VIỆN NGHIÊN CỨU DỮ LIỆU LỚN VINBIGDATA

Ứng dụng mô hình time-series trong dự báo giá chứng khoán FPT

Giảng viên hướng dẫn: TS. Cao Văn Lợi
Thành viên:

1. Trần Văn Tuấn
2. Dương Chí Vinh
3. Trần Thị Ngọc Anh
4. Võ Đức Mẫn

HÀ NỘI, 2020

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Tổng quan

1.1 Giới thiệu

Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Các dự báo chuỗi thời gian có tính ứng dụng cao và được sử dụng rất nhiều lĩnh vực như ngân hàng, thương mại điện tử, bảo hiểm, marketing và không thể thiếu một lĩnh vực rất hấp dẫn đó là chứng khoán. Nếu giá giao dịch chứng khoán có thể được dự báo tốt, điều này sẽ giúp ích cho các nhà đầu tư trong việc quản trị danh mục đầu tư của mình, ra quyết định giữ, mua hay bán cổ phiếu phù hợp, từ đó thu được lợi nhuận cao. Tuy nhiên việc dự báo giá chứng khoán luôn là bài toán rất khó bởi giá cả trên thị trường chịu tác động của rất nhiều yếu tố phức tạp như: các sự kiện kinh tế, xã hội lớn và nhỏ, nhận thức của công chúng, kỳ vọng về sự thay đổi,.. Chính vì vậy, sai số trong các mô hình dự đoán là không tránh khỏi. Khi tiếp cận bài toán này, nhóm chúng tôi cố gắng sử dụng nhiều phương pháp và cải tiến mô hình nhằm đưa ra các dự đoán hợp lý dựa trên giá trị trong quá khứ. Mục tiêu cuối cùng là tìm ra một mô hình khả dĩ nhất để giúp các nhà đầu tư có góc nhìn đa chiều hơn, ra quyết định một cách khôn ngoan, ít cảm tính, từ đó tạo ra lợi nhuận nhiều hơn.

1.2 Dữ liệu

Tiêu chí lựa chọn mã cổ phiếu của nhóm là cổ phiếu nằm trong rổ cổ phiếu VN30 (danh sách 30 cổ phiếu hàng đầu được xếp hạng bởi Sở giao dịch Chứng Khoán – TP. HCM). Cổ phiếu nằm trong rổ VN30 có đặc điểm:

- Có lượng thanh khoản và giá trị vốn hóa cao, khối lượng giao dịch hàng ngày lớn.
- Thường là cổ phiếu của các doanh nghiệp đầu ngành ở các lĩnh vực khác nhau.
- Cổ phiếu có sức hấp dẫn cao đối với các nhà đầu tư nước ngoài và nhà đầu tư dài hạn. Vì đặc thù các doanh nghiệp lớn có sức khỏe tài chính tốt, hoạt động kinh doanh hiệu quả, nên các cổ phiếu này thường được sự quan tâm lớn từ các khối đầu tư.

Trong rổ cổ phiếu VN30, nhóm quyết định sử dụng dữ liệu từ mã cổ phiếu FPT của công ty Cổ phần FPT - một đơn vị hàng đầu trong lĩnh vực Công nghệ thông tin và viễn thông. Chúng tôi sử dụng số liệu liên quan đến giá, khối lượng giao dịch của cổ phiếu FPT được thống kê tại trang cophieu68.com. Số liệu thu thập được kéo dài hơn 14 năm, từ 16/09/2006 đến 12/2020 bao gồm 5 thuộc tính:

- Open: Giá mở cửa (giá khớp lệnh đầu tiên trong ngày).

- High: Giá khớp lệnh cao nhất trong ngày.
- Low: Giá khớp lệnh thấp nhất trong ngày.
- Close: Giá đóng cửa (giá khớp lệnh cuối cùng trong ngày).
- Volume: Khối lượng giao dịch trong ngày

	Ticker	Time	Open	High	Low	Close	Volume
0	FPT	20201216	57.2	57.5	57.0	57.1	1995530
1	FPT	20201215	57.5	57.6	56.6	56.7	2666900
2	FPT	20201214	57.0	58.0	56.9	57.2	1443760
3	FPT	20201211	56.6	57.0	56.3	57.0	1750270
4	FPT	20201210	57.0	57.7	56.5	56.5	2112490
5	FPT	20201209	56.3	57.8	56.2	57.5	3001820
6	FPT	20201208	55.7	56.6	55.5	56.2	2297110

HÌNH 1.1: Mô tả dữ liệu chứng khoán của FPT.

Trong đó, nhóm sẽ sử dụng những dữ liệu trên để dự đoán Giá đóng cửa (Close) của cổ phiếu FPT.

1.3 Phương pháp đánh giá

Nhóm sử dụng $MAPE$ là metric để đánh giá mô hình dự báo. $MAPE$ phản ánh giá trị dự báo sai khác bao nhiêu phần trăm so với giá trị trung bình và được tính theo công thức sau:

$$\mathcal{MAPE} = \frac{1}{n}\sum_{t=1}^{n}|\frac{actual_t - predict_t}{actual_t}| \tag{1.1}$$

Chương 2

Phân tích dữ liệu (EDA)

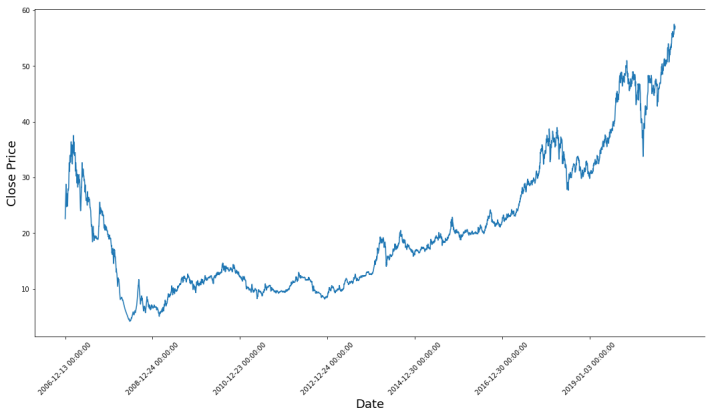
2.1 Yếu tố xu hướng và yếu tố chu kỳ

Dữ liệu chuỗi thời gian thường mang những thành phần đặc trưng như: yếu tố xu hướng (Trend) và yếu tố chu kỳ, thời vụ (Seasonal):

