Dự đoán giá cổ phiếu sử dụng mô hình chuỗi thời gian với BigDL

Trịnh Ngọc Pháp[1,2,*], Trần Nguyễn Anh Khoa[1,2,*], Hà Văn Luân[1,2,*], Trần Trung Hiếu[1,2,*], Đỗ Trọng Hợp[1,2,†]

[1] Trường đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh
[2] Đại học Quốc gia Việt Nam, Thành phố Hồ Chí Minh
[*]{18521227, 18520938, 18521062, 18520754}@gm.uit.edu.vn
[†]{hopdt}@uit.edu.vn

Abstract. Stock market prediction is one of the most important and challenging issues. This is an interesting problem that has attracted the attention of researchers and investors from the past to the present time. In this study, we will build time series models based on BigDL to serve the stock price prediction problem. We approach three methods: AutoTS Pipeline (with TCN, LSTM, Seq2Seq models), Standalone Forecaster Pipeline (with LSTM, TCN, Prophet, Seq2Seq, NBeats models) and self-built model (GRU, LSTM). The methods are tested and compared and evaluated on the stock price data set of Joint Stock Commercial Bank for Foreign Trade of Vietnam (Vietcombank) collected by us, this is a stock with a series of trading days within the range broad and we will use models to predict daily closing prices. The results obtained, the AutoTS Pipeline approach with the LSTM model gives the best results with RMSE = 2.30, MAPE = 1.35. In the future, we will study more ways to improve the performance of the models, approach new methods to solve this problem, and expand the work and solve other related problems.

Keywords: Stock price prediction · Time series model · BigDL · Deep Learning · Compare approaches.

Tóm tắt nội dung Dự đoán thị trường cổ phiếu là một trong những vấn đề quan trọng và nhiều thách thức. Đây là một bài toán thú vị thu hút được sự quan tâm của các nhà nghiên cứu lẫn các nhà đầu tư từ quá khứ cho đến thời điểm hiện tại. Trong nghiên cứu này, chúng tôi sẽ xây dựng các mô hình chuỗi thời gian dựa trên nền tảng BigDL để phục vụ cho bài toán dự đoán giá cổ phiếu, chúng tôi tiếp cận theo ba phương pháp: AutoTS Pipeline (với các mô hình TCN, LSTM, Seq2Seq), Standalone Forecaster Pipeline (với các mô hình LSTM, TCN, Prophet, Seq2Seq, NBeats) và mô hình tự xây dựng (GRU, LSTM). Các phương pháp được thực nghiệm và so sánh, đánh giá trên bộ dữ liệu về giá cổ phiếu của Ngân hàng TMCP Ngoại thương Việt Nam (Vietcombank) do chúng tôi tự thu thập, đây là cổ phiếu có chuỗi ngày giao dịch trong phạm vi rộng lớn và chúng tôi sẽ sử dụng các mô hình để dự đoán giá đóng cửa hàng ngày. Kết quả thu được, phương pháp tiếp cận AutoTS Pipeline với mô hình LSTM cho kết quả tốt nhất với RMSE = 2.30,

MAPE = 1.35. Trong tương lai, chúng tôi sẽ nghiên cứu thêm cách cải thiện hiệu suất các mô hình, tiếp cận các phương pháp mới để giải quyết bài toán này đồng thời mở rộng công việc, giải quyết các bài toán có liên quan khác.

Keywords: Dự đoán giá cổ phiếu · Mô hình chuỗi thời gian · BigDL · Deep Learning · So sánh phương pháp tiếp cận.

