

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN



ĐỒ ÁN MÔN HỌC
ĐỀ TÀI

**Phân tích Tình hình tài chính của ba Ngân hàng TMCP tại
Việt Nam và Xây dựng Mô hình dự đoán giá cổ phiếu**

Môn học: Phân tích dữ liệu kinh doanh

Giảng viên hướng dẫn: Dr. Trần Văn Hải Triều

Lớp: IS403.P24

Nhóm sinh viên thực hiện: Nhóm 1

- | | |
|-------------------------------|-----------------------|
| 1. Lưu Bảo Uyên (Nhóm trưởng) | MSSV: 22521640 |
| 2. Lê Vy | MSSV: 22521703 |
| 3. Trần Lương Vân Nhi | MSSV: 22521044 |
| 4. Trương Nhật Quang | MSSV: 22521207 |
| 5. Nguyễn Anh Hải Ngọc | MSSV: 22520955 |

Mục lục

1	Giới thiệu	6
1.1	Bối cảnh	6
1.2	Định hướng và mục tiêu	6
2	Bộ dữ liệu	7
2.1	Nguồn dữ liệu	7
2.2	Giới thiệu bộ dữ liệu	7
2.3	Mô tả về bộ dữ liệu	7
3	Phân tích các chỉ số vĩ mô và tác động của chúng đến tình hình kinh tế của Việt Nam	9
3.1	Tăng trưởng GDP	9
3.2	Lạm phát (Chỉ số giá tiêu dùng)	10
3.3	Tỷ giá hối đoái	12
4	Phân tích các chỉ số doanh nghiệp	14
4.1	Ngân hàng thương mại cổ phần Ngoại thương Việt Nam (Vietcombank)	14
4.2	Ngân hàng thương mại cổ phần Đầu tư và Phát triển Việt Nam (BIDV)	17
4.3	Ngân hàng thương mại cổ phần Kỹ thương Việt Nam (Techcombank)	20
4.4	Đánh giá tổng quan	23
5	Lựa chọn thuộc tính	24
5.1	Loại bỏ các thuộc tính không liên quan	24
5.2	Loại bỏ các biến trùng lắp thông tin	24
5.3	Phân tích tương quan	25
5.4	Lựa chọn thuộc tính bằng Random Forest	26
6	Lý thuyết mô hình	27
6.1	Seasonal Autoregressive Integrated Average with Exogenous Regressors (SARIMAX)	27
6.2	Machine Learning	28
6.2.1	Linear Regression	28
6.2.2	Ridge Regression	28
6.2.3	Random Forest	29
6.2.4	Support Vector Machine (SVM)	29
6.3	Deep Learning	30
6.3.1	Long Short-term Memory (LSTM)	30
6.3.2	Gated Recurrent Units (GRU)	31
7	Thực nghiệm và phân tích kết quả	32
7.1	Xây dựng thực nghiệm	32
7.2	Kết quả thực nghiệm	32
7.2.1	Kết quả chỉ số đánh giá	32
7.2.1.1	Kết quả một bước	32
7.2.1.2	Kết quả nhiều bước	32
7.2.1.3	Nhận xét chỉ số đánh giá	33
7.2.2	Hiệu suất	34
7.2.2.1	Kết quả một bước	34
7.2.2.2	Kết quả nhiều bước	34
7.2.2.3	Đánh giá	34
7.2.3	Kết quả dự đoán	35
7.2.3.1	SARIMAX	35
7.2.3.2	Linear Regression	36
7.2.3.3	Ridge Regression	37
7.2.3.4	Random Forest	37

7.2.3.5	Support Vector Machine	38
7.2.3.6	LSTM	39
7.2.3.7	GRU	40
7.3	Phân tích kết quả	40
7.3.1	Chỉ số đánh giá	41
7.3.1.1	Dự đoán 1 bước	41
7.3.1.2	Dự đoán nhiều bước	42
7.3.2	Hiệu suất	43
7.3.2.1	Thời gian huấn luyện	44
7.3.2.2	Thời gian tính toán	44
7.4	Đánh giá chung	45
8	Tổng kết	46

Danh sách hình vẽ

1.1	Quy trình thực hiện	7
3.1	Tăng trưởng GDP bình quân theo đầu người tại Việt Nam giai đoạn 2022-2024	10
3.2	Tốc độ tăng chỉ số tiêu dùng của Việt Nam giai đoạn 2022-2024	10
3.3	Biểu đồ thống kê chỉ số tiêu dùng mỗi tháng	11
3.4	Biểu đồ tỷ giá USD thị trường giai đoạn 2022-2024	12
3.5	Biểu đồ tỷ giá USD thị trường năm 2022	12
3.6	Biểu đồ tỷ giá USD thị trường năm 2023	13
3.7	Biểu đồ tỷ giá USD thị trường năm 2024	13
4.1	Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025	14
4.2	Khối lượng giao dịch theo ngày của ngân hàng Vietcombank giai đoạn 2022-2025	14
4.3	Doanh thu của ngân hàng Vietcombank giai đoạn 2022-2025	15
4.4	Biểu đồ tăng trưởng doanh thu của ngân hàng Vietcombank giai đoạn 2022-2025	15
4.5	Biểu đồ tăng trưởng lợi nhuận của ngân hàng Vietcombank giai đoạn 2022-2025	15
4.6	Cổ tức của ngân hàng Vietcombank giai đoạn 2022-2025	16
4.7	Giá trị ROE và ROA của ngân hàng Vietcombank giai đoạn 2022-2025	16
4.8	Vốn hóa của ngân hàng Vietcombank giai đoạn 2022-2025	16
4.9	Giá trị P/Cash của ngân hàng Vietcombank giai đoạn 2022-2025	17
4.10	Giá cổ phiếu của ngân hàng BIDV giai đoạn 2022-2025	17
4.11	Khối lượng giao dịch theo ngày của ngân hàng BIDV giai đoạn 2022-2025	17
4.12	Doanh thu của ngân hàng BIDV giai đoạn 2022-2025	18
4.13	Biểu đồ tăng trưởng doanh thu của ngân hàng BIDV giai đoạn 2022-2025	18
4.14	Biểu đồ tăng trưởng lợi nhuận của ngân hàng BIDV giai đoạn 2022-2025	18
4.15	Cổ tức của ngân hàng BIDV giai đoạn 2022-2025	19
4.16	Giá trị ROE và ROA của ngân hàng BIDV giai đoạn 2022-2025	19
4.17	Vốn hóa của ngân hàng BIDV giai đoạn 2022-2025	19
4.18	Giá trị P/Cash của ngân hàng BIDV giai đoạn 2022-2025	20
4.19	Giá cổ phiếu của ngân hàng Techcombank giai đoạn 2022-2025	20
4.20	Khối lượng giao dịch theo ngày của ngân hàng Techcombank giai đoạn 2022-2025	20
4.21	Doanh thu của ngân hàng Techcombank giai đoạn 2022-2025	21
4.22	Biểu đồ tăng trưởng doanh thu của ngân hàng Techcombank giai đoạn 2022-2025	21
4.23	Biểu đồ tăng trưởng lợi nhuận của ngân hàng Techcombank giai đoạn 2022-2025	21
4.24	Cổ tức của ngân hàng Techcombank giai đoạn 2022-2025	22
4.25	Giá trị ROE và ROA của ngân hàng Techcombank giai đoạn 2022-2025	22
4.26	Vốn hóa của ngân hàng Techcombank giai đoạn 2022-2025	22
4.27	Giá trị P/Cash của ngân hàng Vietcombank giai đoạn 2022-2025	23
4.28	Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025	23
4.29	Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025	23
4.30	Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025	24
5.1	Ví dụ 1: Biểu đồ tỷ giá EURO của Vietcombank	24
5.2	Ví dụ 2: Biểu đồ tỷ giá Yên Nhật của BIDV	25
5.3	Ví dụ: Heatmap về tương quan giữa các biến của bộ dữ liệu Vietcombank	25
5.4	Ví dụ: Tương quan giữa các biến với biến mục tiêu của Vietcombank	26
5.5	Ví dụ: Độ quan trọng của từng biến của ngân hàng Vietcombank	27
6.1	SARIMAX	27
6.2	Linear Regression	28
6.3	Ridge Regression	28
6.4	Random Forest	29
6.5	SVM	30
6.6	LSTM	31
6.7	GRU	31
7.1	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình SARIMAX	35

7.2	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình SARIMAX	35
7.3	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình SARIMAX	35
7.4	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Linear Regression	36
7.5	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Linear Regression	36
7.6	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Linear Regression	36
7.7	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Ridge Regression	37
7.8	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Ridge Regression	37
7.9	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Ridge Regression	37
7.10	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Random Forest	37
7.11	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Random Forest	38
7.12	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Random Forest	38
7.13	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình SVM	38
7.14	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình SVM	38
7.15	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình SVM	39
7.16	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình LSTM	39
7.17	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình LSTM	39
7.18	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình LSTM	39
7.19	Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình GRU	40
7.20	Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình GRU	40
7.21	Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình GRU	40
7.22	ANOVA và Tukey's HSD với MSE của các chỉ số đánh giá với các mô hình dự đoán một bước . .	41
7.23	ANOVA và Tukey's HSD với RMSE của các chỉ số đánh giá với các mô hình dự đoán một bước . .	41
7.24	ANOVA và Tukey's HSD với MAE của các chỉ số đánh giá với các mô hình dự đoán một bước . .	42
7.25	ANOVA và Tukey's HSD với MAPE của các chỉ số đánh giá với các mô hình dự đoán một bước . .	42
7.26	ANOVA và Tukey's HSD với MSE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước . .	42
7.27	ANOVA và Tukey's HSD với RMSE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước . .	43
7.28	ANOVA và Tukey's HSD với MAE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước . .	43
7.29	ANOVA và Tukey's HSD với MAPE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước . .	43
7.30	ANOVA và Tukey's HSD với thời gian huấn luyện của các mô hình dự đoán một bước	44
7.31	ANOVA và Tukey's HSD với thời gian huấn luyện của các mô hình dự đoán nhiều bước	44
7.32	ANOVA và Tukey's HSD với thời gian tính toán của các mô hình dự đoán một bước	45
7.33	ANOVA và Tukey's HSD với thời gian tính toán của các mô hình dự đoán nhiều bước	45

Danh sách bảng

2.1	Mô tả chi tiết thuộc tính	9
3.1	Tăng trưởng GDP theo cơ cấu khu vực tại Việt Nam giai đoạn 2022-2024	9
7.1	So sánh chỉ số MSE giữa các mô hình dự đoán một bước	32
7.2	So sánh chỉ số RMSE giữa các mô hình dự đoán một bước	32
7.3	So sánh chỉ số MAE giữa các mô hình dự đoán một bước	32
7.4	So sánh chỉ số MAPE giữa các mô hình dự đoán một bước	32
7.5	So sánh chỉ số MSE giữa các mô hình dự đoán nhiều bước	33
7.6	So sánh chỉ số RMSE giữa các mô hình dự đoán nhiều bước	33
7.7	So sánh chỉ số MAE giữa các mô hình dự đoán nhiều bước	33
7.8	So sánh chỉ số MAPE giữa các mô hình dự đoán nhiều bước	33
7.9	So sánh thời gian huấn luyện giữa các mô hình dự đoán một bước	34
7.10	So sánh thời gian tính toán (trên tập test) giữa các mô hình dự đoán một bước	34
7.11	So sánh thời gian huấn luyện giữa các mô hình dự đoán nhiều bước	34
7.12	So sánh thời gian tính toán (trên tập test) giữa các mô hình dự đoán nhiều bước	34

1 Giới thiệu

1.1 Bối cảnh

"Tuần giao dịch từ 31/3 đến 4/4/2025 là một tuần đầy biến động đối với nhóm cổ phiếu ngân hàng. Đầu tuần, nhóm cổ phiếu ngân hàng diễn biến khá tích cực, đặc biệt trong phiên ngày 1/4, sắc xanh bao trùm toàn ngành. Tuy nhiên, diễn biến tích cực không kéo dài. Đến hai phiên giao dịch cuối tuần, nhóm cổ phiếu ngân hàng lao dốc mạnh, chịu ảnh hưởng nặng nề từ tâm lý tiêu cực trên thị trường. Ngày 3/4, sau thông tin Tổng thống Mỹ Donald Trump công bố áp thuế đối ứng với 180 nền kinh tế, bao gồm Việt Nam, VN-Index giảm kỷ lục 88 điểm (6,68%), đánh dấu phiên giao dịch tệ nhất trong lịch sử thị trường chứng khoán Việt Nam. Nhóm cổ phiếu ngân hàng không thoát khỏi làn sóng bán tháo, với hàng loạt mã giảm kịch biên độ, như VCB, BID, CTG, VPB, và TCB,... đều giảm 7%, chạm mức sàn. Sang phiên giao dịch cuối tuần ngày 4/4, cổ phiếu ngân hàng tiếp tục biến động mạnh. Buổi sáng, sắc đỏ bao trùm với nhiều mã giảm gần hết biên độ. Tuy nhiên, lực cầu bắt đáy xuất hiện trong phiên chiều đã giúp thị trường phục hồi nhẹ. Một số mã chuyển sang sắc xanh. Hầu hết các mã ngân hàng thu hẹp mức giảm so với buổi sáng." [1]

Có thể thấy, cổ phiếu là một trong những loại hình đầu tư phổ biến trên thị trường tài chính. Đặc biệt, cổ phiếu ngân hàng ngày càng thu hút sự quan tâm lớn từ các nhà đầu tư, không chỉ tại Việt Nam mà còn trên toàn thế giới. Đây là một trong những loại cổ phiếu dễ tiếp cận đối với các nhà đầu tư mới bởi tính dễ hiểu và dễ theo dõi trên qua các chỉ số và báo cáo tài chính được cập nhật thường xuyên và công khai. Hơn thế, cổ phiếu của ngân hàng cũng phản ánh rất rõ tình hình của thị trường kinh tế. Tuy là lĩnh vực rất thu hút vì khả năng sinh lời cao đầy hấp dẫn, song nó cũng chứa đựng rất nhiều rủi ro tiềm ẩn. Thị trường chứng khoán nói chung hay cổ phiếu ngân hàng nói riêng đều có những biến động đột ngột. Sự biến động này rất phức tạp, bởi phải chịu tác động của nhiều yếu tố khác nhau như các chính sách tiền tệ, các thay đổi của thị trường kinh tế, các sự kiện chính trị xã hội - khủng hoảng tài chính, các đặc thù của ngành và doanh nghiệp, tâm lý của nhà đầu tư và còn nhiều yếu tố khác. Có thể thấy, để đầu tư thành công, các nhà đầu tư không chỉ cần phải có kiến thức về thị trường tài chính, hiểu rõ các yếu tố tác động mà còn phải có khả năng dự đoán và phân tích rủi ro. Chính vì sự đa dạng và khó đoán này mà việc đưa ra quyết định đầu tư gặp rất nhiều khó khăn.

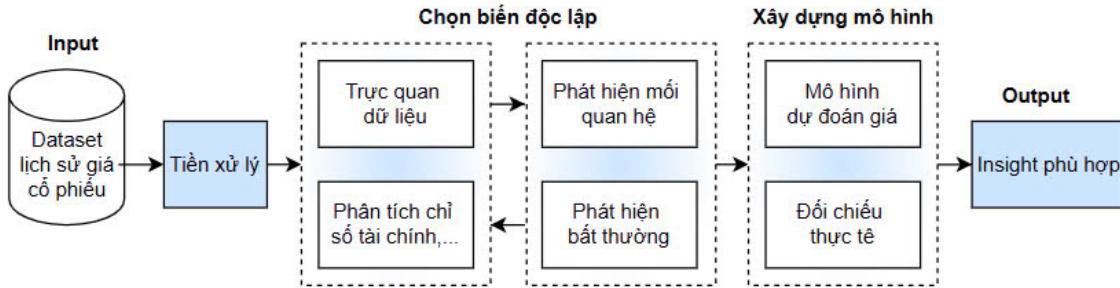
1.2 Định hướng và mục tiêu

Chúng tôi đề xuất ý tưởng thực hiện phân tích tình hình tài chính của Việt Nam cùng với tình hình giá cổ phiếu để có thể đánh giá mức độ tác động của các chỉ số kinh tế đối với doanh nghiệp nhằm xây dựng mô hình dự đoán sự biến động về giá cổ phiếu trên thị trường và đưa ra được các lý giải phù hợp. Đây là một đề tài kết hợp giữa lĩnh vực công nghệ thông tin và tài chính, với mong muốn sự hiệu quả của mô hình có thể mang lại những đóng góp có ý nghĩa cho các nhà đầu tư cũng như doanh nghiệp trong việc đưa ra các quyết định, chiến lược phù hợp, từ đó nâng cao lợi nhuận và giảm thiểu rủi ro. Ngoài lợi ích của các nhà đầu tư, chúng tôi cũng mong muốn đề tài này góp phần đưa ra các thông tin hữu ích cho các doanh nghiệp và thị trường chứng khoán. Tuy nhiên kết quả dự đoán của mô hình chỉ là giá trị mang tính chất tham khảo vì giá cổ phiếu rất dễ bị tác động bởi nhiều yếu tố khó kiểm soát. Vậy nên, sau khi có kết quả chúng tôi sẽ kết hợp phân tích kết quả dự đoán của mô hình cùng các yếu tố ngoại sinh trong thực tế để có thể đưa ra insight chính xác nhất.

Với mong muốn hiểu rõ hơn về sự tác động của hoạt động tài chính với sự biến động của giá cổ phiếu nên thay vì trải rộng toàn ngành, chúng tôi sẽ lựa chọn và phân tích chuyên sâu một vài doanh nghiệp cụ thể. Để phù hợp với đề tài đưa ra, chúng tôi đề xuất lựa chọn ba ngân hàng tiêu biểu với những quy mô, định hướng và xu hướng phát triển tài chính nổi bật trong ngành ngân hàng hiện nay, bao gồm:

- Ngân hàng thương mại cổ phần Ngoại thương Việt Nam (Vietcombank), đây là ngân hàng lớn nhất trên thị trường chứng khoán Việt Nam tính theo vốn hóa. Hiện tại, Ngân hàng Nhà nước Việt Nam nắm giữ 75% cổ phần và là cổ đông lớn nhất.
- Ngân hàng thương mại cổ phần Đầu tư và Phát triển Việt Nam (BIDV), ngân hàng liên tiếp 7 năm được đánh giá là ngân hàng Small and Medium Enterprises (SME) hàng đầu Việt Nam với nhiều chính sách phù hợp thu hút phân khúc khách hàng.
- Ngân hàng thương mại cổ phần Kỹ thương Việt Nam (Techcombank), ngân hàng tiên phong trong việc áp dụng công nghệ, đầu tư mạnh mẽ các chiến lược số hóa mang lại hiệu quả tài chính tốt.

Chúng tôi sẽ thu thập dữ liệu về giá cổ phiếu trong quá khứ của các ngân hàng, thực hiện tiền xử lý dữ liệu để có một bộ dữ liệu hoàn chỉnh. Tiếp theo đó, chúng tôi sử dụng các thư viện sklearn, matplotlib, plotly để trực quan dữ liệu, kết hợp phân tích các chỉ số thực tế để phát hiện mối quan hệ giữa các đặc trưng, đánh giá những bất thường. Bên cạnh đó, chúng tôi cũng kết hợp đánh giá các yếu tố ngoại sinh của thị trường và cụ thể từng ngân hàng để tìm ra các điểm nổi bật. Dựa vào kết quả của hai quy trình trên, kết hợp với Random Forest để chọn lựa những thuộc tính có ảnh hưởng đến mục tiêu. Sau đó, vận dụng các kiến thức đã học để xây dựng mô hình machine learning, deep learning để dự đoán giá cổ phiếu trong tương lai, thực nghiệm và đánh giá dựa trên các thông số kiểm định để chọn ra những mô hình phù hợp nhất. Cuối cùng cần phân tích lại kết quả đã dự đoán, đổi chiều so với thực tế và đưa ra những luận giải và insight phù hợp với bài.



Mục tiêu của bài toán là xây dựng được mô hình machine learning, deep learning hiện đại dự báo giá cổ phiếu với mức độ chính xác cao nhất có thể, diễn giải và phân tích được tất cả insight liên quan đến kết quả.

Định nghĩa bài toán:

- Đầu vào của bài toán: Dữ liệu quá khứ của cổ phiếu ngân hàng.
- Đầu ra của bài toán: Giá cổ phiếu ngày tiếp theo.

