# STOCK PRICE OF TECH COMPANIES

Time Series Analyst and Predict

| DA101 2021|

Học viên thực hiện:

Đào Hoàng Minh

Nguyễn Hoàng Khởi

# MỤC LỤC

# DOANH MỤC HÌNH ẢNH

# 1. GIỚI THIỆU

## 1.1 Giới thiệu:

Mục đích của bài báo cáo là sử dụng các nguồn dữ liệu về giá cổ phiếu, chủ yếu tập trung vào nhóm cổ phiếu của những công ty công nghệ (Facebook, Google, Apple,...) nhằm tìm ra hướng để phân tích và dự đoán được 1 cách chính xác nhất có thể về giá cổ phiếu.

Giới thiệu về các mã cổ phiếu: Gồm các cổ phiếu các công ty công nghệ:

- Google (Mã: GOOG)
- Facebook (Mã: FB)
- Apple (Mã AAPL)
- Netflix (Mã: NFLX)

Với sự phát triển của các công ty công nghệ trong thập kỷ vừa qua, với xu hướng tăng mạnh, cổ phiếu các nhóm ngành này đã chiếm tỷ trọng khá lớn trong xu hướng giá của thị trường.

Điển hình như Google, là một cổ phiếu được phát hành từ năm 2004. Và kể từ đó có xu hướng tăng trưởng ổn định

Các cổ phiếu nhóm công nghệ này có đặc điểm là thuộc các công ty có tuổi đời khá trẻ so với thị trường, vì vậy data set sẽ lấy giá cổ phiếu từ năm 2014 trở đi.

Các insight trong bài có được dựa vào cách nghiên cứu dữ liệu đã có (historical data) và chạy các phép toán như Decomposition, Moving Average,...

## 2. Xử lý dữ liệu

### 2.1 Tập dữ liệu

Dữ liệu được lấy từ dữ liệu chứng khoán của Yahoo Finance (một siêu data khổng lồ chứa dữ liệu của các công ty được niêm yết trên thị trường chứng khoán Mỹ). [W2]

Mỗi tệp sẽ cung cấp dữ liệu gồm các biến sau:

| Tên biến | Ý nghĩa |
|---|---|
| Date | Đơn vị thời gian theo ngày |
| High | Giá trần: Giá cao nhất trong ngày cổ phiếu có thể đạt. |
| Low | Giá sàn: Giá thấp nhất trong ngày cổ phiếu có thể hạ giá. |
| Open | Giá mở cửa: Giá tại thời điểm bắt đầu phiên giao dịch |
| Close | Giá đóng cửa: Giá tại thời điểm cuối phiên giao dịch |
| Volume | Tổng khối lượng cổ phiếu được giao dịch trên thị trường |
| Adj. Close | Giá đóng cửa điều chỉnh: Giá đóng cửa của cổ phiếu được điều chỉnh để phản ánh chính xác giá trị của cổ phiếu đó |

### 2.2 Xử lý dữ liệu

Phần này sử dụng một số phương pháp để biến dữ liệu thô thành dữ liệu đã chứng hoá như:

- Dưa dữ liệu về dạng dataframe phù hợp với các đối tượng dữ liệu dạng time series

| Index<br>Date | Open | High | Low | Close | Volume | Adj Close |
|---|---|---|---|---|---|---|

Phần chỉ số index nên là datetime với ngày bắt đầu và ngày kết thúc cụ thể

- Trích xuất các cột từ bảng

Sử dụng các phương pháp để trích xuất các cột “Date”, “Close”, “Open”… trong bản để tính toán một số giá trị như EMA, SMA,…

- Tính toán một số biến thêm vào như: EMA, SMA, UP, DOWN, RSI, MACD

Trong bài toán phân tích ta cần tính toán thêm một vài giá trị từ dữ liệu thô để thực hiện các phương pháp phân tích chuyên sau như EMA (exploitation mean avarage), SMA (simple mean avarage),…

- Trong quá trình thực hiện ta nên tạo một số bản sao của dữ liệu để không bị trôi dữ liệu thô

Việc sử dụng một sử dụng dữ liệu thô trong suốt quá trình phân tích sẽ dẫn đến không kiểm soát tốt dữ liệu và nghiêm trọng hơn là thây đổi giá trị thô, dẫn đến các đối tượng sau này sẽ làm việc với các giá trị thiếu chính xác.

