Sử dụng time series forecast dự đoán giá cổ phiếu

Trần Thị Mỹ Nhung   
IS304.N21.HTTT  
Đại học Công nghệ   
Thông tin  
[20520267@gm.uit.edu.vn](mailto:20520267@gm.uit.edu.vn)

Nguyễn Anh Kiệt   
IS304.N21.HTTT  
Đại học Công nghệ  
 Thông tin  
[20521498@gm.uit.edu.vn](mailto:20521498@gm.uit.edu.vn)

Nguyễn Thành Phát   
IS304.N21.HTTT  
Đại học Công nghệ  
 Thông tin[20520270@gm.uit.edu.vn](mailto:20520270@gm.uit.edu.vn)

Nguyễn Hoài Linh  
IS304.N21.HTTT  
Đại học Công nghệ   
Thông tin20521534@gm.uit.edu.vn

Nguyễn Đạt  
IS304.N21.HTTT  
Đại học Công nghệ  
 Thông tin20520434@gm.uit.edu.vn

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction

Cổ phiếu được coi là hình thức đầu tư trọng điểm trong ngành tài chính, đại diện cho quyền sở hữu một phần trong tổ chức phát hành. Sự kiện đáng chú ý trong lịch sử là việc Công ty Đông Ấn Hà Lan phát hành cổ phiếu đầu tiên vào năm 1602 tại Sở Giao dịch Chứng khoán Amsterdam, đồng thời cũng là công ty đầu tiên phát hành cổ phiếu và trái phiếu.[STT]

Định giá cổ phiếu là quy trình xác định giá trị thị trường thực sự của cổ phiếu tại một thời điểm nhất định, nhằm hiểu rõ tiềm năng của cổ phiếu để đưa ra quyết định đầu tư phù hợp. Đối với doanh nghiệp, việc định giá cổ phiếu được coi là một trong những bước tiên quyết khi công ty cổ phần dự định phát hành cổ phiếu, huy động vốn và tăng cường ảnh hưởng của mình trên thị trường. Từ góc độ của nhà đầu tư, việc định giá cổ phiếu giúp họ xác định cổ phiếu nào đáng để đầu tư và có tiềm năng mang lại lợi nhuận tối đa.

Một phương pháp tiếp cận sơ bộ trong việc định giá cổ phiếu là đánh giá giá trị cổ phiếu. Nếu giá cổ phiếu hiện tại thấp hơn giá trị đã định giá, nhà đầu tư có thể xem xét mua cổ phiếu. Ngược lại, nếu giá cổ phiếu vượt quá giá trị đã định giá và nhà đầu tư hiện đang sở hữu cổ phiếu, họ có thể bán cổ phiếu để thu về lợi nhuận.

Thực tế cho thấy có nhiều thuật toán và kỹ thuật hỗ trợ việc dự báo giá cổ phiếu. Trong nghiên cứu này, chúng tôi sẽ áp dụng 12 mô hình: GRU, ARIMA, LSTM, Linear Regression, ES, KF, GBT, DNN, XGBoost, DNN, TCA, RNN để dự báo giá cổ phiếu của một số công ty trong 30 ngày tiếp theo. Sau cùng, sẽ tiến hành đánh giá và so sánh giữa các mô hình để xem mô hình nào hiệu quả nhất.

# Related Work

## Giới thiệu chung

## Giới thiệu từng thuật toán

Nhóm tác giả Vaishnavi Gururaj, Shriya V R và Dr. Ashwini K [STT] đã nghiên cứu về thị trường chứng khoán bằng mô hình Linear Regression. Dataset mà nhóm tác giả sử dụng là một năm dữ liệu cổ phiếu của Công ty Coca-Cola, từ tháng 1 năm 2017 đến năm 2018. Các kết quả về độ đo bao gồm: 3.22 (RMSE), 2.53 (MAE), 10.37 (MSE) và 0.73 (R-Squared).

[GRU]

Box và Jenkins đã giới thiệu mô hình ARIMA vào năm 1970. Đây còn được gọi là phương pháp Box-Jenkins, bao gồm một tập hợp các hoạt động để xác định, ước lượng và chẩn đoán các mô hình ARIMA với dữ liệu chuỗi thời gian[STT]. Các mô hình ARIMA đã chứng minh khả năng tạo ra dự đoán ngắn hạn hiệu quả. ARIMA liên tục vượt trội so với các mô hình phức tạp khác trong dự đoán ngắn hạn [STT].