- Yếu tố xu hướng thể hiện sự tăng giảm, hướng thay đổi của dữ liệu trong dài hạn.
- Yếu tố chu kỳ thể hiện những đặc điểm lặp đi lặp lại theo một tần suất cố định, ví dụ như lượng tiêu thụ điện trong các hộ gia đình thường giảm vào buổi trưa và tăng cao vào buổi tối.

2.1.1 Yếu tố xu hướng

Biểu diễn giá đóng cửa cổ phiếu FPT trong cả giai đoạn 2008-2020, ta quan sát hình 2.1 thấy một số điểm đáng chú ý như sau:



HÌNH 2.1: Biểu đồ giá tiền của FPT.

Giai đoạn 12/2016 - 6/2018: giá có xu hướng giảm liên tục và chạm đáy vào giữa năm 2008 với mức 4.1599 đồng/cổ phiếu. Năm 2008 thị trường chứng khoán Việt Nam chứng kiến sự ảnh hưởng sâu sắc bởi thị trường thế giới. Trong bối cảnh cuộc khủng hoảng tài chính bùng nổ tại Mỹ, kế tiếp là cuộc khủng hoảng kinh tế toàn cầu, hàng loạt cổ phiếu trên sàn chứng khoán bao gồm cả cổ phiếu FPT đều bị sụt giảm mạnh.

Giai đoạn 06/2008 - 12/2017: giá cổ phiếu tăng trưởng tốt trong một thập kỷ tiếp theo. Mặc dù có những nhịp giảm giá , tuy nhiên cổ phiếu của công ty này vẫn được đánh giá là một trong những lựa chọn tốt nhất trên thị trường chứng khoán.

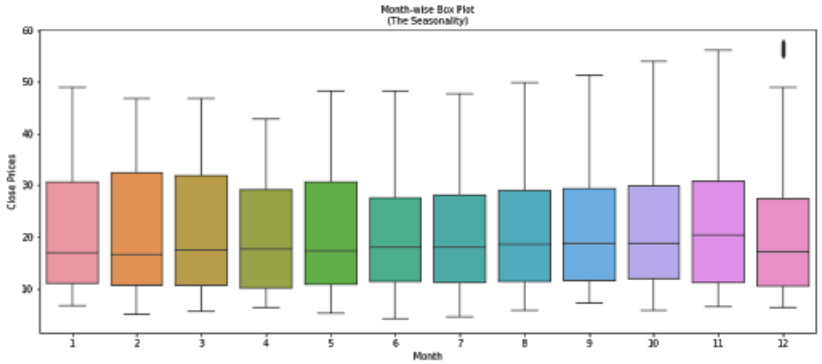
Giai đoạn 01/2018 - 12/2020: có 2 thời điểm giá giảm đáng chú ý.

- Một là, cổ phiếu chạm đáy ở mức giao dịch 27.6926 đồng/ cổ phiếu vào tháng 07/2018. Thời điểm này, việc FED nhiều lần tăng lãi suất cùng sự căng thẳng của cuộc chiến thương mại Mỹ - Trung đã ảnh hưởng mạnh đến dòng vốn vào thị trường chứng khoán toàn cầu cũng như Việt Nam nên cổ phiếu FPT bị ảnh hưởng khá nhiều. Sau đó giá cổ phiếu dần tăng trưởng trở lại.
- Hai là, ngày 30/03/2020, giá cổ phiếu FPT giao dịch ở mức 33.7632 đồng/cổ phiếu, đây là mức giá thấp nhất của FPT trong vòng 1 năm qua. Theo các chuyên gia chứng khoán, sở dĩ giá cổ phiếu FPT giảm mạnh là do các nhà đầu tư bán tháo cổ phiếu này trong vòng xoáy dịch COVID-19.

Tuy nhiên đây chỉ là sự giảm ngắn hạn, còn xét trung hạn cổ phiếu FPT vẫn duy trì được xu hướng giá tăng và khối lượng giao dịch lớn, bằng chứng là giai đoạn nửa cuối 2020, khối lượng giao dịch tăng 50% so với cùng kỳ 2019 đồng thời giá giao dịch dao động quanh 57.000 đồng/CP, tăng gấp đôi so với mức đáy năm 2020.

2.1.2 Yếu tố chu kỳ, thời vụ

Tiếp tục biểu diễn dữ liệu theo các tháng để xem dữ liệu có bị ảnh hưởng bởi yếu tố mùa vụ (Seasonal) hay không bị ảnh hưởng bởi yếu tố mùa vụ.