1 Giới thiệu

Trong sự phát triển của xã hội hiện đại, thị trường cổ phiếu có một vai trò rất quan trọng. Ngày nay, tất cả các quốc gia phát triển và hầu hết các nước đang phát triển đều có thị trường chứng khoán, một thị trường không thể thiếu với mọi nền kinh tế muốn phát triển vững mạnh. Chúng cho phép triển khai các nguồn lực kinh tế. Sự thay đổi giá cổ phiếu phản ánh những thay đổi trên thị trường. 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. Qua đó, sẽ giúp các nhà quản lý doanh nghiệp, các nhà đầu tư và các cá nhân muốn tham gia thị trường chứng khoán có thông tin đầy đủ và rõ ràng hơn về giá cổ phiếu để họ có thể đưa ra quyết định tham gia thị trường chứng khoán như thế nào để có lợi nhuận cao và bền vững trong quá trình giao dịch cổ phiếu. Từ đó nâng cao lòng tin, độ tin cậy cao vào thị trường cổ phiếu, việc đem lại sự thỏa mãn tối đa cho các nhà đầu tư trên thị trường cũng là làm cho thị trường chứng khoán ngày một hiệu quả và hoạt động tốt hơn. Chính vì lẽ đó dự đoán thị trường chứng khoán là một nhu cầu cấp thiết và có ý nghĩa thực tiễn. Bài toán dự đoán giá cổ phiếu của chúng tôi với đầu vào là danh sách các dữ liệu giá trị đóng cửa của cổ phiếu trong N ngày và đầu ra chính là giá trị đóng cửa của cổ phiếu từ ngày N+1 trở đi.

Dựa trên nền tảng BigDL, chúng tôi tiếp cận ba phương pháp để giải quyết bài toán: AutoTS Pipeline (với các mô hình TCN, LSTM, Seq2Seq), Standalone Forecaster Pipeline (với các mô hình LSTM, TCN, Prophet, Seq2Seq, NBeats) và mô hình tự xây dựng (GRU, LSTM). Chúng tôi sẽ tiến hành thực nghiệm các mô hình, tinh chỉnh, so sánh và đánh giá chúng dựa trên bộ dữ liệu về giá cổ phiếu của Ngân hàng TMCP Ngoại thương Việt Nam (Vietcombank) do chúng tôi tự thu thập thông qua API của VNDIRECT. Mục tiêu của đề tài này là từ các phương pháp tiếp cận sẽ thông qua quá trình thực nghiệm, tinh chỉnh để tìm ra được phương pháp cho hiệu suất cao nhất, phù hợp để giải quyết bài toán này.

Phần còn lại của bài viết này được tổ chức như sau. Các công trình liên quan được trình bày trong Phần 2. Tiếp theo ở Phần 3, chúng tôi sẽ giới thiệu về bộ dữ liệu. Trong Phần 4, các phương pháp tiếp cận, mô hình được chúng tôi trình bày. Quá trình thực nghiệm cùng kết quả sẽ được đánh giá, phân tích ở Phần 5. Cuối cùng, Phần 6 sẽ là kết luận và hướng phát triển trong tương lai.

2 Công trình liên quan

Các bài toán dự đoán chuỗi thời gian xuất hiện từ rất sớm và nhiều phương pháp được áp dụng như các phương pháp thống kê, học máy, đặc biệt là các phương pháp học sâu phát triển mạnh mẽ gần đây. Vì vậy, chúng tôi đã tham khảo những nghiên cứu được công bố gần đây với đa phần là các mô hình áp dụng học sâu.

Năm 2019, Peter T. Yamak và các cộng sự [7] công bố nghiên cứu so sánh mức độ hiệu quả đối bài toán dự đoán chuỗi thời gian của ba mô hình là Arima, GRU và LSTM dựa trên dữ liệu giá bitcoin. Theo đó, mô hình Arima cho thấy hiệu suất tốt nhất, kế đến lần lượt là GRU và LSTM.

Năm 2020, Vinay Kumar Reddy Chimmula và LeiZhang [9] dựa trên thời điểm bùng phát đại dịch Covid-19 đã xây dựng mô hình dự đoán chuỗi thời gian để dự đoán các thời điểm bùng phát hoặc chấm dứt đợt bùng phát cùng với dự đoán số ca nhiễm vào các ngày trong tương lai. Mạng Recurrent Long short-term memory (RNN-LSTM) được các tác giả nhận xét là cho hiệu suất tốt hơn các phương pháp thống kê và dự đoán chuỗi thời gian truyền thống. Cụ thể, trên bộ dữ liệu Covid-19 của đại học John Hopkins và cơ quan y tế Canada cung cấp, mô hình LSTM cho hiệu suất 93,4%.

