Dự đoán giá chứng khoán trực tuyến

Nguyễn Xuân Vĩnh Phú, Đỗ Nhật Kha, Trần Cao Khánh Ngọc,
Ngô Quang Bảo

Đại Học Công Nghệ Thông Tin,
TP. Hồ Chí Minh
{phunxv.14, khadn.14, ngoctck.14, baonq.13}@grad.uit.edu.vn

Abstract. Dự đoán cổ phiếu là một trong những bài toán quan trọng trong lĩnh vực khoa học dữ liệu giúp các công ty đưa ra chiến lược đầu tư tốt hơn. Có hai cách phân tích và đưa ra dự đoán giá cổ phiếu thường được sử dụng đó chính là phân tích cơ bản và phân tích kĩ thuật, trong đó phân tích kĩ thuật là xu hướng hiện tại mang lại hiệu quả rất lớn và được áp dụng rất rộng rãi. Hướng tiếp cận này sử dụng máy học để phân tích dữ liệu trong lịch sử và đưa ra các dự đoán trong tương lai. Trong báo cáo này, nhóm em đã đề xuất áp dụng các mô hình máy học khác nhau để dự đoán giá cổ phiếu từ dữ liệu truyền trực tuyến theo thời gian thực. Dữ liệu truyền trực tuyến là một nguồn dữ liệu quan trọng để dự đoán cổ phiếu bao gồm các luồng dữ liệu liên tục có thông tin từ nhiều nguồn khác nhau như trang web mạng xã hội, nhật ký máy chủ, ứng dụng điện thoại di động, sàn giao dịch, ... Nhóm em đã sử dụng công nghệ Spark streaming để phân tích dữ liệu trực tuyến từ sàn giao dịch S&P500 cho cổ phiếu Microsoft, và áp dụng các mô hình máy học (LSTM, GRU, ARIMA, Facebook prophet) để dự đoán giá cổ phiếu trong tương lai, các đầu ra được thể hiện qua giao diện web. Kết quả thu được là một ứng dụng web hoàn chỉnh có chức năng hỗ trợ người dùng đưa ra các quyết định hợp lý và đúng đắn nhất.

Keywords: Dự đoán giá chứng khoán, stock prediction.

1 Giới thiệu

Thị trường cổ phiếu ngày nay đã trở thành một phần quan trọng trong sự phát triển của xã hội hiện đại. Chúng phản ánh những thay đổi trên thị trường và cho phép khai thác các nguồn lực kinh tế. Với khả năng xử lý dữ liệu mạnh mẽ trên nhiều lĩnh vực, học sâu cũng được sử dụng một cách rộng rãi trong lĩnh vực tài chính như: dự đoán thị trường cổ phiếu, đầu tư tối ưu, xử lý thông tin tài chính và thực hiện các chiến lược giao dịch tài chính. Do đó, dự đoán thị trường cổ phiếu được xem là một trong những lĩnh vực khá phổ biến và quý giá nhất trong lĩnh vực tài chính và bài toán dự đoán giá chứng khoán cũng trở thành bài toán quan trọng trong xã hội.

Trong khoa học máy tính, Xử lý phân tích trực tuyến (tiếng Anh: Online Analytical Processing, viết tắt: OLAP) là một phương pháp để xử lý các truy vấn về phân tích khối lượng dữ liệu lớn, nhiều chiều mà nếu cho thực thi các truy vấn này trong hệ thống cơ sở dữ liệu thông thường sẽ không thể cho kết quả hoặc sẽ mất rất nhiều thời gian. OLAP được đặt ra để giải quyết các bài toán liên quan đến khai phá dữ liệu phục vụ cho các báo cáo về tài chính, bán hàng, tiếp thị, quản trị, dự báo.

Những công cụ sử dụng OLAP cho phép phân tích dữ liệu nhiều chiều từ các khung nhìn trực quan khác nhau do người dùng lựa chọn. Ba tác vụ phân tích cơ bản của OLAP là thống nhất hóa (roll-up), chi tiết hóa (drill-down) và xúc xắc và nhát cắt dữ liệu. Thống nhất hóa là quá trình tập hợp lại dữ liệu từ một hay nhiều chiều.