## 3. PHÂN TÍCH DỮ LIỆU

### 3.1 PHÂN TÍCH DỰA VÀO ĐỒ THỊ

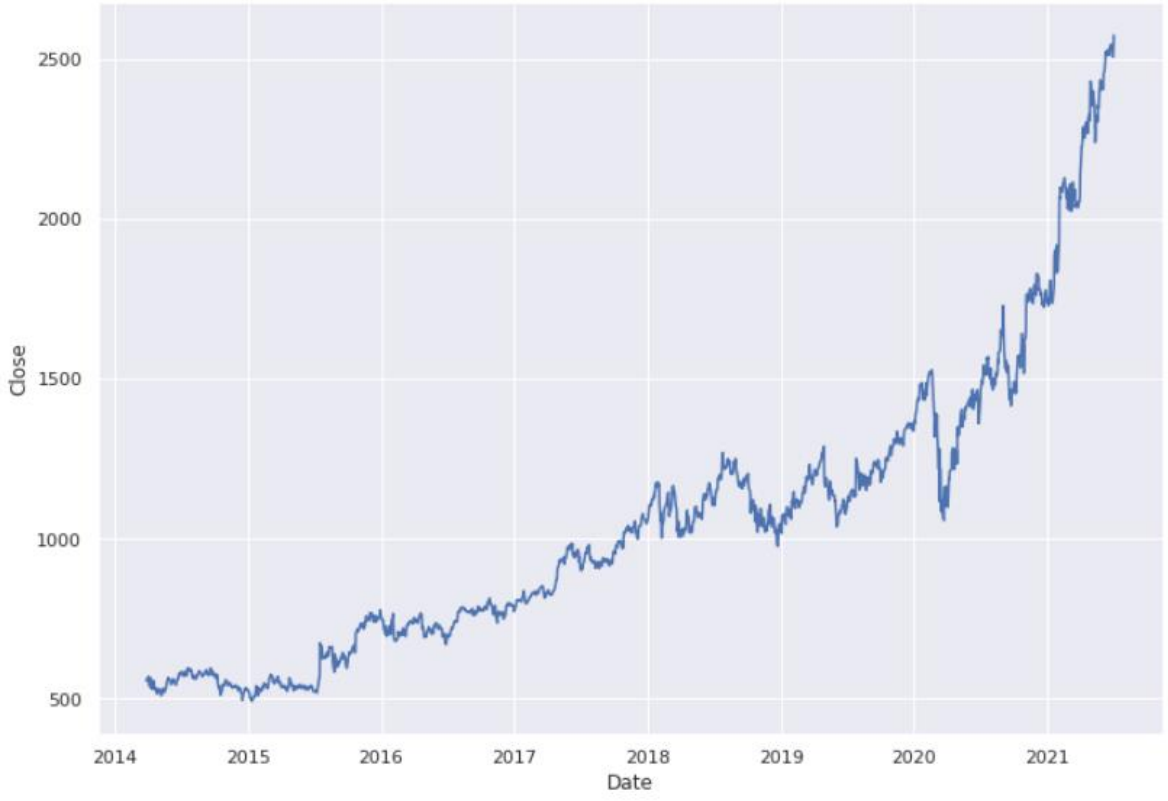
Mục đích việc analyze là để tìm được những insight thông qua dữ liệu trong quá khứ của data set. Với hy vọng bằng việc thực hiện những thao tác như cleaning data, visualize, technical analyze, có thể đưa ra những insight hợp lý cho data.

Tạm thời sẽ sử dụng dataset của mã cổ phiếu Google (GOOG) để phân tích và đưa ra dự đoán cho nhóm ngành Công Nghệ.

Để dự đoán và phân tích giá cổ phiếu

i. Time Series

Times Series cho thấy giá đóng cửa của thị trường theo thời gian. Tổng quát giá có sự tăng trưởng qua các năm và tăng mạng kể từ năm 2018.




Hình 3.1: Biểu đồ time series của dữ liệu GOOG Stocks từ năm 2014 - 2021


Hình 3.2: Biểu đồ time series của dữ liệu GOOG Stocks từ năm 2018 - 2019

Biểu đồ dạng combo kết hợp đường và cột với biểu đồ đường thể hiện giá đóng cửa của mỗi ngày và biểu đồ cột thể hiện khối lượng cổ phiếu được giao dịch trên thị trường của ngày hôm đó.

Với cột màu đỏ thể hiện volume của những ngày giá giảm, và ngược lại cột màu xanh thể hiện volume của những ngày giá tăng.