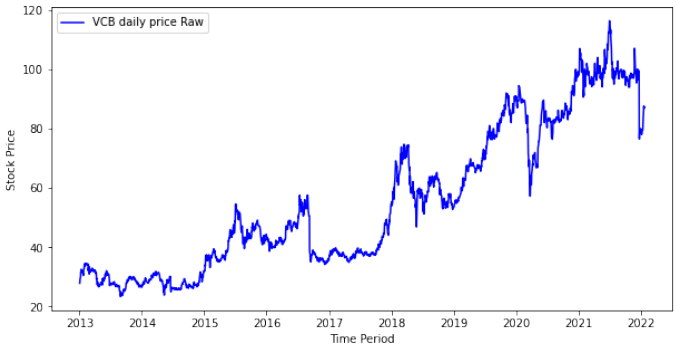
Cũng trong năm 2020, giữa bối cảnh kinh tế-tài chính thế giới biến động vì đại dịch, Ioannis E. Livieris và các cộng sự [4] công bố nghiên cứu về các mô hình dự đoán chuỗi thời gian dự đoán giá vàng. Mục tiên là dự đoán sự biến động giá vàng, xu hướng và động lực. Kết quả đạt được chỉ ra rằng mô hình CNN-LSTM cho hiệu tốt nhất đối với bài toán này.

Đến năm 2021, Dawei Cheng và cộng sự [1] đã đề xuất một mạng nơ-ron đồ thị đa phương thức (Multi-modality Graph Neural Network) để đưa ra dự đoán cho giá cổ phiếu. Dựa vào các thông tin bao gồm: chuỗi giá cổ phiếu trong lịch sử, tin tức truyền thông, sự kiện liên quan, ... họ đã xây dựng một mạng đồ thị không đồng nhất (Heterogeneous Graph Network) với các đối tượng công ty, sự kiện, tin tức dưới dạng các nút và các mối quan hệ chứa đựng kiến thức tài chính của chúng dưới dạng các cạnh. Từ đồ thị trên, khai thác các thông tin thu được để từ đó đưa ra dự đoán cho giá cổ phiếu. Mô hình đã chứng minh hiệu suất vượt trội trong việc dự đoán thị trường tài chính ở Trung Quốc. Bên cạnh đó, nó cũng hỗ trợ các nhà đầu tư trong việc đưa ra các quyết định đầu tư tiềm năng.

3 Bộ dữ liệu

Ở nghiên cứu này, chúng tôi sử dụng bộ dữ liệu về thị trường chứng khoán, cụ thể là giá cổ phiếu của Ngân hàng TMCP Ngoại thương Việt Nam (Vietcombank), là công ty lớn nhất trên thị trường chứng khoán Việt Nam tính theo vốn hóa. Chúng tôi thu thập thông tin về giá cổ phiếu VCB (mã cổ phiếu của Vietcombank) theo từng ngày với mốc thời gian từ năm 2013 cho đến thời điển hiện tại, cụ thể từ ngày 1/1/2013 đến ngày 20/1/2022. Chúng tôi thu thập dữ liệu thông qua API[3] của VNDIRECT cung cấp, dữ liệu trả về ở dạng tập tin JSON với nhiều trường dữ liệu về thông tin chứng khoán. Trong đề tài này, chúng tôi muốn dự đoán giá cổ phiểu theo từng ngày. Do đó, chúng tôi sử dụng hai trường dữ liệu bao gồm: "date" cho biết thông tin về ngày cập nhật giá chứng khoán và "close" cho biết giá đóng cửa của cổ phiếu. Cuối cùng, chúng tôi thu được bộ dữ liệu với 2,259 dòng dữ liệu tương ứng với 2,259 ngày giao dịch gồm trường "date" là đầu vào của mô hình và trường "close" là giá trị cần dự đoán.