Khi nhìn vào biểu đồ giá của bất kỳ một cổ phiếu nào và tự hỏi rằng làm cách nào để tìm được điểm vào lệnh khi có những thời điểm giá đi ngang, có thời điểm giá tăng rất mạnh và cũng có lúc giá giảm tưởng như vô tận. Đó chính là lúc ta cần phải biết đến phân tích kỹ thuật chứng khoán, phương pháp giúp chúng ta xác định được xu hướng giá, thời điểm vào lệnh và thời điểm thoát khỏi thị trường để bảo toàn lợi nhuận. Vì vậy việc kết hợp phân tích dữ liệu trực tuyến để dự báo chính xác, kịp thời các thông tin và diễn biến của thị trường chứng khoán đóng vai trò quan trọng trong việc hoạch định chính sách kinh tế vĩ mô của nhà quản lý và quyết định góp vốn của các nhà đầu tư.

2 Công trình nghiên cứu liên quan

Ở bất kỳ quốc gia nào, thị trường chứng khoán cũng là một trong những thành phần quan trọng trong nền kinh tế. Do đó, việc tìm hiểu xu hướng của thị trường này là rất cần thiết. Dự đoán xu hướng, mà cụ thể là dự đoán giá cổ phiếu, đã trở thành một chủ đề thú vị, thu hút sự quan tâm của các chuyên gia, nhà nghiên cứu ở Việt Nam và trên khắp thế giới.Ở bất kỳ quốc gia nào, thị trường chứng khoán cũng là một trong những thành phần quan trọng trong nền kinh tế. Do đó, việc tìm hiểu xu hướng của thị trường này là rất cần thiết. Dự đoán xu hướng, mà cụ thể là dự đoán giá cổ phiếu, đã trở thành một chủ đề thú vị, thu hút sự quan tâm của các chuyên gia, nhà nghiên cứu ở Việt Nam và trên khắp thế giới.
Dự báo giá cổ phiếu là việc không dễ dàng vì thị trường này chịu ảnh hưởng nhiều bởi các quy luật kinh tế và các quy luật này lại khác nhau tại mỗi quốc gia, vùng miền.

2.1 Trong nước

Thị trường chứng khoán tại Việt Nam tuy chưa phát triển vượt trội so với mặt bằng chung của thế giới, tuy nhiên ta cũng có những nghiên cứu đang chú ý như:

- **Ứng dụng mạng nơ-ron nhân tạo (ANNs) trong dự báo giá đóng cửa các mã cổ phiếu niêm yết trên sàn chứng khoán** của ThS. Lê Thị Thu Giang và ThS. Vũ Thị Huyền Trang thuộc trường ĐH Thương mại: tuy có những kết quả đáng khích lệ nhưng nghiên cứu này chỉ sử dụng duy nhất dữ kiện giá đóng cửa trong quá khứ để dự báo giá trong tương lai. Do vậy mà giá trị thực tiễn của nghiên cứu vẫn chưa cao.
- **Dự đoán xu thế chỉ số chứng khoán Việt Nam sử dụng phân tích hồi quy quá trình Gauss và mô hình tự hồi quy trung bình động** của tác giả Huỳnh Quyết Thắng, Phùng Đình Vũ, Tống Văn Vinh thuộc trường ĐH Bách khoa Hà Nội: trong nghiên cứu này, tác giả đã áp dụng mô hình tự hồi quy trung bình động (ARMA: Autoregressive moving average) để dự đoán thành phần thời gian ngẫu nhiên ở một bước kế tiếp, phân tích hồi quy quá trình Gauss (GPR: Gaussian process regression) để dự đoán thành phần thời gian xu thế. Cuối cùng, kết quả dự đoán các thành phần riêng lẻ được tổng hợp lại để đưa ra kết quả dự đoán cuối cùng cho phương pháp kết hợp GPR-ARMA.

2.2 Nước ngoài

Một số công trình nghiên cứu ở nước ngoài có thể kể đến như:

- **Research on Market Stock Index Prediction Based on Network Security and Deep Learning** của tác giả Chi-Hua Chen: Trong bài báo này, tác giả sử dụng mô hình CNN để đào tạo mô hình dự đoán.
- **Deep Learning for Stock Market Prediction** của các tác giả M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana: Trong nghiên cứu này, tác giả đã sử dụng nhiều mô hình đào tạo khác nhau như ANN, RNN, LSTM,…và kết luận rằng LSTM cho ra kết quả tốt nhất.