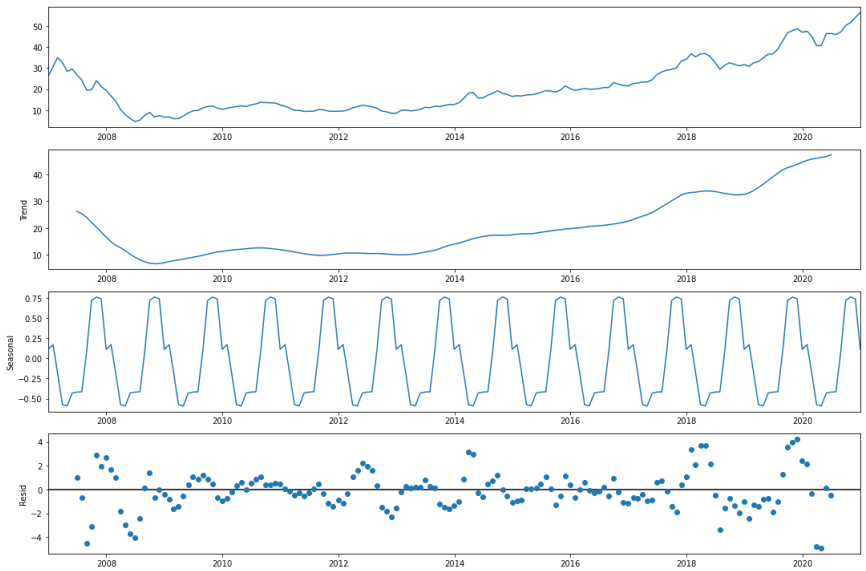
HÌNH 2.2: Biểu đồ hộp theo tháng của FPT.

Từ biểu đồ 3.5 ta thấy giá đóng cửa ít bị ảnh hưởng bởi yếu tố mùa vụ. Giá trung bình các tháng cùng xấp xỉ ở mức 20.000 và chỉ tăng nhẹ vào tháng 11.

HÌNH 2.3: Biểu đồ mùa vụ của FPT.

Tuy nhiên, đây là chỉ là xu hướng khi gộp chung 14 năm để xét. Nếu cẩn thận quan sát cụ thể biến động giá theo tháng trong tất cả các năm trong biểu đổ 2.3, sự tăng giảm của giá cổ phiếu qua các tháng cũng không cho thấy một quy luật chung nào: lúc tăng mạnh vào đầu năm, lúc tăng mạnh vào giữa năm, lúc tăng trưởng vào cuối năm. Điều này phản ánh tính chất biến động liên tục, khó lường trước của giá cổ phiếu.

Bên cạnh việc trực quan hóa dữ liệu giá cổ phiếu theo năm, theo tháng để tìm kiếm xu hướng dài hạn và các quy luật theo mùa vụ của giá đóng cửa cổ phiếu FPT, nhóm chúng tôi tiến hành phân rã các yếu tố Trend, Seasonal, Residual như hình 2.4 để biết được mức độ quan trọng của các yếu tố này trong giá cổ phiếu:

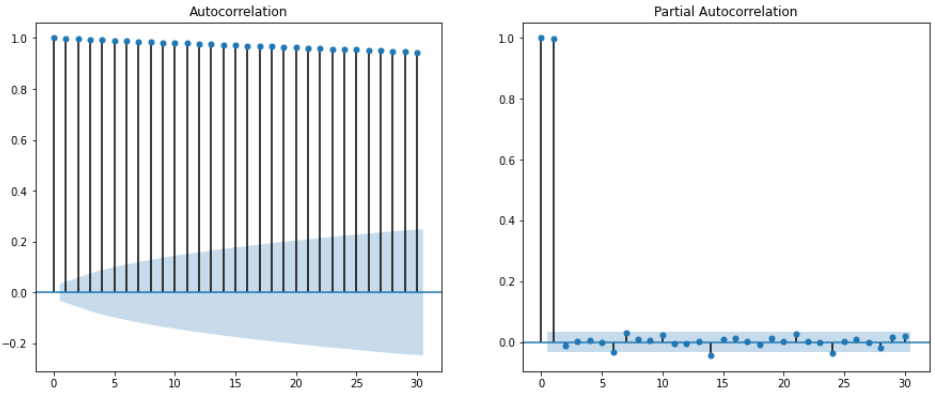


HÌNH 2.4: Biểu đồ phân rã Trend, Seasonal, Residual của FPT.

Kết quả phân rã cho thấy thành phần Seasonal trong giá là rất nhỏ, độ lớn trong khoảng -750 đến 750 đồng. Điều này cho thấy giá đóng cửa của cổ phiếu FPT bị tác động chủ yếu bởi yếu tố Trend, tiếp theo là yếu số Residual.

2.2 Tự tương quan và tự tương quan một phần

Tự tương quan (Autocorrelation function - ACF) và tự tương quan riêng phần (Partial autocorrelation function - PACF) là các khái niệm quan trọng trong chuỗi thời gian. Sự tự tương quan đo lường mối quan hệ tuyến tính giữa các giá trị trễ của một chuỗi thời gian. Quan sát biểu đồ dưới đây, ta thấy giá đóng cửa có sự tương quan rất mạnh với giá cổ phiếu của 1 ngày, 2 ngày,... , 30 ngày trước đó thể hiện ở chỉ số tự tương quan rất cao (gần bằng 1) với khoảng $lag = 1, 2, ..., 30$.



HÌNH 2.5: Biểu đồ sự tương quan giá tiền của ngày hiện tại và các ngày trước đó của FPT.

Tuy nhiên cần lưu ý là điều này không có nghĩa là giá cổ phiếu hôm nay chịu sự tác động mạnh của giá cổ phiếu 1 tháng trước đó.

Ví dụ đơn giản: khi giá cổ phiếu hôm qua ảnh hưởng tới giá cổ phiếu hôm nay, tức là giá cổ phiếu hôm kia cũng phải ảnh hưởng tới giá cổ phiếu hôm nay. Nếu tồn tại sự tương quan cao giữa giá cổ phiếu hôm nay và giá cổ phiếu hôm kia, điều này không có nghĩa là giá cổ phiếu hôm kia có thể mang lại thông tin gì giá trị đáng kể trong việc dự đoán giá cổ phiếu hôm nay, mà chỉ là vì chúng cùng có liên hệ với giá cổ phiếu hôm qua! Để giải quyết vấn đề này, ta dùng thêm chỉ số Tự tương quan riêng phần nhằm đánh giá sự tương quan giữa giá trị hôm nay với giá trị của k ngày trước đó sau khi đã loại bỏ đi sự ảnh hưởng tất cả các ngày nằm ở giữa.