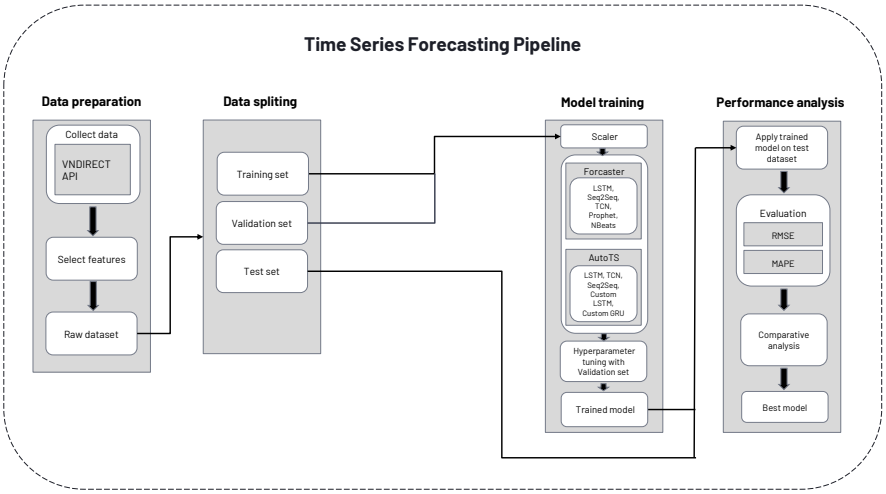
[3] VNDIRECT API https://finfo-api.vndirect.com.vn.



Hình 1. Phân phối bộ dữ liệu.

4 Phương pháp tiếp cận

Các phương pháp được chúng tôi cài đặt và thí nghiệm với framework BigDL. Trong đó, hai mô hình tự xây dựng là LSTM và GRU kết hợp với AutoTS, ngoài ra còn sử dụng các mô hình Forecaster và AutoTS có sẵn từ thư viện Chronos. Dưới đây là minh hoạ tổng quan pipeline dự báo giá cổ phiếu do chúng tôi xây dựng trong đề tài này.



Hình 2. Tổng quan pipeline cho bài toán dự báo giá cổ phiếu.

4.1 Các mô hình sử dụng

LSTM Mạng Long Short Term Memory (LSTM) được giới thiệu bởi Hochreiter & Schmidhuber (1997) [3], là một dạng đặc biệt của RNN, nó có thể giải quyết vấn đề

vanishing gradient một cách rất rõ ràng. Chìa khóa của LSTM là trạng thái tế bào (cell state), nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) nên các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi. Một mô hình LSTM bao gồm ba cổng: cổng quên, cổng đầu vào và cổng đầu ra. Cổng quên đưa ra quyết định lưu giữ / xóa thông tin hiện có, cổng đầu vào chỉ định mức độ thông tin mới sẽ được thêm vào bộ nhớ và cổng đầu ra kiểm soát xem giá trị hiện có trong ô có đóng góp vào đầu ra hay không. Mô hình LSTM hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau, trong đó có bài toán dự đoán dữ liệu chuỗi thời gian.

GRU Được đề xuất bởi Kyunghyun Cho và cộng sự vào năm 2014 [2], Gated Recurrent Unit (GRU) tương tự như LSTM nhưng được đơn giản hóa để có thể hỗ trợ tính toán và triển khai tốt hơn. Nó sử dụng các kết nối thông qua một chuỗi các nút để thực hiện các tác vụ học máy liên quan đến bộ nhớ và phân cụm. Điều này giúp điều chỉnh trọng số đầu vào của mạng nơ-ron để giải quyết vấn đề vanishing gradient vốn là một vấn đề phổ biến với mạng RNN. Thêm vào đó, GRU còn sử dụng cổng cập nhật (update gate) và cổng đặt lại (reset gate) để tinh chỉnh kết quả đầu ra bằng cách kiểm soát luồng thông tin đi qua mô hình.

Prophet Prophet [8] là một phương pháp dự đoán dữ liệu chuỗi thời gian được phát triển bởi Core Data Science team của Facebook. Prophet được xây dựng dựa trên một additive model phù hợp với các dữ liệu có tính thời vụ theo năm, tháng, tuần và hoạt động tốt đối với dữ liệu có tính thời vụ cao. Prophet cũng có khả năng xử lý tốt các trường hợp dữ liệu khuyết, thay đổi xu hướng hay các ngoại lệ. Kết quả dự đoán của mô hình đến từ việc tổng hợp các ảnh hưởng của trend, season và event. Tốc độ xử lý cao cũng là ưu điểm của mô hình này.