2 Bộ dữ liệu

2.1 Nguồn dữ liệu

Bộ dữ liệu cho đề tài được thu thập từ ba nguồn. Thông tin liên quan đến các chỉ số doanh nghiệp và cổ phiếu của ba ngân hàng Vietcombank, BIDV, Techcombank được thu thập từ thư viện VnStock [2] trên Python. Đây là bộ giải pháp mã nguồn mở toàn diện cho phân tích và tự động hóa đầu tư chứng khoán. Bên cạnh đó, nhóm còn sử dụng công cụ Selenium và API để thu thập dữ liệu từ các trang web. Trong đó, các chỉ số kinh tế vĩ mô như chỉ số tiêu dùng, chỉ số thu nhập bình quân đầu người được lấy trên trang VietstockFinance [3], đây là website cung cấp thông tin đầu tư chứng khoán hàng đầu, dựa trên nền tảng dữ liệu tài chính chính xác và toàn diện. Còn giá trị tỷ giá hối đoái ngoại tệ sẽ được thu thập trên các trang chủ chính thức của các ngân hàng.

2.2 Giới thiệu bộ dữ liệu

Dữ liệu được thu thập từ ngày 01/01/2022 đến 31/03/2025. Bộ dữ liệu lưu trữ các thông tin về ba ngân hàng Vietcombank, BIDV, Techcombank, cụ thể bao gồm: giá cả cao, thấp, mở cửa, đóng cửa trong ngày của cổ phiếu, các thông tin vốn hóa, lợi nhuận của cổ phiếu, kết quả kinh doanh của doanh nghiệp, hiệu quả tài chính. Ngoài ra, bộ dữ liệu còn có các chỉ số vĩ mô của thị trường như chỉ số tiêu dùng, chỉ số thu nhập bình quân đầu người và tỷ giá hối đoái ngoại tệ. Bộ dữ liệu sau khi quá trình tổng hợp bao gồm 58 thuộc tính, phản ánh đầy đủ các khía cạnh kinh tế từ vi mô đến vĩ mô có ảnh hưởng đến giá cổ phiếu.

Link drive: https://drive.google.com/drive/folders/1mxejhZZimoHlkHn-GFO4StmHz_4qL8dm

2.3 Mô tả về bộ dữ liệu

Các thuộc tính và mô tả về thuộc tính được thể hiện trong bảng sau:

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
-----	------------	--------------	-------

1	Time	Datetime	Ngày ghi nhận dữ liệu
2	Open	Float	Giá mở cửa
3	High	Float	Giá cao nhất trong ngày
4	Low	Float	Giá thấp nhất trong ngày
5	Close	Float	Giá đóng cửa
6	Volume	Float	Khối lượng giao dịch trong ngày
7	Previous_GDP	Float	Giá trị tổng sản phẩm quốc nội kỳ trước
8	Previous_CPI	Float	Chỉ số tiêu dùng kỳ trước
9	Previous_AUD_Cash	Float	Tỷ giá mua của Dollar Úc
10	Previous_AUD_Sell	Float	Tỷ giá bán của Dollar Úc
11	Previous_AUD_Transfer	Float	Tỷ giá chuyển khoản của Dollar Úc
12	Previous_CAD_Cash	Float	Tỷ giá mua của Dollar Canada
13	Previous_CAD_Sell	Float	Tỷ giá bán của Dollar Canada
14	Previous_CAD_Transfer	Float	Tỷ giá chuyển khoản của Dollar Canada
15	Previous_CHF_Cash	Float	Tỷ giá mua của Franc Thụy Sĩ
16	Previous_CHF_Sell	Float	Tỷ giá bán của Franc Thụy Sĩ
17	Previous_CHF_Transfer	Float	Tỷ giá chuyển khoản của Franc Thụy Sĩ
18	Previous_EUR_Cash	Float	Tỷ giá mua của Euro
19	Previous_EUR_Sell	Float	Tỷ giá bán của Euro
20	Previous_EUR_Transfer	Float	Tỷ giá chuyển khoản của Euro
21	Previous_GBP_Cash	Float	Tỷ giá mua của Pound Anh
22	Previous_GBP_Sell	Float	Tỷ giá bán của Pound Anh
23	Previous_GBP_Transfer	Float	Tỷ giá chuyển khoản của Pound Anh
24	Previous_JPY_Cash	Float	Tỷ giá mua của Yên Nhật
25	Previous_JPY_Sell	Float	Tỷ giá bán của Yên Nhật
26	Previous_JPY_Transfer	Float	Tỷ giá chuyển khoản của Yên Nhật
27	Previous_SGD_Cash	Float	Tỷ giá mua của Dollar Singapore
28	Previous_SGD_Sell	Float	Tỷ giá bán của Dollar Singapore
29	Previous_SGD_Transfer	Float	Tỷ giá chuyển khoản của Dollar Singapore
30	Previous_THB_Cash	Float	Tỷ giá mua của Baht Thái
31	Previous_THB_Sell	Float	Tỷ giá bán của Baht Thái
32	Previous_THB_Transfer	Float	Tỷ giá chuyển khoản của Baht Thái
33	Previous_USD_Cash	Float	Tỷ giá mua của Dollar Mỹ
34	Doanh thu (đồng)	Float	Tổng doanh thu của doanh nghiệp trong kỳ
35	Tăng trưởng doanh thu (%)	Float	Tỷ lệ tăng/giảm doanh thu so với kỳ trước
36	Lợi nhuận sau thuế của Cổ đông công ty mẹ (đồng)	Float	Lợi nhuận sau thuế của Cổ đông công ty mẹ
37	Tăng trưởng lợi nhuận (%)	Float	Tỷ lệ tăng/giảm lợi nhuận so với kỳ trước
38	LN trước thuế	Float	Lợi nhuận trước thuế
39	Lợi nhuận thuần	Float	Lợi nhuận sau khi trừ các phí liên quan
40	Cổ tức đã nhận	Float	Số tiền cổ tức đã nhận được trong kỳ
41	Biên lợi nhuận ròng (%)	Float	Tỷ suất lợi nhuận ròng trên doanh thu

42	ROE (%)	Float	Tỷ suất sinh lời trên vốn chủ sở hữu
43	ROIC (%)	Float	Tỷ suất sinh lời trên vốn đầu tư
44	ROA (%)	Float	Tỷ suất sinh lời trên tổng tài sản
45	Tỷ suất cổ tức (%)	Float	Lợi suất cổ tức theo giá thị trường
46	Vốn hóa (Tỷ đồng)	Float	Giá trị thị trường của công ty
47	Số CP lưu hành (Triệu CP)	Float	Số lượng cổ phiếu đang lưu hành
48	P/E	Float	Hệ số giá trên lợi nhuận
49	P/B	Float	Hệ số giá trên giá trị sổ sách
50	P/S	Float	Hệ số giá trên doanh thu
51	P/Cash Flow	Float	Hệ số giá trên dòng tiền
52	EPS (VND)	Float	Lợi nhuận trên mỗi cổ phiếu
53	BVPS (VND)	Float	Giá trị sổ sách trên mỗi cổ phiếu
54	Tổng cộng tài sản (đồng)	Float	Tổng tài sản của doanh nghiệp
55	Tổng cộng nguồn vốn (đồng)	Float	Tổng nguồn vốn huy động
56	Nợ phải trả (đồng)	Float	Tổng nợ của doanh nghiệp
57	Vốn chủ sở hữu (đồng)	Float	Vốn thuộc về cổ đông công ty
58	Lợi ích của cổ đông	Float	Lợi ích kinh tế dành cho cổ đông

Bảng 2.1: Mô tả chi tiết thuộc tính

Bộ dữ liệu sẽ được tiền xử lý và lựa chọn những thuộc tính phù hợp với mục tiêu tạo ra một tập dữ liệu rõ ràng và phù hợp để phục vụ cho bài toán.

3 Phân tích các chỉ số vĩ mô và tác động của chúng đến tình hình kinh tế của Việt Nam

3.1 Tăng trưởng GDP

Gross Domestic Product (GDP) [4] là tổng sản phẩm nội địa hay tổng sản phẩm quốc nội. Đây là một chỉ tiêu dùng để đo lường tổng giá trị thị trường của tất cả các hàng hoá và dịch vụ cuối cùng được sản xuất ra trong phạm vi một lãnh thổ quốc gia trong một thời kỳ nhất định.

	2022	2023	2024
Tổng GDP	8.12	5.05	7.09
Nông nghiệp	3.48	3.83	3.27
Công nghiệp	7.87	3.74	8.24
Dịch vụ	10.11	6.82	7.38

Bảng 3.1: Tăng trưởng GDP theo cơ cấu khu vực tại Việt Nam giai đoạn 2022-2024

Trong năm 2022, nền kinh tế đã dần khôi phục mạnh mẽ sau đại dịch, với GDP ước tính tăng 8.02% so với năm 2021. Cụ thể, trong các quý của năm, tăng trưởng lần lượt đạt 5.05% (quý I), 7.83% (quý II), 13.71% (quý III), và 5.92% (quý IV), mức tăng cao nhất trong giai đoạn 2011-2022. Xét theo các khu vực, khu vực nông, lâm nghiệp và thủy sản tăng 3.36%, đóng góp 5.11% vào tăng trưởng tổng giá trị tăng thêm của nền kinh tế. Khu vực công nghiệp và xây dựng tăng 7.78%, đóng góp 38.24%, trong khi khu vực dịch vụ đạt mức tăng 9.99%, đóng góp 56.65%.

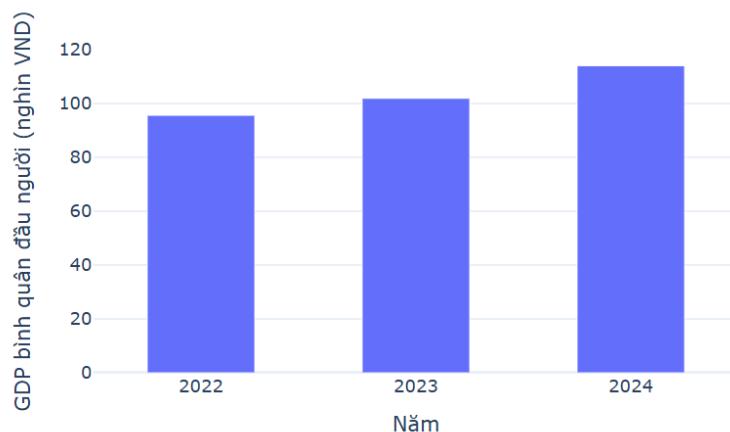
Trong lĩnh vực ngân hàng và chứng khoán, các chỉ số cũng có sự cải thiện so với cuối năm 2021. Tổng phuơng tiện thanh toán tăng 3.85%, huy động vốn của các tổ chức tín dụng tăng 5.99%, và tăng trưởng tín dụng của nền kinh tế đạt 12.87%. Tuy nhiên, trên thị trường chứng khoán, giá trị giao dịch bình quân trong năm 2022 giảm so

với năm trước: giá trị giao dịch bình quân trên thị trường cổ phiếu đạt 20.410 tỷ đồng/phíên, giảm 23.3%, trong khi trên thị trường trái phiếu, con số này là 7.737 tỷ đồng/phíên, giảm 32.2%.

Bước sang năm 2023, nền kinh tế tiếp tục tăng trưởng tuy nhiên không quá rõ sự vượt trội, với GDP ước tính tăng 5.05% so với năm 2022. Tốc độ này chỉ cao hơn mức tăng của năm 2020 và 2021 (2.87% và 2.55%, tương ứng). Cơ cấu tăng trưởng GDP trong năm 2023 có sự thay đổi, với khu vực nông, lâm nghiệp và thủy sản tăng 3.83% (đóng góp 8.84%), khu vực công nghiệp và xây dựng tăng 3.74% (đóng góp 28.87%), và khu vực dịch vụ tăng 6.82% (đóng góp 62.29%). Trong lĩnh vực ngân hàng, tổng thương mại thanh toán tăng 10.03%, huy động vốn của các tổ chức tín dụng tăng 10.85%, và tăng trưởng tín dụng của nền kinh tế đạt 11.09%.

Đến năm 2024, GDP ước tính tăng 7.09% so với năm trước, thấp hơn mức tăng của các năm 2018, 2019 và 2022. Trong cơ cấu tăng trưởng GDP, khu vực nông, lâm nghiệp và thủy sản ước tính tăng 3.27%, đóng góp 5.37%, khu vực công nghiệp và xây dựng tăng 8.24%, đóng góp 45.17%, và khu vực dịch vụ tăng 7.38%, đóng góp 49.46%.

Trong hoạt động ngân hàng, bảo hiểm và thị trường chứng khoán, tính đến tháng 12/2024, tổng thương mại thanh toán đã tăng 9.42% so với cuối năm 2023, huy động vốn của các tổ chức tín dụng tăng 9.06%, và tăng trưởng tín dụng của nền kinh tế đạt 13.82%.

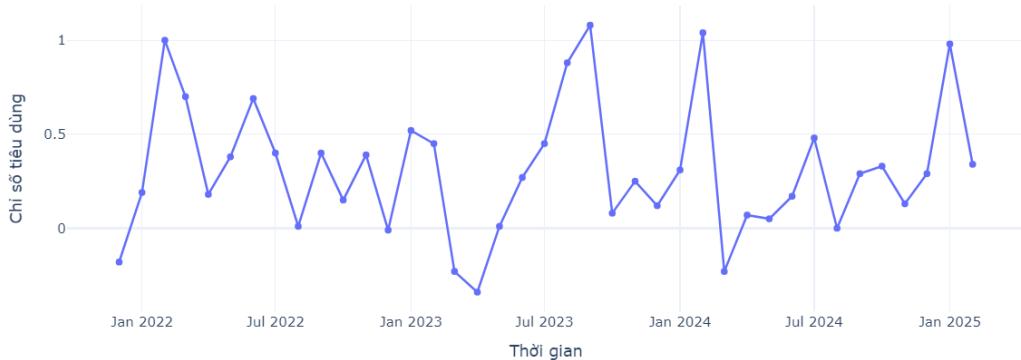


Hình 3.1: Tăng trưởng GDP bình quân theo đầu người tại Việt Nam giai đoạn 2022-2024

GDP bình quân đầu người của Việt Nam năm 2024 duy trì đà tăng trưởng ổn định, đạt mức 114 triệu/người, tăng 12 triệu so với năm 2023.

3.2 Lạm phát (Chỉ số giá tiêu dùng)

Chỉ số giá tiêu dùng (Consumer Price Index - CPI) [5] là chỉ số tính theo phần trăm để phản ánh mức thay đổi tương đối của giá hàng tiêu dùng theo thời gian. Đây là chỉ số quan trọng trong việc đánh giá sự tăng giảm của chi phí sinh hoạt và sức mua của người tiêu dùng. Đồng thời chỉ số CPI sẽ được dùng để đo tỷ lệ lạm phát của một quốc gia trong khoảng thời gian nhất định. Chỉ số CPI biến động sẽ giúp các nhà kinh tế xác định về tỷ lệ lạm phát tăng hay giảm.



Hình 3.2: Tốc độ tăng chỉ số tiêu dùng của Việt Nam giai đoạn 2022-2024

Năm 2022, dù tình hình dịch Covid-19 phần nào đã được kiểm soát, song nền kinh tế Việt Nam vẫn còn đối mặt với nhiều khó khăn đặc biệt là việc chuỗi cung ứng sản xuất, tiêu dùng tiếp tục bị đứt gãy trong suốt thời gian xảy ra đại dịch. Xung đột giữa Nga - Ukraina và nhiều yếu tố khác khiến Việt Nam phải đổi mới với những thách thức về giá cả năng lượng và hàng hóa tăng cao; chính sách tài khóa, tiền tệ được nhiều nền kinh tế điều chỉnh theo hướng thắt chặt để kiềm chế lạm phát đã tác động tới khả năng phục hồi và tăng trưởng kinh tế, một số nền kinh tế có dấu hiệu suy thoái.

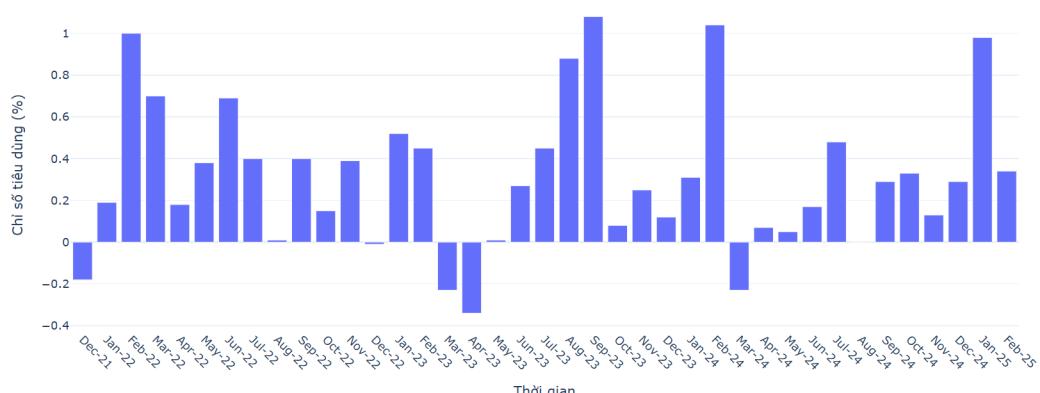
Trong bối cảnh đó tình hình lạm phát thế giới tiếp tục tăng cao. Việt Nam thuộc nhóm các nước có mức lạm phát thấp so với mặt bằng chung khi CPI tháng 12/2022 tăng 4,55% so với cùng kỳ năm trước nhưng vẫn cao hơn mức lạm phát của Nhật Bản và Trung Quốc. Trong nước, kinh tế đang dần phục hồi. Tuy nhu cầu sản xuất hàng hóa phục vụ tiêu dùng và xuất khẩu cùng với tác động của giá hàng hóa thế giới đã đẩy giá hàng hóa và dịch vụ thiết yếu tăng, nhưng Chính phủ đã có những chỉ đạo kịp thời, thực hiện nhiều chính sách và đồng bộ các giải pháp nhằm kiểm soát mặt bằng giá cơ bản, hạn chế những tác động tiêu cực đến phát triển kinh tế – xã hội. Một số chính sách hiệu quả nhằm giảm áp lực lạm phát như: giảm thuế giá trị gia tăng với một số nhóm hàng hóa và dịch vụ, giảm 50% mức thuế bảo vệ môi trường đối với nhiên liệu bay và xăng dầu... Chỉ số giá tiêu dùng bình quân năm 2022 tăng 3,15% so với năm trước, đạt mục tiêu Quốc hội đề ra trong bối cảnh một năm nhiều biến động khó lường.

Thị trường hàng hóa thế giới năm 2023 có nhiều biến động và chịu ảnh hưởng bởi các yếu tố kinh tế, chính trị, xã hội. Xung đột quân sự Nga – Ukraina vẫn tiếp diễn cùng với bất ổn giá tăng tại Trung Đông, nhiều quốc gia duy trì chính sách tiền tệ thắt chặt khiến kinh tế tăng trưởng chậm, cùng lúc đó thị trường tài chính tiền tệ, bất động sản tại một số nước tiềm ẩn nhiều rủi ro. Lạm phát toàn cầu có xu hướng giảm dần sau thời gian áp dụng các chính sách kiềm chế lạm phát của các nước. Nhưng so với mục tiêu dài hạn, mức lạm phát hiện tại vẫn ở mức cao đối với nhiều quốc gia. Việt Nam tiếp tục thuộc nhóm các nước kiểm soát tốt lạm phát khi CPI tháng 12/2023 tăng 3,58% so với cùng kỳ năm trước.

Chính phủ, Thủ tướng Chính phủ Việt Nam đã chủ động, quyết liệt, sát sao chỉ đạo triển khai nhiều giải pháp nhằm tháo gỡ khó khăn, thúc đẩy tăng trưởng, giữ vững ổn định kinh tế vĩ mô, kiểm soát lạm phát, đảm bảo các cân đối lớn của nền kinh tế. Nhiều giải pháp được tích cực triển khai như: giảm mặt bằng lãi suất cho vay; thúc đẩy giải ngân vốn đầu tư công; giảm thuế giá trị gia tăng với một số nhóm hàng hóa và dịch vụ... Nhờ đó, thị trường các mặt hàng thiết yếu không có biến động bất thường, nguồn cung được bảo đảm, giá hàng hóa tăng giảm đan xen. Bình quân CPI năm 2023 tăng 3,25% so với năm 2022.

Sang đến năm 2024, thị trường hàng hóa thế giới tiếp tục có nhiều biến động do ảnh hưởng bởi các yếu tố chính trị, kinh tế, xã hội của các quốc gia. Xung đột quân sự, biến động chính trị tại một số nước khiến cho kinh tế, thương mại toàn cầu phục hồi chậm, thiếu vững chắc, tỷ giá, lãi suất biến động khó lường. Xu hướng cắt giảm lãi suất của một số ngân hàng trung ương lớn trên thế giới tiếp tục mở rộng do lạm phát ngày một cao.