Nhận xét:

- Với những cột volume cao đều là điểm báo hiệu cho thị trường sắp có chuyển biến ngược lại với khoảng thời gian trước đó. Nếu giá đang đi xuống nhưng ngày hôm đó có khối lượng cao thì sẽ có khả năng giá sẽ tăng vào khoảng thời gian các ngày tiếp theo.
- Hầu hết các cột volume cao đều là cột màu đỏ. Và ngay sau đó là giá liền đi lên chứng tỏ thị trường vẫn còn sức mua mạnh, mỗi khi giá giảm vẫn có đủ tiềm lực mua vào để giá đi lên.

ii. CandleStick

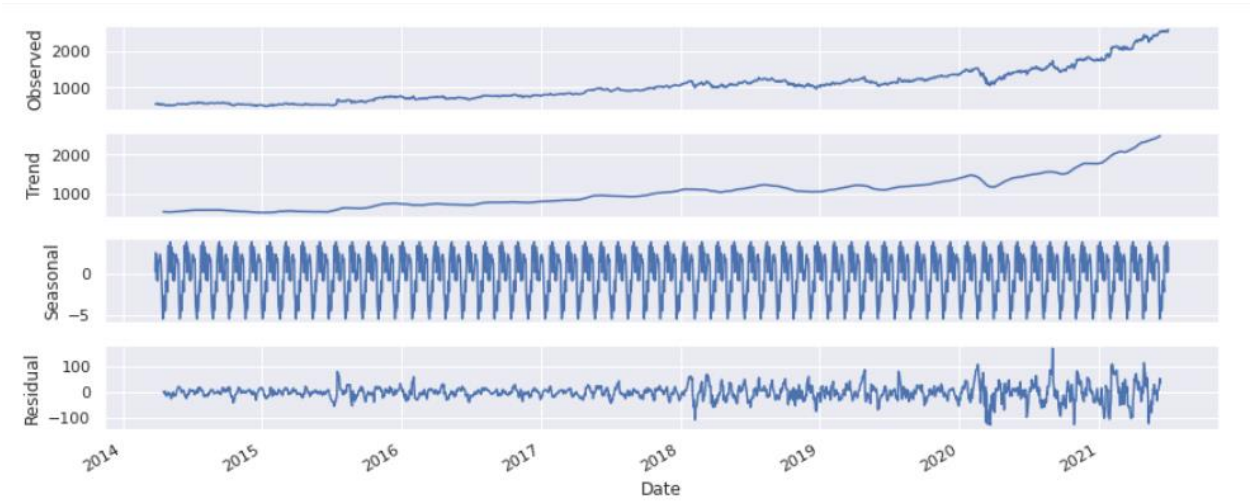
Dùng các biến: Open, Close để vẽ biểu đồ dạng nến, với Open sẽ là mức giá nơi nến bắt đầu được vẽ và Close sẽ là mức giá nơi nến được kết thúc vẽ.


Hình 3.3: Biểu đồ CandleStick của dữ liệu GOOG Stocks từ năm 2014 - 2021

Đồ thị candlestick trong 3 tháng đầu năm 2018 (01-01-2018 : 01-03-2018)

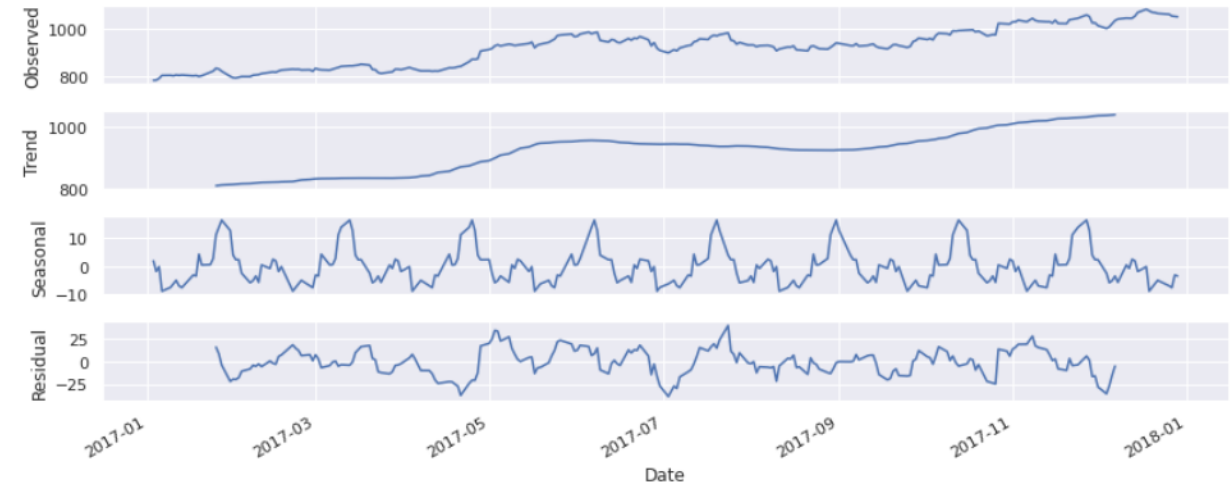

Hình 3.4: Đồ thị candlestick trong 3 tháng đầu năm 2018 (01-01-2018 : 01-03-2018)