TCN Temporal Convolutional Networks [5] là một mạng nơ-ron sử dụng kiến trúc tích chập thay vì recurrent networks. Nó hỗ trợ các trường hợp nhiều bước và nhiều biến thể. Nhờ vào Causal Convolutions cung cấp khả năng tính toán một cách song song trên quy mô lớn, mô hình TCN cần ít thời gian tính toán, suy luận hơn so với các mô hình dựa trên recurrent neural network như LSTM. Sự cải thiện về mặt thời gian tính toán cũng như độ chính xác trên một số tác vụ, TCN đang dần nổi lên như một phương pháp được sử dụng để thay thế cho LSTM.

Seq2Seq Trong đề tài này, mô hình Seq2Seq[4] nhận vào một chuỗi dữ liệu thời gian và cho ra một chuỗi khác bằng bộ mã hóa và giải mã LSTM. Bộ mã hóa LSTM xử lý từng token trong chuỗi đầu vào, và nó cố gắng nhồi nhét toàn bộ thông tin đầu vào vào một vector có độ dài cố định, tức là "vector trung gian". Sau đó bộ mã hóa sẽ chuyển vector này sang bộ giải mã. Vector này có chức năng gói gọn toàn bộ ý nghĩa của chuỗi đầu vào và giúp bộ giải mã đưa ra được quyết định chính xác. Đây là trạng thái ẩn nằm cuối chuỗi và được tính bởi bộ mã hóa, vector này sau đó cũng hoạt động như trạng thái ẩn đầu tiên của bộ giải mã. Bộ giải mã LSTM sử dụng vector trung gian và cố gắng dự đoán chuỗi đích.

[4] Seq2Seq https://bigdl.readthedocs.io/en/latest/doc/PythonAPI/Chronos/forecasters.html#seq2seqforecaster

N-Beats N-Beats là một mô hình học sâu được Boris N. Oreshkin và các cộng sự giới thiệu năm 2020 [6] cho thấy khả năng đạt hiệu quả cao của các kiến trúc Deep Learning thuần túy trong việc dự đoán chuỗi thời gian. N-Beats là mô hình "có khả năng diễn giải" dựa trên các liên kết backforward và forward được xử lý bởi một ngăn xếp sâu gồm nhiều lớp kết nối đầy đủ với nhau (block). Mỗi block gồm các lớp được kết nối đầy đủ với nhau và chia ra hai nhánh, một nhánh cố gắng xây dựng lại chuỗi backhorizon và một nhánh sẽ cố gắng dự đoán chuỗi horizon.

4.2 BigDL

BigDL[5] là một framework học sâu phân tán được xây dựng cho nền tảng dữ liệu lớn sử dụng hệ thống Apache Spark. Nó được mô phỏng theo Torch giúp cung cấp khả năng hỗ trợ toàn diện cho các cách tiếp cận học sâu, bao gồm tính toán số (thông qua Tensor) và mạng nơ-ron cấp cao. Ngoài ra, BigDL còn sử dụng Intel MKL / Intel MKL-DNN và lập trình đa luồng trong mỗi tác vụ Spark. Do đó giúp cải thiện đáng kể thời gian tính toán, suy luận so với các nền tảng khác như Caffe, Torch hoặc TensorFlow. Mặt khác, BigDL có thể giúp mở rộng quy mô hoạt động một cách hiệu quả để thực hiện phân tích trên dữ liệu lớn bằng cách tận dụng Apache Spark. Với các ưu điểm trên, nó là một phương tiện hữu ích dùng để lưu trữ dữ liệu, xử lý và khai thác dữ liệu, triển khai các mô hình máy học truyền thống cũng như các mô hình học sâu phức tạp khác trên nền tảng dữ liệu lớn, góp phần hỗ trợ mạnh mẽ trong việc xây dựng các ứng dụng AI phân tán.

