Triển khai hệ thống phân tích và xử lý dữ liệu lớn

Case study: Stock Price BigData (Dữ liệu chứng khoán Việt Nam)

Hoàng Minh Quyền
MSSV: 23020421
Lớp: K68A-AI1

October 20, 2025

Abstract

Báo cáo mô tả quá trình triển khai và thực hiện hệ thống xử lý dữ liệu lớn cho bài toán phân tích giá chứng khoán. Dự án được xây dựng dựa trên hệ thống Hadoop, Spark và Docker, tái hiện pipeline phân tích dữ liệu chứng khoán Việt Nam. Mã nguồn thực thi trong notebook (`Stock_price_demo_for_VN.ipynb`) minh họa quá trình kết nối HDFS, tiền xử lý dữ liệu, phân tích thống kê và mô hình dự báo LSTM.

Contents

1 Giới thiệu **3**

2 Kiến trúc hệ thống **3**
2.1 Thành phần hệ thống . . . 3
2.2 Cấu hình Docker Compose . . . 3

3 Dữ liệu **3**
3.1 Chuẩn bị dữ liệu . . . 3
3.2 Tải dữ liệu lên HDFS . . . 4

4 Xử lý dữ liệu với PySpark **4**
4.1 Đọc dữ liệu từ HDFS . . . 4
4.2 Tiền xử lý . . . 4
4.3 Chuyển đổi sang Pandas để trực quan hóa . . . 5
4.4 Trực quan hóa dữ liệu . . . 5
4.5 Nhận xét từng biểu đồ . . . 7
4.5.1 Phân tích biểu đồ theo diễn biến giá lịch sử của 5 mã . . . 7
4.5.2 Phân tích phân phối lợi nhuận của 5 mã . . . 7
4.6 Kết luận chung . . . 8

5 Huấn luyện mô hình dự báo LSTM **8**
5.1 Chuẩn bị dữ liệu . . . 8
5.2 Cấu trúc mô hình . . . 9
5.3 Kết quả huấn luyện . . . 9

6 Phân tích kết quả từ mô hình dự đoán giá của mã CII **10**

7 Kết luận và hướng phát triển **11**

1 Giới thiệu

Mục tiêu project: Vận dụng lại hệ thống Big Data mô phỏng quá trình lưu trữ và xử lý dữ liệu chứng khoán bằng Hadoop và Spark. Phân tích dữ liệu chứng khoán Việt Nam và áp dụng mô hình học sâu (LSTM) để dự báo giá.

Ý nghĩa: Việc ứng dụng Big Data trong lĩnh vực tài chính giúp khai thác dữ liệu lớn, đưa ra quyết định đầu tư thông minh và giảm rủi ro.

2 Kiến trúc hệ thống

2.1 Thành phần hệ thống

Hệ thống được xây dựng bằng Docker Compose sử dụng các image Hadoop và Spark của Big Data Europe (bde2020) cùng với image `pyspark-notebook` để thao tác qua Jupyter.

Các thành phần chính:

- 1 Namenode + 4 Datanode (HDFS)
- 1 Spark Master + 4 Spark Worker (Spark Cluster)
- 1 Jupyter container (để chạy notebook)

2.2 Cấu hình Docker Compose

Cấu hình mô tả trong file `docker-compose.yml` gồm các dịch vụ: namenode, datanode, yarn, spark-master, spark-worker và jupyter.
[**TODO: Bổ sung đường dẫn file và thông tin cấu hình cụ thể của nhóm**]

3 Dữ liệu

Nguồn dữ liệu: Dữ liệu được thay đổi từ chứng khoán Mỹ sang dữ liệu chứng khoán Việt Nam. Dữ liệu chính thống từ API vnstock với hơn 1600 mã ,bao gồm các mã cổ phiếu nổi bật như `VIC`, `VNM`, `VCB`, `HPG`, `SSI`, `CII`.

3.1 Chuẩn bị dữ liệu

Các tập dữ liệu gốc được tải từ nguồn api vnstock dưới định dạng CSV, gồm các cột:

- `Time`: ngày giao dịch
- `Open`, `High`, `Low`, `Close`: giá mở cửa, cao nhất, thấp nhất, và đóng cửa
- `Volume`: khối lượng giao dịch

Các file được đặt trong thư mục nội bộ `/data/vn_stock/`.