iii. Additive Decomposition

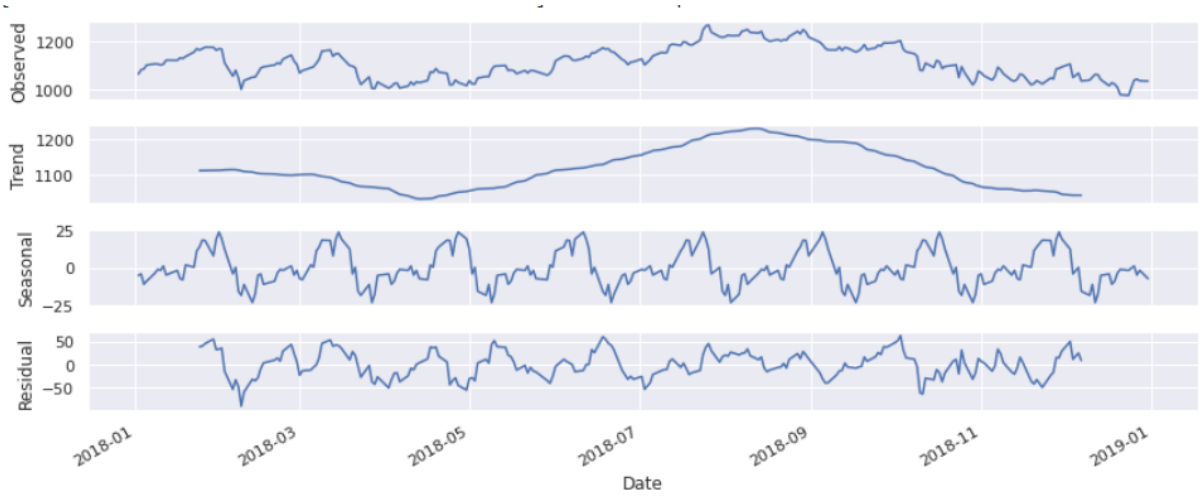



Hình 3.5: Biểu đồ Additive Decomposition của dữ liệu GOOG Stocks từ năm 2014 - 2021