Trong nước, Thủ tướng Chính phủ vẫn quyết liệt triển khai nhiều giải pháp nhằm thúc đẩy tăng trưởng, giữ vững ổn định kinh tế vĩ mô, kiểm soát lạm phát như: đảm bảo thông suốt hoạt động cung ứng, lưu thông, phân phối hàng hóa, dịch vụ; giảm mặt bằng lãi suất cho vay, ổn định thị trường ngoại hối; giảm thuế giá trị gia tăng đối với một số nhóm hàng hóa và dịch vụ... Nhìn chung, giá hàng hóa và dịch vụ trên thị trường không có biến động quá bất thường, lạm phát trong tầm kiểm soát. Bình quân năm 2024, CPI tăng 3,63% so với năm 2023; lạm phát cơ bản tăng 2,71%.



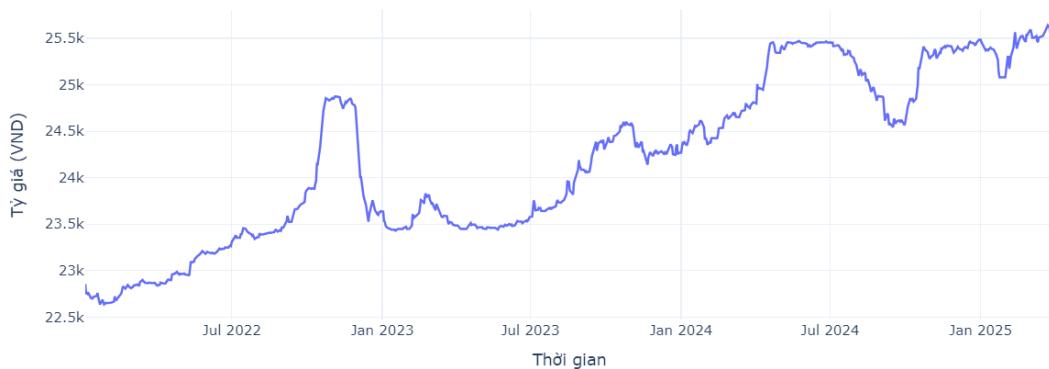
Hình 3.3: Biểu đồ thống kê chỉ số tiêu dùng mỗi tháng

Dựa trên biểu đồ thống kê CPI mỗi tháng trong khoảng từ tháng 1 năm 2022 đến tháng 3 năm 2025, có thể nhận thấy một số xu hướng và đặc điểm đáng chú ý về sự biến động của chỉ số giá tiêu dùng (CPI) như sau:

Trong khoảng thời gian này, CPI có sự biến động khá lớn giữa các tháng. Điều này phản ánh sự thay đổi không đồng đều, nền kinh tế bị ảnh hưởng bởi nhiều yếu tố như biến động về giá cả hàng hóa, năng lượng, các chính sách kinh tế siết chặt và tình hình chính trị của các nước đang trong thời kỳ thiếu ổn định.

Điểm nổi bật của dữ liệu có thể kể đến là sự ảnh hưởng của tính mùa vụ trong sự thay đổi CPI. Nhìn chung trong giai đoạn này, CPI có sự dao động lớn theo tháng, nhưng tổng thể vẫn có thể cho thấy mức độ lạm phát khá thấp. CPI thấp trong nhiều tháng cho thấy nhu cầu tiêu dùng và giá cả của các mặt hàng trong những tháng này thay đổi không quá nhiều. CPI âm cũng chỉ kéo dài trong những giai đoạn cụ thể, không kéo dài quá lâu. Đan xen với đó, CPI dương không quá cao, các tháng cao vượt bậc rất hiếm và không kéo dài, giá trị thường tăng trong phạm vi có thể chấp nhận được. Đây là các yếu tố chứng minh nền kinh tế trong mức ổn định, duy trì và kiểm soát được mức lạm phát, không có sự bùng nổ quá mức về giá cả.

3.3 Tỷ giá hối đoái



Hình 3.4: Biểu đồ tỷ giá USD thị trường giai đoạn 2022-2024

Trong giai đoạn 2022, ngân hàng trung ương của các quốc gia phải ứng phó các cú sốc lạm phát gia tăng hậu Covid-19. Nguyên nhân chủ yếu đến từ việc giá cả hàng hóa, năng lượng gia tăng do việc tái tổ chức chuỗi giá trị cung ứng toàn cầu, căng thẳng địa chính trị, biến đổi khí hậu và chủ nghĩa bảo hộ lương thực... Tỷ giá hối đoái được xem là một công cụ đệm nhằm hỗ trợ nền kinh tế thực hiện mục tiêu kiểm soát lạm phát. Đối với tình hình tại Việt Nam, Ngân hàng nhà nước cũng đã trải qua hành trình đầy khó khăn và vất vả.

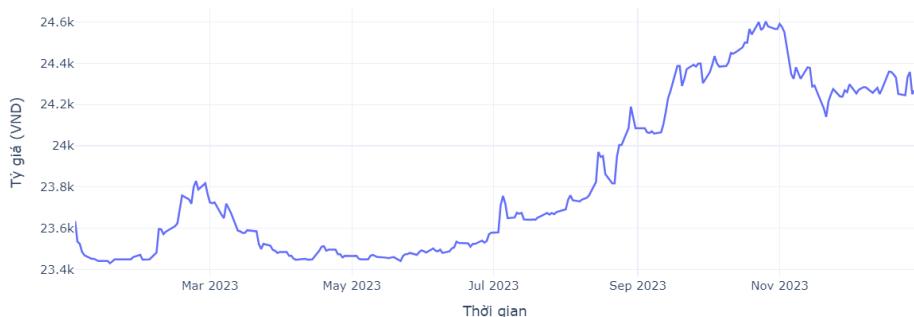


Hình 3.5: Biểu đồ tỷ giá USD thị trường năm 2022

Theo số liệu được thể hiện trong biểu đồ, từ cuối Quý I, Cục Dự trữ Liên bang Mỹ (Fed) đã kích hoạt đợt tăng giá mạnh và liên tục không ngừng nghỉ, đưa chỉ số USD Index lên mức cao nhất trong 2 thập kỷ qua. Trước sức ép liên tục gia tăng, NHNN đã phải bán ra lượng lớn ngoại tệ từ dự trữ ngoại hối để ổn định thị trường. Dù NHNN đã triển khai nhiều công cụ hỗ trợ như sử dụng Quỹ dự trữ ngoại hối, tăng lãi suất, chuyển phương thức giao dịch ngoại tệ từ bán kỳ hạn 3 tháng sang phương thức bán giao ngay... nhưng tỷ giá trong nước vẫn liên tục leo thang.

Đỉnh điểm là vào trong quý III, do FED thắt chặt chính sách tiền tệ và chênh lệch lãi suất giữa USD và VND lớn khiến giá trị đồng USD tăng lên. Giá USD tại các ngân hàng đã tăng thêm khoảng 600 đồng, cao hơn cả mức tăng lũy kế của cả 6 tháng đầu năm. Chưa đầy 1 tháng sau đó, giá USD đã leo lên mức kỷ lục gần 24.900 đồng, đưa mức mất giá của VND lên 8,6% - cao nhất trong nhiều năm qua. Qua tháng 10, NHNN đã quyết định nới biên độ tỷ giá từ mức $+/-3\%$ lên $+/-5\%$ sau khi biên độ tỷ giá được nới rộng ra, tỷ giá thị trường đã tiếp tục biến động thêm một số ngày rồi dần dần tìm được điểm cân bằng và ổn định trong tháng 11. Sang tháng 12, FED đã giảm mức độ tăng lãi suất, làm chậm lại khả năng trượt giá của VND, đồng thời giúp hạ nhiệt tỷ giá.

Bước sang năm 2023, tỷ giá đã có một khởi đầu tích cực.



Hình 3.6: Biểu đồ tỷ giá USD thị trường năm 2023

Nhìn chung, từ tháng 1 cho đến hết tháng 7 năm 2023, tỷ giá có biến động tương đối thấp, dao động trong khoảng 23.700 đến dưới 24.000 (tỷ giá bán ra của NHTM). Tuy vào giữa tháng 2, tỷ giá cũng có lúc tăng vọt khi NHNN có động thái hút ròng mạnh tay nhưng về gần cuối tháng, lãi suất đã được cân chỉnh, chênh lệch lãi suất liên ngân hàng lại nghiêng về phía VND, giúp tỷ giá được ổn định. Ngoài ra, tỷ giá trên thị trường chợ đen cũng thấp hơn so với thị trường chính thức. Những dấu hiệu này cho thấy nhu cầu đồng USD thấp trong quý I. Trong quý II, tỷ giá không có nhiều biến động lớn. Trong khoảng nửa đầu năm, tỷ giá có phần ổn định. Kể từ giữa tháng 7, lãi suất liên ngân hàng kỳ hạn qua đêm đã giảm gần như bằng không, tương tự như giai đoạn đại dịch COVID. Điều này đặt ra vấn đề cần cảnh báo khi thanh khoản vào thời điểm này đang trong trạng thái dư thừa khi tăng trưởng tín dụng ở mức thấp. Bước sang tháng 8, tỷ giá USD ghi nhận nhiều biến động mạnh khi FED tiếp tục nâng cao lãi suất. Vào giữa tháng 9, Chủ tịch Fed Jerome Powell lại làm thị trường dậy sóng khi bình luận rằng lãi suất sẽ phải duy trì ở mức cao hơn, trong thời gian dài hơn, lần đầu tiên trong lịch sử, NHNN phải đưa tỷ giá trung tâm vượt mốc 24.000 VND/USD. Ngoài bối cảnh kinh tế không thuận lợi tại những khu vực như Liên minh châu Âu (EU) và Trung Quốc, khủng hoảng Ukraine và xung đột mới tại Trung Đông cũng là những yếu tố hỗ trợ USD mạnh lên. Điều này đã kéo chênh lệch lãi suất USD - VND lên hơn 5%, Việt Nam đứng trước áp lực lớn về tỷ giá. NHNN đã có những nỗ lực đã giúp lãi suất liên ngân hàng tăng lên, thu hẹp chênh lệch với USD. Trong nửa đầu quý IV, lạm phát và thị trường lao động của Mỹ đã hạ nhiệt, FED cũng ngừng tăng lãi suất. Tháng 12/2023, FED dự báo sẽ cắt giảm lãi suất 3 lần trong năm 2024. Tỷ giá nhanh chóng giảm xuống và duy trì mức ổn định đến cuối năm.

Năm 2024, ngân hàng nhà nước Việt Nam đã trải qua một năm đầy khó khăn với rất nhiều bất thường quanh cặp tỷ giá USD-VND do những nhiễu động trước thềm bầu cử Tổng thống Mỹ và các căng thẳng địa chính trị khác.



Hình 3.7: Biểu đồ tỷ giá USD thị trường năm 2024

Trong 2 tháng đầu quý II, tỷ giá USD/VND đã tăng mạnh lên mức cao nhất năm 25.460, tương đương mức mất giá hơn 3%. Trong giai đoạn đó, FED vẫn chưa thực hiện kỳ vọng cắt giảm lãi suất, đưa chênh lệch lãi suất VND – USD tăng cao. NHNN đã có những can thiệp thông qua nghiệp vụ như bán USD từ dự trữ ngoại hối hay phát hành tín phiếu ngắn hạn nhằm giảm đà tăng của tỷ giá. Bước sang quý III, nhờ vào các biện pháp hỗ trợ thị trường từ cơ quan điều hành, cũng như việc FED hạ lãi suất đưa chỉ số dollar hạ nhiệt đã góp phần giúp thu hẹp chênh lệch lãi suất VND-USD. Tỷ giá USD/VND lại giảm trở lại về mức 24.600, tại thời điểm cuối tháng 9 mức mất giá chỉ còn 1,3%. Tuy nhiên, bước sang quý IV, áp lực tỷ giá tăng quay trở lại. Trong tháng 10, tiền tệ châu Á trung bình đã giảm hơn một nửa mức tăng trong quý III so với đồng USD. Cho tới khi ông Donald Trump đắc cử chức Tổng thống Mỹ đã đưa ra nhiều chính sách và quyết định khiến cho đồng USD giữ ở mức cao và không có dấu hiệu giảm xuống. Kỳ vọng về việc giảm lãi suất của FED cũng không thể thực hiện khiến mặt bằng lãi suất trên thị trường liên ngân hàng liên tục đứng trước áp lực tăng. Đến cuối năm 2024, chênh lệch tỷ giá vẫn tiếp tục tăng đến chạm mức 25.485, giá trị cao nhất trong cả năm qua.

4 Phân tích các chỉ số doanh nghiệp

4.1 Ngân hàng thương mại cổ phần Ngoại thương Việt Nam (Vietcombank)

Trong giai đoạn này giá cổ phiếu khá ổn định, từ năm 2022 có xu hướng hơi giảm, thấp nhất là đầu tháng 10. Tuy nhiên từ cuối năm 2022, giá cổ phiếu đã có xu hướng tăng trở lại, sau đó ổn định và đi ngang đến 2025.



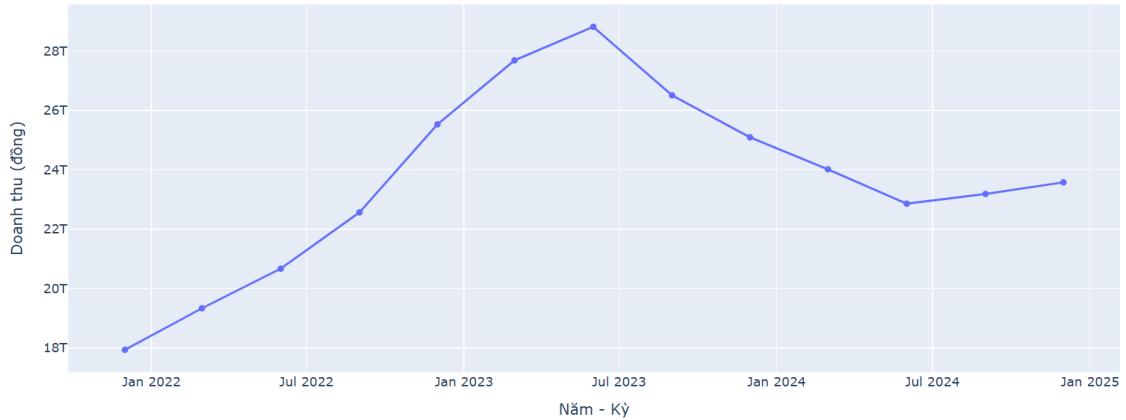
Hình 4.1: Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025

Theo biểu đồ về khối lượng giao dịch, có thể thấy giá trị không ổn định, chênh lệch rất nhiều ở các giai đoạn. Theo trực quan đánh giá, dữ liệu có tính chu kỳ trong khoảng 3 tháng



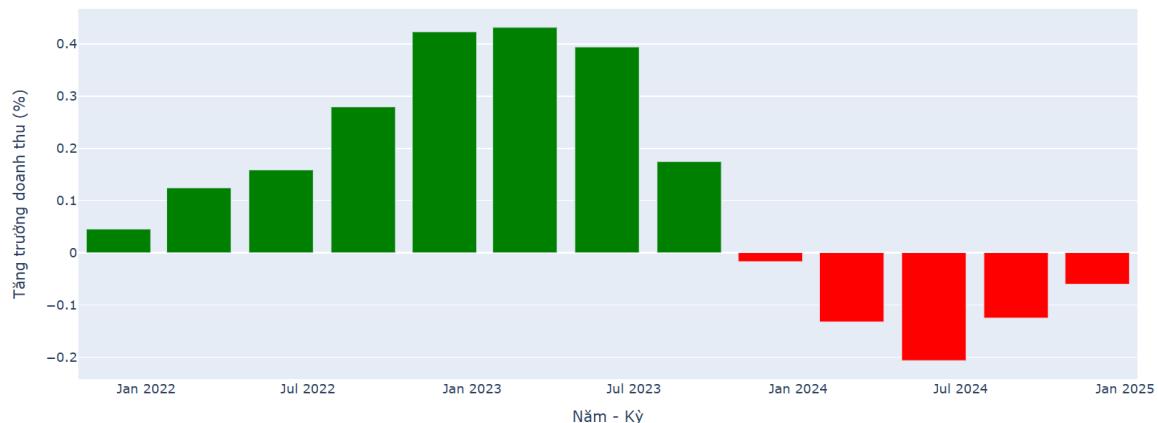
Hình 4.2: Khối lượng giao dịch theo ngày của ngân hàng Vietcombank giai đoạn 2022-2025

Từ đầu đầu năm 2022 đến quý 2 năm 2023, doanh thu của ngân hàng có xu hướng tăng và đạt đỉnh ở quý 2 năm 2023. Tuy nhiên, từ sau quý 2 năm 2023, doanh thu giảm dần, nhưng đã có dấu hiệu phục hồi dần từ quý 2 năm 2024.



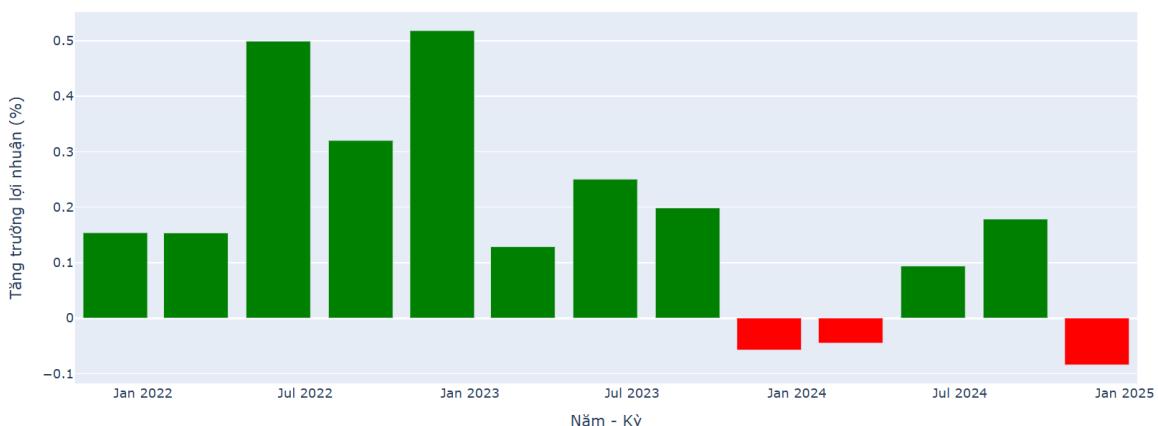
Hình 4.3: Doanh thu của ngân hàng Vietcombank giai đoạn 2022-2025

Dựa theo số liệu trên, sẽ có biểu đồ tăng trưởng doanh thu tương ứng. Từ đầu năm 2022 đến quý 3 năm 2023, tăng trưởng doanh thu dương. Trong đó, cuối năm 2022 tăng trưởng doanh thu đạt đỉnh. Tuy nhiên xu hướng tăng doanh thu chỉ kéo dài trong năm 2022, sang đến năm 2023, doanh thu của ngân hàng đã có xu hướng giảm. Trong đó từ cuối năm 2023 tăng trưởng doanh thu bắt đầu âm và chạm đáy vào quý 2 năm 2024.



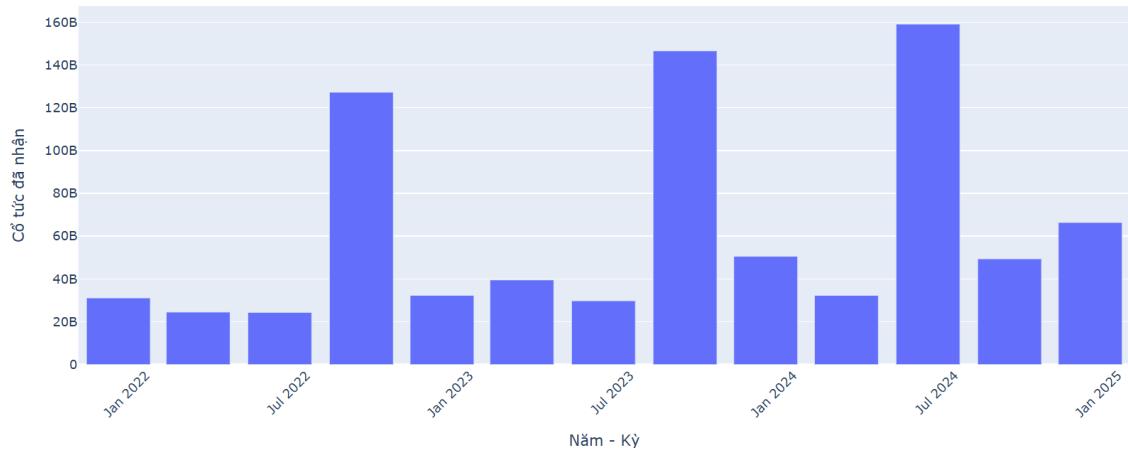
Hình 4.4: Biểu đồ tăng trưởng doanh thu của ngân hàng Vietcombank giai đoạn 2022-2025

Kéo theo đó, từ năm 2022 đến quý 3 năm 2023, tăng trưởng lợi nhuận dương, trong đó quý 4 năm 2022 tăng trưởng đạt đỉnh. Từ năm 2023, tăng trưởng lợi nhuận cũng có xu hướng giảm, tăng trưởng âm kéo dài từ cuối năm 2023 đến đầu năm 2025, có giai đoạn chạm đáy vào cuối năm 2024.