Hai tác giả F. Sidqi và I. D. Sumitra [STT] đã nghiên cứu và áp dụng hai phương pháp Single Exponential Smoothing (SES) và Double Exponential Smoothing (DES) để dự báo doanh số bán hàng sản phẩm. Kết quả cho thấy MAPE của SES là 20% và MAPE của DES khoảng 24%. Phương pháp Single Exponential Smoothing có sai số thấp hơn.

[LSTM]

[RNN]

[GBT]

[HMM]

[XGBoost]

[DFNN]

Y. Lin và các cộng sự [STT] đã nghiên cứu và áp dụng mô hình Temporal Convolutional Attention Neural Networks cho dự báo năng lượng mặt trời. Kết quả cho thấy TCAN cho kết quả tốt trên cả dự báo điểm (0.5) và dự báo xác suất (0.9). Các kết quả 0.5/0.9 bao gồm: 0.062/0.031 (Sanyo), 0.068/0.035 (Hanery) và 0.209/0.081 (Solar).

Identify applicable funding agency here. If none, delete this text box.

# Materials

## Giới thiệu về Dataset

Bài báo sử dụng dataset được lấy từ dữ liệu chứng khoán của 3 công ty *CTLT* (Dữ liệu của công ty Catalent được lấy về từ trang web Investing.com có…), *INTC* (Dữ liệu của công ty Intel Corporation được lấy về từ trang web Investing.com có… ) và *NTR*(Dữ liệu của công ty Nutrien được lấy về từ trang web Investing.com có…). Ba tập dữ liệu này được lấy từ mốc thời gian là từ ngày 1/1/2019 đến ngày 20/6/2023 và được tải về vào ngày 20/6/2023. Ở cả 3 dataset đều có các cột thuộc tính liên quan đến vấn đề chứng khoán như Date, Open, High, Low, Close, Adj Close và Volume.

## Giới thiệu về các công cụ sử dụng

*Google Colab:* Colaborator hay còn gọi là Google Colab, là một sản phẩm từ Google Research, nó cho phép chạy các dòng lệnh code python qua trình duyệt. Có thể sử dụng tài nguyên máy tính để chạy nhanh hơn như CPU tốc độ cao, GPUs, TPUs [..]. Nhóm đã sử dụng Google Colab để hỗ trợ chạy các mô hình cần đến sử dụng GPUs để có thể tăng tốc độ chạy. *Visual Studio Code:* Visual Studio Code (VS Code) là một trình soạn thảo mã nguồn mở gọn nhẹ nhưng có khả năng vận hành mạnh mẽ trên 3 nền tảng là Windows, Linux và macOS được phát triển bởi Microsoft [..]. Nhóm đã sử dụng VS Code để dễ dàng trong việc sửa đổi code bởi sự linh hoạt của nó. *Các thư viện sử dụng:* numpy, pandas, matplotlib, scikit-learn, keras, minmaxscaler,…

# Method

[Giới thiệu]

## Linear Regression

Trong lĩnh vực thống kê, Linear Regression là một phương pháp dùng để xác định mối quan hệ giữa một biến phụ thuộc có giá trị số và một hoặc nhiều biến độc lập.

Khi chỉ có một biến độc lập, chúng ta gọi là hồi quy tuyến tính đơn giản (Simple Linear Regression). Trong trường hợp có nhiều biến độc lập, ta gọi là hồi quy tuyến tính đa biến (Multiple Linear Legression) [STT].

Simple Linear Regression được mô tả qua công thức:

Trong đó :

: biến phụ thuộc (dependent variable) cần dự đoán.

: biến độc lập (independent variable) được sử dụng để dự đoán giá trị của .

: hệ số góc (intercept) của đường hồi quy, đại diện cho giá trị dự đoán của khi .

: hệ số hồi quy (regression coefficient), đại diện cho mức độ thay đổi của dựa trên mỗi đơn vị thay đổi của .

: lỗi ngẫu nhiên (random error), biểu thị sự không thể tránh khỏi của mô hình trong việc mô phỏng dữ liệu thực tế.