4.3 Áp dụng các mô hình

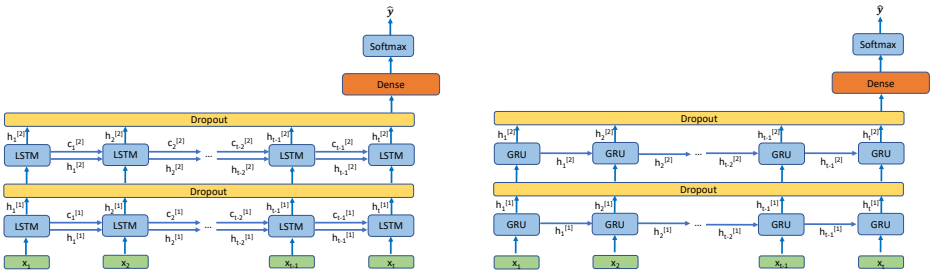
Standalone Forecaster Pipeline Chronos cung cấp các mô hình dự báo chuỗi thời gian độc lập mà không cần sự hỗ trợ của AutoML, bao gồm các mô hình học máy, học sâu và những mô hình thống kê truyền thống. Trong đề tài này, chúng tôi sử dụng các mô hình dự báo sau:

- **LSTMForecaster**: là mô hình dự báo dựa trên mô hình LSTM truyền thống, thích hợp cho các bài toán dự báo chuỗi thời gian đơn biến.
- **Seq2SeqForecaster**: là mô hình từ chuỗi sang chuỗi (sequence to sequence model) dự trên mô hình LSTM, thích hợp cho các bài toán dự báo chuỗi thời gian đơn biến và đa biến.
- **TCNForecaster**: là mô hình dự báo dựa trên mạng nơ-ron tích chập.
- **ProphetForecaster**: là mô hình dự báo dựa trên mô hình Prophet.
- **NBeatsForecaster**: là mô hình báo dựa trên mô hình N-Beats.

AutoTS Pipeline AutoTSEstimator (Automated TimeSeries Estimator) là một mô hình tự động cho phép người dùng huấn luyện các mô hình dự báo chuỗi thời gian một cách dễ dàng với ba mô hình được xây dựng sẵn (LSTM, TCN và Seq2Seq) và có thể khai báo các thông số để mô hình tự động tìm ra bộ tham số tốt nhất. Trong đề tài này, chúng tôi sử dụng cả ba mô hình được xây dựng sẵn trong AutoTSEstimator là LSTM, TCN và Seq2Seq.

[5] BigDL https://bigdl.readthedocs.io/en/latest/

AutoTS với mô hình tùy chỉnh Ngoài ba mô hình được xây dựng sẵn, AutoTSEstimator còn cho phép người dùng truyền vào một mô hình tự xây dựng tùy ý. Ở thời điểm hiện tại, AutoTSEstimator chỉ cho phép xây dựng mô hình tùy chỉnh dựa vào thư viện Pytorch, chúng tôi đã xây dựng thêm hai mô hình LSTM và GRU và truyền vào AutoTSEstimator để tìm ra kiến trúc mô hình và bộ tham số tốt nhất đối với từng loại mô hình mà chúng tôi xây dựng. Dưới đây là kiến trúc hai mô hình do chúng tôi xây dựng.



Hình 3. Kiến trúc mô hình LSTM và GRU.

5 Thực nghiệm và đánh giá

5.1 Xử lý dữ liệu

Chúng tôi tiến hành phân chia bộ dữ liệu gốc gồm 2,259 dòng thành ba tập: tập huấn luyện (tập training), tập kiểm thử (tập validation) và tập đánh giá (tập test) với tỉ lệ 8:1:1. Tập huấn luyện dùng để huấn luyện các mô hình, tập kiểm thử dùng để tinh chỉnh tham số mô hình nhằm tìm ra bộ tham số tốt nhất và tập kiểm tra dùng để đánh giá hiệu suất dự báo của mô hình tốt nhất. Chúng tôi sử dụng chuẩn hóa Min-Max và chuẩn hóa Chuẩn để chuẩn hóa giá trị trường "close" làm đầu vào để huấn luyện các mô hình.

5.2 Độ đo đánh giá

Chúng tôi sử dụng hai độ đo chính là RMSE và MAPE để giá hiệu suất dự báo giá cổ phiếu theo thời gian của các mô hình.