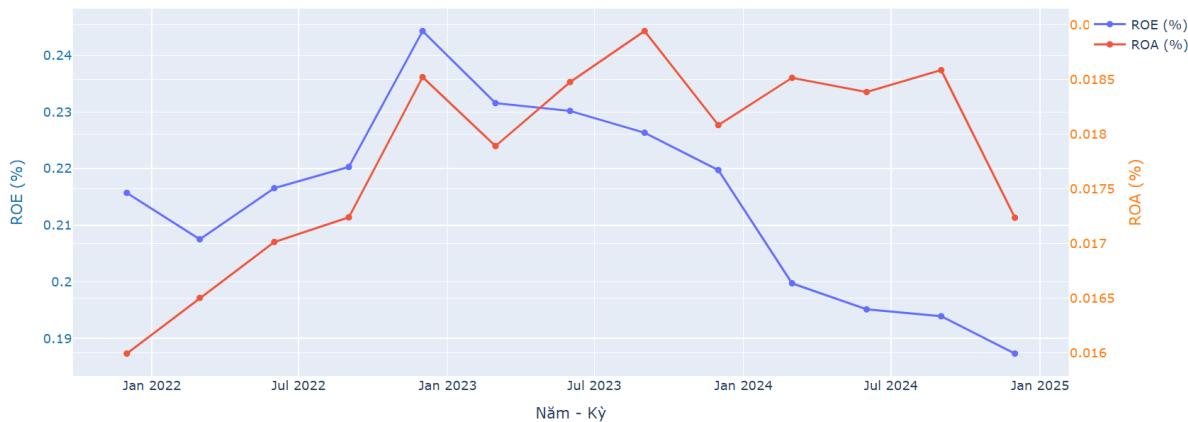
Hình 4.5: Biểu đồ tăng trưởng lợi nhuận của ngân hàng Vietcombank giai đoạn 2022-2025

Cổ tức đã nhận có tính chu kỳ, luôn tăng cao vào khoảng cuối năm.



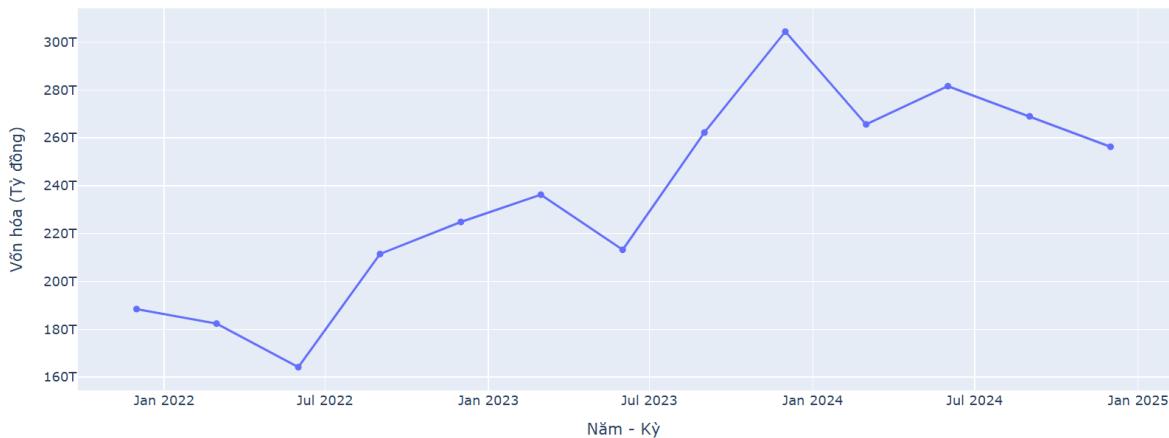
Hình 4.6: Cổ tức của ngân hàng Vietcombank giai đoạn 2022-2025

ROE có xu hướng tăng trưởng đáng kể trong suốt năm 2022, đạt đỉnh vào khoảng cuối năm 2022. Nhưng sau khi đạt đỉnh, ROE lại có xu hướng giảm mạnh vào 2023 đến 2024. Tương tự với ROA, giá trị này cũng cho thấy sự tăng trưởng nhẹ vào đầu năm 2022, rồi lại giảm mạnh cho đến 2024. Tuy nhiên ở giai đoạn này, ROE luôn cao hơn ROA đáng kể, điều này cho thấy công ty đang sử dụng đòn bẩy tài chính (leverage) để khuếch đại lợi nhuận trên vốn chủ sở hữu.



Hình 4.7: Giá trị ROE và ROA của ngân hàng Vietcombank giai đoạn 2022-2025

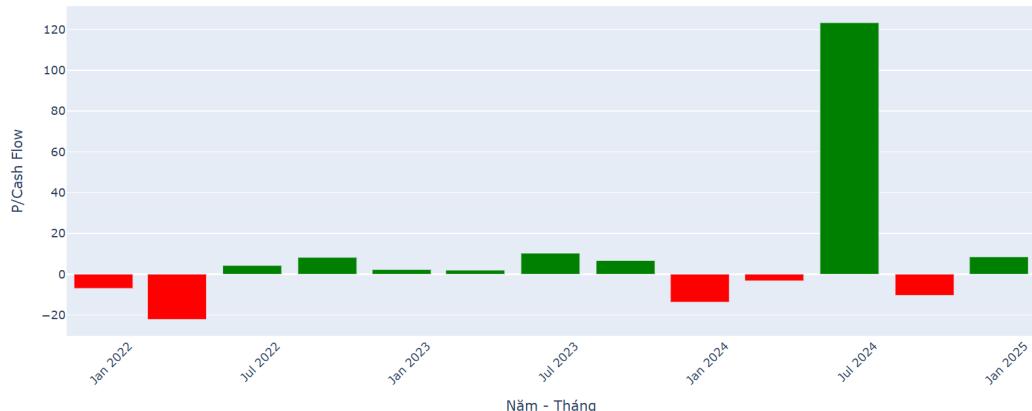
Vốn hóa có xu hướng tăng dần từ năm 2022 - 2024, tuy nhiên đến khoảng giữa năm 2024 bắt đầu giảm dần và chưa có dấu hiệu phục hồi.



Hình 4.8: Vốn hóa của ngân hàng Vietcombank giai đoạn 2022-2025

Đặc biệt, chỉ số P/cash, điểm bất thường được thấy rõ rệt. Vào tháng 6 năm 2024, tỷ số P/Cash Flow tăng vọt lên mức 123.35. Đây là một giá trị rất cao, cho thấy nhà đầu tư sẵn sàng trả một mức giá rất cao cho mỗi đơn vị dòng tiền

của công ty vào thời điểm đó. Trong khi ở các giai đoạn thời gian còn lại, P/Cash không có một xu hướng tăng giảm rõ rệt nào. Vậy nên điểm giá trị vượt trội này có thể là điểm bất thường.



Hình 4.9: Giá trị P/Cash của ngân hàng Vietcombank giai đoạn 2022-2025

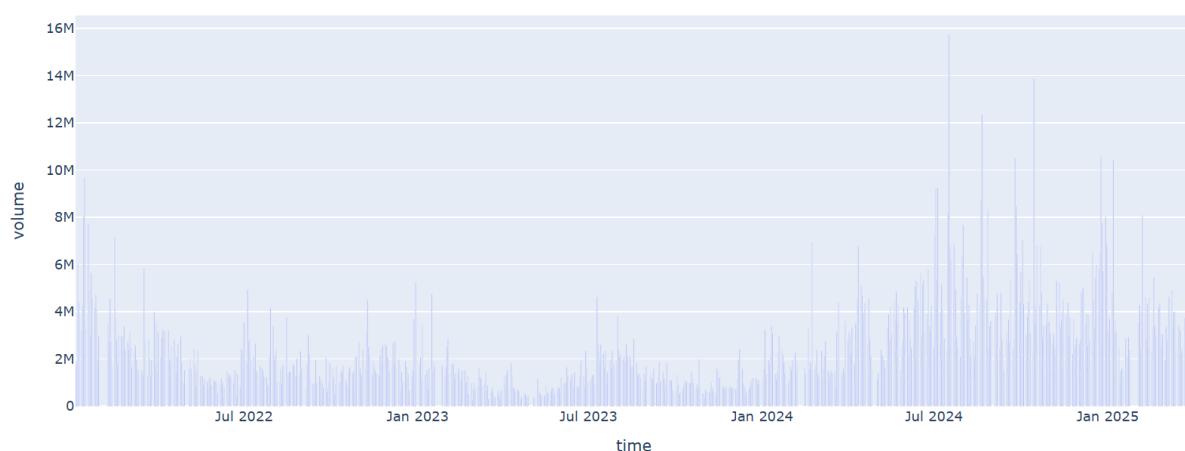
4.2 Ngân hàng thương mại cổ phần Đầu tư và Phát triển Việt Nam (BIDV)

Giá cổ phiếu của ngân hàng BIDV trong năm 2022 có điểm không ổn định, từ đầu năm 2022 đến giữa năm 2022, giá xu hướng giảm dần, tuy có thời điểm có dấu hiệu tăng nhẹ nhưng sau đó lại giảm. Đến những tháng cuối năm 2022, giá cổ phiếu bắt đầu tăng, sau đó đi ngang đến tháng 9/2023. Tháng 10/2023, giá cổ phiếu có sự giảm nhẹ nhưng sau đó đã tăng cao đến tháng 3/2024 rồi đi ngang.



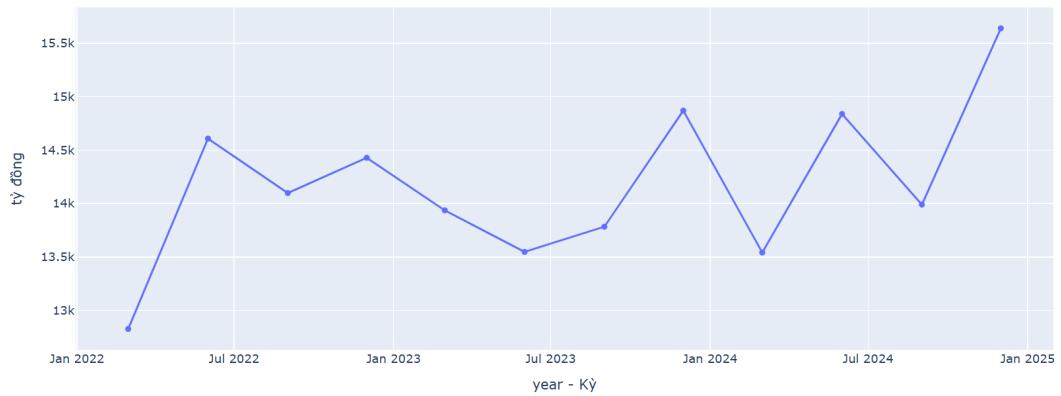
Hình 4.10: Giá cổ phiếu của ngân hàng BIDV giai đoạn 2022-2025

Đầu năm 2022, khối lượng giao dịch bắt đầu giảm, sau đó đi ngang cho đến tận năm 2024. Từ năm 2024, khối lượng giao dịch bắt đầu tăng dần, thậm chí từ tháng 6/2024 trở đi, khối lượng giao dịch tăng rất cao.



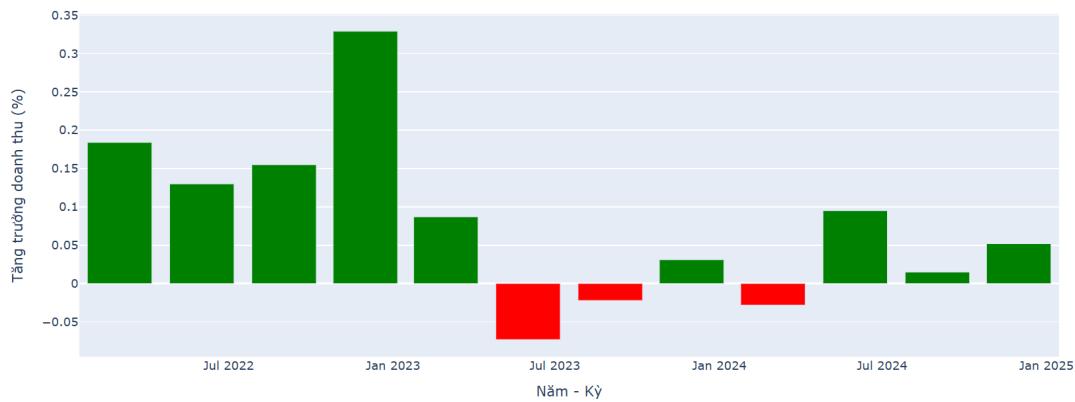
Hình 4.11: Khối lượng giao dịch theo ngày của ngân hàng BIDV giai đoạn 2022-2025

Kết quả kinh doanh của Ngân hàng BIDV có sự tăng trưởng đáng kể từ đầu năm 2022 đến giữa năm 2022. Nhưng sau đó là giai đoạn giảm doanh thu kéo dài đến tận giữa năm 2023. Đến cuối năm 2023, doanh thu mới có sự phục hồi và tăng trưởng trở lại. Doanh thu năm 2024 lại có biến động rất lớn, sự tăng giảm được thể hiện rõ ở mỗi quý. Đầu năm 2025 cho thấy một sự tăng trưởng mạnh mẽ về lợi nhuận, đạt mức cao nhất trong giai đoạn.



Hình 4.12: Doanh thu của ngân hàng BIDV giai đoạn 2022-2025

Đầu năm 2022 đến 2023, tăng trưởng doanh thu khá cao, giữa năm 2023 có sự giảm nhẹ nhưng đã nhanh chóng phục hồi. Nhìn chung, sự tăng trưởng doanh thu của ngân hàng BIDV trong năm 2024 cũng khá biến động, tăng trưởng âm, dương luôn đan xen nhau.



Hình 4.13: Biểu đồ tăng trưởng doanh thu của ngân hàng BIDV giai đoạn 2022-2025

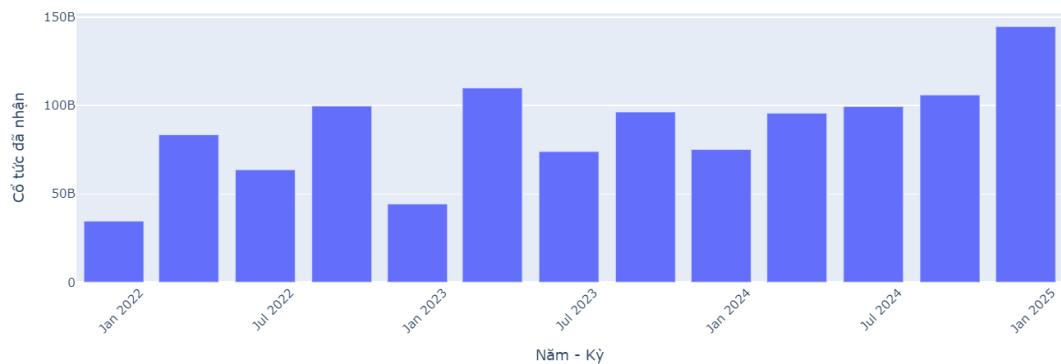
Phần lớn các giai đoạn đều cho thấy sự tăng trưởng lợi nhuận dương. Nhưng sự tăng trưởng có biến động mạnh, không theo quy tắc. Trong giai đoạn năm 2022, tăng trưởng lợi nhuận luôn ở mức dương nhưng sang đến năm 2024 đã có sự sụt giảm đến âm ở một số tháng.



Hình 4.14: Biểu đồ tăng trưởng lợi nhuận của ngân hàng BIDV giai đoạn 2022-2025

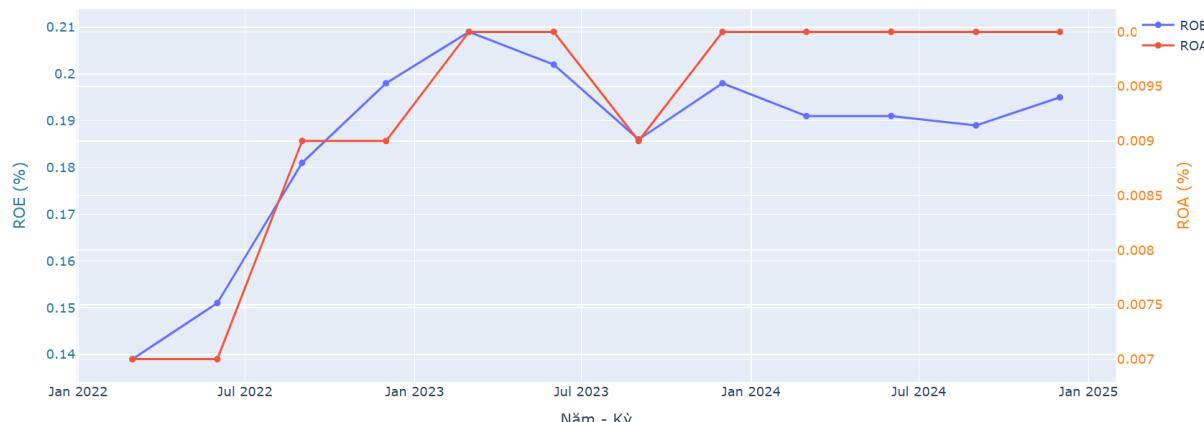
Biểu đồ cho thấy có sự tăng trưởng đáng kể về giá trị cổ tức đã nhận qua các năm, đặc biệt là sự tăng mạnh vào

đầu năm 2025.



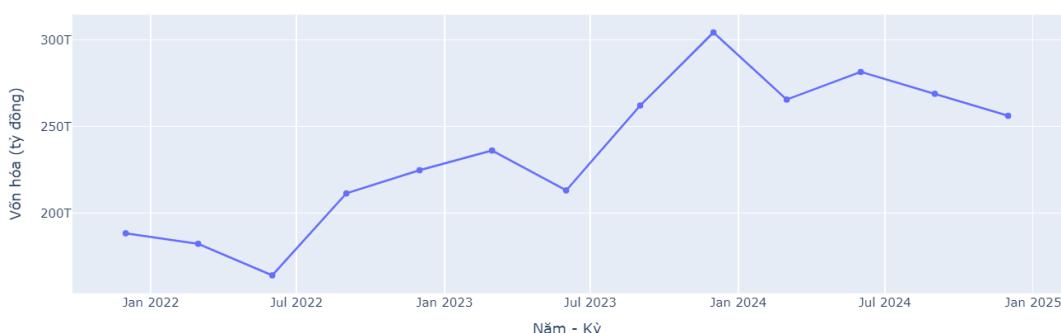
Hình 4.15: Cổ tức của ngân hàng BIDV giai đoạn 2022-2025

Từ năm 2022 đến năm 2023, giá trị ROE tăng mạnh. Đến năm 2023, ROE có xu hướng giảm dần và đi ngang ở năm 2024, đặc biệt là từ giữa năm 2024 trở đi ROE duy trì ở mức khá cao, dao động quanh 19% - 20%. Đối với ROA, từ năm 2022 đến giữa năm 2023 vẫn tăng mạnh nhưng sau đó có sự điều chỉnh nhẹ ở quý 3 năm 2023 rồi quay lại mức cũ. Sang đến năm 2024, ROA duy trì ở mức tương đối ổn định quanh 0.01. Có sự chênh lệch lớn giữa ROA và ROE cho thấy ngân hàng đang sử dụng đòn bẩy tài chính rất cao.



Hình 4.16: Giá trị ROE và ROA của ngân hàng BIDV giai đoạn 2022-2025

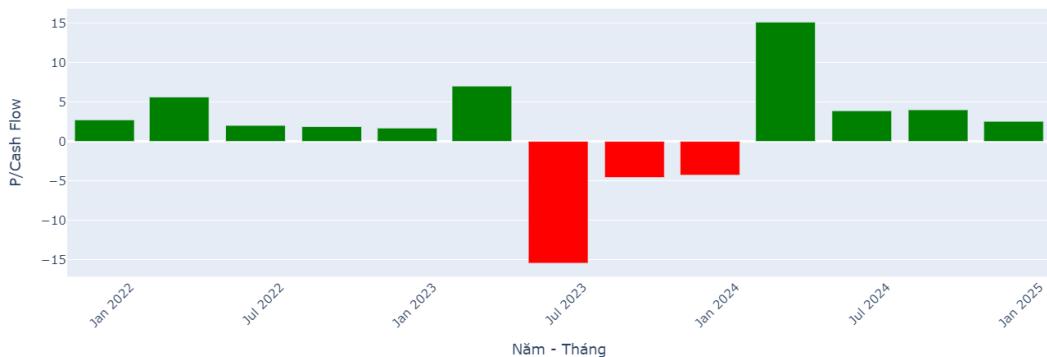
Vốn hóa có xu hướng tăng dần từ năm 2022 - 2024, tuy nhiên đến khoảng giữa năm 2024 bắt đầu giảm dần và chưa có dấu hiệu phục hồi



Hình 4.17: Vốn hóa của ngân hàng BIDV giai đoạn 2022-2025

Đầu năm 2022 đến 2023, tăng trưởng doanh thu khá cao, giữa năm 2023 có sự giảm nhẹ nhưng đã nhanh chóng phục hồi. Nhìn chung, sự tăng trưởng doanh thu của ngân hàng BIDV trong năm 2024 cũng khá biến động, tăng trưởng âm, dương luân đan xen nhau. Giá trị P/Cash của ngân hàng BIDV cũng có nhiều điểm bất thường. Vào

tháng 7/2023, tỷ số P/Cash giảm mạnh đến mức âm và kéo dài giá trị âm đến hết năm 2023. Sang đến đầu năm 2024, giá trị P/Cash tăng cao lên mức dương và tăng mạnh đột ngột.