Biểu đồ trong năm 2017 và 2018



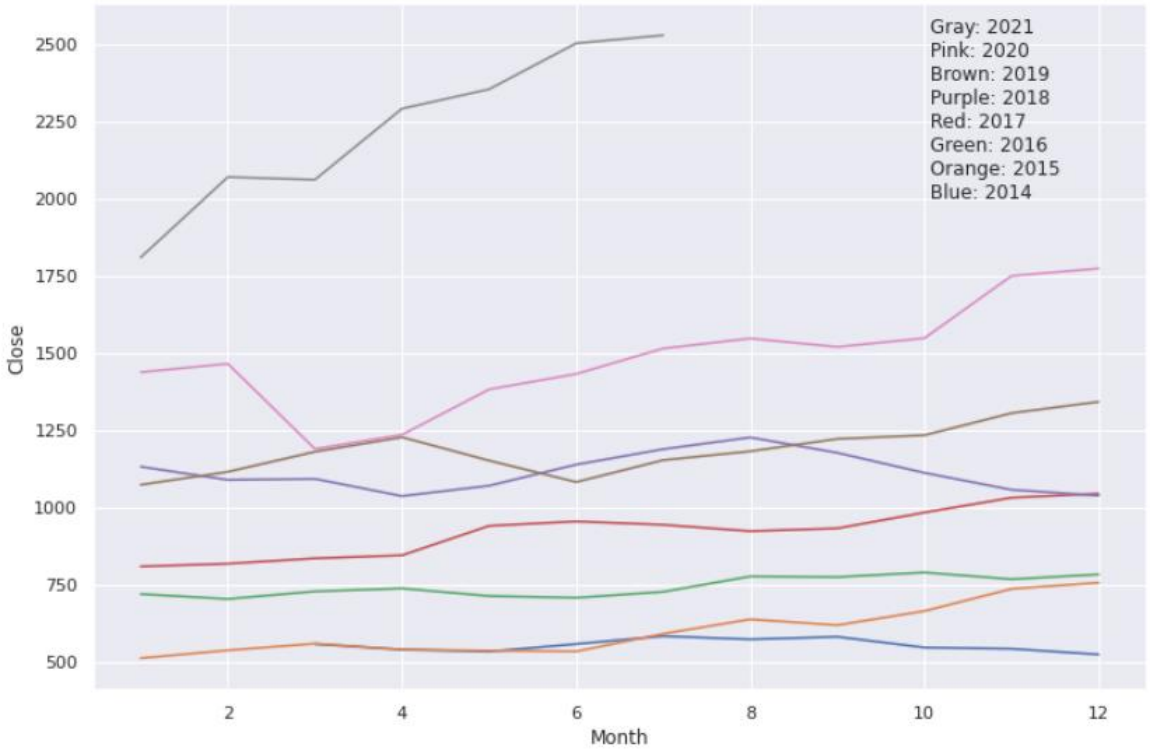

Hình 3.6: Biểu đồ Additive Decomposition của dữ liệu GOOG Stocks năm 2017




Hình 3.7: Biểu đồ Additive Decomposition của dữ liệu GOOG Stocks năm 2018

iv. Seasonality

Biểu đồ Seasonality nhằm phân tích chu kỳ mùa vụ trong một năm để tìm được chu kỳ giá của cổ phiếu



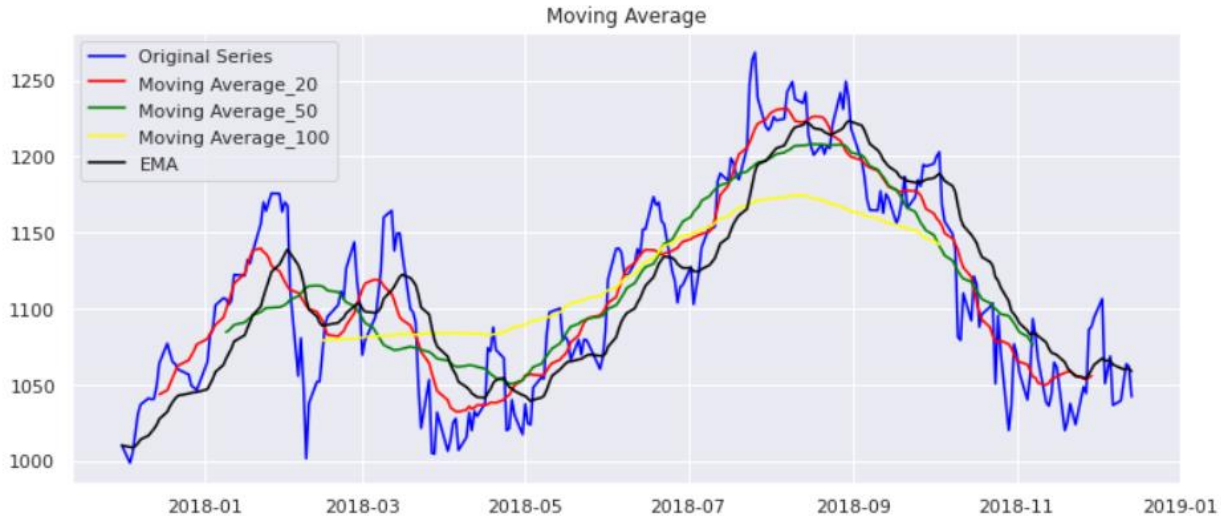

Hình 3.8: Biểu đồ seasonality

Nhận xét:

- Nhìn vào độ thị chúng ta có một ngoại lệ là giá cổ phiếu trong năm 2021 tăng liên tục và đang có xu hướng ổn đinh tại giá 2015 USD/Share
- Hầu hết giá cổ phiếu GOOG giảm vào tháng 3-4 và tăng đều từ tháng 5-12, đạt tỉ lệ tối đa ở cuối năm
- Về đầu tư dài hạn giá cổ phiếu GOOG tăng đều sau mỗi năm và tăng gấp 5 lần từ năm 2014 - 2021

### 3.2 Sử dụng các công thức phân tích

i. Moving Average




Hình 3.9: Biểu đồ time series với các đường trung bình

Ta có Moving Average (MA) là đường trung bình động của giá trong 1 khoảng thời gian nhất định. Với MA(20) là đường màu đỏ, biểu hiện cho trung bình động của giá trong 20 ngày giao dịch vừa rồi (hay còn tính là 1 tháng). MA(50) là đường màu xanh lá cây, đại diện cho trung bình động của 1 quý và cuối cùng là MA(100) đại diện cho 2 quý.

