Tổng hợp kết quả

Tổng hợp kết quả - Tổng quan

Mô hình	MSE	MAPE	MAE
EMA(baseline)	2.9135	0.0206	0.9587
ARIMA	58.9363	0.117	6.307
SVR	2.424	0.0184442	0.85362
LSTM	319.8838	0.1175	9.0102

Bảng kết quả trung bình của 5 loại cổ phiếu.

Tổng hợp kết quả - các công ty

Mô hình	MSE	MAPE	MAE
EMA(baseline)	2.7125	0.0211	1.1681
ARIMA	23.5579	0.067222	3.863489
SVR	2.985	0.018941	1.0569
LSTM	19.6027	0.0643	3.4929

Bảng kết quả của BVH.

Mô hình	MSE	MAPE	MAE
EMA(baseline)	0.6615	0.0191	0.4985
ARIMA	49.9821	0.240708	5.693627
SVR	0.5135	0.016484	0.4347
LSTM	1.4204	0.0343	0.8929

Bảng kết quả trung bình của ACB.

Mô hình	RMSE	MAPE	MAE
EMA(baseline)	1.1850	0.0219	0.8046
ARIMA	56.4126	0.18482	6.438573
SVR	0.969	0.020564	0.7497
LSTM	3.0530	0.0375	1.3989

Bảng kết quả của BID.

Mô hình	RMSE	MAPE	MAE
EMA(baseline)	0.6313	0.0225	0.5996
ARIMA	61.0393	0.266443	6.860772
SVR	0.4982	0.019964	0.5247
LSTM	2.0949	0.0435	1.2302

Bảng kết quả trung bình của CTG.

Mô hình	MSE	MAPE	MAE
EMA(baseline)	9.3772	0.0186	1.7228
ARIMA	129.7215	0.092189	8.677041
SVR	7.1543	0.016268	1.5021
LSTM	1573.2482	0.4081	38.0363

Bảng kết quả của FPT.

Kết luận

Như vậy, trong bài tập lớn này, nhóm đã nghiên cứu về 4 mô hình cho việc dự đoán time series gồm Exponential Moving Average (EMA) (baseline), ARIMA, SVR và LSTM. Trong đó, tuy EMA là một kỹ thuật cổ điển nhưng lại là 1 baseline rất tốt ít nhất trong bài tập lớn này. ARIMA, tuy có phức tạp hơn, nhưng do yêu cầu về dữ liệu phải ổn định (stationary) nên có thể là 1 nguyên nhân dẫn đến mô hình không tốt. Về SVR, tuy là 1 kỹ thuật machine learning, nhưng lại hiệu quả hơn so với LSTM, một mô hình deep learning được nhóm kỳ vọng sẽ đạt kết quả tốt nhất.

Trong tương lai, nhóm sẽ thực hiện chuẩn hóa và cho các mô hình deep learning như LSTM nhiều dữ liệu hơn thay vì chỉ sử dụng 1 mô hình cho 1 loại cổ phiếu. Đồng thời, nhóm sẽ tìm hiểu và áp dụng các mô hình, kỹ thuật khác trong bài toán dự đoán trong time series như Transformer, 1D Dilated Causal Convolution Network,...

Tài liệu tham khảo

- [1] Machine learning cơ bản, *Kernel Support Vector Machine*
<https://machinelearningcoban.com/2017/04/22/kernelsmv/>
- [2] Paul Paisitkriangkrai, *Linear Regression and Support Vector Regression*
https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf
- [3] Hyndman, R.J., Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 2022-08-11.
- [4] Thushan Ganegedara, *LStock Market Predictions with LSTM in Python*
<https://www.datacamp.com/tutorial/lstm-python-stock-market>
- [5] Jordi Corbilla , *Stock prediction using deep neural learning*
<https://jordicorbilla.github.io/stock-prediction-deep-neural-learning/>