3 Phương pháp

3.1 Thu thập dữ liệu

Dữ liệu được thu thập từ trang Yahoo Finance với mốc thời gian là từ 13.03.1986 đến 13.09.2021 với các thông tin giá cả theo từng ngày gồm các cột:

- **Date**: Ngày tháng
- **Open**: giá lúc mở cửa
- **High**: giá cao nhất trong ngày
- **Low**: giá thấp nhất trong ngày
- **Close**: giá lúc đóng cửa phiên giao dịch của ngày hôm đó
- **Adj close**: giá lúc đã được điều chỉnh
- **Volumn**: khối lượng giao dịch

3.2 Tiền xử lý dữ liệu

Đối với giá của cổ phiếu, nhóm chỉ sử dụng gía trị lúc đóng cửa của cổ phiếu để đưa vào mô hình dự đoán (cột Close). Trong đó, trước khi đưa vào mô hình dữ liệu được xử lí thông qua Min max scale từ -1 đến 1, và được chia thành 2 phần: 1 để test và 1 để train, theo tỷ lệ 8:2 nối tiếp nhau, với tập train bao gồm 7134 giá trị và tập test gồm 1784 giá trị (riêng LSTM và GRU còn 1 tham số là số bước nhìn lại (lookback) quy định lượng dữ liệu dùng cho bộ nhớ để dự đoán với giá trị là 20).

	Date	Open	High	Low	Close	Adj Close	Volume	ds	y
0	1986-03-13	0.088542	0.101563	0.088542	0.097222	0.061491	1031788800	1986-03-13	0.061491
1	1986-03-14	0.097222	0.102431	0.097222	0.100694	0.063687	308160000	1986-03-14	0.063687
2	1986-03-17	0.100694	0.103299	0.100694	0.102431	0.064785	133171200	1986-03-17	0.064785
3	1986-03-18	0.102431	0.103299	0.098958	0.099826	0.063138	67766400	1986-03-18	0.063138
4	1986-03-19	0.099826	0.100694	0.097222	0.098090	0.062040	47894400	1986-03-19	0.062040
5	1986-03-20	0.098090	0.098090	0.094618	0.095486	0.060393	58435200	1986-03-20	0.060393
6	1986-03-21	0.095486	0.097222	0.091146	0.092882	0.058746	59990400	1986-03-21	0.058746
7	1986-03-24	0.092882	0.092882	0.089410	0.090278	0.057099	65289600	1986-03-24	0.057099
8	1986-03-25	0.090278	0.092014	0.089410	0.092014	0.058197	32083200	1986-03-25	0.058197
9	1986-03-26	0.092014	0.095486	0.091146	0.094618	0.059844	22752000	1986-03-26	0.059844

Hình 3.1: Dữ liệu mẫu

	Open	High	Low	Close	Adj Close	Volume
count	8949.000000	8949.000000	8949.000000	8949.000000	8949.000000	8.949000e+03
mean	37.283290	37.677558	36.886772	37.295919	32.263043	5.915600e+07
std	50.043699	50.514658	49.569799	50.075782	50.239185	3.856527e+07
min	0.088542	0.092014	0.088542	0.090278	0.057099	2.304000e+06
25%	3.921875	3.968750	3.875000	3.921875	2.480494	3.513600e+07
50%	26.700001	26.969999	26.400000	26.687500	18.772810	5.241310e+07
75%	37.930000	38.230000	37.490002	37.889999	28.143866	7.299320e+07
max	305.019989	305.839996	302.000000	304.649994	304.649994	1.031789e+09

Hình 3.2: Tổng quan về dữ liệu

Hình 3.3: Biểu đồ thể hiện sự biến động của giá cổ phiếu

3.3 Sơ lược về các mô hình ARIMA, Prophet, LSTM, GRU

3.3.1. Mô hình dự báo ARIMA

ARIMA là phương pháp dự báo yếu tố nghiên cứu một cách độc lập (dự báo theo chuỗi thời gian). Bằng các thuật toán sử dụng độ trễ sẽ đưa ra mô hình dự báo thích hợp.