Kết quả tính PACF cho thấy, giá cổ phiếu tương quan rất mạnh của giá cổ phiếu ngay ngày hôm trước, còn sự tương quan với giá cổ phiếu tại thời điểm 2,3... ngày trước đó là không đáng kể.

2.3 Tính dừng và tính không dừng

Trong phân tích dữ liệu chuỗi thời gian, một mô hình tốt được đưa ra khi phân tích trên các dữ liệu dừng (Stationary). Một chuỗi thời gian là dừng khi giá trị trung bình, phương sai, hiệp phương sai (tại các độ trễ khác nhau) không đổi cho dù chuỗi được xác định vào thời điểm nào đi nữa.

Nói cách khác, một chuỗi thời gian không dừng sẽ không có các yếu tố Trend và Seasonal. Theo Ramanathan (2002) hầu hết các chuỗi thời gian về kinh tế là không dừng vì chúng thường có một xu hướng tuyến tính hoặc mũ theo thời gian. Số liệu chứng khoán cũng không ngoại lệ. Hơn nữa, bằng các biểu đồ trực quan tại mục 2.1.2, giá đóng cửa của cổ phiếu FPT có yếu tố Trend chiếm phần lớn nên khả năng cao đây là dữ liệu không có tính dừng (non-stationary). Để khẳng định chắc chắn suy luận trên, nhóm sử dụng 2 kiểm định Unit root test phổ biến là: KPSS và ADF

để kiểm định tính dừng của dữ liệu. Kết quả của 2 kiểm định đều cho kết quả dữ liệu Không có tính dừng, như vậy cần biến đổi dữ liệu về có tính dừng.

Kiểm định	Dickey và Fuller mở rộng (ADF)	Kwiatkowski, Phillips, Schmidt and Shin (KPSS)
Giả thuyết	H0: non-stationary	H0: stationary
Đối thuyết	H1: stationary	H1: non-stationary
Kết quả	ADF Statistic: 0.763 p-value: 0.991023 >0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	KPSS Statistic: 1.9823 p-value: 0.01 <0.05 Critical Values: 1%: 0.216 2.5%: 0.175 5%: 0.146 10%: 0.119

BẢNG 2.1: Bảng kiểm định tính dừng của giá chứng khoán FPT.

Chương 3

Mô hình lựa chọn và kết quả dự đoán

3.1 Autoregressive Integrated Moving Average (ARIMA)

3.1.1 Mô hình ARIMA

ARIMA là một mô hình cơ bản trong dự đoán dữ liệu chuỗi thời gian. Tên mô hình là từ viết tắt của Autoregressive Integrated Moving Average. Mô hình đầy đủ có thể được viết là:

$$y^{'} = c + \phi_1 y^{'}_{t-1} + ... + \phi_p y^{'}_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t \tag{3.1}$$

ARIMA hoạt động dựa trên giả thuyết chuỗi dừng và phương sai sai số không đổi. Mô hình sử dụng đầu vào chính là những tín hiệu quá khứ của chuỗi được dự báo để dự báo nó. Các tín hiệu đó bao gồm: chuỗi tự hồi quy AR (Autoregression) và chuỗi trung bình trượt MA (Moving Average). Mô hình được đặc tả bởi 3 tham số ARIMA(p, d, q) là:

- p: bậc của phần Autoregression.
- d: bậc sai phân (số lần lấy sai phân).
- q: bậc của phần Moving Average

Hầu hết các chuỗi thời gian sẽ có xu hướng tăng hoặc giảm theo thời gian, do đó yếu tố chuỗi dừng thường không đạt được. Trong trường hợp chuỗi không dừng thì ta sẽ cần biến đổi sang chuỗi dừng trước khi đưa vào mô hình. Quá trình biến đổi dữ liệu trở thành có tính dừng có thể sử dụng một số cách như: transformation (ví dụ: lấy log), differencing (lấy sai phân bậc 1 hoặc bậc 2).

3.1.2 Sử dụng các tham số từ việc phân tích và khai phá dữ liệu

Để lựa chọn các tham số cho mô hình ARIMA, nhóm lần lượt xác định các bộ giá trị (p, d, q).

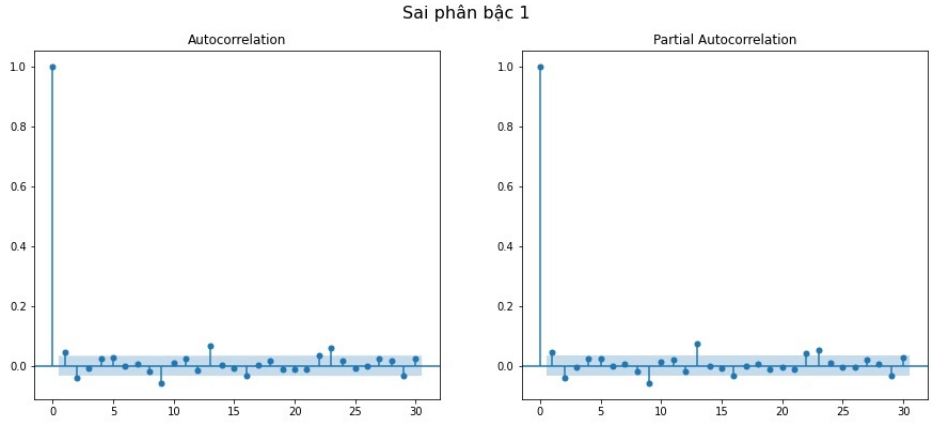
Biến đổi dữ liệu	Dùng hàm logarit	Sai phân bậc 1	Sai phân bậc 2
Kiểm tra tính dừng	ADF Statistic: -2.080686 p-value: 0.252351 >0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	ADF Statistic: -12.054887 p-value: 0.000000 <0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	ADF Statistic: -19.637767 p-value: 0.000000 <0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567
Kết luận	Dữ liệu không có tính dừng, chuyển sang dùng sai phân	Dữ liệu có tính dừng, chọn d = 1	Dữ liệu có tính dừng, chọn d = 2

BẢNG 3.1: Bảng kiểm tra tính dừng và chọn d.