3.2 Tải dữ liệu lên HDFS

Dữ liệu được nạp vào hệ thống HDFS bằng lệnh:

```
hdfs dfs -mkdir -p /user/root/vn_stock/
hdfs dfs -put stock_data_VIC.csv /user/root/vn_stock/
hdfs dfs -put stock_data_VNM.csv /user/root/vn_stock/
...
```

Sau khi tải lên, kiểm tra bằng:

```
hdfs dfs -ls /user/root/vn_stock/
```

Kết quả hiển thị xác nhận các file CSV đã nằm trên HDFS, sẵn sàng cho Spark đọc.

4 Xử lý dữ liệu với PySpark

4.1 Đọc dữ liệu từ HDFS

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .appName("Stock_Price_Vietnam") \
    .master("spark://spark-master:7077") \
    .getOrCreate()

df_vic = spark.read.csv("hdfs://namenode:9000/user/root/vn_stock/
   stock_data_VIC.csv",
                        header=True, inferSchema=True)
```

4.2 Tiền xử lý

Dữ liệu đại diện: Ở đây chúng ta sẽ chỉ chọn ra 5 loại cổ phiếu nổi tiếng và nổi bật nhất:

- VinGroup JIC (mã VIC)
- Vinamilk (mã VNM)
- Vietcombank (mã VCB)
- Hòa Phát Group (HPG)
- Chứng Khoán SSI (SSI)

Các bước xử lý chính trong notebook:

- Loại bỏ dòng thiếu dữ liệu.
- Tạo cột `Mean = (High + Low)/2`.
- Tính cột `Change = (Close/Close.shift(1)) - 1` để đánh giá biến động hàng ngày.

```
from pyspark.sql.functions import col, expr
df_VICmean = df_VIC.withColumn("Mean", expr('(High+Low)/2'))
df_VNMmean = df_VNM.withColumn("Mean", expr('(High+Low)/2'))
df_VCBmean = df_VCB.withColumn("Mean", expr('(High+Low)/2'))
df_HPGmean = df_HPG.withColumn("Mean", expr('(High+Low)/2'))
df_SSImean = df_SSI.withColumn("Mean", expr('(High+Low)/2'))
```

4.3 Chuyển đổi sang Pandas để trực quan hóa

```
df_VICmean = df_VICmean.toPandas()
df_VNMmean = df_VNMmean.toPandas()
df_VCBmean = df_VCBmean.toPandas()
df_HPGmean = df_HPGmean.toPandas()
df_SSImean = df_SSImean.toPandas()
```

4.4 Trực quan hóa dữ liệu

Sử dụng matplotlib để vẽ:

- Biểu đồ giá trung bình theo thời gian.

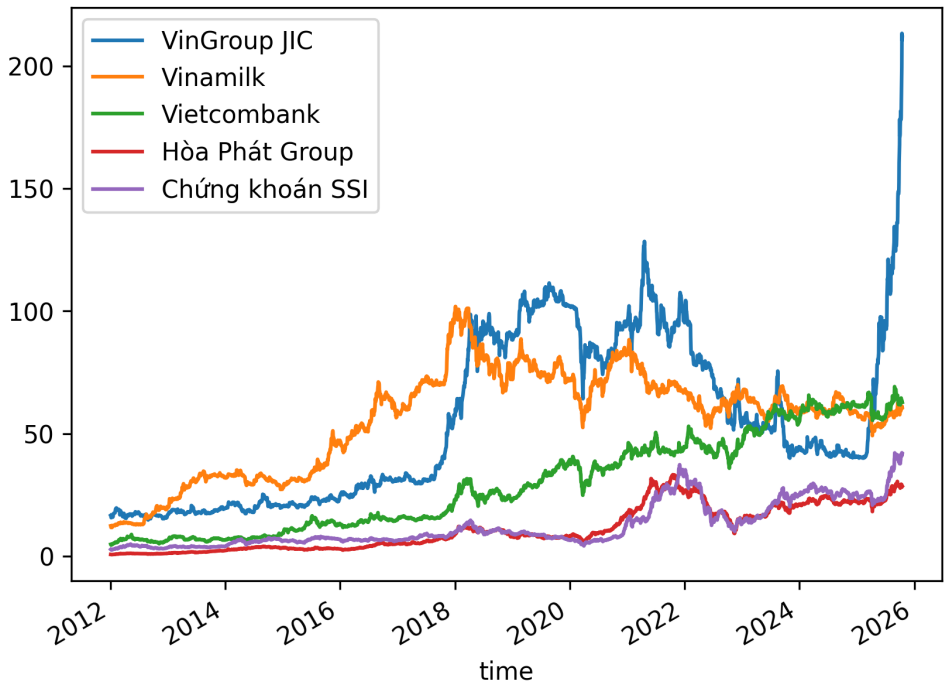

Figure 1: Biểu đồ giá trung bình theo thời gian của 5 mã cổ phiếu.

- Histogram phân phối lợi suất.