## GRU

## ARIMA

Trong một mô hình ARIMA, giá trị tương lai của một biến được giả định là một hàm tuyến tính của một số quan sát quá khứ cộng với các lỗi ngẫu nhiên. Hàm tuyến tính này dựa trên ba thành phần tham số: tự hồi quy (AR), tích hợp sai phân (I) và trung bình trượt (MA) [STT]. Mô hình ARIMA có thể được ký hiệu là ARIMA(p, d, q), trong đó p là số lượng thành phần tự hồi quy, d là số lượng sự khác biệt không mùa và q là số lượng lỗi dự báo trễ trong phương trình dự đoán [STT]. Giá trị tương lai của một biến trong ARIMA được biểu thị như sau:

Trong đó:

: Giá trị thực tế.

: Lỗi ngẫu nhiên tại thời điểm t.

và : Là các hệ số.

## ETS

Exponential Smoothing (ETS) là một phương pháp dự báo trong phân tích chuỗi thời gian, nơi các chuỗi thời gian quan sát được gán trọng số không đều [STT].

Simple Exponential Smoothing (SES) gán các trọng số giảm dần theo thời gian. Cụ thể, phương pháp này tính toán dự báo bằng cách lấy trọng số lớn nhất cho quan sát gần nhất và trọng số nhỏ hơn cho các quan sát xa hơn.

Công thức cập nhật dự đoán của phương pháp SES như sau:

trong đó, là giá trị dự đoán tại thời điểm t, là giá trị quan sát ở thời điểm t, là giá trị dự đoán tại thời điểm t – 1, là một tham số trơn (smoothing parameter) nằm trong khoảng từ 0 đến 1 và nó quyết định mức độ ảnh hưởng của mỗi quan sát.

## LSTM

LSTM là viết tắt của "Long Short-Term Memory", một loại mạng nơ-ron học sâu hay còn được biến đến là một loại đặc biệt của mạng Recurrent Neural Network (RNN)[..].

Một LSTM layer bao gồm một tập hợp các khối nhớ được kết nối theo chu kỳ. Mỗi khối chứa một hoặc nhiều ô nhớ được kết nối theo chu kỳ thông qua ba cổng nhân tích - cổng đầu vào, cổng đầu ra và cổng quên. Chúng cung cấp các phép ghi, đọc và đặt lại liên tục cho ô nhớ.

Sự ra đời của LSTM đã giúp hạn chế phần nào vấn đề phụ thuộc xa mà RNN mắc phải nhờ khả năng học các phụ thuộc dài hạn.

Cấu trúc của LSTM:

A picture containing screenshot, diagram

Description automatically generated

Hình: Kiến trúc LSTM

Công thức tính toán các cổng trong LSTM:

*Forget gate:*

*Input gate:*

*Cell gate:*

*Ouput gate:*

*Cell state:*

Trong đó:

## RNN

Recurrent Neural Network (RNN) là một mô hình mạng neural hồi quy gồm 3 thành phần chính là Input layer, Hidden layer và Output layer.

Mô hình tổng hợp và lan truyền thông tin theo trình tự xâu chuỗi. Mô hình được gọi là hồi quy vì thực hiện tính toán ở thời điểm hiện tại sẽ phụ thuộc vào các kết quả tính toán ở những thời điểm trước đó. Do đó, RNN phải nhớ thông tin đã được đưa vào để tính toán trước đó để phục vụ cho các bước tính toán sau này.

Sau nhiều bước biến đổi, các thông tin ở những bước đầu tiên bị biến đổi nhiều lần dẫn đến bị triệt tiêu. Do đó, RNN chỉ có nhớ được một vài bước trước đó để mô hình được tối ưu nhất.

Cấu trúc của RNN:

A picture containing screenshot, diagram, text, line

Description automatically generated

RNN được biểu diễn bằng công thức sau:

và

Trong đó:

: giá trị đầu vào tại thời điểm t

: giá trị đầu ra tại thời điểm t

: giá trị kích hoạt

: các ma trận trọng số

: vector độ lệch

: các hàm kích hoạt

## GBT

Grandient Boosted Tree (GBT) là một thuật toán học máy tiên tiến, được sử dụng trong các tác vụ hồi quy và phân loại cùng với các tác với nguồn gốc từ khái niệm gradient boosting. Gradient boosting được diễn giải như một thuật toán tối ưu hóa dựa trên một hàm chi phí phù hợp, dựa trên quan sát của Leo Breiman[STT]. Cụ thể hơn, các thuật toán Gradient Boosting đã được phát triển dưới góc nhìn hạ gradient trong không gian hàm số - nghĩa là các thuật toán tối ưu hóa một hàm chi phí qua không gian hàm số bằng cách lựa chọn lặp đi lặp lại một hàm (giả thuyết yếu) chỉ vào hướng gradient âm. Phương pháp này đã được Jerome H. Friedman áp dụng cho việc phát triển Gradient Boosted Trees​.