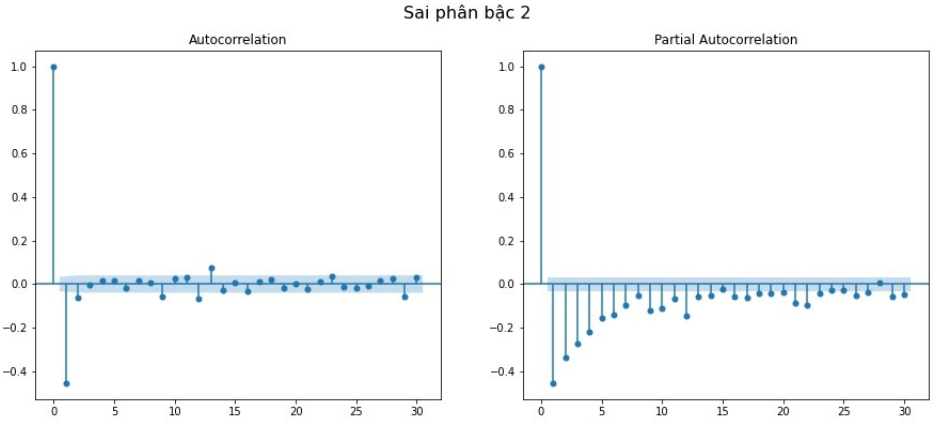
Chọn giá trị d: Vì dữ liệu ban đầu không có tính dừng, ta thử biến đổi dữ liệu bằng cách dùng logarit để kiểm tra tính dừng.

Chọn giá trị p: Từ biểu đồ tương quan riêng phần (Partial Autocorrelation) của các sai phân, ta có thể chọn giá trị bậc p của quá trình tự hồi quy AR chính là giá trị đầu tiên nằm ngoài khoảng tin cậy: p = 1.

Chọn giá trị q: Từ biểu đồ tự tương quan (Autocorrelation) của các sai phân, vì bậc q không nên quá lớn, ta chọn được hai giá trị nằm ngoài khoảng tin cậy: q = 0, q = 1.



HÌNH 3.1: Biểu đồ sai phân bậc 1.



HÌNH 3.2: Biểu đồ sai phân bậc 2.

Kết hợp các bậc của (p, d, q), ta thu được một số mô hình ARIMA phù hợp với khảo sát dữ liệu: (p, d, q) = (1, 1, 1), (1, 1, 0), (1, 2, 1), (1, 2, 0).

3.1.3 Chỉ số AIC

Một trong những tiêu chí thường được sử dụng để lựa chọn mô hình là chỉ số AIC (Akaike Information Criteria), được tính theo công thức:

$$AIC = Tlog(\frac{SSE}{T}) + 2(k+2) \tag{3.2}$$

Trong đó, T là số lượng quan sát trên tập dữ liệu, k là số lượng tham số trong mô hình. Tiêu chí thông tin này là một công cụ phạt lỗi dự báo (SSE) và phạt số lượng tham số của mô hình. Có thể nói rằng giá trị của AIC càng nhỏ thì mô hình của chúng ta càng phù hợp.

Từ các mô hình ARIMA tương ứng với các bộ tham số (p, d, q) ở trên, ta chọn được mô hình có giá trị AIC nhỏ nhất là (p, d, q) = (1, 1, 1).

3.1.4 Kết quả mô hình

MAPE trên tập kiểm tra: 3.7088%

HÌNH 3.3: Biểu đồ dự đoán giá tiền của mô hình ARIMA.

3.2 Ensemble Learning

3.2.1 Giới thiệu

Khi kết hợp nhiều mô hình độc lập lại với nhau ta được một ensemble. Một ensemble có thể đưa ra những dự đoán tốt hơn các mô hình đơn nhờ việc triệt tiêu các lỗi ngẫu nhiên (random errors), từ đó giảm thiểu sự phân tán của kết quả dự đoán, cải thiện hiệu suất chung.

3.2.2 Kết hợp bằng trung bình cộng có trọng số (weighted average ensemble)

Kết quả của phép dự báo sẽ là một trung bình cộng có trọng số từ kết quả dự báo của các mô hình trong ensemble, với wi được xác định bằng công thức: $w_i = 1 - MAPE(M_i)$, là kết quả đánh giá trên tập kiểm định (validation set) của mô hình thứ i.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.3}$$

3.2.3 Một số mô hình được đưa vào Ensemble

Theta model

Là một mô hình có hiệu suất cao, được chứng nhận qua nhiều nghiên cứu và các cuộc thi về dự báo.

Mô hình dựa trên ý tưởng thay đổi độ cong của một chuỗi thời gian thông qua hệ số θ

Ta thu được các đường Theta-lines bằng cách lấy đạo hàm cấp hai của dữ liệu, đại diện cho các tính chất dài hạn của chuỗi thời gian hoặc tính chất ngắn hạn (tùy thuộc vào giá trị). Việc phân tách thành các Theta-lines sẽ giúp đơn giản hóa quá trình dự báo.

Exponential Smoothing (Holt-Winters)

Exponential weighted average: Giá trị của một điểm cần dự đoán sẽ được tính bằng trung bình cộng có trọng số của các quan sát trước đó, các trọng số này sẽ giảm dần ngược chiều thời gian và tuân theo hàm mũ.