RMSE Độ đo RMSE (Root Mean Square Error) là độ lệch chuẩn của phần dư (lỗi dự đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy. RMSE là thước đo mức độ lan truyền của những phần dư này, thể hiện mức độ tập trung của dữ liệu xung quanh dòng phù hợp nhất, đo lường sự khác biệt giữa các giá trị dự đoán và giá trị thực tế . RMSE là thước đo mức độ hiệu quả của các mô hình tuyến tính

RMSE càng nhỏ tức là sai số càng bé đồng nghĩa với độ tin cậy của mô hình càng cao. RMSE được tính theo công thức:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\overline{y_i} - y_i)^2}{n}}$$

Trong đó n là tổng số lượng quan sát; $\overline{y_i}$ là giá trị dự đoán; y_i là giá trị thực tế.

MAPE Độ đo MAPE (Mean Absolute Percentage Error) là phần trăm sai số trung bình tuyệt đối. MAPE là độ đo thống kê mức độ chính xác của một hệ thống dự báo dưới dạng tỉ lệ phần trăm. MAPE càng bé thì mô hình có độ chính xác càng cao. MAPE được tính theo công thức:

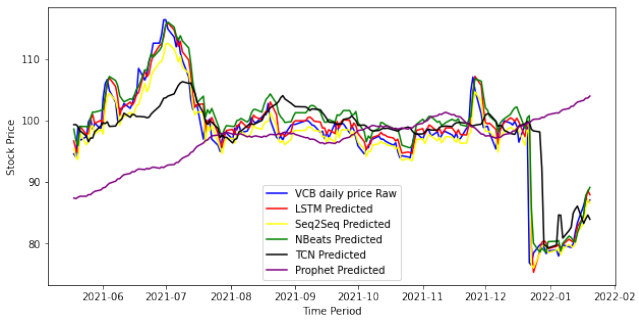
$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \overline{y_i}}{y_i} \right|$$

Trong đó n là tổng số lượng quan sát; $\overline{y_i}$ là giá trị dự đoán; y_i là giá trị thực tế.

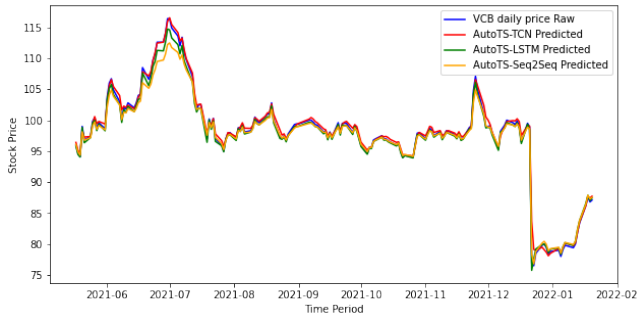
5.3 Kết quả

Bảng 1. Độ đo đánh giá kết quả dự đoán tập kiểm tra.

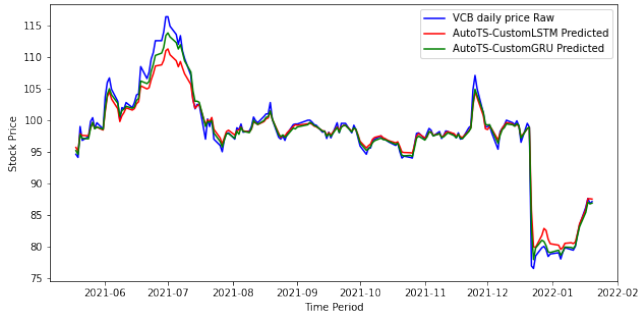
Model	RMSE	MAPE
LSTMForcaster	2.40	1.47
Seq2SeqForecaster	2.52	1.60
NBeatsForecaster	3.11	2.31
ProphetForecaster	10.60	7.82
TCNForecaster	5.06	3.46
AutoTS-TCN	2.80	1.96
AutoTS-LSTM	**2.30**	**1.35**
AutoTS-Seq2Seq	2.39	1.40
AutoTS-CustomLSTM	2.63	1.67
AutoTS-CustomGRU	2.46	1.48



Hình 4. Trực quan chuỗi giá trị dự đoán của các mô hình Forecaster và chuỗi giá trị của tập test.