GBT hoạt động dựa trên ba thành phần chính

1. Hàm mất mát: Hàm mất mát được sử dụng trong GBT phụ thuộc vào loại bài toán được giải quyết. Hàm mất mát phải là một hàm có thể vi phân.
2. Học viên yếu: GBT sử dụng cây quyết định như là học viên yếu. Cụ thể hơn, GBT sử dụng cây hồi quy, mà đầu ra của chúng có thể được cộng lại, cho phép đầu ra của các mô hình sau có thể cải thiện sai số của các dự đoán hiện tại.
3. Mô hình phụ thuộc: Trong GBT, các cây được thêm vào một cách tuần tự, và các cây đã tồn tại trong mô hình không được thay đổi. Quá trình hạ greadient được sử dùng

## HMM

Hidden Markov Model (HMM) là một mô hình thống kê được xử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên, nhận dạng giọng nói, phân tích chuỗi DNA, xử lý ảnh. Trong đó, mô hình cho phép dự đoán trạng thái ẩn dựa trên các quan sát được từ thực tế.

Mô hình gồm 5 thành phần chính: số lượng trạng thái ẩn (N), số lượng quan sát đối với mỗi trạng thái ẩn (M), ma trận sác xuất chuyển đổi trạng thái (A), ma trận sác xuất quan sát (B) và phân phối xác suất ban đầu của các trạng thái ẩn ().

Để phù hợp với bài toán về dữ liệu có tính liên tục, mô hình được mở rộng thành HMM liên tục (Continuous HMM – CHMM) và thuộc vào mô hình HMM liên tục bậc 2.

## XGBoost

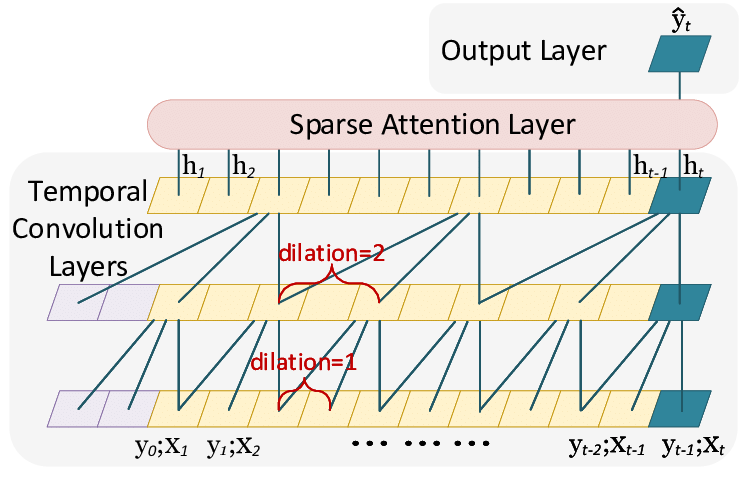
XGBoost là từ viết tắt của “eXtreme Gradient Boosting”. Là một thư viện tối ưu hóa cho phép thực hiện phương pháp gradient boosting phân tán, được thiết kế để đạt hiệu suất, linh hoạt và có thể di chuyển cao.

Nó triển khai các thuật toán học máy trong framework Gradient Boosting. XGBoost cung cấp một thuật toán boosting cây song song (còn được gọi là GBDT, GBM) giải quyết nhiều nhiều vấn đề trong khoa học dữ liệu một cách nhanh chóng và chính xác.

## DFNN

## TCAN

Temporal Convolutional Attention Neural Networks (TCAN) gồm 3 phần: Các lớp tích chập thời gian (Temporal Convolution Layers), lớp chú ý thưa thớt (Sparse Attention Layer) và lớp đầu ra (Output Layer) [STT].