```
df_VICmean['Change']= (df_VICmean['close']/df_VICmean['
    close'].shift(1)) - 1
df_VNMmean['Change']= (df_VNMmean['close']/df_VNMmean['
    close'].shift(1)) - 1
```

```
df_VCBmean['Change']= (df_VCBmean['close']/df_VCBmean['
    close'].shift(1)) - 1
df_HPGmean['Change']= (df_HPGmean['close']/df_HPGmean['
    close'].shift(1)) - 1
df_SSImean['Change']= (df_SSImean['close']/df_SSImean['
    close'].shift(1)) - 1
```

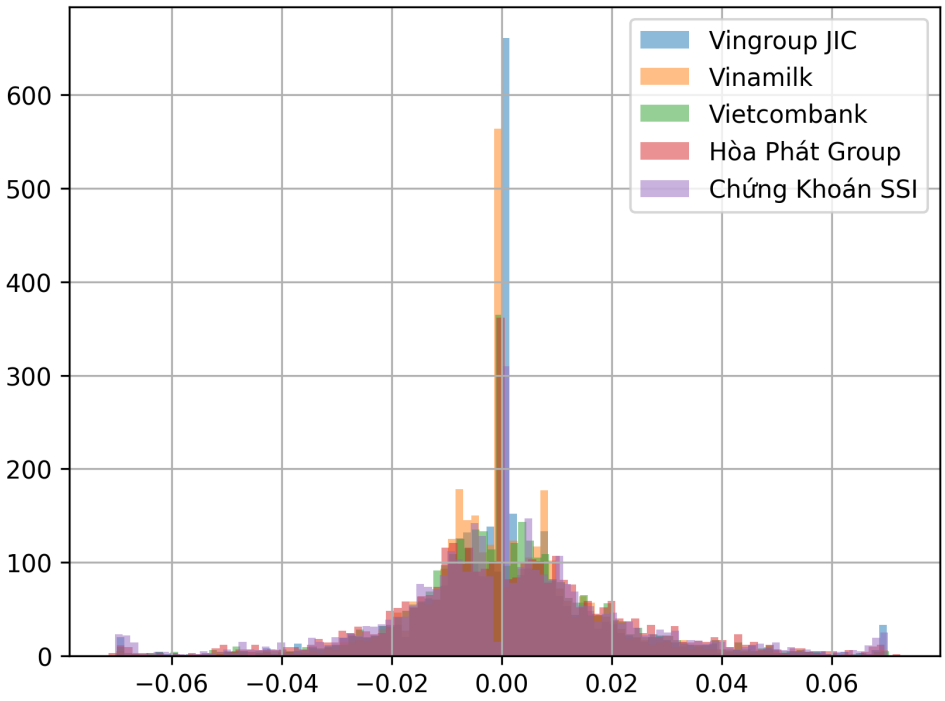

Figure 2: Biểu đồ phân phối lợi suất theo thời gian của 5 mã cổ phiếu.

- So sánh giữa các mã VIC, VNM, VCB, HPG, SSI. (miêu tả rõ hơn sự biến động)

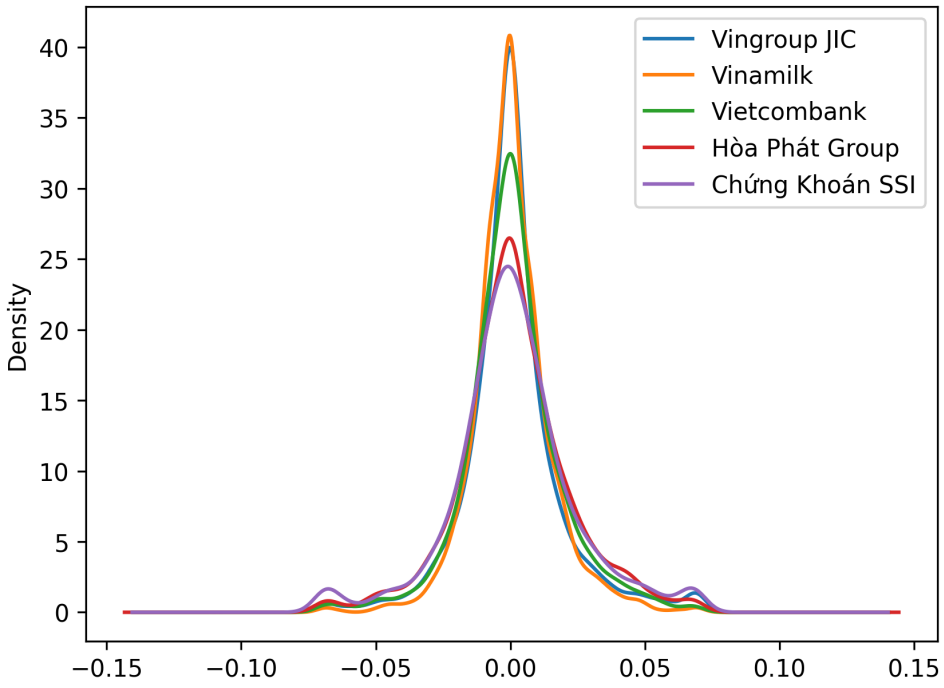

Figure 3: Biểu đồ sự biến động về giá theo thời gian của 5 mã cổ phiếu.

4.5 Nhận xét từng biểu đồ

4.5.1 Phân tích biểu đồ theo diễn biến giá lịch sử của 5 mã

Biểu đồ này cho thấy bức tranh dài hạn về sự tăng trưởng của các cổ phiếu từ năm 2012 đến nay.

- Vingroup JIC (VIC - xanh dương): Đây là cổ phiếu có biến động mạnh nhất trong dài hạn. Nó có những giai đoạn tăng trưởng bùng nổ nhưng cũng có những đợt điều chỉnh rất sâu. Đặc biệt, có một cú tăng giá dựng đứng và cực kỳ mạnh mẽ ở giai đoạn cuối của biểu đồ (khoảng 2025-2026), giải thích cho sự tồn tại của "đuôi dày" bên phải trong biểu đồ phân phối lợi nhuận.
- Vinamilk (VNM - cam): Từng là cổ phiếu "vàng" với sự tăng trưởng rất ấn tượng cho đến khoảng năm 2018, sau đó bước vào một giai đoạn đi ngang và suy giảm kéo dài. Gần đây mới có dấu hiệu phục hồi.
- Vietcombank (VCB - xanh lá): Cho thấy một xu hướng tăng trưởng ổn định và bền vững nhất trong nhóm. Mặc dù có những biến động, nhưng đường giá của VCB có xu hướng đi lên một cách vững chắc.
- Hòa Phát Group (HPG - đỏ) & Chứng khoán SSI (tím): Cả hai cổ phiếu này đều thể hiện rõ tính chu kỳ, gắn liền với chu kỳ của ngành thép và của thị trường chứng khoán nói chung. Chúng có những con sóng tăng trưởng rất mạnh nhưng cũng xen kẽ những giai đoạn điều chỉnh sâu.

4.5.2 Phân tích phân phối lợi nhuận của 5 mã

Cả hai biểu đồ Histogram và KDE (biểu đồ mật độ) đều mô tả cùng một thứ: phân phối của tỷ suất sinh lợi hàng ngày. Biểu đồ KDE là một phiên bản được làm mịn của

Histogram.