Hình 4.18: Giá trị P/Cash của ngân hàng BIDV giai đoạn 2022-2025

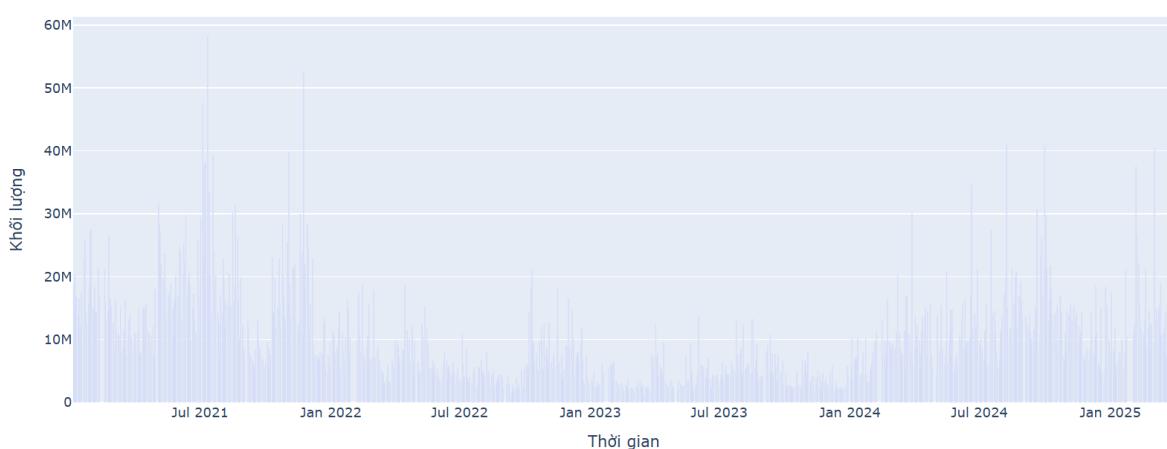
4.3 Ngân hàng thương mại cổ phần Kỹ thương Việt Nam (Techcombank)

Từ đầu năm 2022 - gần cuối năm 2022, giá của cổ phiếu có xu hướng giảm mạnh và chạm đáy vào tháng 10. nhưng sau đó đã phục hồi và đi ngang đến gần cuối năm 2023. Từ đầu năm 2024, giá của cổ phiếu có xu hướng tăng mạnh đến tháng 3/2024 rồi tiếp tục đi ngang đến 2025.



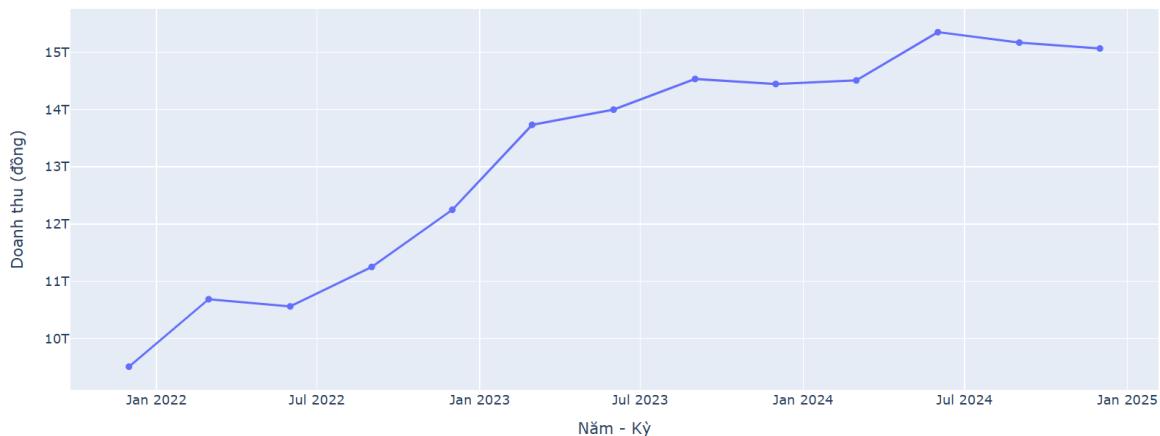
Hình 4.19: Giá cổ phiếu của ngân hàng Techcombank giai đoạn 2022-2025

Khối lượng giao dịch ổn định trong giai đoạn 2022 đến đầu năm 2024. Sang đến năm 2024, khối lượng giao dịch có dấu hiệu tăng mạnh, chứng tỏ thị trường bắt đầu sôi động hơn. Khối lượng giao dịch của Techcombank có xu hướng chu kỳ trong khoảng 6 tháng.



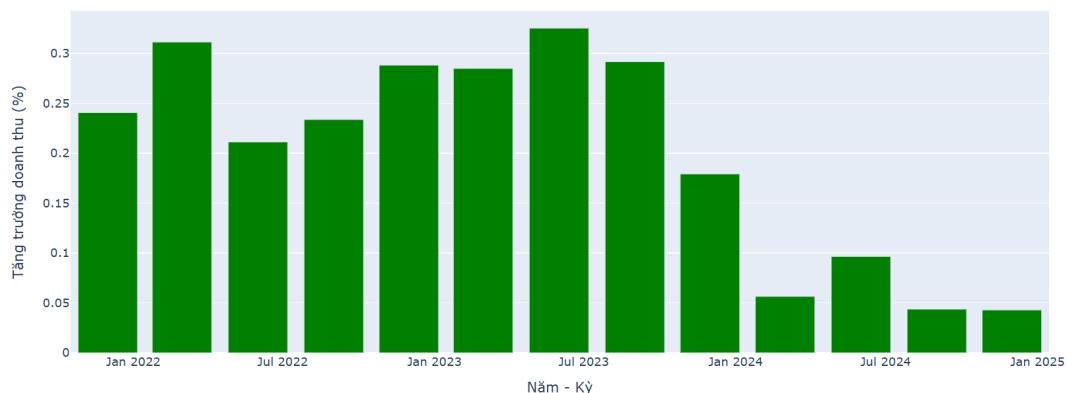
Hình 4.20: Khối lượng giao dịch theo ngày của ngân hàng Techcombank giai đoạn 2022-2025

Từ đầu năm 2022 đến năm 2025, doanh thu có xu hướng tăng và đạt đỉnh ở quý 2 năm 2024. Sau đó, doanh thu có xu hướng giảm nhẹ.



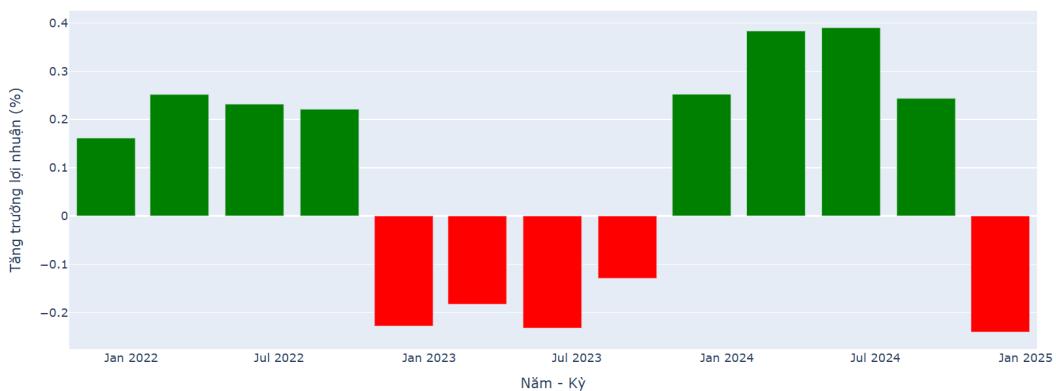
Hình 4.21: Doanh thu của ngân hàng Techcombank giai đoạn 2022-2025

Kéo theo đó, từ đầu năm 2022 đến năm 2024, tăng trưởng doanh thu luôn dương. Trong đó, quý 2 năm 2023 tăng trưởng doanh thu đạt đỉnh. Từ quý 2 năm 2023, tăng trưởng doanh thu có xu hướng giảm và chạm mức thấp nhất tại quý 4/2024.



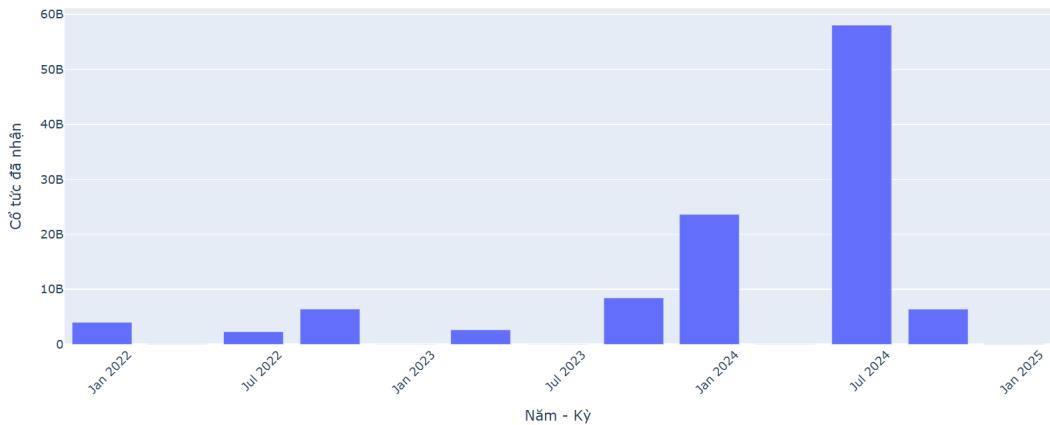
Hình 4.22: Biểu đồ tăng trưởng doanh thu của ngân hàng Techcombank giai đoạn 2022-2025

Từ năm 2022 đến quý 3 năm 2023, tăng trưởng lợi nhuận luôn dương nhưng từ cuối năm 2023 đã bắt đầu có xu hướng giảm thậm chí chạm mức tăng trưởng âm và chạm đáy vào cuối năm 2024.



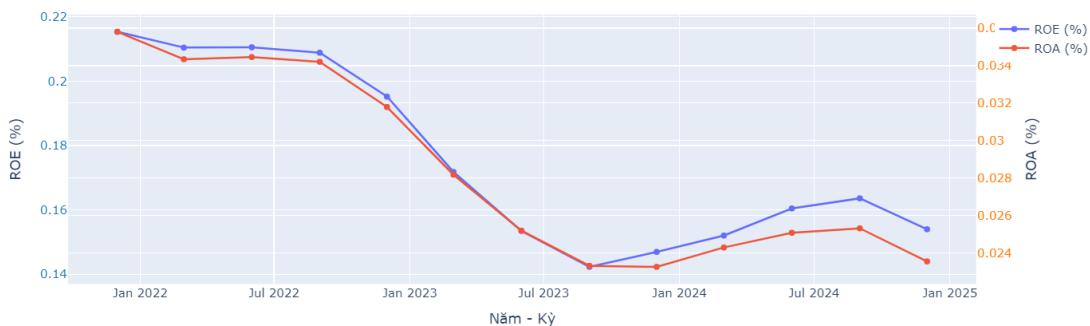
Hình 4.23: Biểu đồ tăng trưởng lợi nhuận của ngân hàng Techcombank giai đoạn 2022-2025

Giá trị cổ tức của ngân hàng Techcombank không ổn định.



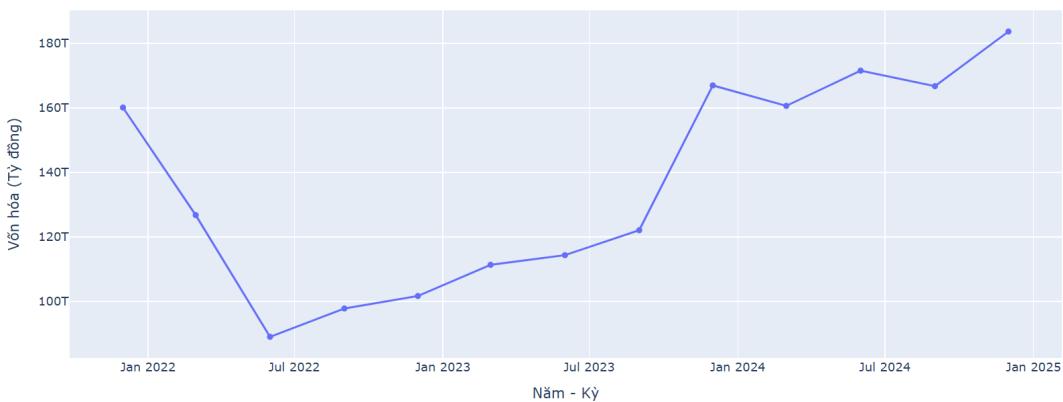
Hình 4.24: Cổ tức của ngân hàng Techcombank giai đoạn 2022-2025

Giá trị ROE và ROA của ngân hàng Techcombank có xu hướng giống nhau, cùng tăng và cùng giảm trong suốt giai đoạn. Từ giai đoạn 2022 đến giữa năm 2023, ROE và ROA giảm mạnh. Đến giữa 2023 trở đi mới có xu hướng tăng lại. Giá trị ROE của ngân hàng Techcombank vẫn duy trì lớn hơn giá trị ROA, tuy nhiên điều này được thể hiện không rõ ràng so với hai ngân hàng trước.



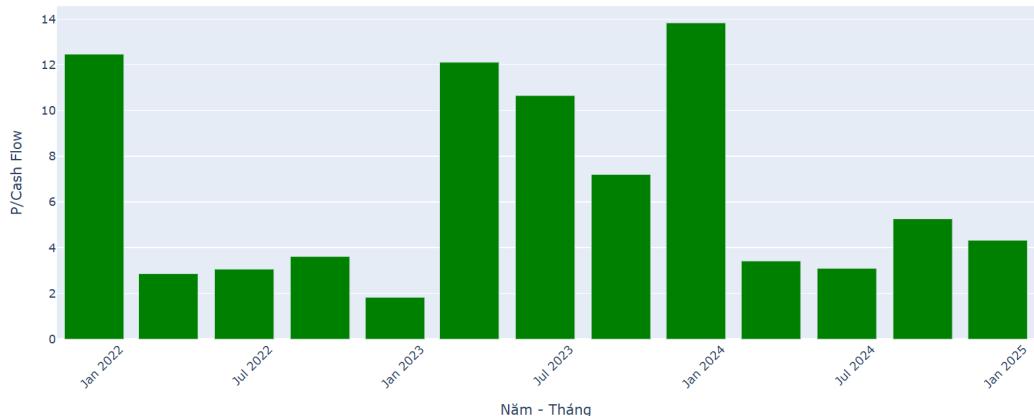
Hình 4.25: Giá trị ROE và ROA của ngân hàng Techcombank giai đoạn 2022-2025

Trong hai quý đầu năm 2022, Vốn hóa giảm mạnh từ khoảng 126 nghìn tỷ đồng xuống mức thấp nhất là khoảng 89 nghìn tỷ đồng. Đây là giai đoạn giảm sâu nhất trong toàn bộ chu kỳ. Từ sau quý 2 năm 2022, vốn hoá có xu hướng tăng. Trong đó giai đoạn từ quý 2/2022 đến quý 3/2023, vốn hoá tăng đều từ khoảng 89 nghìn tỷ lên 122 nghìn tỷ, rồi tăng mạnh kèm theo biến động mạnh.



Hình 4.26: Vốn hóa của ngân hàng Techcombank giai đoạn 2022-2025

Tỷ lệ P/Cash Flow có xu hướng giảm nhẹ. Đạt mức thấp nhất là khoảng 1.8 trong năm 2022. Ngược lại, tỷ lệ P/Cash Flow có sự tăng trưởng vượt bậc và duy trì quanh mức cao, đạt đỉnh vào cuối năm 2023 khoảng 14.0. Sau khi đạt đỉnh, tỷ lệ P/Cash Flow giảm mạnh xuống khoảng 3.4, tuy sau đó có sự phục hồi nhẹ lên khoảng 5.2, nhưng lại giảm xuống khoảng 3.0 và chỉ tăng trở lại lên khoảng 4.4 vào cuối kỳ.



Hình 4.27: Giá trị P/Cash của ngân hàng Vietcombank giai đoạn 2022-2025

4.4 Đánh giá tổng quan

Mặc dù giá trị các chỉ số tài chính giữa ba ngân hàng có sự chênh lệch nhất định, nhưng nhìn chung, các chỉ số này vẫn phản ánh xu hướng vận động chung của thị trường. Đối với các chỉ tiêu như doanh thu, lợi nhuận,... cả ba ngân hàng đều ghi nhận sự biến động tăng hoặc giảm rõ rệt qua từng giai đoạn, cho thấy sự tương đồng trong phản ứng trước các biến động kinh tế vĩ mô hoặc điều kiện thị trường.



Hình 4.28: Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025



Hình 4.29: Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025



Hình 4.30: Giá cổ phiếu ngân hàng Vietcombank giai đoạn 2022-2025

Tuy xu hướng tăng giảm của giá cổ phiếu giữa ba ngân hàng có sự khác biệt nhất định do phụ thuộc vào tình hình hoạt động, chiến lược kinh doanh riêng biệt, cũng như mức độ hiệu quả trong quản trị rủi ro và phản ứng với các yếu tố nội tại, nhưng nhìn chung vẫn tồn tại một mức độ tương quan nhất định giữa các biến động giá cổ phiếu này. Cụ thể, trong một số giai đoạn, cả ba cổ phiếu đều ghi nhận mức giá cao hoặc thấp tương đồng, phản ánh sự ảnh hưởng mang tính hệ thống từ các yếu tố kinh tế vĩ mô như lãi suất điều hành, chính sách tiền tệ, hoặc biến động trên thị trường tài chính quốc tế. Điều này cho thấy, mặc dù mỗi ngân hàng có đặc điểm riêng, nhưng giá cổ phiếu của các ngân hàng thương mại lớn vẫn bị chi phối bởi những yếu tố chung của môi trường kinh tế và tâm lý thị trường.

5 Lựa chọn thuộc tính

Đối với một bộ dữ liệu quá nhiều thuộc tính, nếu số lượng các dữ liệu không phù hợp thì rất dễ ảnh hưởng đến kết quả cũng như tốc độ thực hiện trong quá trình làm việc với dữ liệu. Đối với những trường hợp này, thì việc lựa chọn thuộc tính để loại bỏ các dữ liệu không phù hợp là hết sức cần thiết. Chọn lựa thuộc tính là một quá trình để tìm ra một tập các thuộc tính phù hợp nhất theo một tiêu chí nào đó. Lựa chọn thuộc tính phù hợp đóng vai trò quan trọng trong việc cải thiện hiệu năng của quá trình tiền xử lý, khai phá dữ liệu cũng như thực thi mô hình machine learning mà không làm thay đổi bản chất của dữ liệu.

5.1 Loại bỏ các thuộc tính không liên quan

Để dự đoán xu hướng giá cổ phiếu, nhóm sẽ sử dụng thuộc tính Close để đánh giá sự đồng thuận của giá trị cuối cùng trong phiên. Các biến Open, High, Low, Volume không sử dụng trong quá trình huấn luyện mô hình vậy nên nhóm sẽ loại bỏ các biến này.

5.2 Loại bỏ các biến trùng lặp thông tin

Trong tập dữ liệu gốc, các biến tỷ giá hối đoái được chia thành ba loại là Cash, Sell và Transfer. Tuy nhiên, quá trình trực quan hóa dữ liệu cho thấy ba thuộc tính này có xu hướng biến động gần như tương đồng với nhau. Vậy nên nhóm chỉ giữ lại thuộc tính Cash và loại bỏ hai thuộc tính còn lại.