George Box và Gwilym Jenkins (1976) đã nghiên cứu mô hình ARIMA (Autoregressive Integrated Moving Average – Tự hồi qui tích hợp Trung bình trượt), và tên của họ thường được dùng để gọi tên các quá trình ARIMA tổng quát, áp dụng vào việc phân tích và dự báo các chuỗi thời gian. Phương pháp Box-Jenkins với bốn bước: nhận dạng mô hình thử nghiệm, ước lượng, kiểm định bằng chẩn đoán, và dự báo.

Có nhiều phương pháp dự báo, ví dụ PP sử dụng hồi quy bội (yêu cầu nhiều biến, nhiều dữ liệu và người nghiên cứu phải có lý thuyết tốt). Nhưng mô hình ARIMA sẽ giúp dự báo với độ tin cậy cao hơn từ các PP lập mô hình kinh tế lượng truyền thống, đặc biệt đối với dự báo ngắn hạn. Số quan sát tối thiểu để dùng được ARIMA là 50, môi trường dự báo trong tương lai ít có sự biến động. ARIMA được sử dụng khá phổ biến trong dự báo ngắn hạn, từ ARIMA có thể mở rộng PP dự báo ARCH và GARCH (các mô hình ARCH, mô hình GARCH, GARCH-M, GJR-GARCH và một số mô hình biến thể khác khi có xét tới các yếu tố rủi ro hay các cú sốc trong thị tường).

3.3.2. Mô hình Prophet

Facebook Prophet là một thuật toán mã nguồn mở để tạo ra các mô hình chuỗi thời gian sử dụng một số ý tưởng cũ với một số điểm mới. Nó đặc biệt tốt trong việc lập mô hình chuỗi thời gian có nhiều thời vụ và không gặp phải một số nhược điểm ở trên của các thuật toán khác. Cốt lõi của nó là tổng của ba hàm số của thời gian cộng với một thuật ngữ lỗi: tăng trưởng g(t), thời vụ s(t), ngày lễ h(t)và lỗi e_t

3.3.3. Mô hình LSTM

Bộ nhớ ngắn hạn dài (LSTM) là một kiến trúc mạng nơ-ron lặp lại nhân tạo (RNN) được sử dụng trong lĩnh vực học sâu. Không giống như các mạng nơron truyền thẳng tiêu chuẩn, LSTM có kết nối phản hồi. Nó không chỉ có thể xử lý các điểm dữ liệu đơn lẻ (ví dụ: hình ảnh), mà còn toàn bộ chuỗi dữ liệu (chẳng hạn như đầu vào giọng nói

hoặc video).

Các mô hình LSTM có thể lưu trữ thông tin trong một khoảng thời gian. Đặc tính này cực kỳ hữu ích khi chúng ta xử lý Chuỗi thời gian hoặc Dữ liệu tuần tự. Khi sử dụng mô hình LSTM, chúng tôi được tự do và có thể quyết định thông tin nào sẽ được lưu trữ và thông tin nào sẽ bị loại bỏ.

3.3.4. Mô hình GRU

Gated Recurrent Unit (GRU), tương tự như LSTM, là một dạng tiến hóa khác của RNN. GRU giúp giảm thiểu vấn đề gradient biến mất, với việc sử dụng các cổng, điều chỉnh luồng thông tin qua chuỗi trình tự. Việc sử dụng LSTM và GRU cho kết quả đáng kể trong các ứng dụng như nhận dạng giọng nói, tổng hợp giọng nói, hiểu ngôn ngữ tự nhiên,…

3.4 Phương pháp đánh giá

3.4.1. RMSE

Lỗi bình phương gốc (RMSE) là một trong những số liệu đánh giá phổ biến nhất cho các vấn đề hồi quy. RMSE được tính bằng cách lấy căn bậc hai của MSE:

$$RMSE = \sqrt{\frac{\sum_i^n (f_i - y_i)^2}{n}}$$

Giá trị kết quả có thể được diễn giải theo cùng đơn vị với giá trị mà chúng tôi đang cố gắng dự đoán, điều này giúp dễ hiểu hơn so với một số chỉ số khác. Tuy nhiên, điều quan trọng cần nhớ là giá trị RMSE chỉ có thể được so sánh giữa các mô hình đo lỗi sử dụng cùng một đơn vị.