- Lợi nhuận không tuân theo phân phối chuẩn: Đây là đặc điểm kinh điển của dữ liệu tài chính. Thay vì có hình chuông hoàn hảo, phân phối của các cổ phiếu này có đặc tính "đỉnh nhọn" (Leptokurtosis) và "đuôi dày" (Fat Tails).

  – Đỉnh nhọn: Phần lớn các ngày giao dịch, giá cổ phiếu chỉ thay đổi rất ít, tạo ra một đỉnh rất cao và nhọn quanh mức 0.

  - Đuôi dày: Các sự kiện biến động mạnh (tăng hoặc giảm hơn 4-6%) xảy ra thường xuyên hơn so với lý thuyết phân phối chuẩn dự đoán. Điều này có nghĩa là rủi ro về các cú sốc bất ngờ là có thật và cần được quản lý.

- So sánh độ biến động (Volatility):

  - Vinamilk (màu cam) có đỉnh cao và nhọn nhất. Điều này cho thấy đây là cổ phiếu có độ biến động thấp nhất trong nhóm. Phần lớn lợi nhuận hàng ngày của nó co cụm rất gần với 0.

  - Hòa Phát Group (đỏ) và Chứng khoán SSI (tím) có vẻ có đỉnh thấp hơn và phân phối bè ra hai bên hơn một chút. Điều này ngụ ý chúng có độ biến động cao hơn so với Vinamilk và Vietcombank.

  - Vingroup (xanh dương) và Vietcombank (xanh lá) nằm ở mức trung bình trong nhóm.

4.6 Kết luận chung

Kết luận & Liên kết các Biểu đồ:

- Mối liên hệ giữa rủi ro và tăng trưởng: Biểu đồ giá lịch sử đã giải thích hoàn hảo cho biểu đồ phân phối lợi nhuận. Cổ phiếu Vingroup có những cú tăng/giảm sốc trên biểu đồ giá, và đó chính là nguyên nhân tạo ra "đuôi dày" (nguy cơ biến động mạnh) trên biểu đồ phân phối.

- Sự ổn định của Vietcombank: Đường giá tăng trưởng bền vững của VCB tương ứng với một phân phối lợi nhuận tương đối gọn gàng hơn.

- Tính chu kỳ: Các con sóng lên xuống của HPG và SSI trên biểu đồ giá tạo ra một phân phối lợi nhuận rộng hơn, cho thấy mức độ biến động hàng ngày cao hơn so với các cổ phiếu ổn định như VCB hay VNM (trong giai đoạn hoàng kim).

5 Huấn luyện mô hình dự báo LSTM

5.1 Chuẩn bị dữ liệu

Dữ liệu sử dụng: cổ phiếu của CTCP Đầu tư Hạ tầng Kỹ thuật Thành phố Hồ Chí Minh (HOSE: CII) với dữ liệu gồm gần 3500 lần biến động được ghi nhận từ 2012 - 2025. Từ đây chúng ta chia dữ liệu đó ra làm 2 phần:

- Train data: dữ liệu từ 2012-2023 (2996 hàng dữ liệu)

- Test data: dữ liệu năm 2024 và 2025

Trước khi huấn luyện, dữ liệu được scale về [0,1] bằng MinMaxScaler, chia thành tập huấn luyện và kiểm thử:

```
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler(feature_range=(0,1))
training_set_scaled = sc.fit_transform(training_set)
```

5.2 Cấu trúc mô hình

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dropout, Dense
model = Sequential([
    LSTM(50, return_sequences=True, input_shape=(X_train.shape[1],
        1)),
    Dropout(0.2),
    LSTM(50, return_sequences=True),
    Dropout(0.2),
    LSTM(50, return_sequences=True),
    Dropout(0.2),
    LSTM(50),
    Dropout(0.2),
    Dense(1)
])
model.compile(optimizer='adam', loss='mean_squared_error')
```

5.3 Kết quả huấn luyện

Mô hình được huấn luyện 100 epochs, batch size 32. Sau khi dự báo, kết quả được đảo chuẩn hóa để đưa về đơn vị gốc:

```
predicted_price = model.predict(X_test)
predicted_price = sc.inverse_transform(predicted_price)
```

```
Model summary:
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 60, 50)            10400

 dropout (Dropout)           (None, 60, 50)            0

 lstm_1 (LSTM)               (None, 60, 50)            20200

 dropout_1 (Dropout)         (None, 60, 50)            0

 lstm_2 (LSTM)               (None, 60, 50)            20200

 dropout_2 (Dropout)         (None, 60, 50)            0

 lstm_3 (LSTM)               (None, 50)                20200

 dropout_3 (Dropout)         (None, 50)                0

 dense (Dense)               (None, 1)                 51

=================================================================
Total params: 71051 (277.54 KB)
...
Sample inputs (scaled) first 5: [0.19485774 0.20542939 0.20451579 0.21143305 0.19511877]
1/1 [==============================] - 0s 25ms/step
raw predict (first 5): [0.1842798  0.18801367 0.18922436 0.19088528 0.18762112]
after inverse transform: [13.679759  13.8228035 13.869185  13.932815  13.807765 ]
```

Figure 4: Thông số cụ thể của mô hình sau huấn luyện.