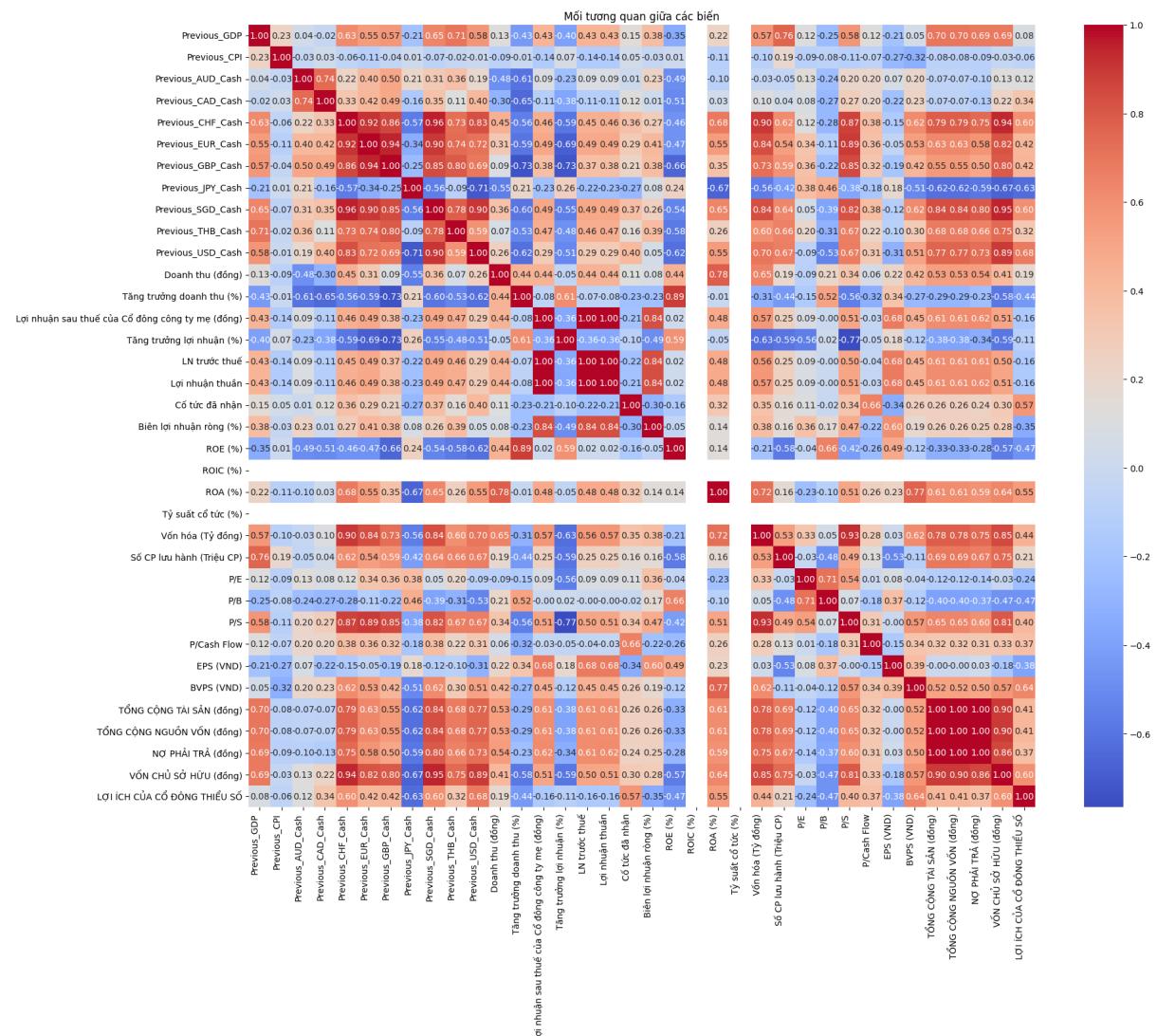
Hình 5.1: Ví dụ 1: Biểu đồ tỷ giá EURO của Vietcombank



Hình 5.2: Ví dụ 2: Biểu đồ tỷ giá Yên Nhật của BIDV

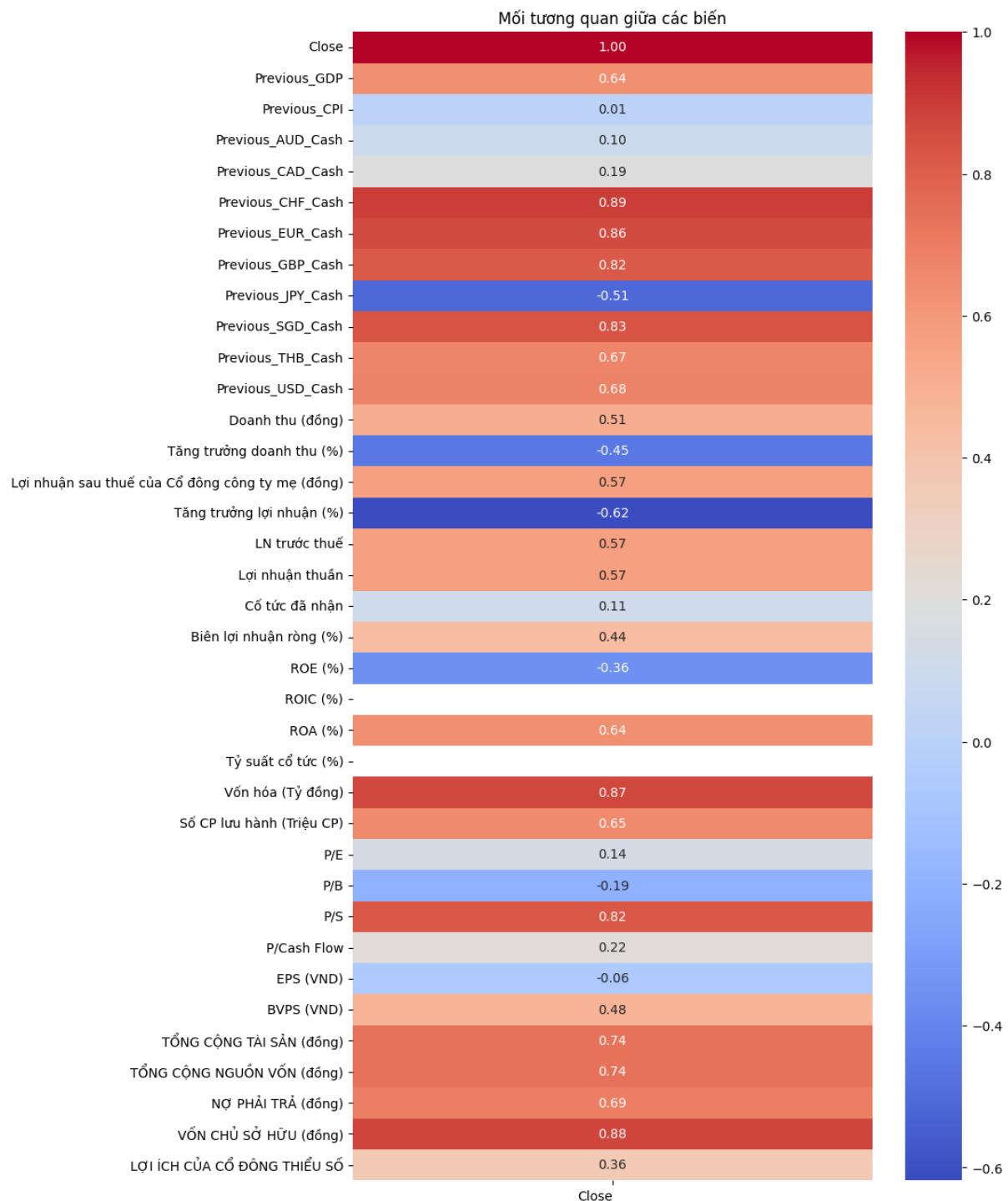
5.3 Phân tích tương quan

Bước tiếp theo, xây dựng ma trận tương quan giữa các biến độc lập để kiểm tra hiện tượng đa cộng tuyến. Những cặp biến có hệ số tương quan cao được xem xét loại bỏ một biến trong cặp để đảm bảo mô hình không bị lạm thuẫn vào các biến mang thông tin trùng lặp.



Hình 5.3: Ví dụ: Heatmap về tương quan giữa các biến của bộ dữ liệu Vietcombank

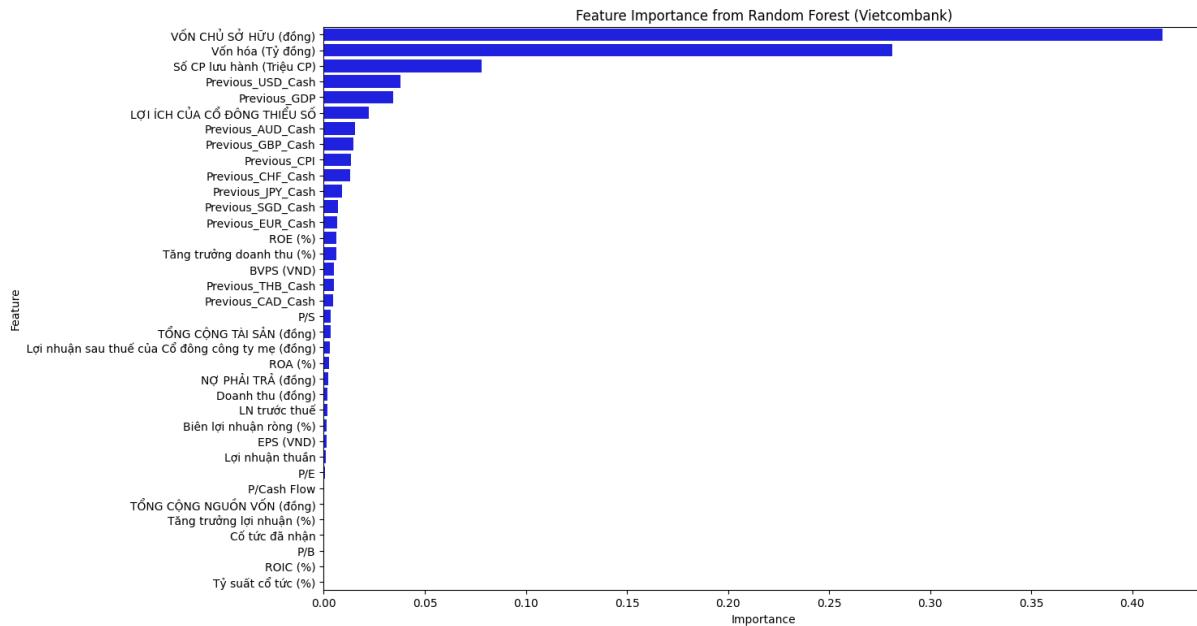
Sau khi loại bỏ các biến có tính tương quan cao với nhau, các thuộc tính còn lại được đánh giá quan hệ tuyến tính với biến mục tiêu thông qua hệ số tương quan. Những đặc trưng có tương quan rất thấp với Close, tức là những biến này không ảnh hưởng đáng kể cũng như không diễn giải được Close sẽ bị loại bỏ.



Hình 5.4: Ví dụ: Tương quan giữa các biến với biến mục tiêu của Vietcombank

5.4 Lựa chọn thuộc tính bằng Random Forest

Không chỉ là một thuật toán mạnh để dự đoán, Random Forest có thể được sử dụng như một công cụ đánh giá độ quan trọng của các thuộc tính trong bộ dữ liệu. Chúng tôi sử dụng Random Forest để xác định các đặc trưng có ảnh hưởng lớn nhất đến biến mục tiêu, từ đó lựa chọn những thuộc tính quan trọng nhất và loại bỏ các thuộc tính không cần thiết.



Hình 5.5: Ví dụ: Độ quan trọng của từng biến của ngân hàng Vietcombank

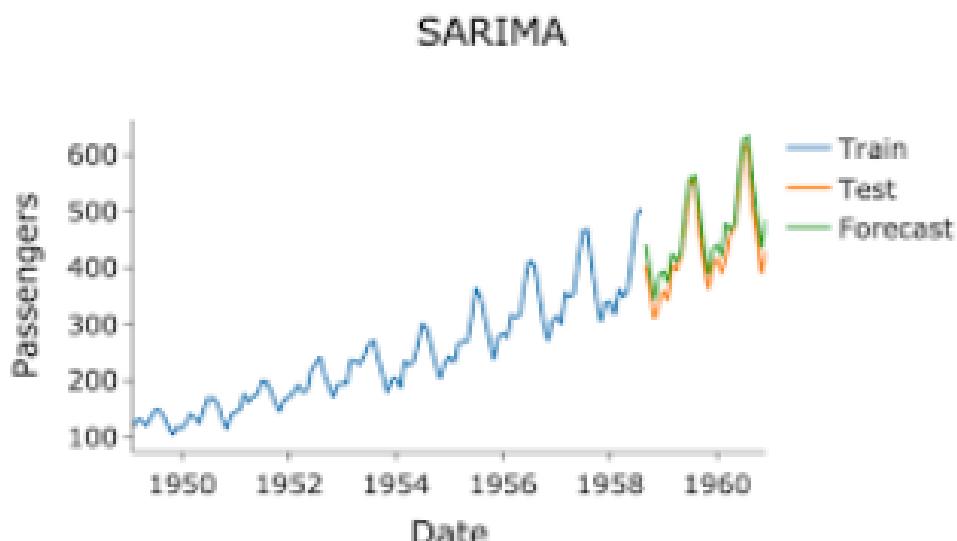
Dựa trên chỉ số feature_importance do mô hình cung cấp, hai đặc trưng có độ quan trọng bằng 0 là ROIC và tỷ số cổ tức bị loại bỏ. Việc này nhằm giữ lại các biến thực sự mang thông tin hữu ích cho mô hình, đồng thời giảm thiểu độ phức tạp và nguy cơ overfitting.

6 Lý thuyết mô hình

6.1 Seasonal Autoregressive Integrated Average with Exogenous Regressors (SARIMAX)

SARIMAX (Seasonal Autoregressive Integrated Average with Exogenous Regressors) [6] là một mô hình dự báo chuỗi thời gian nâng cấp từ mô hình ARIMA **arima** bằng cách kết hợp các thành phần theo mùa và các biến bên ngoài (yếu tố ngoại sinh).

Quá trình lựa chọn mô hình phù hợp bắt đầu bằng việc trực quan hóa dữ liệu nhằm phát hiện các điểm bất thường và thực hiện các biến đổi cần thiết để ổn định phương sai. Sau đó, kiểm tra tính dừng thông qua biểu đồ chuỗi thời gian ACF và PACF. Nếu dữ liệu chưa dừng, thực hiện phân biệt dữ liệu theo mùa hoặc không theo mùa và đánh giá lại tính dừng. Khi dữ liệu đạt được tính dừng, biểu đồ ACF và PACF sẽ giúp gợi ý cấu trúc của mô hình. Nếu các mẫu mùa vụ xuất hiện tại độ trễ theo mùa, yếu tố thời vụ cần được đưa vào mô hình. Trong trường hợp không rõ ràng, có thể sử dụng mô hình hỗn hợp.



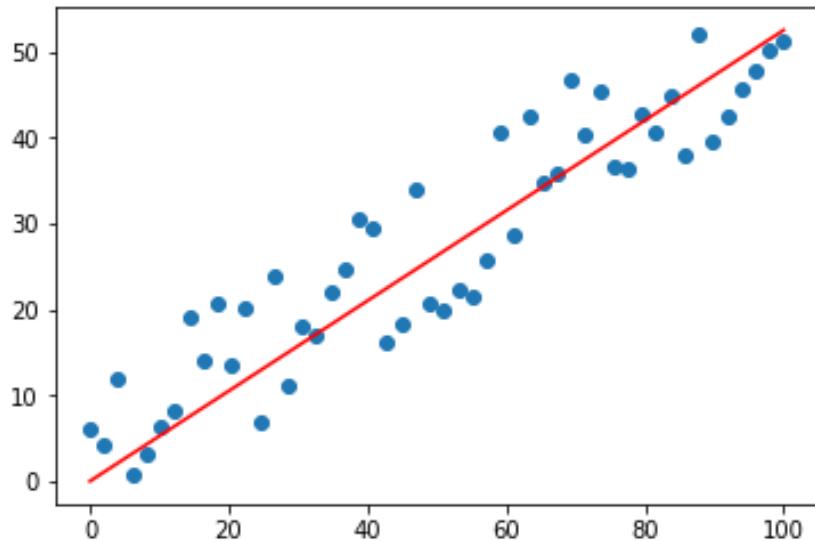
Hình 6.1: SARIMAX

6.2 Machine Learning

6.2.1 Linear Regression

Linear Regression là một thuật toán học có giám sát. Đây là một phương pháp truyền thống dùng thống kê để ước lượng mối quan hệ giữa các biến độc lập và biến phụ thuộc.

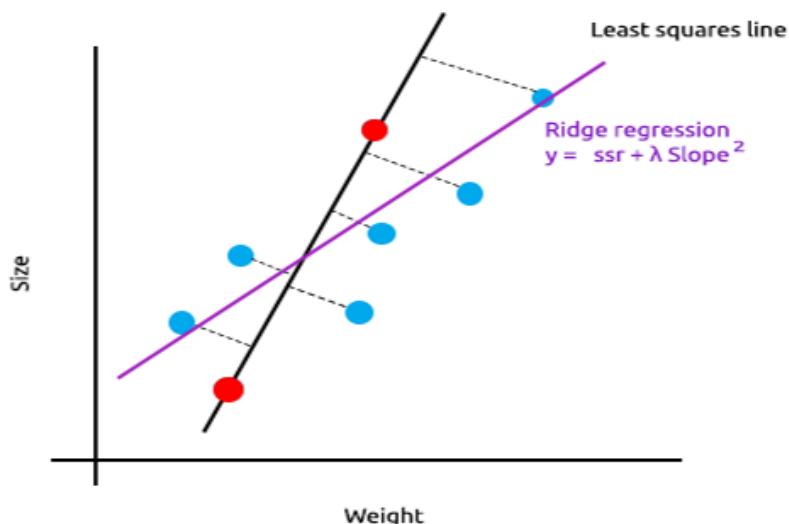
Linear Regression dự báo giá trị của biến mục tiêu từ các giá trị của các biến đầu vào. Thuật toán này giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất, giá trị dự đoán gần đúng nhất với sai số nhỏ nhất để biểu diễn mối quan hệ này.



Hình 6.2: Linear Regression

6.2.2 Ridge Regression

Ridge Regression [7] là một kỹ thuật để phân tích dữ liệu hồi quy nhiều lần. Đây là phiên bản cải tiến của Linear Regression bằng cách thêm mức độ chênh lệch sai số vào hàm mất mát. Trong trường hợp nhiều biến đầu vào có quan hệ chặt chẽ với nhau, Ridge Regression được sử dụng để xem mức độ chêch được thêm vào các ước tính hồi quy với mong muốn làm giảm các sai số tiêu chuẩn và ngăn mô hình học quá mức vào dữ liệu huấn luyện.

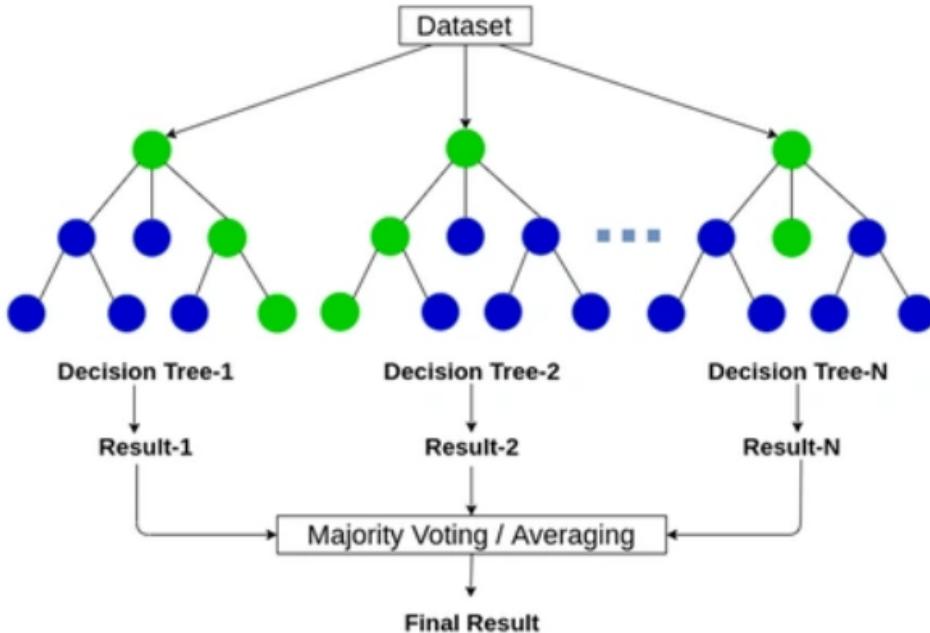


Hình 6.3: Ridge Regression

6.2.3 Random Forest

Random Forest [8] là một mô hình thuộc nhóm mô hình machine learning học có giám sát, thường dùng cho cả bài toán phân loại và hồi quy. Thuật toán này được xây dựng dựa trên việc tạo ra một tập hợp các cây quyết định (Decision Trees), với mỗi cây được xây dựng một cách ngẫu nhiên và độc lập từ nhau. Ý tưởng cơ bản của Random Forest là kết hợp dự đoán từ nhiều cây quyết định khác nhau để tạo ra một dự đoán tổng hợp.