Ta có thể thấy mỗi khi 2 đường MA đang cách xa nhau gặp tại 1 điểm, hướng đi của đường MA có giá trị nhỏ hơn sau khi gặp đường MA lớn sẽ góp phần quyết định giá của thị trường trong thời gian sắp tới.

Ví dụ: MA(20) sau khi chạm MA(50) thì đi lên, giống với xu hướng của giá tăng.

Nhận xét: Dựa vào các đường MA, khi 2 đường gặp nhau có thể là dữ liệu chuẩn bị để cho thị trường có xu hướng đảo chiều.

Khi thị trường có xu hướng tăng giá (uptrend) giá sẽ nằm trên các đường MA, ngược lại khi thị trường có xu hướng giảm giá (downtrend) giá sẽ nằm dưới các đường MA.

ii. Exponential Moving Average

Đường trung bình động lũy thừa chính xác hơn đường trung bình động vì có gán nhiều trọng số trong phiên giao dịch gần nhất vào hơn. Tuy nhiên do chính xác hơn đồng nghĩa cũng nhạy hơn với việc biến động của thị trường và dễ bị gây nhiễu.

Sử dụng đường EMA(50) để kéo dài mốc thời gian tính trung bình động lũy thừa ra, hạn chế việc gây nhiễu trong dự báo.


Hình 3.10: Biểu đồ time series với EMA

Nhận xét:

- Nhìn vào đồ thị ta thấy giá cổ phiếu đang có xu hướng đi lên và có lẽ chưa đạt đỉnh trong hiện tại
- Về đầu tư lâu dài giá cổ phiếu GOOG tăng gấp 5 lần trong 7 năm 2014-2021

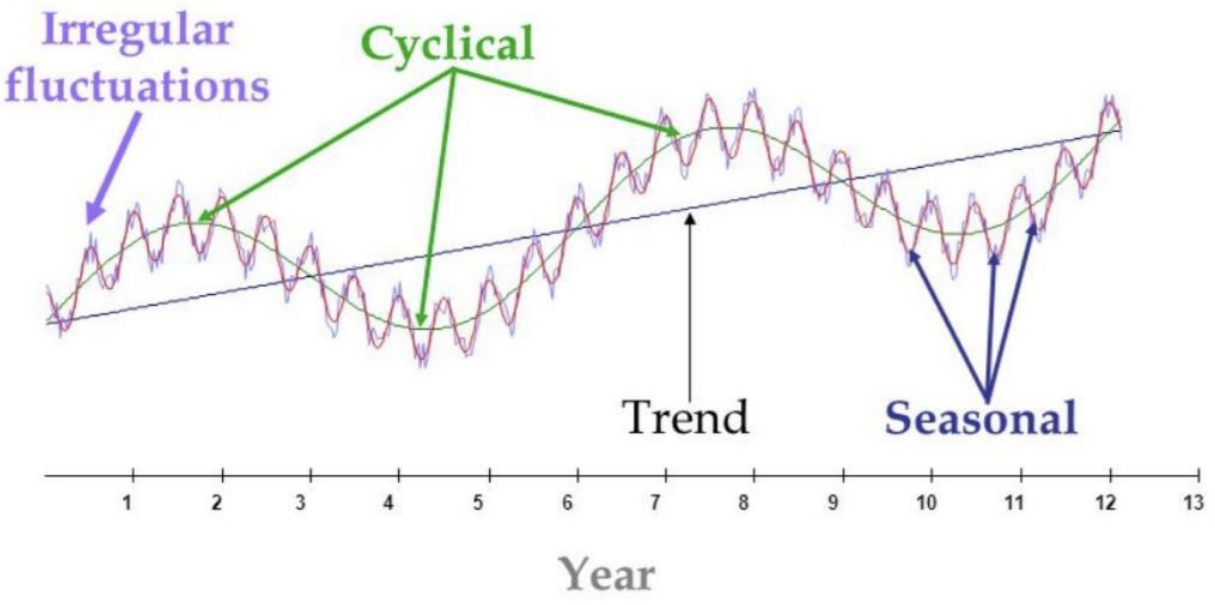
## 4. DỰ ĐOÁN

Dự đoán time series là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Sở dĩ chúng ta gọi lớp mô hình này là time series là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Time series thường dự đoán dựa trên giả định rằng các qui luật trong quá khứ sẽ lặp lại ở tương lai.