CẢM ƠN THẦY VÀ CÁC
BẠN ĐÃ THEO DÕI



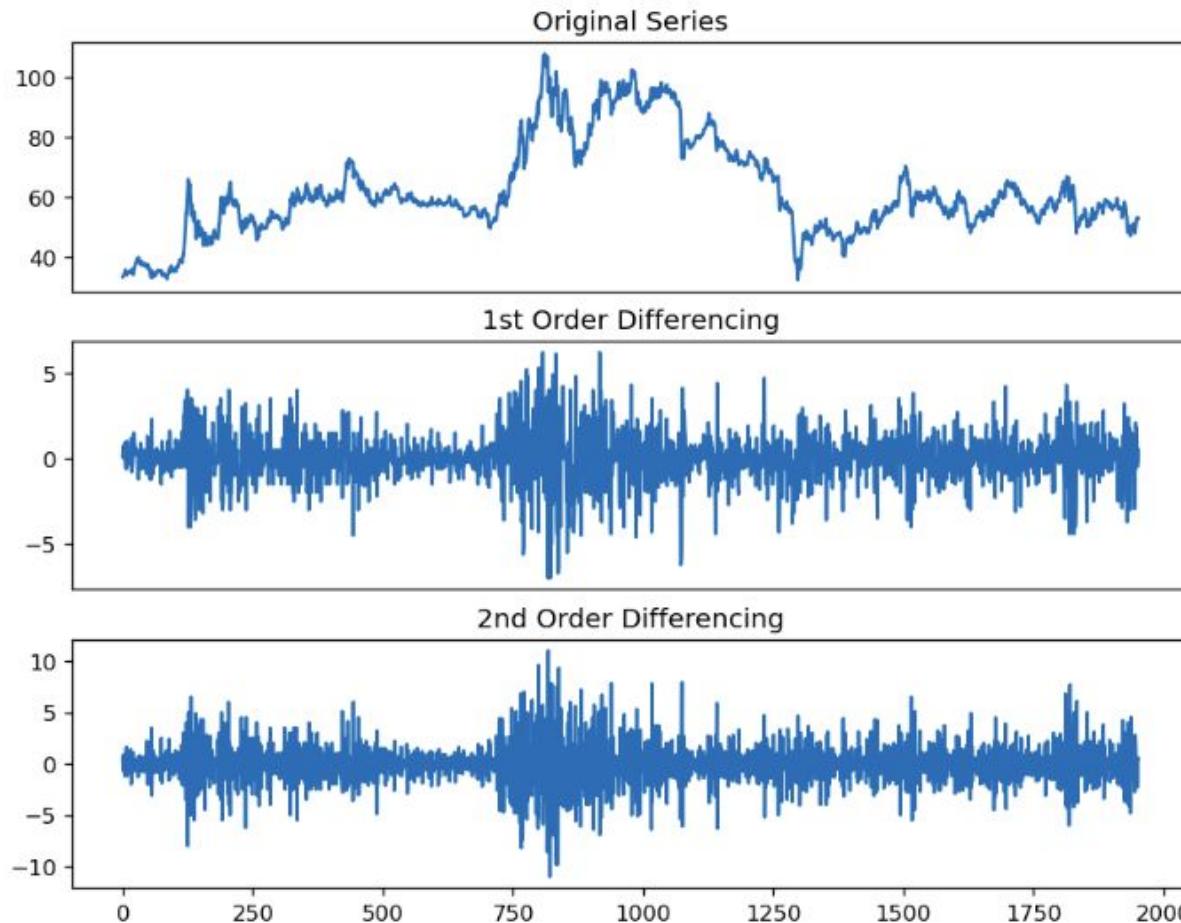
Appendix

Ví dụ cách chọn tham số cho mô hình ARIMA trong việc dự đoán giá cổ phiếu của công ty BVH.