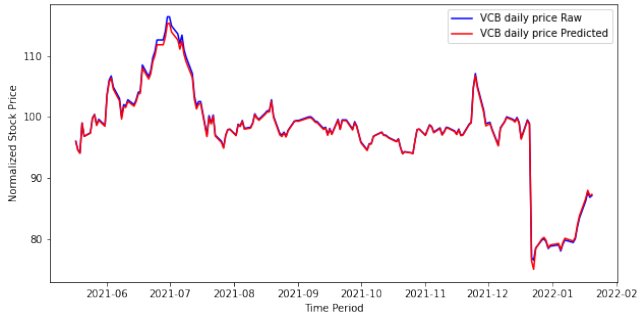


Hình 5. Trực quan chuỗi giá trị dự đoán của các mô hình AutoTS và chuỗi giá trị của tập test.



Hình 6. Trực quan chuỗi giá trị dự đoán của các mô hình AutoTS tuỳ chỉnh và chuỗi giá trị của tập test.

Trong ba phương pháp tiếp cận thì phương pháp AutoTS Pipeline với mô hình LSTM cho kết quả dự đoán tốt nhất, kết quả dự đoán được thể hiện ở hình 7.



Hình 7. Trực quan chuỗi giá trị dự đoán của mô hình AutoTS-LSTM và chuỗi giá trị của tập test.

5.4 Đánh giá

Hầu hết các mô hình đều cho kết quả dự báo rất tốt, đặt biệt là các mô hình sử dụng AutoTS. Mô hình AutoTS-LSTM cho kết quả tốt nhất ở cả hai độ đo, điều đó chứng tỏ việc kết hợp một mô hình hiệu quả trên bài toán dự báo chuỗi thời gian là LSTM và mô hình tự động AutoTS sẽ cho ra một mô hình với hiệu suất cao nhất. Điều này cũng tương ứng với kết quả của các công trình liên quan đã nêu.

Các mô hình ProphetForecaster và TCNForecaster cho kết quả thấp nhất. Bộ dữ liệu không có tính thời vụ cao là nguyên nhân quan trọng khiến mô hình Prophet không hiệu quả và hạn chế ở khả năng tinh chỉnh tham số là nguyên nhân khiến mô hình TCNForecaster cho hiệu suất thấp.

6 Kết luận và hướng phát triển

Qua đề tài này, chúng tôi đã cài đặt thành công các mô hình dự đoán dữ liệu chuỗi thời gian với BigDL để giải quyết bài toán dự đoán giá chứng khoán. Trong đó, mô hình AutoTS-LSTM đạt được hiệu suất cao nhất với RMSE và MAPE lần lượt là 2.30 và 1.35. Qua đó cho thấy các phương pháp tiếp cận của chúng tôi mang lại kết quả rất khả quan và có khả năng để ứng dụng vào lĩnh vực tài chính ngoài thực tế.

Trong tương lai, chúng tôi sẽ tiếp tục phát triển các mô hình bằng cách kết hợp các nguồn thông tin khác liên quan như: tình hình tài chính công ty, tin tức liên quan, sự kiện xảy ra, nhằm thu được nhiều giá trị hữu ích hơn để cải thiện hiệu suất cho quá trình đưa ra dự đoán.

Tài liệu

1. Cheng, D., Yang, F., Xiang, S., Liu, J.: Financial time series forecasting with multi-modality graph neural network. Pattern Recognition **121**, 108218 (2022)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
4. Ioannis E. Livieris, Emmanuel Pintelas, P.P.: A cnn–lstm model for gold price time-series forecasting. Neural Computing and Applications **32**, 17351–17360 (2020)
5. Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: A unified approach to action segmentation. In: European Conference on Computer Vision. pp. 47–54. Springer (2016)
6. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-beats: Neural basis expansion analysis for interpretable time series forecasting (2020)
7. Peter T. Yamak, Li Yujian, P.K.G.: A comparison between arima, lstm, and gru for time series forecasting. In: International Conference on Algorithms, Computing and Artificial Intelligence. p. 49–55. ACAI (2020)
8. Taylor, S.J., Letham, B.: Forecasting at scale. PeerJ Preprints **5**, e3190v2 (Sep 2017). https://doi.org/10.7287/peerj.preprints.3190v2, https://doi.org/10.7287/peerj.preprints.3190v2
9. Vinay Kumar ReddyChimmula, L.: Time series forecasting of covid-19 transmission in canada using lstm networks. Chaos, Solitons Fractals **135**, 109864 (2020)