Forecasting equation:

$$\hat{y}_{t+h|t} = l_t \tag{3.4}$$

Smoothing equation:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1} \tag{3.5}$$

Mô hình hóa yếu tố xu hướng (trend):

Forecast equation:

$$\hat{y}_{t+h|t} = l_t + h b_t \tag{3.6}$$

Level equation:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{3.7}$$

Trend equation:

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1} \tag{3.8}$$

Mô hình hóa yếu tố mùa vụ (seasonality):

$$\hat{y}_{t+h|t} = l_t + h b_t + s_{t+h-m(k+1)} \tag{3.9}$$

$$l_t = \alpha(y_t - st - m) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{3.10}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*) b_{t-1} \tag{3.11}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma) s_{t-m} \tag{3.12}$$

3.2.4 Auto ARIMA (SARIMAX)

Mô hình ARIMAX: Là một dạng mở rộng của ARIMA. Mô hình cũng dựa trên giả định về quan hệ tuyến tính giữa giá trị trong quá khứ nhằm dự báo tương lai, nhưng có thêm yếu tố tự tương quan được biểu diễn trong phần dư của mình, và được xem như một mô hình hồi quy động.

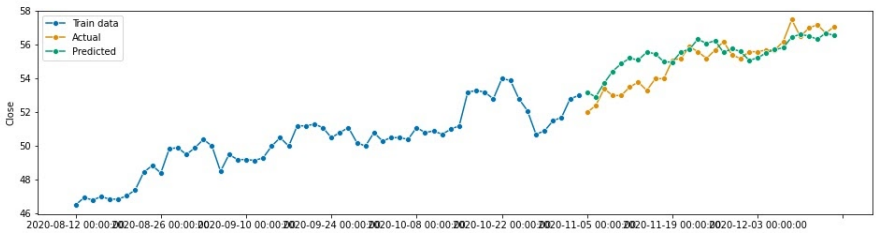
Mô hình SARIMA: Là mô hình ARIMA được điều chỉnh đặc biệt cho những chuỗi thời gian có yếu tố mùa vụ, giúp tìm ra chu kỳ và quy luật của yếu tố mùa vụ, loại bỏ nó ra khỏi chuỗi để có kết quả dự báo tốt hơn.

Các bậc của mô hình được tìm kiếm vét cạn và chọn lấy mô hình có AIC tương ứng thấp nhất.

3.2.5 Kết quả mô hình

Mô hình	Trọng số
Theta model	0.9012
Exponential Smoothing	0.7696
Auto ARIMA	0.7933

BẢNG 3.2: Trọng số của các mô hình



HÌNH 3.4: Biểu đồ dự đoán giá tiền của mô hình Ensemble.

MAPE trên tập kiểm tra: 1.3429%.

3.3 Prophet

3.3.1 Giới thiệu Prophet

Prophet là một thư viện mở do Core data science team của Facebook xây dựng cho phép các nhà kinh tế, kỹ thuật có thể phân tích, dự đoán các dữ liệu Time series ở tương lai mà không đòi hỏi người thực hiện phải có quá nhiều kiến thức sâu rộng về lĩnh vực đang dự đoán hay các kĩ thuật phân tích và xử lý dữ liệu chuyên sâu.

3.3.2 Mô hình hóa chuỗi thời gian

Dữ liệu chuỗi thời gian sẽ được Prophet mô hình thành tổng các thành phần sau:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3.13}$$

Trong đó:

- $g(t)$: hàm mô hình hóa Trend của dữ liệu theo thời gian.
- $s(t)$: hàm mô hình hóa tính seasonal của dữ liệu theo thời gian.
- $h(t)$: hàm mô hình hóa tính holiday của dữ liệu theo thời gian.
- ϵ_t: thành phần đặc trưng của dữ liệu theo thời gian.

Mô hình hóa trend của dữ liệu theo thời gian

Prophet có hai tùy chọn hàm để mô tả trend của dữ liệu. Gồm hàm Linear và Logistic. Trong đó để mô hình hóa tốt các dữ liệu có sự biến động về trend theo thời gian, Prophet thực hiện việc hiệu chỉnh hàm trend tại nhiều điểm có sự thay đổi trend lớn (được gọi là changepoints). Tại mỗi điểm này hàm trend sẽ được hiệu chỉnh lại cho

phép fit tốt với các chuỗi dữ liệu có trend biến động. Khi qua mỗi changepoints thì sẽ tiến hành hiệu chỉnh growth rate bằng cách cộng dồn thêm vào một lượng biến thiên. Ta có hàm Linear và Logistic sau khi hiệu chỉnh như sau:

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma) \tag{3.14}$$

$$g(t) = \frac{C(t)}{1 + exp(-(k + a(t)^T\delta)(t - (m + a(t)^T\gamma)))} \tag{3.15}$$

Prophet cho biết Growth rate được rút ra từ phân bố $Laplace(0, T)$. Vì vậy ta có thể hiệu chỉnh tham số T này để điều chỉnh sự fitting dữ liệu trên hàm trend. Khi T lớn, phân phối Laplace sẽ dãn rộng ra dẫn đến hệ số điều chỉnh lớn, có thể fit rất tốt với dữ liệu và có thể gây overfitting. Ngược lại khi T nhỏ có thể gây ra hiện tượng underfitting.

Mô hình hóa Seasonal của dữ liệu theo thời gian

Dữ liệu seasonal của dữ liệu sẽ được phân tách vào một hàm số theo thời gian sử dụng kiến thức về chuỗi Fourier.

$$s(t) = \sum_{t=1}^{n}(a_n cos(\frac{2\pi nt}{P}) + b_n sin(\frac{2\pi nt}{P})) \tag{3.16}$$

Trong đó P là chu kỳ của dữ liệu biến thiên theo seasonal. Seasonal có thể được tiếp tục phân tách thành tổng hoặc tích các hàm theo năm (yearly), tuần (weekly) và theo ngày (daily). Với hàm theo năm sẽ thường lấy $P = 365.25$ và theo tuần sẽ lấy $P = 7$. N là số bậc của các sòng hài, thường $N = 10$ với yearly, $N = 3$ với weekly. Với $[a_n, b_n] \approx N(0, sigma)$. Ta có thể hiệu chỉnh sigma để thay đổi biên độ của các hàm seasonal.