*Hình: Kiến trúc TCAN*

Các lớp tích chập thời gian áp dụng phép tích chập để trích xuất đặc trưng từ chuỗi thời gian đầu vào, sử dụng cửa sổ và bước nhảy khác nhau để mã hóa mẫu thời gian quan trọng. Lớp chú ý thưa thớt tập trung vào các điểm dữ liệu quan trọng trong chuỗi thời gian và tạo ra trọng số chú ý cho từng điểm dữ liệu để giải thích kết quả dự đoán. Lớp đầu ra kết hợp đặc trưng và trọng số chú ý từ các lớp trước đó để tính toán đầu ra cho bài toán chuỗi thời gian.

TCAN sử dụng kiến trúc tích chập phân cấp để mã hóa chuỗi đầu vào và trích xuất mẫu thời gian dưới dạng biến tiềm ẩn (latent variables).

Bằng việc sử dụng lớp chú ý thưa thớt, TCAN có khả năng truy cập vào tất cả các bước lịch sử của chuỗi đầu vào mà không phụ thuộc vào độ dài của chuỗi. Điều này cho phép TCAN tập trung vào những bước quan trọng nhất của chuỗi đầu vào và cung cấp một hình ảnh trực quan về kết quả, giúp dễ dàng hiểu và giải thích kết quả dự đoán.

## Metrics

Sử dụng RMSE(Root Mean Squared Error) để đánh giá hiệu suất dựa trên sai số bình phương trung bình. Giá trị RMSE càng thấp, hiệu suất mô hình càng cao. RMSE phạt

# Experiment

Bảng chia train test

Kết quả dự đoán giá

## Linear Regression

## GRU

## ETS

## LSTM

## RNN

## GBT

## HMM

## XGBoost

## DFNN

## TCAN

Kết luận thuật toán nào tốt nhất

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | | CTLT | | INTC | | NTR | |
| Tỉ lệ chia | | 7:2:1 | 6:2:2 | 7:2:1 | 6:2:2 | 7:2:1 | 6:2:2 |
| Mô hình | Độ đo |  | | | | | |
| LR | MAPE | 0.61 | 0.2 | 0.47 | 0.2 | 0.295 | 0.35 |
| RMSE | 56.22 | 26.6 | 18.44 | 10.2 | 29.28 | 33.4 |
| MSLE | 0.25 | 0.05 | 0.18 | 0.04 | 0.14 | 0.23 |
| GRU | MAPE | 184.94 |  |  |  |  |  |
| RMSE | 16.86 |  |  |  |  |  |
| MSLE | 96.00 |  |  |  |  |  |
| ETS | MAPE | 0.027 | 0.02 | 0.02 | 0.02 | 0.025 | 0.02 |
| RMSE | 3.18 | 2.79 | 1.02 | 1.03 | 2.85 | 2.22 |
| MSLE | 0.001 | 4e-4 | 6e-4 | 4e-4 | 0.001 | 6e-4 |
| LSTM | MAPE |  |  |  |  |  |  |
| RMSE |  |  |  |  |  |  |
| MSLE |  |  |  |  |  |  |
| RNN | MAPE |  |  |  |  |  |  |
| RMSE |  |  |  |  |  |  |
| MSLE |  |  |  |  |  |  |
| GBT | MAPE | 7.52 | 3.34 | 40.59 | 2.77 | 11.84 | 28.72 |
| RMSE | 7.38 | 4.85 | 13.5 | 1.67 | 12.14 | 30.9 |
| MSLE | 0.01 | 0.001 | 0.128 | 0.001 | 0.02 | 0.14 |
| HMM | MAPE |  |  |  |  |  |  |
| RMSE |  |  |  |  |  |  |
| MSLE |  |  |  |  |  |  |
| XGBoost | MAPE |  |  |  |  |  |  |
| RMSE |  |  |  |  |  |  |
| MSLE |  |  |  |  |  |  |
| DFNN | MAPE |  |  |  |  |  |  |
| RMSE |  |  |  |  |  |  |
| MSLE |  |  |  |  |  |  |
| TCAN | MAPE | 0.194 | 0.55 |  |  | 0.032 | 0.03 |
| RMSE | 0.058 | 0.12 | 0.13 | 0.16 | 0.022 | 0.025 |
| MSLE | 0.002 | 0.008 | 0.013 | 0.02 | 1e-4 | 2e-4 |

# Conclusion

##### Acknowledgment

##### References