Random Forest tạo ra nhiều cây quyết định, mỗi cây được huấn luyện trên một tập con ngẫu nhiên của bộ dữ liệu. Với điểm dữ liệu mới, mỗi cây quyết định trong Random Forest sẽ đưa ra dự đoán một cách độc lập. Với bài toán hồi quy, kết quả dự đoán cuối cùng thường được tính bằng giá trị trung bình của các kết quả nhỏ.



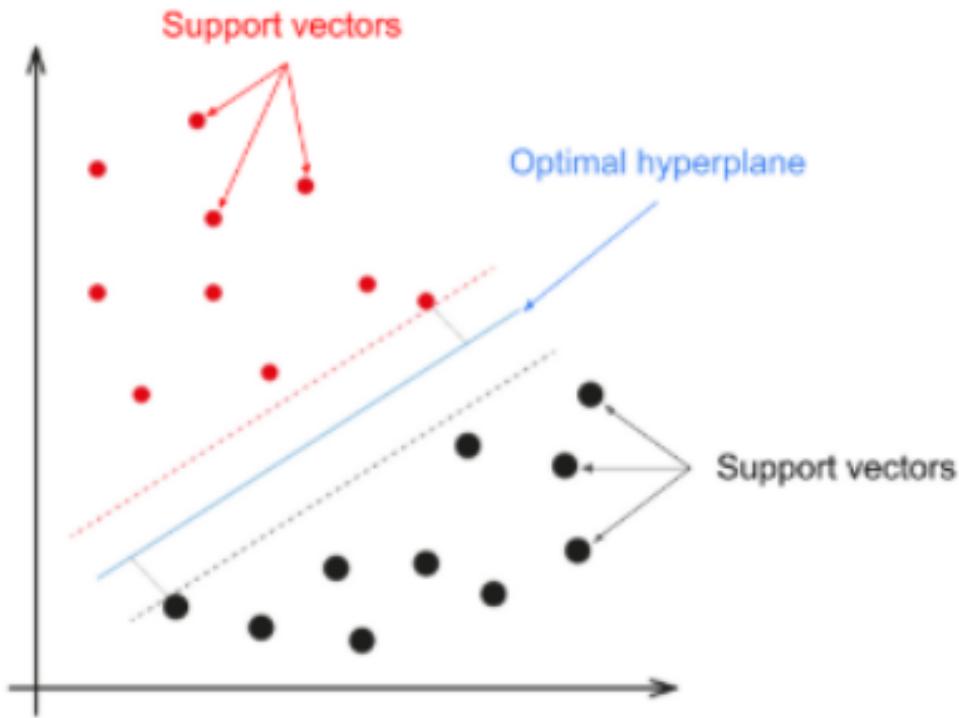
Hình 6.4: Random Forest

6.2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) [9] là một thuật toán khá hiệu quả được ứng dụng cho các bài toán dự báo học có giám sát. SVM nổi bật với việc có thể hoạt động tốt trên các mẫu dữ liệu lớn, dữ liệu tuyến tính và cả phi tuyến tính.Thêm vào đó, SVM cho phép linh hoạt điều chỉnh theo bản chất dữ liệu thông qua việc lựa chọn kernel phù hợp, từ đó tăng khả năng thích ứng.

Thuật toán SVM hoạt động bằng cách tìm kiếm một siêu mặt phẳng (hoặc nhiều siêu mặt phẳng) để phân tách các lớp dữ liệu trong không gian nhiều chiều. Mục tiêu là tìm ra siêu mặt phẳng tối ưu, sao cho khoảng cách từ siêu mặt phẳng đó đến các điểm dữ liệu gần nhất của mỗi lớp là lớn nhất – khoảng cách này được gọi là biên (margin). Siêu mặt phẳng tối ưu, hay còn gọi là "biên cứng" (hard margin), là đường phân chia giúp tạo ra sự tách biệt rõ ràng nhất giữa hai lớp, bằng cách tối đa hóa biên an toàn. Việc này giúp mô hình trở nên ổn định và có khả năng tổng quát hóa tốt hơn khi áp dụng với dữ liệu mới.

SVM được áp dụng cho các bài toán hồi quy thông qua Support Vector Regression (SVR). SVR sử dụng các nguyên tắc tương tự như SVM nhưng tập trung vào việc dự đoán đầu ra liên tục hơn là phân loại các điểm dữ liệu. Ý tưởng cơ bản của SVR là tìm siêu mặt phẳng trong không gian tính năng tối đa hóa biên độ giữa các giá trị dự đoán của biến đáp ứng và các giá trị quan sát được của biến đáp ứng trong dữ liệu đào tạo.



Hình 6.5: SVM

6.3 Deep Learning

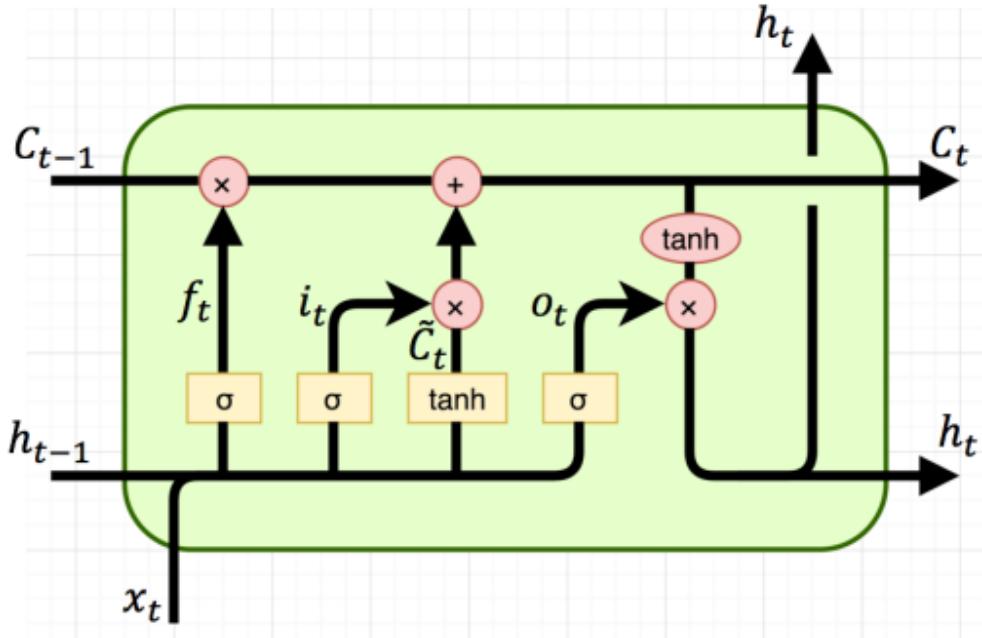
6.3.1 Long Short-term Memory (LSTM)

Recurrent Neural Network (RNN) [10] là một mô hình nổi bật và hiệu quả trong việc xử lý dữ liệu dạng chuỗi, tuy nhiên khi huấn luyện trên các chuỗi dài RNN có thể gặp phải vấn đề liên quan đến gradient. Gradient là thuật ngữ chỉ hiện tượng độ dốc trong dự đoán, giá trị này dùng để cập nhật trọng số trong mạng neural thông qua cơ chế lan truyền ngược. Gradient thể hiện mức độ ảnh hưởng của từng trọng số đến sai số đầu ra để giúp mô hình điều chỉnh dự đoán cho chính xác hơn. Khi dữ liệu huấn luyện quá dài, có thể gặp trường hợp gradient bị quá nhỏ hoặc quá lớn. Điều này gây ảnh hưởng đến RNN trong quá trình học dữ liệu và dự đoán.

Để khắc phục vấn đề trên, chúng tôi đề xuất sử dụng thêm mô hình LSTM [11]. LSTM là một biến thể cải tiến của RNN, được thiết kế đặc biệt để ghi nhớ thông tin trong thời gian dài hơn. Điểm nổi bật của LSTM là kiến trúc ba cổng (gates) của từng đơn vị (LSTM cell), bao gồm:

- Forget Gate: Quyết định thông tin nào nên được đi tiếp, thông tin nào nên bỏ đi khỏi cell.
- Input Gate: Sau khi quyết định được thông tin liên quan, thông tin sẽ chuyển đến cổng đầu vào, cổng đầu vào thêm thông tin mới vào thông tin hiện.
- Output Gate: Quyết định thông tin cần xuất ra ngoài tại thời điểm hiện tại.

Nhờ cơ chế này, LSTM giữ được thông tin quan trọng xuyên suốt nhiều bước thời gian, đồng thời tránh được hiện tượng gradient biến mất. Hầu hết mọi mô hình SOTA hiện đại dựa trên RNN đều tuân theo mạng LSTM để dự đoán.



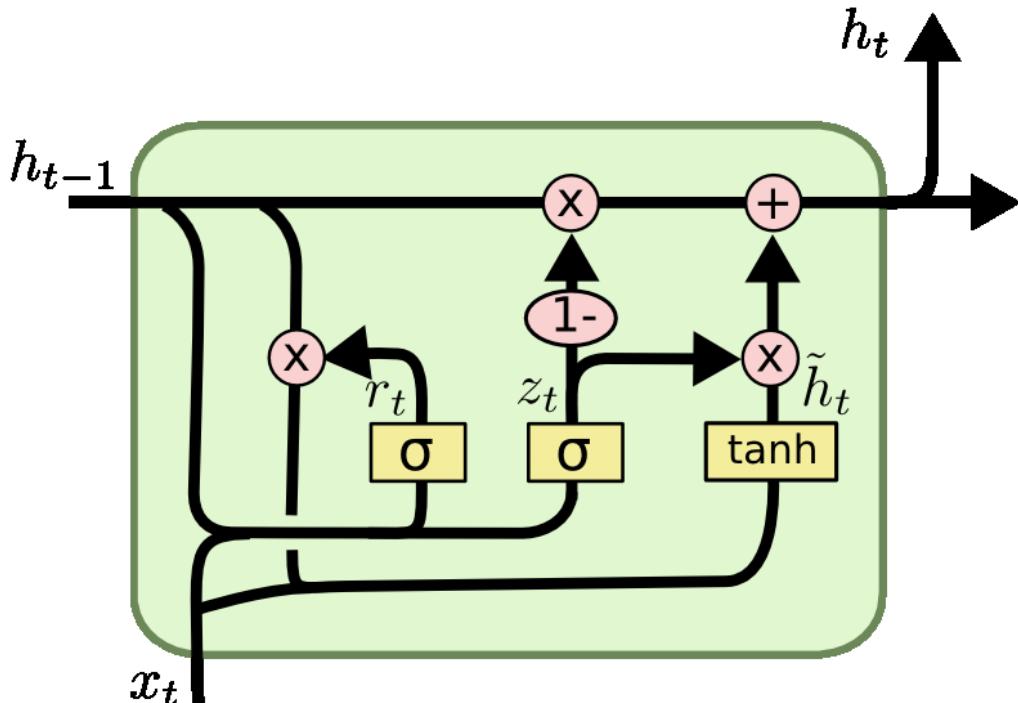
Hình 6.6: LSTM

6.3.2 Gated Recurrent Units (GRU)

Tương tự với LSTM, GRU [12] được sử dụng để khắc phục nhược điểm của RNN. GRU không chứa trạng thái tế bào, mô hình này sử dụng trạng thái ẩn để vận chuyển thông tin. GRU có hai cổng thông tin, gồm:

- Update Gate: Kết hợp giữa Forget Gate và Input Gate, cổng này có nhiệm vụ quyết định thông tin nào sẽ bỏ qua và thông tin nào cần thêm vào bộ nhớ.
- Reset Gate: Cổng này sẽ đặt lại thông tin trong quá khứ để loại bỏ những sự cố liên quan đến gradient. Reset Gate xác định các thông tin trong quá khứ nên được quên.

Chính nhờ cơ chế 2 cổng và có kèm hoạt động của tensor mà GRU có thể ghi nhớ thông tin dài hạn tốt hơn, huấn luyện nhanh hơn so với mô hình có cấu trúc phức tạp như LSTM.



Hình 6.7: GRU

7 Thực nghiệm và phân tích kết quả

7.1 Xây dựng thực nghiệm

Bộ dữ liệu thành ba tập train, validation, test. Tập train sẽ gồm dữ liệu từ ngày 01/01/2022 đến ngày 30/09/2024, tập validation gồm dữ liệu từ ngày 01/10/2024 đến ngày 31/12/2024 và tập test sẽ gồm dữ liệu từ ngày 01/01/2025 đến ngày 31/03/2025. Nhóm tiến hành xây dựng thực nghiệm mô hình dự đoán trên ba bộ dữ liệu VCB_dataset, BIDV_dataset và TCB_dataset và sử dụng 7 mô hình trong đó có 1 mô hình chuỗi thời gian SARIMAX, 4 mô hình machine learning Linear Regression, Ridge Regression, Random Forest, SVM và hai mô hình deep learning LSTM, GRU để dự đoán. Kết quả sẽ được đánh giá dựa trên các chỉ số MSE, RMSE, MAE, MAPE và hiệu suất (thời gian huấn luyện và thời gian tính toán), sau đó sử dụng ANOVA [13] và Tukey HSD [14] để kiểm định lại kết quả.

7.2 Kết quả thực nghiệm

7.2.1 Kết quả chỉ số đánh giá

7.2.1.1 Kết quả một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	56.1094	0.009711	0.013312	1.139314	0.009711	4.283530	4.000556
BIDV	3.917424	0.005869	0.008437	0.048895	0.005869	4.409818	3.153586
Techcombank	2.556011	0.013332	0.013871	1.717241	0.013332	4.946814	12.100565

Bảng 7.1: So sánh chỉ số MSE giữa các mô hình dự đoán một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	7.490622	0.098544	0.115376	1.067386	0.098544	2.069669	2.000139
BIDV	1.979248	0.076612	0.091853	0.221123	0.076612	2.099957	1.775834
Techcombank	1.598753	0.115464	0.117776	1.310436	0.115464	2.224144	3.478587

Bảng 7.2: So sánh chỉ số RMSE giữa các mô hình dự đoán một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	6.636755	0.070450	0.085435	0.514801	0.070450	1.646892	1.854924
BIDV	1.348229	0.054630	0.066016	0.110583	0.054630	1.988704	1.532414
Techcombank	1.224325	0.077033	0.079518	0.886202	0.077033	2.011590	3.363729

Bảng 7.3: So sánh chỉ số MAE giữa các mô hình dự đoán một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.105922	0.001107	0.001346	0.007835	0.001107	0.025601	0.029275
BIDV	0.034094	0.001378	0.001665	0.002766	0.001378	0.049191	0.038398
Techcombank	0.046239	0.002990	0.003087	0.032734	0.002990	0.074065	0.124808

Bảng 7.4: So sánh chỉ số MAPE giữa các mô hình dự đoán một bước

7.2.1.2 Kết quả nhiều bước

Vì SARIMAX không dự đoán được nhiều bước nên không có kết quả chỉ số đánh giá ở phần này.

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.636331	0.395026	2.726460	0.493373	5.325199	4.690114
BIDV	0.504201	0.393619	0.291791	0.359869	2.829083	4.236629
Techcombank	0.382141	0.307270	2.887087	0.248385	6.915676	5.255990

Bảng 7.5: So sánh chỉ số MSE giữa các mô hình dự đoán nhiều bước

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.793108	0.560506	1.611143	0.617842	2.296565	2.149609
BIDV	0.697677	0.551421	0.509360	0.523709	1.655054	1.918162
Techcombank	0.600519	0.491839	1.663269	0.445754	2.550973	2.148964

Bảng 7.6: So sánh chỉ số RMSE giữa các mô hình dự đoán nhiều bước

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.671295	0.442577	0.988774	0.494317	1.804450	1.981444
BIDV	0.529171	0.408248	0.359937	0.387493	1.519515	1.826001
Techcombank	0.511410	0.395984	1.232931	0.350596	2.403468	2.002800

Bảng 7.7: So sánh chỉ số MAE giữa các mô hình dự đoán nhiều bước

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.010717	0.007025	0.015279	0.007864	0.028088	0.031581
BIDV	0.013344	0.010296	0.008987	0.009772	0.037955	0.045150
Techcombank	0.019821	0.015357	0.046020	0.013578	0.088806	0.073909

Bảng 7.8: So sánh chỉ số MAPE giữa các mô hình dự đoán nhiều bước

7.2.1.3 Nhận xét chỉ số đánh giá

Dựa trên chỉ số MSE, các mô hình Linear Regression và SVM cho kết quả dự đoán tốt nhất trên cả ba ngân hàng, với sai số rất thấp và ổn định. Ngược lại, các mô hình deep learning như LSTM và GRU có sai số, đặc biệt GRU thể hiện không tốt trên dữ liệu của Techcombank. Trong dự đoán một bước, mô hình SARIMAX cũng cho kết quả kém hiệu quả hơn so với các mô hình hồi quy tuyến tính.

Chỉ số RMSE cho thấy các mô hình tuyến tính như Linear Regression, Ridge Regression và SVM hoạt động rất hiệu quả trên cả ba ngân hàng. Ngược lại, các mô hình deep learning như LSTM và GRU cho kết quả RMSE cao hơn, đặc biệt là với dữ liệu Techcombank. Riêng với mô hình Random Forest cho kết quả không đồng đều, cho thấy mô hình có thể phù hợp hơn với một số tập dữ liệu nhất định. Riêng với SARIMAX chỉ có kết quả cho BIDV và Techcombank nhưng RMSE cũng cao hơn so với các mô hình hồi quy tuyến tính.

Linear Regression và SVM tiếp tục là hai mô hình tốt nhất với sai số trung bình tuyệt đối nhỏ nhất trên cả ba ngân hàng, thể hiện độ chính xác và độ ổn định cao. Ridge Regression cũng cho kết quả khá tốt, tuy có sai số cao hơn. Random Forest thể hiện kết quả không ổn định, sai số trên tập dữ liệu Techcombank khá cao so với hai tập dữ liệu còn lại. Mô hình SARIMAX, LSTM và GRU có MAE cao hơn so với mô hình hồi quy tuyến tính.

Các mô hình Linear Regression, SVM và Ridge Regression tiếp tục vượt trội với chỉ số MAPE cực thấp cho thấy độ chính xác rất cao và phù hợp với dữ liệu. Cũng như các chỉ số MAE, RMSE, Random Forest thể hiện tính không ổn định trên tập dữ liệu Techcombank. SARIMAX có MAPE ở mức trung bình cao hơn hẳn các mô hình hồi quy. LSTM và GRU một lần nữa cho kết quả yếu nhất, với MAPE đặc biệt cao.

7.2.2 Hiệu suất

7.2.2.1 Kết quả một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	3.123368	0.000999	0.025892	5.558147	1.803368	2.420768	0.528394
BIDV	3.858862	0.001005	0.025099	2.225699	1.901575	1.022396	3.950188
Techcombank	4.053464	0.001026	0.024276	1.577517	0.906087	3.263742	0.262314

Bảng 7.9: So sánh thời gian huấn luyện giữa các mô hình dự đoán một bước

	SARIMAX	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.007007	0.001006	0.000000	0.015829	0.001003	0.083346	0.073400
BIDV	0.000000	0.000997	0.000000	0.012505	0.001022	0.126664	0.077997
Techcombank	0.007133	0.001215	0.001451	0.008198	0.000000	0.059171	0.058320

Bảng 7.10: So sánh thời gian tính toán (trên tập test) giữa các mô hình dự đoán một bước

7.2.2.2 Kết quả nhiều bước

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.004603	0.088840	11.75292	2.663372	1.108885	3.371456
BIDV	0.003079	0.080017	8.865192	2.874597	1.509137	0.332205
Techcombank	0.004096	0.102535	6.836975	1.639043	1.336780	1.506414

Bảng 7.11: So sánh thời gian huấn luyện giữa các mô hình dự đoán nhiều bước

	Linear R	Ridge R	Random Forest	SVM	LSTM	GRU
Vietcombank	0.000000	0.001004	0.036214	0.003151	0.098945	0.136719
BIDV	0.000000	0.001003	0.047267	0.003027	0.077362	0.082265
Techcombank	0.000000	0.000000	0.029622	0.002014	0.057753	0.061801

Bảng 7.12: So sánh thời gian tính toán (trên tập test) giữa các mô hình dự đoán nhiều bước

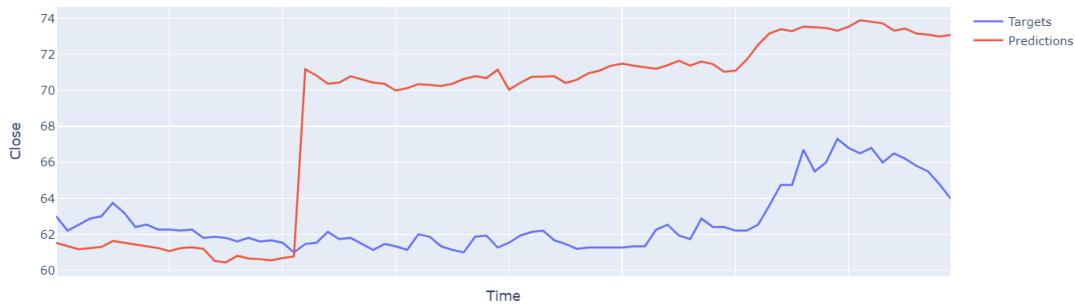
7.2.2.3 Đánh giá

Mô hình Linear và Ridge mang lại hiệu suất huấn luyện tốt nhất, trong khi Random Forest huấn luyện với thời gian lâu nhất ở cả một bước và nhiều bước. GRU ở một bước và nhiều bước đều có hiệu suất biến động lớn, báo hiệu cần xem xét lại mô hình. Ngược lại, mô hình có hiệu quả về mặt thời gian cho cả 3 ngân hàng là SVM.