Đầu tiên cần kiểm tra dữ liệu có stationary hay không bằng cách thử $d=0,1,2$

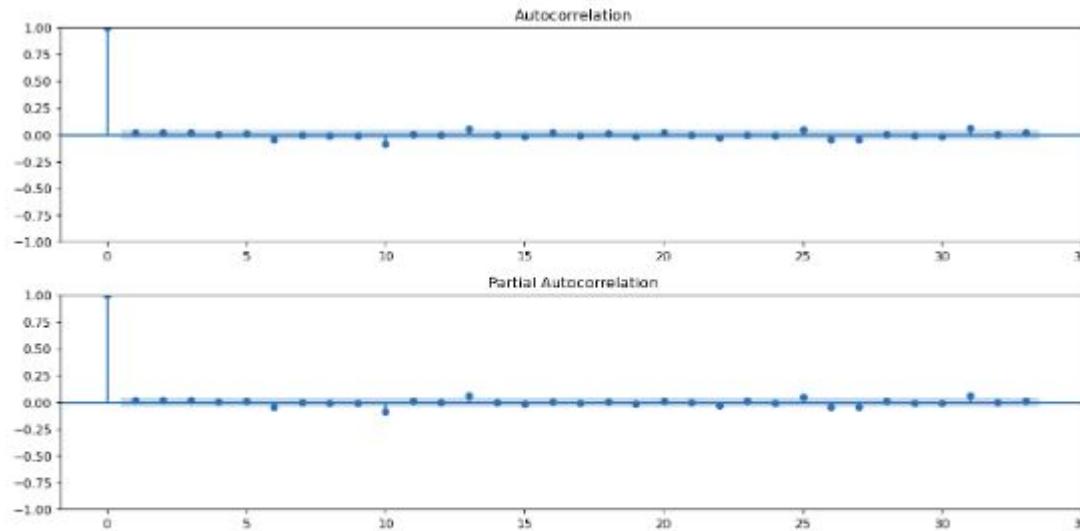
$d=0$	$d=1$	$d=2$
p-value=0.136442	p-value=0	p-value=0

Từ đây chọn $d=1$ và $d=2$



Hình 16: Dữ liệu gốc, dữ liệu khi lấy hiệu 1 lần và dữ liệu khi lấy hiệu 2 lần

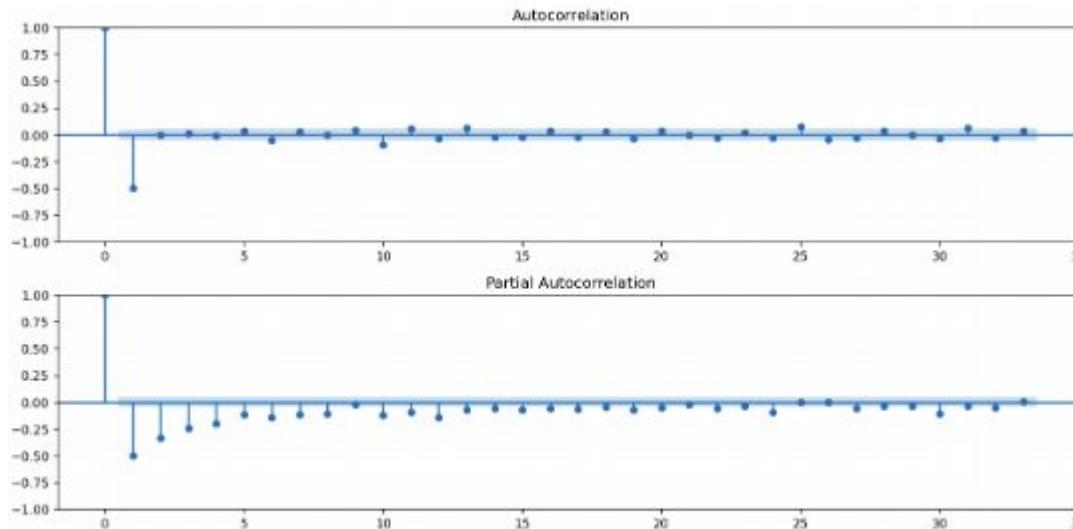
Sau đó, vẽ ACF và PACF để xác định p và q.



(a) ACF, PACF khi $d=1$

Nhìn vào hình này, khi $d=1$, ACF và PACF cho ta giá trị ban đầu của p có thể là 0 và q có thể là 0. Tuy nhiên ta sẽ lấy thêm 1 giá trị nữa cho mỗi p và q quanh giá trị ban đầu.
→ Tóm lại $p = \{0, 1\}$; $q = \{0, 1\}$

Sau đó, vẽ ACF và PACF để xác định p và q.



(b) ACF, PACF khi $d=2$

Nhìn vào hình này, khi $d=2$, ACF và PACF cho ta giá trị ban đầu của p có thể là 8 và q có thể là 1. Tuy nhiên ta sẽ lấy thêm 2 giá trị nữa cho mỗi p và q quanh giá trị ban đầu.
→ Tóm lại $p = \{6, 7, 8\}$; $q = \{0, 1, 2\}$