RMSE cũng đại diện cho độ lệch chuẩn của các phần dư do mô hình của chúng tôi tạo ra. Vì giá trị trung bình của các phần dư luôn bằng 0, nên việc biết độ lệch chuẩn của các phần dư cho chúng ta biết các phần dư của chúng ta tập trung gần như thế nào xung quanh 0.

Giá trị RMSE thấp hơn cho thấy hiệu suất mô hình tốt hơn.

3.4.2. Thời gian thực thi

Đo và so sánh thời gian huấn luyện của các model với cùng tập dữ liệu. Giá trị của thời gian thực thi thấp cho thấy mô hình ít tốn tài nguyên vài chi phí tính toán.

4 Tổng quan hệ thống xử lý

Apache Spark Streaming là một hệ thống xử lý dữ liệu trực tuyến có khả năng xử lý luồng dữ liệu trực tiếp và mở rộng, thông lượng cao, chịu được lỗi. Spark Streaming là một phần mở rộng của Spark API cho phép người dùng xử lý dữ liệu thời gian thực từ nhiều nguồn khác nhau như Kafka, Flume và Amazon Kinesis và có thể được xử lý bằng cách sử dụng các thuật toán phức tạp được thể hiện bằng các high-level function như reduce, join and window. Cuối cùng, dữ liệu đã xử lý có thể được đẩy ra filesystems, databases và dashboard.

Hình 4.1: Mô tả hệ thống Spark streaming

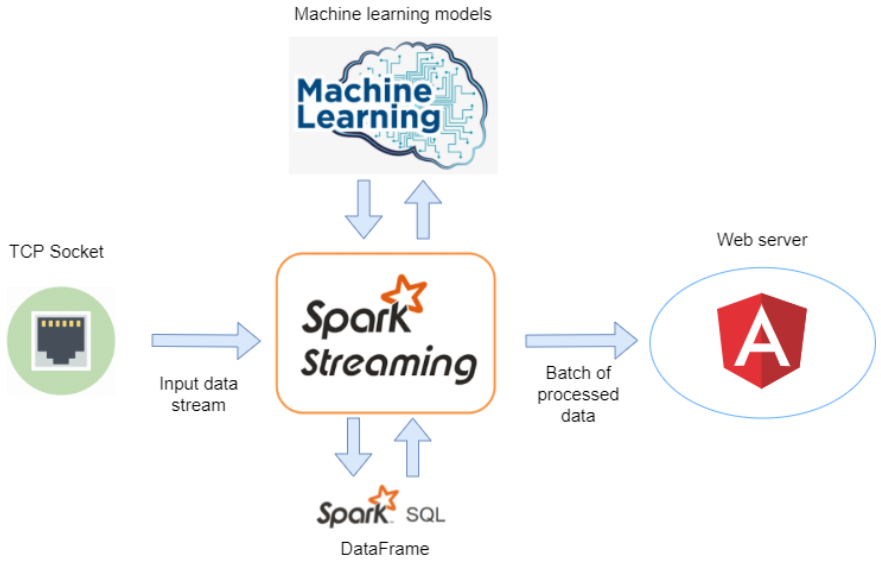
Về cách thức hoạt động, Spark Streaming nhận các luồng dữ liệu đầu vào trực tiếp và chia dữ liệu thành các batch, sau đó được xử lý bởi công cụ Spark để tạo ra stream theo các batch.

Hình 4.2: Cách thức hoạt động Spark streaming

Spark có 4 tính chất quan trọng sau:

- Phục hồi nhanh chóng từ các sự cố và lỗi
- Cân bằng tải và sử dụng tài nguyên tốt hơn
- Kết hợp dữ liệu truyền trực tuyến với tập dữ liệu tĩnh và các truy vấn tương tác
- Tích hợp gốc với các thư viện xử lý nâng cao (SQL, máy học, xử lý đồ thị)

Với khả năng xử lý dữ liệu khác nhau là lý do chính đằng sau việc Spark Streaming được áp dụng rộng rãi.