Do đó xây dựng mô hình chuỗi thời gian là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Một mô hình time series thường phân tích trên một quá quy luật sau

- Tính mùa vụ (seasonal)
- Xu hướng (trend)
- Chu kỳ (Cyclical)
- Nhiễu (irregular)




Hình 4.1: phân tích một time series căn bản

Để dự đoán biến phụ thuộc từ biến đầu vào của một mô hình time series người ta thường sử dụng mô hình hồi quy tuyến tính.

Trong đề tài nhóm sẽ trình bày 2 mô hình tuyến tính thường được sử dụng

- Mô hình hồi quy tuyến tính một biến và đa biến
- Mô hình ARIMA/ARIMAS

### 4.1 Hồi quy tuyến tính

Bài toán hồi quy là một trong những bài toán cơ bản nhất của học máy (machine learning). Nội dung bài toán là tìm các trọng số tốt nhất của một phương trình đường thẳng đối với mô hình một biến và phương trình mặt phẳng trong không gian đa chiều. Biến phụ thuộc sẽ tuân theo một quy luật tuyến tính

Một số ứng dụng như dự đoán giá nhà tại thành phố, giá cổ phiếu,…

Biểu diễn của một phương trình hồi quy dưới dạng

$$Y = W*X$$

Ví dụ: y = a*x + b

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.43 | 0.11 | 0.18 | 174 |
| 1 | 0.52 | 0.86 | 0.65 | 192 |
| accuracy | | | 0.51 | 366 |
| macro avg | 0.48 | 0.49 | 0.42 | 366 |
| weighted avg | 0.48 | 0.51 | 0.43 | 366 |

Nhận xét:

- Độ chính xác của mô hình hồi quy tuyến tính không cao khoảng 50%

### 4.2. ARIMA (p,d,q) vs SARIMA

ARIMA là viết tắt của cụm từ Autoregressive Intergrated Moving Average là một mô hình hồi qui tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) gồm 3 thành phần chính:

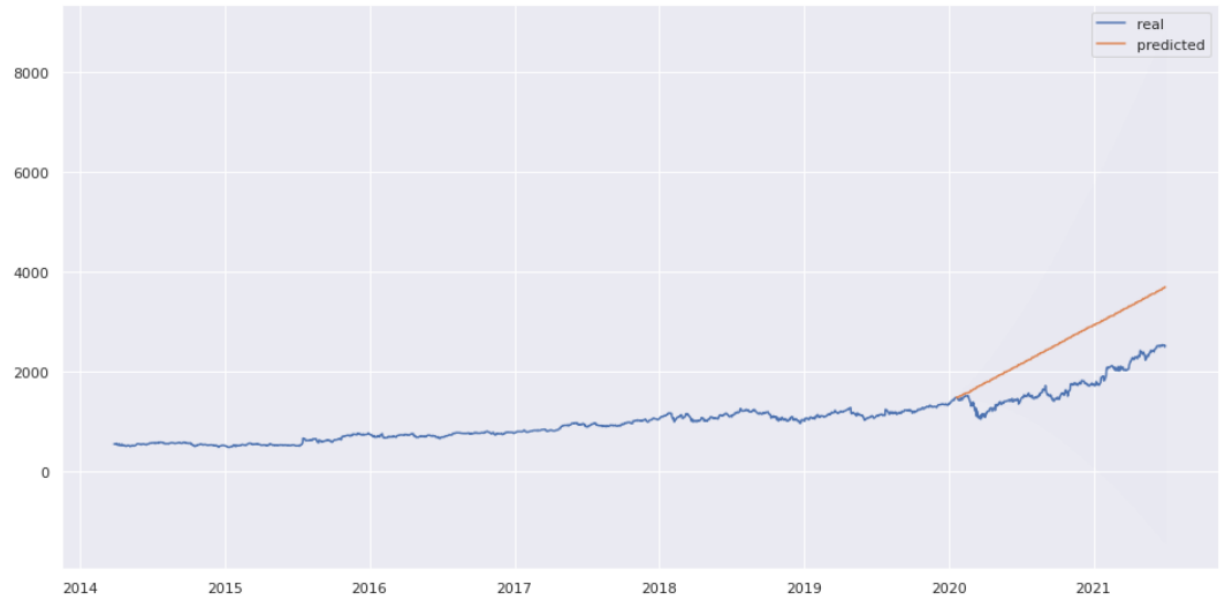
ARIMA = AutoRegressive (AR) + Integrated (I) + Moving Average (MA)

AR (Auto regression): là thành phần tuyến tính của mô hình ARIMA và có khả năng tự hội quy thể hiện qua thông số p

MV (Moving average): Quá trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Thể hiện dữ liệu có tính trung bình thể hiện ở thông số p