Mô hình hóa Holiday của dữ liệu theo thời gian

Trong đó K được rút ra từ phân phối chuẩn $N(0, \gamma)$. Nhưng trong bài toán chứng khoán, vì các ngày holiday các sàn không giao dịch nên ta sẽ bỏ qua mô hình này.

3.3.3 Hiệu chỉnh mô hình sử dụng Grid Search

Ta sẽ tiến hành hiệu chỉnh khoảng đặt các changepoints, T trong phân phối Laplace của trend, Sigma trong phân phối chuẩn của seasonal. Các thành phần này trong Prophet có các tên tương ứng sau:

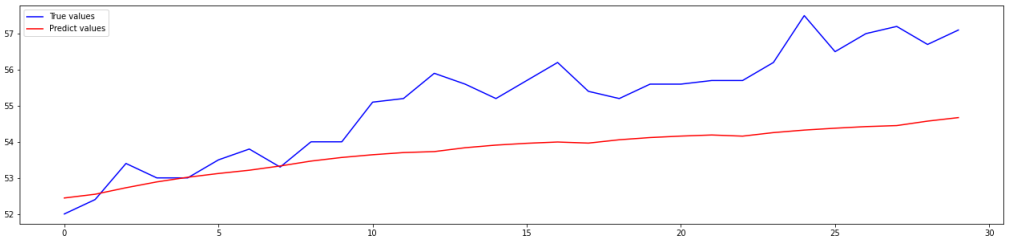
- changepoint range: khoảng hiệu chỉnh $[0.8, 0.99]$.
- changepoint prior scale: khoảng hiệu chỉnh $[0.05, 0.8]$.
- seasonality prior scale: khoảng hiệu chỉnh $[0.01, 10]$

3.3.4 Kết quả mô hình

Sau khi tinh chỉnh mô hình trên tập dữ liệu validation gồm 100 sample, ta thu được bộ tham số phù hợp như sau:

changepoint_range	changepoint_prior_scale	seasonality_prior_scale	MAPE
0.99	0.2	10	2.4516 %

BẢNG 3.3: Kết quả mô hình prophet.

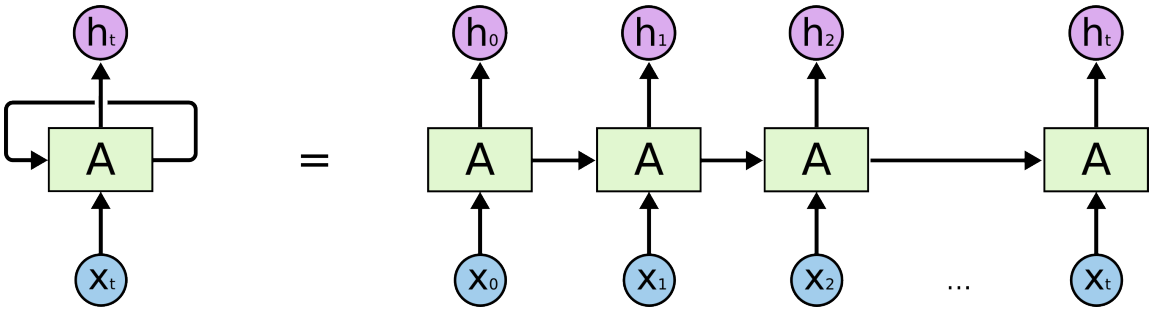


HÌNH 3.5: Kết quả forecasting sử dụng Prophet sau khi hiệu chỉnh mô hình.

3.4 Long Short Term Memory Networks (LSTM)

3.4.1 Recurrent Neural Network (RNN)

Mô hình RNN được sinh ra để giải quyết được các vấn đề sử dụng các thông tin đã được học trước đó để cho ra kết quả của mô hình. Trong mô hình mạng chứa các vòng lặp cho phép lưu trữ các thông tin



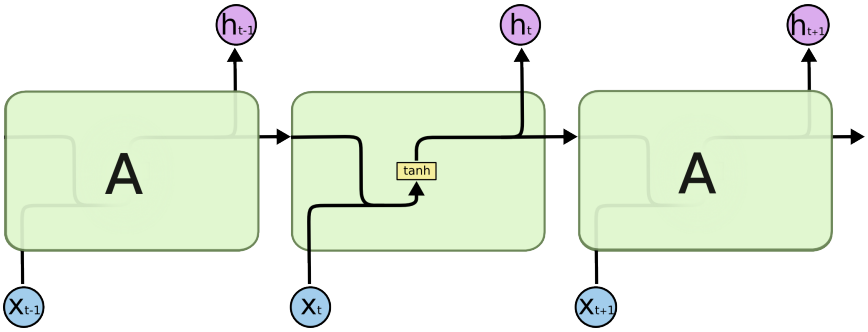
HÌNH 3.6: Mô hình RNN.

Hình vẽ trên mô tả một đoạn của mạng nơ-ron hồi quy A với đầu vào là x_t và đầu ra là h_t. Một vòng lặp cho phép thông tin có thể được truyền từ bước này qua bước này qua bước khác của mạng nơ-ron.