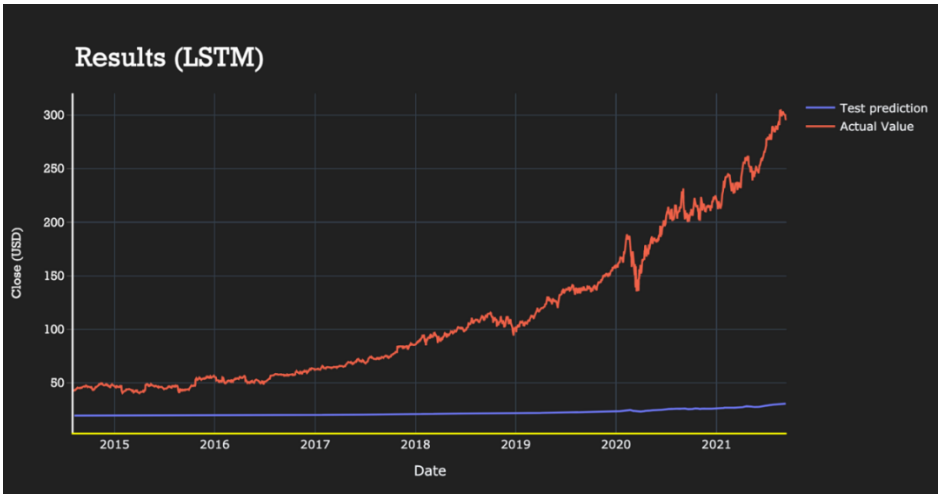
Hình 4.3: Mô hình xử lý dữ liệu

Hình trên mô tả hệ thống xử lý dữ liệu. Đầu tiên dữ liệu được lấy từ sàn giao dịch được ghi vào TCP socket. Tuy nhiên, do khó khăn trong quá trình thu thập dữ liệu nên trong báo cáo này nhóm em sẽ giả lập bằng cách đọc dữ liệu lên từ file chứa các thông số về chứng khoán và ghi lại vào socket. Tiếp theo nhóm em sẽ sử dụng Spark Streaming để đọc luồng dữ liệu từ TCP socket này, qua một số bước tiền xử lý và lưu lại vào Dataframe. Các mô hình máy học sẽ sử dụng dữ liệu này để đưa ra các dự đoán, đầu ra sẽ là giá và xu hướng tăng hoặc giảm với phiên trước đó. Kết quả dự đoán sẽ được sử dụng bởi ứng dụng web hiển thị giá của các phiên giao dịch trước đó và các dự đoán trong tương lai.

5 Thực nghiệm

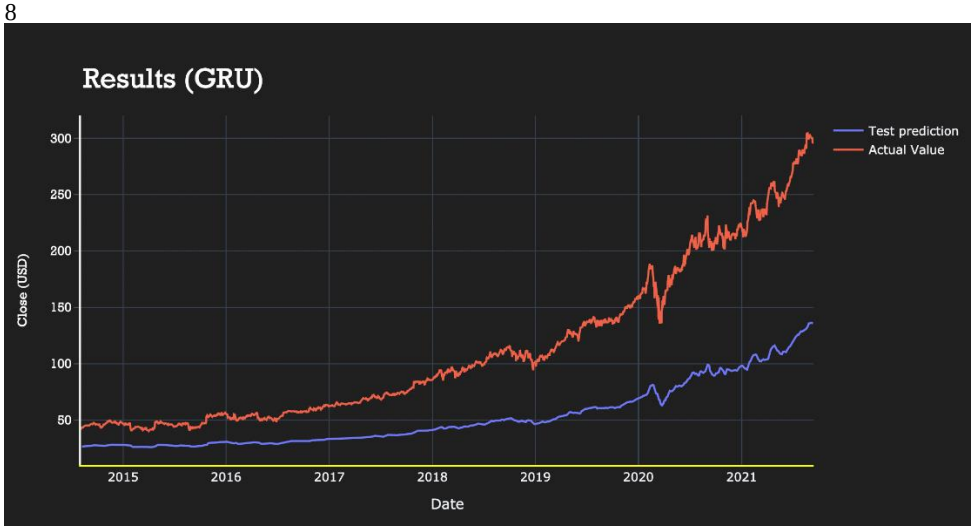
5.1 Kết quả

Sau 100 epochs cho quá trình train LSTM và GRU ta thu được:

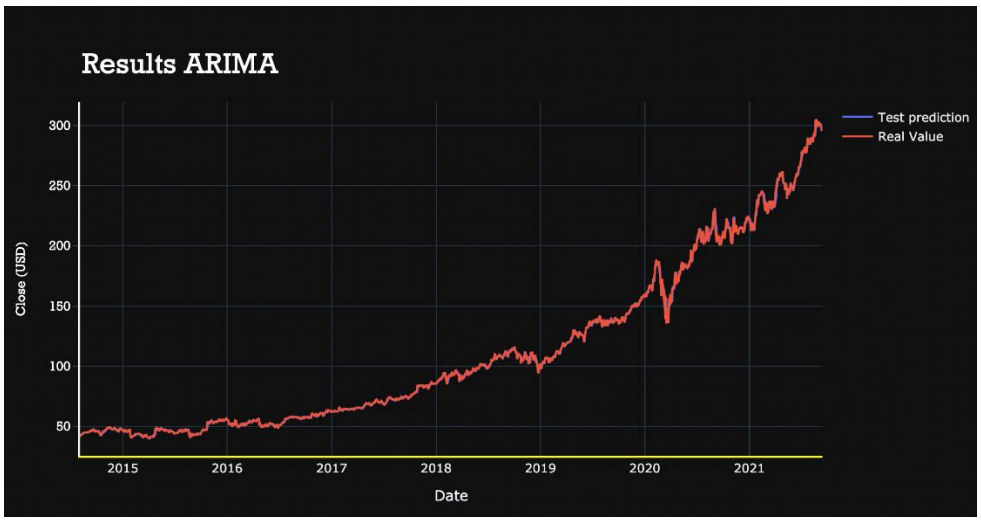


Hình 5.1: Kết quả của mô hình LSTM

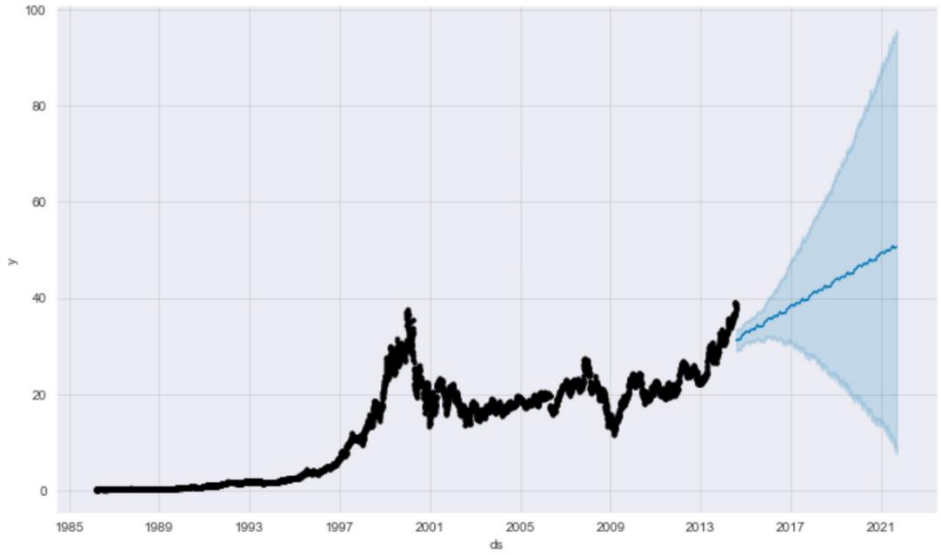
8



Hình 5.2: Kết quả của mô hình GRU



Hình 5.3: Kết quả của mô hình ARIMA



Hình 5.4: Kết quả của mô hình ARIMA

5.2 So sánh về thời gian thực thi, độ lỗi

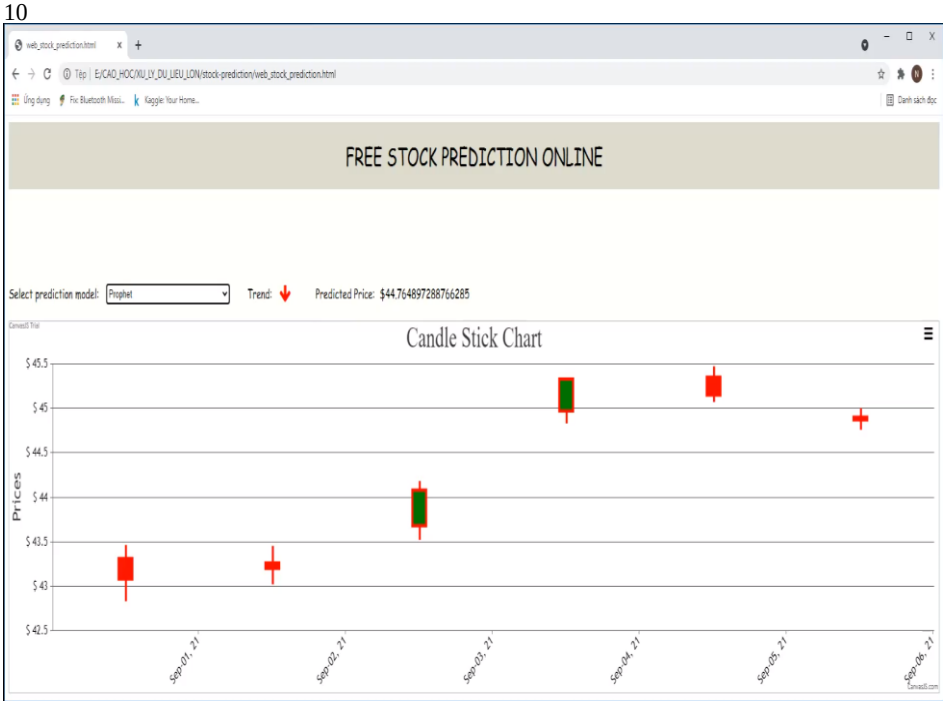
Model	Thời gian thực thi (s)	Độ lỗi (RMSE)
LSTM	59.9	110.05
GRU	44.24	12.2
ARIMA	279.77	65.3
Prophet	9.8	93.6

Hình 5.1: Bảng so sánh về thời gian thực thi, độ lỗi

5.3 Mô tả demo

Để có thể dự báo giá trị cổ phiếu một cách sinh động, nhóm đã tiến hành xây dựng một ứng dụng web đơn giản, bao gồm các thành phần như sau:

- Select prediction model list: combobox chứa các mô tên model đã được sử dụng để đào tạo mô hình.
- Trend: Xu hướng tăng (mũi tên xanh hướng lên) hoặc xu hướng giảm (mũi tên đỏ xuống dưới).
- Predicted Price: Giá cổ phiếu dự báo.
- Candle Stick Chart: Biểu đồ nến thể hiện giá trị chứng khoán thực.



Hình 5.5: Ứng dụng web dự đoán giá cổ phiếu theo các mô hình

Tại combobox Select prediction model, ta tiến hành thực hiện chọn model và khi đó, dữ liệu dự báo dựa trên model được chọn sẽ được load và hiển thị bao gồm thông tin về Trend và Prediction Price.

Tài liệu tham khảo

1. Umadevi, K. S., et al. "Analysis of stock market using streaming data framework." 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.
2. Menon, Vijay Krishna, et al. "Bulk price forecasting using spark over nse data set." International Conference on Data Mining and Big Data. Springer, Cham, 2016.
3. Behera, Ranjan Kumar, et al. "Comparative Study of Real Time Machine Learning Models for Stock Prediction through Streaming Data." J. Univers. Comput. Sci. 26.9 (2020): 1128-1147.
4. Shakva, Abin, et al. "Real-Time Stock Prediction Using Neural Network." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.
5. Javed Awan, Mazhar, et al. "Social media and stock market prediction: a big data approach." MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua 67.2 (2021): 2569-2583.
6. Peng, Zhihao. "Stocks analysis and prediction using big data analytics." 2019 international conference on intelligent transportation, big data & smart City (ICITBS). IEEE, 2019.