(I) Integrated: Là quá trình đồng tích hợp hoặc lấy sai phân đảm bảo tính dừng cho thuật toán thể hiện trên thông số d

Ta giải bài toán ARIMA là tìm trọng số tốt nhất cho ba thông số p,d,q




Hình 4.2: Biểu đồ dự đoán giá cổ phiếu dựa vào mô hình ARIMA

| | r2_score | mean_absolute_error | mean_squared_error | mean_absolute_percentage_error |
|---|---|---|---|---|
| 0 | -1.166829 | 544.478099 | 328177.269009 | 32.445238 |

Nhận xét:

- Mô hình dự đoán giá thị trường từ năm 2020 đến cuối năm 2021 có xu hướng tăng gấp đôi
- Thực tế giá cổ phiêu tăng nhưng tăng 1/2 phần trăm
- Đánh giá do thị trường bất ổn bởi dịch COVID đo đó các mặc hàng kinh doanh của GOOG đã giảm nhưng bởi vì lợi ích là công ty TECH nên GOOG vẫn dự được đà tăng trưởng tốt
- Mô hình ARIMA có độ chính xác cao hơn so với các mô hình khác 100 - 32.4 = 67.6%

### 4.3 Decision Tree

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.48 | 0.47 | 0.47 | 163 |
| 1 | 0.57 | 0.57 | 0.57 | 198 |
| accuracy | | | 0.53 | 361 |
| macro avg | 0.52 | 0.52 | 0.52 | 361 |
| weighted avg | 0.53 | 0.53 | 0.53 | 361 |

Đôi với mô hình Decision Tree mô hình có khả năng dự đoán chỉ 50% không đủ độ chính xác nhóm mong muốn

## 5. TÀI LIỆU THAM KHẢO


*Tài liệu website*

[W1] https://viblo.asia/p/xay-dung-bieu-do-phan-tich-chung-khoan-su-dung-python-p1-vyDZOaXO5wj?fbclid=IwAR0Xeuw4g5K6llqyCHkPCgJRFTB3ofAyXNnzuOAKP-oQQB60hmK-XKPS31E

[W2] https://ichi.pro/vi/xay-dung-mot-bo-phan-loai-chuyen-dong-co-phieu-don-gian-bang-cach-su-dung-may-hoc-va-python-4381850998960?fbclid=IwAR0bXVOB8u2ilVf9VlMraexxxeJqd__63_4hc4qzvfJRUiZQ1Ss6fQBdE_w

[W3]https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html?fbclid=IwAR2jDY4Yq0xijfG94W75twAcIgMPUzewuxKRRT3r6d4dGx6Y56EmvVXSpGQ#3-%E1%BB%A9ng-d%E1%BB%A5ng-vnquant-trong-thu-th%E1%BA%ADp-d%E1%BB%AF-li%E1%BB%87u

[W4] https://finance.yahoo.com/quote/GOOG/history?period1=1512100800&period2=1544932799&interval=1d&frequency=1d&filter=history

*Tài liệu code*

[C1] Hands-on 3:

https://colab.research.google.com/drive/1R7m-SxtB74MdZTMiPufWHkYLw11bztmS#scrollTo=tZyEsbdERiYF

[C2] Pre_process - data -time

https://colab.research.google.com/drive/1jOZ_t4hRHNtRCJ0H-0Gg0vhl__4SuBcn

*Tài liệu dạng lecture*

[L1] Lecture 10: Times series VEF

https://piazza.com/class_profile/get_resource/knaaokinerw5mq/kpb9zj7ka6x6cd

[L2] DA resource VEF

https://piazza.com/vef_academy/spring2021/da2101/resources

*Final project*

[F1] Phân tích dữ liệu

https://colab.research.google.com/drive/1JIQOgBcZPmrB07L4jW7rdJxhUnaaZQU_?usp=sharing

[F2] dự đoán

https://colab.research.google.com/drive/1RrNN8DKyi1dOcgJ9q9htZSJUug0D0T-x?usp=sharing

## 6. PHỤ LỤC


Hình 6.1: Biểu đồ giá cổ phiếu của 4 công ty TECH lớn


Hình 6.2: mô hình dự đoán ARIMA từ năm 2014-2019 mã GOOG

| | r2_score | mean_absolute_error | mean_squared_error | mean_absolute_percentage_error |
|---|---|---|---|---|
| **0** | -2.075474 | 102.200292 | 13664.745942 | 9.206664 |

Nhận xét:

- Với thị trường ít biết động mô hình ARIMA có độ chính xác cao khoảng 90.8%
- Mô hình có xu hướng tăng trưởng phù hợp với dữ liệu thực tế