Tuy mô hình RNN sử dụng được các thông tin đã được học trước đó để cho ra kết quả hiện tại, nhưng nếu các thông tin đã được học quá lâu sẽ dẫn đến không nhớ được. Ví dụ: dự đoán chữ cuối cùng trong đoạn: "Tôi đã sống ở Việt Nam rất lâu ... tôi nói tiếng việt rất chuẩn.". Rõ ràng là các thông tin gần ("tôi nói ... rất chuẩn") chỉ có phép biết được đằng sau nó sẽ là tên của một ngôn ngữ nào đó, còn không thể nào biết được đó là tiếng gì. Muốn biết là tiếng gì, thì cần phải có thêm ngữ cảnh "Tôi đã sống ở Việt Nam" nữa mới có thể suy luận được. Rõ ràng là khoảng cách thông tin lúc này có thể đã khá xa rồi. Do đó mô hình RNN không thể học được

3.4.2 LSTM

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của mô hình, không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

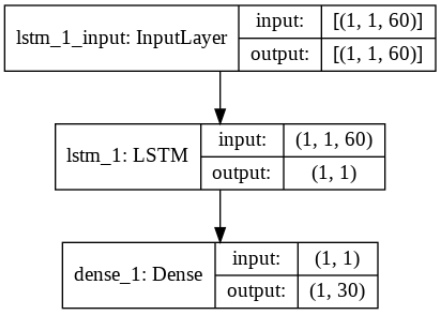


HÌNH 3.7: Mô hình LSTM.

Do đó nhóm sẽ dùng mô hình LSTM để dữ liệu đầu vào là các ngày trước đó và đầu ra sẽ là n-ngày sau các ngày trước đó.

Ví dụ: Lấy 21 ngày để dự đoán 7 ngày tiếp theo

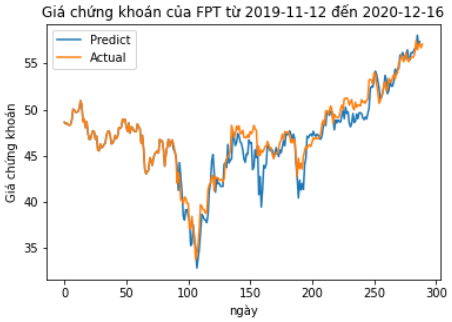
Nhóm sẽ xây dựng một mô hình LSTM 1 lớp đơn giản nhất để có thể dự đoán giá tiền chứng khoáng như hình 3.8



HÌNH 3.8: Mô hình LSTM được thiết kế.

3.4.3 Kết quả

Mô hình được đánh giá với độ đo MAPE: 8.7472%



HÌNH 3.9: Biểu đồ dự đoán của mô hình LSTM.

Chương 4

Tổng kết

4.1 Kết luận

Model	Performance (MAPE) (%)
ARIMA	3.7088
Ensemble model	1.3429
Prophet	2.4516
LSTM	8.7472

BẢNG 4.1: Bảng kết quả của từng mô hình.

Bảng 4.1 so sánh kết quả các mô hình cho thấy việc sử dụng ensemble (kết hợp Theta model, Exponential Smoothing, auto ARIMA) đang cho kết quả tốt nhất, sau đó là Prophet. Mô hình ARIMA dù khá cơ bản cũng cho kết quả ổn. Mô hình LSTM với các tham số được sử dụng đang cho kết quả thấp nhất. Điều này gợi ý cho nhóm nhiều hướng phát triển tiếp theo để cải thiện hiệu quả của các mô hình.

4.2 Hướng nghiên cứu tiếp theo

Từ kết quả của dự án, nhóm chúng tôi nhận thấy một số hướng nghiên cứu nên thực hiện tiếp theo như sau:

Các mô hình thống kê hoạt động khá tốt nhưng cần hiệu chỉnh các siêu tham số phù hợp để đạt kết quả tốt. Việc hiệu chỉnh này có thể được thực hiện bằng AutoML.

Bổ sung các dữ liệu về Khối lượng giao dịch, Lợi tức trên cổ phiếu,... để sử dụng mô hình dự báo đa biến. Mô hình nên kết hợp với các yếu tố constraints (giá giao dịch chỉ được chênh lệch 7% so với giá tham chiếu tại sàn HOSE) và loại trừ đi yếu tố lạm phát qua các năm.

Nhóm nhận thấy giai đoạn tiếp theo, bên cạnh việc dự đoán giá trị đầu ra, cần bổ sung thêm Confidence intervals (khoảng tin cậy) cho dự đoán của mình.

Tiềm năng của các mô hình Deep learning là rất cao, nhóm có thể phát triển tiếp với các mô hình thường dùng cho Sequence: Attention, Transformers hoặc Dilated CNN.

Để mô hình được có thể đóng gói thành sản phẩm và có ứng dụng cao hơn trong thực tế, nhóm mong muốn trong tương lai có thể phát triển hệ thống tự động crawl dữ liệu, mô hình hóa, và đưa ra dự báo nhanh hỗ trợ các nhà đầu tư tại Việt Nam.

4.3 Lời cảm ơn

Nhóm xin gửi lời cảm ơn tới TS. Cao Văn Lợi, TS. Nguyễn Xuân Hoài và TS. Nguyễn Quang Uy đã nhiệt tình hướng dẫn nhóm trong quá trình thực hiện dự án này.

4.4 Tham khảo

1. Hyndman, R.J., Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.
2. TS. Cao Văn Lợi (2020) Bài giảng Mô hình Time Series cơ bản và ứng dụng trong bài toán dự báo.
3. Phạm Đình Khánh (2019) Mô hình ARIMA trong time series, đường dẫn: https://phamdinhkhan
4. Francesca Lazzeri (2020) Machine Learning for Time Series Forecasting with Python
5. Taylor SJ, Letham B. (2017) Forecasting at scale
6. Hyndman, Rob J., and Baki Billah. "Unmasking the Theta method." International Journal of Forecasting 19.2 (2003): 287-290.
7. Assimakopoulos, Vassilis, and Konstantinos Nikolopoulos. "The theta model: a decomposition approach to forecasting." International journal of forecasting 16.4 (2000): 521-530.
8. Do Minh Hai, [RNN] LSTM là gì?, đường dẫn: https://dominhhai.github.io/vi/2017/10/what-is-lstm/1-m%E1%BA%A1ng-h%E1%BB%93i-quy-rnn