Một vài thông số của mô hình sau huấn luyện:

6 Phân tích kết quả từ mô hình dự đoán giá của mã CII

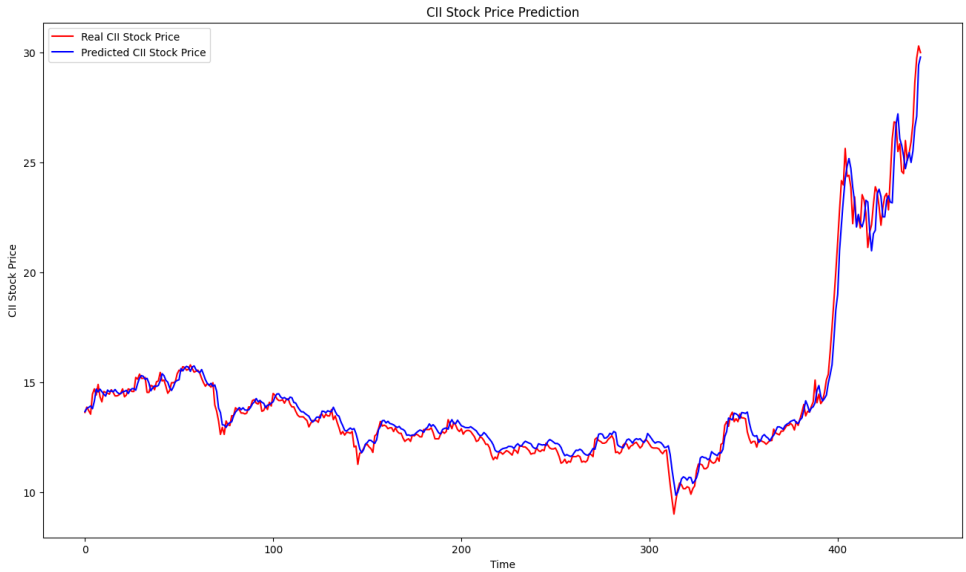

Figure 5: Biểu đồ giá so sánh giá dự đoán và thực tế của CII

Từ biểu đồ ta nhận thấy:

- Bám sát xu hướng chung: Đường dự đoán (màu xanh) bám rất sát với đường giá thực tế (màu đỏ) trong suốt quá trình. Mô hình đã nắm bắt thành công xu hướng chung của giá cổ phiếu, từ những giai đoạn đi ngang (sideways) cho đến giai đoạn tăng giá mạnh ở cuối biểu đồ

- Phản ứng tốt với biến động mạnh: Điểm ấn tượng nhất là mô hình đã dự đoán được cú tăng giá đột biến (từ khoảng thời gian 370 trở đi). Điều này cho thấy mô hình không chỉ học được các quy luật trong giai đoạn ổn định mà còn có khả năng nhận diện và phản ứng với các thay đổi lớn của thị trường.

- Có độ trễ nhỏ (Slight Lag): Nếu quan sát kỹ, có thể thấy đường dự đoán màu xanh thường có một độ trễ rất nhỏ, đi sau đường màu đỏ một chút. Đây là một đặc điểm rất phổ biến của các mô hình dự đoán chuỗi thời gian (như LSTM, GRU), vì chúng thường dự đoán giá trị của ngày mai dựa trên dữ liệu của ngày hôm nay và những ngày trước đó.

- Làm mượt các đỉnh/đáy nhọn: Trong những phiên có biến động mạnh, mô hình có xu hướng "làm mượt" các đỉnh và đáy, tức là giá dự đoán không hoàn toàn vươn tới mức cao nhất hoặc thấp nhất như giá thực tế. Tuy nhiên, mức độ chênh lệch này không lớn.

7 Kết luận và hướng phát triển

Kết luận: Đây là một mô hình dự báo rất khả quan, cho thấy khả năng ứng dụng thực tế. Kết quả này chứng tỏ mô hình đã học được các mẫu (patterns) phức tạp từ dữ liệu lịch sử để đưa ra dự đoán với độ tin cậy cao. Hệ thống Hadoop + Spark hoạt động ổn định trong môi trường Docker. Pipeline xử lý dữ liệu được tái hiện hoàn chỉnh. Mô hình LSTM dự báo giá cổ phiếu hoạt động tốt trên dữ liệu Việt Nam.

GitHub của tôi

Mã nguồn và notebook được công bố tại:
`https://github.com/EddyTryToCode/ASM-Assignment---BIG_DATA_UET`

Tài liệu tham khảo

1. Big Data Europe. *Docker Images for Hadoop and Spark*. Truy cập tại: `https://github.com/big-data-europe/docker-hadoop`
2. Thviet79. *Stock Price BigData - Hadoop Spark Project*. Truy cập tại: `https://github.com/thviet79/Stock-Price`
3. Apache Spark Official Documentation: `https://spark.apache.org/docs/latest/`
4. VN Stock PyPI: `https://pypi.org/project/vnstock/`
5. Jason Brownlee (2020). *Time Series Forecasting with LSTM in Keras*. Machine Learning Mastery. Truy cập tại: `https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-`