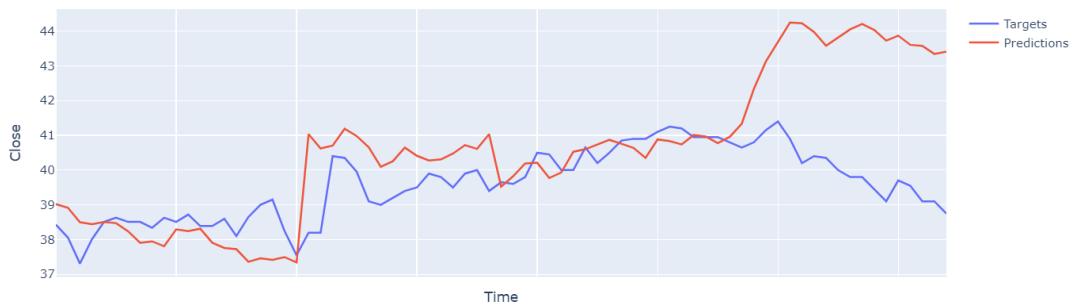
Thời gian tính toán của SVM rất thấp và đều ổn định cho nhu cầu một bước lẫn nhiều bước, phù hợp cho ứng dụng thời gian thực. Hiệu suất tính toán của Linear và Ridge là thấp nhất nhưng sẽ không ổn định trên cả ba ngân hàng. Khi áp dụng mô hình deep learning vào tính toán, GRU tốt hơn khi dự đoán một bước, nhưng LSTM sẽ nhanh và ổn định hơn khi dự đoán nhiều bước.

7.2.3 Kết quả dự đoán

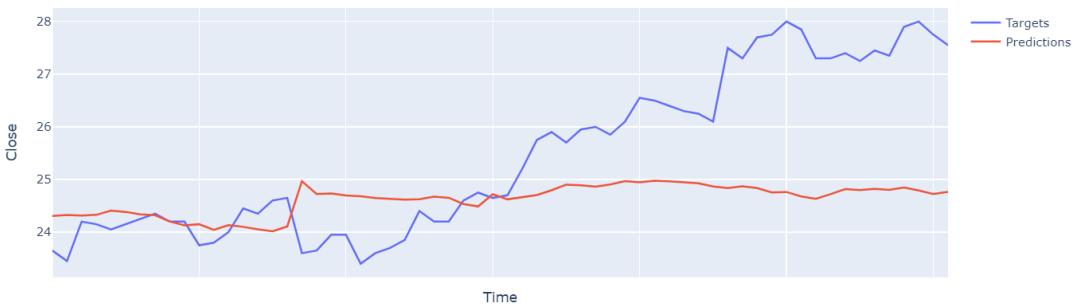
7.2.3.1 SARIMAX



Hình 7.1: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình SARIMAX



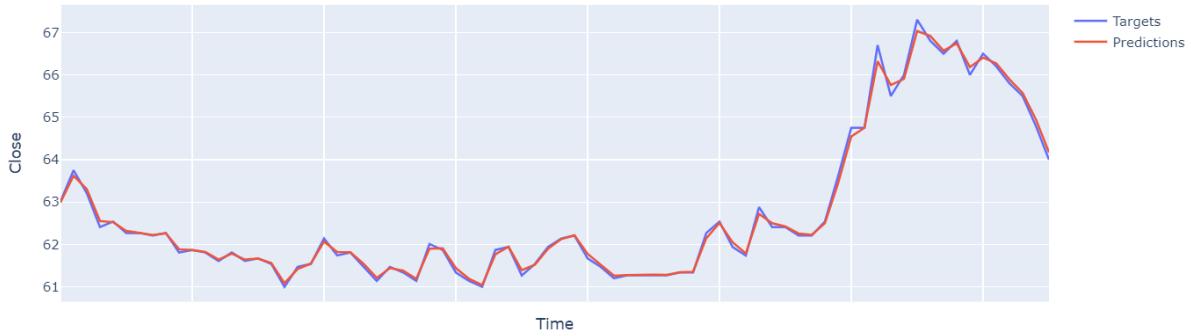
Hình 7.2: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình SARIMAX



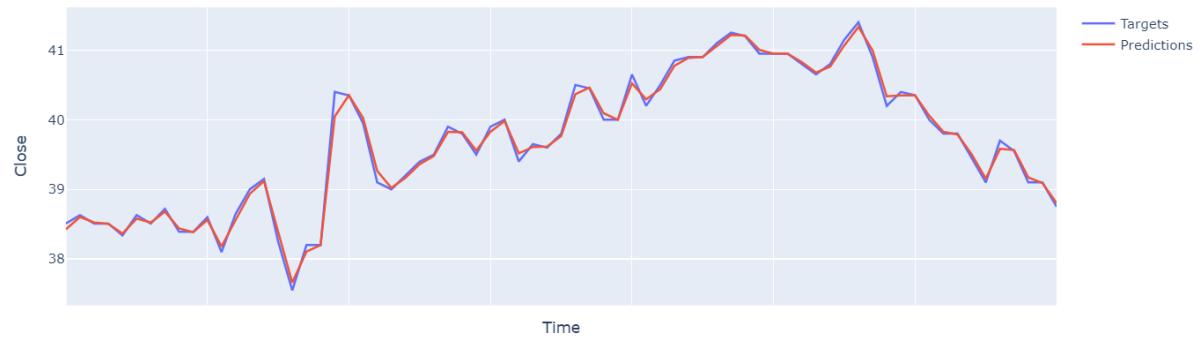
Hình 7.3: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình SARIMAX

Mặc dù SARIMAX là một mô hình mạnh mẽ cho chuỗi thời gian, nhưng với dữ liệu này, mô hình có vẻ không thể nắm bắt được các đặc điểm quan trọng, dẫn đến dự đoán không chính xác. Với hai bộ dữ liệu Vietcombank và BIDV, tại thời điểm dự đoán giá trị mới, kết quả dự đoán bị cao lên bất thường. Dù sau đó cũng có xu hướng đi ngang giống dữ liệu thực tế nhưng vẫn bị chênh lệch khoảng cách khá nhiều so với dữ liệu thực.

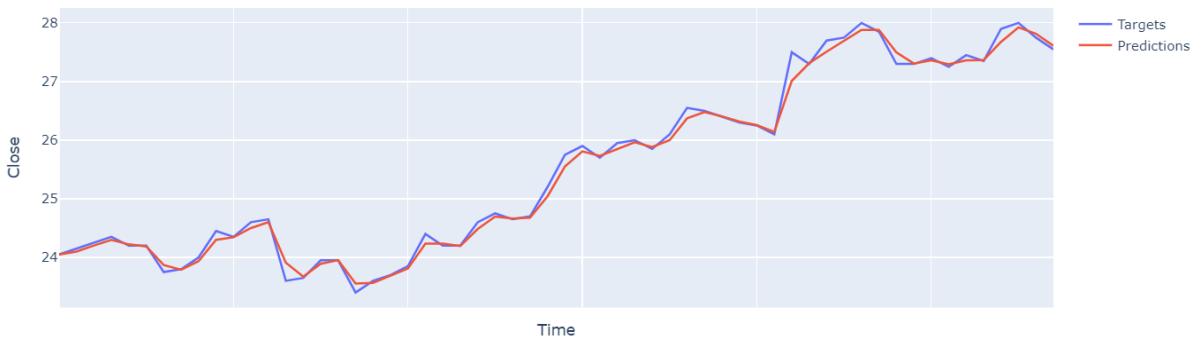
7.2.3.2 Linear Regression



Hình 7.4: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Linear Regression



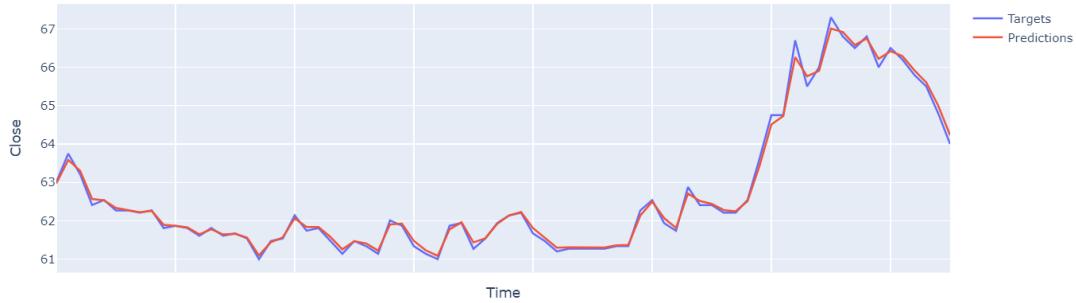
Hình 7.5: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Linear Regression



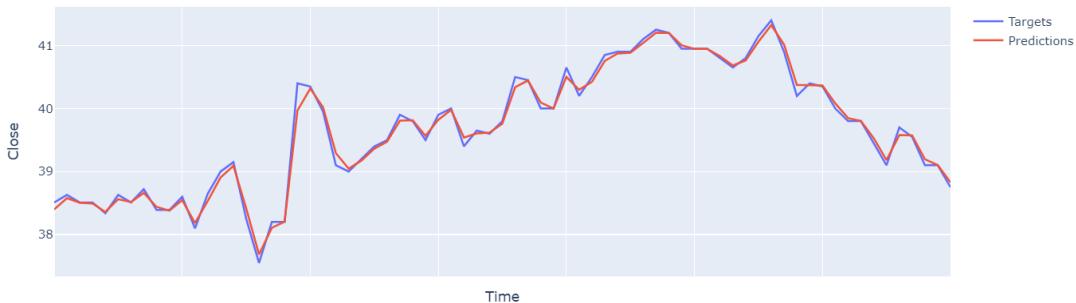
Hình 7.6: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Linear Regression

Linear Regression mặc dù là một mô hình đơn giản, nhưng lại thể hiện được sự chính xác cao trong việc dự đoán kết quả. Mô hình Linear Regression đã cho kết quả dự đoán rất chính xác, gần với thực tế, với sai số cực thấp trong các chỉ số MAE, RMSE, và MAPE.

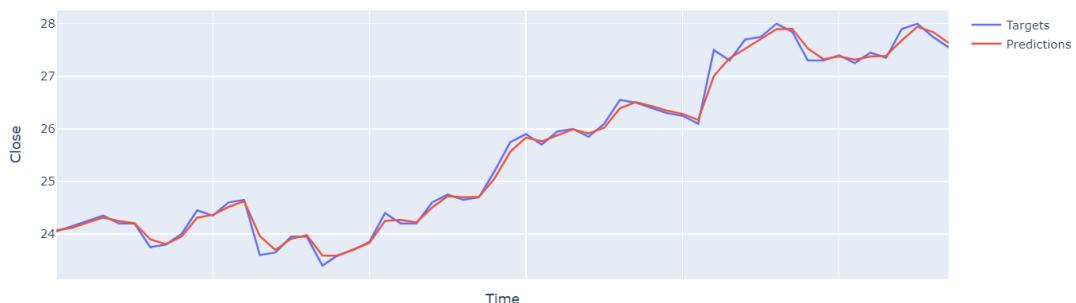
7.2.3.3 Ridge Regression



Hình 7.7: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Ridge Regression



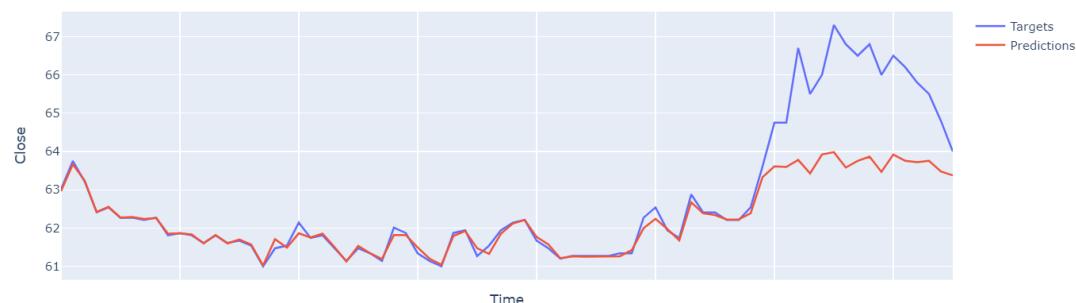
Hình 7.8: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Ridge Regression



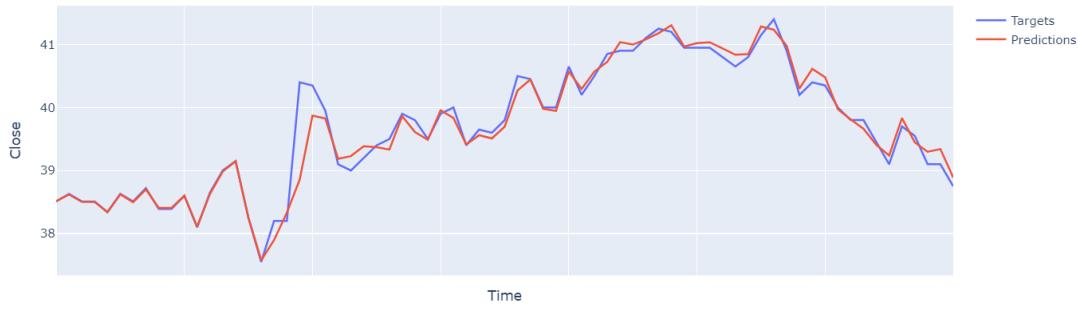
Hình 7.9: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Ridge Regression

Tương tự với Linear Regression, Ridge Regression cũng đã cho kết quả dự đoán gần đúng với dữ liệu thực tế.

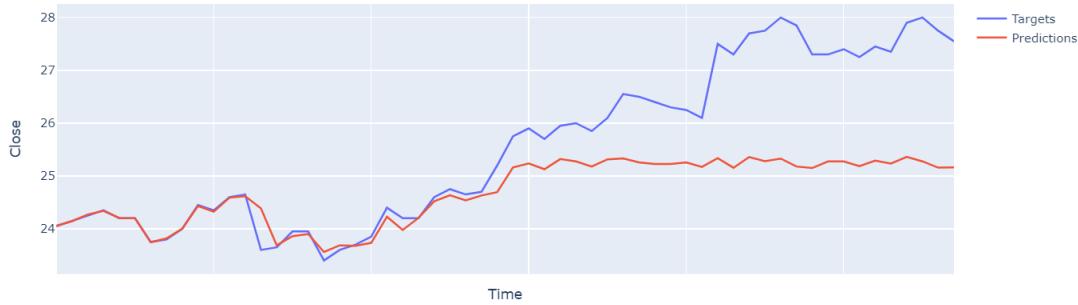
7.2.3.4 Random Forest



Hình 7.10: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình Random Forest



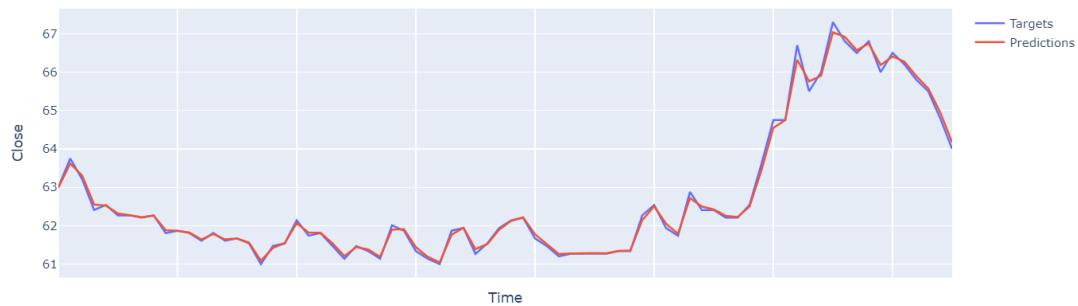
Hình 7.11: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình Random Forest



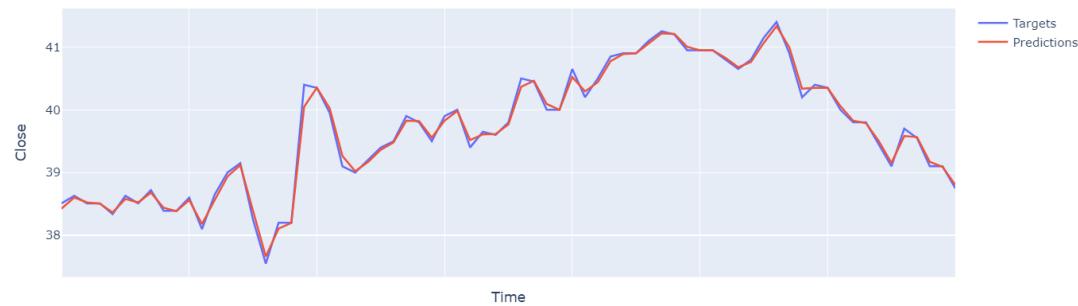
Hình 7.12: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình Random Forest

Mô hình Random Forest cho kết quả dự đoán có độ chính xác cao trên tập dữ liệu BIDV, tuy nhiên lại không hoạt động tốt trên tập dữ liệu Vietcombank và Techcombank. Điều này có thể cho thấy mô hình Random Forest này chỉ phù hợp với một vài bộ dữ liệu với những đặc trưng riêng, không nhạy với toàn bộ các dữ liệu.

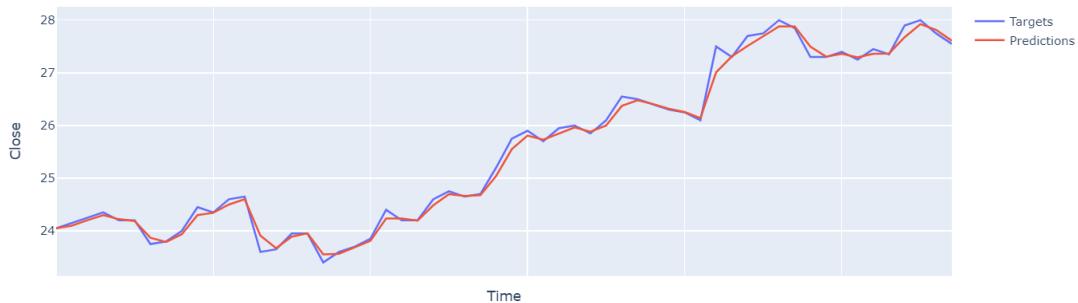
7.2.3.5 Support Vector Machine



Hình 7.13: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình SVM



Hình 7.14: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình SVM



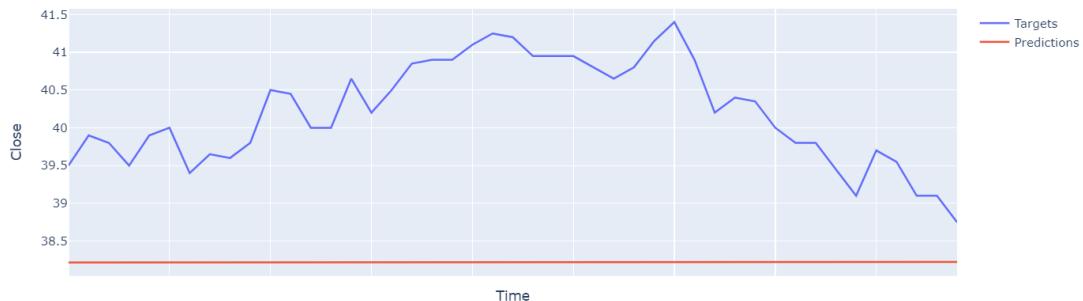
Hình 7.15: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình SVM

Mô hình SVM đã cho kết quả dự đoán khá chính xác trên tất cả các ngân hàng. Mô hình này cho thấy độ chính xác cao, với sai số thấp trong các chỉ số như MAE, RMSE và MAPE.

7.2.3.6 LSTM



Hình 7.16: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình LSTM



Hình 7.17: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình LSTM



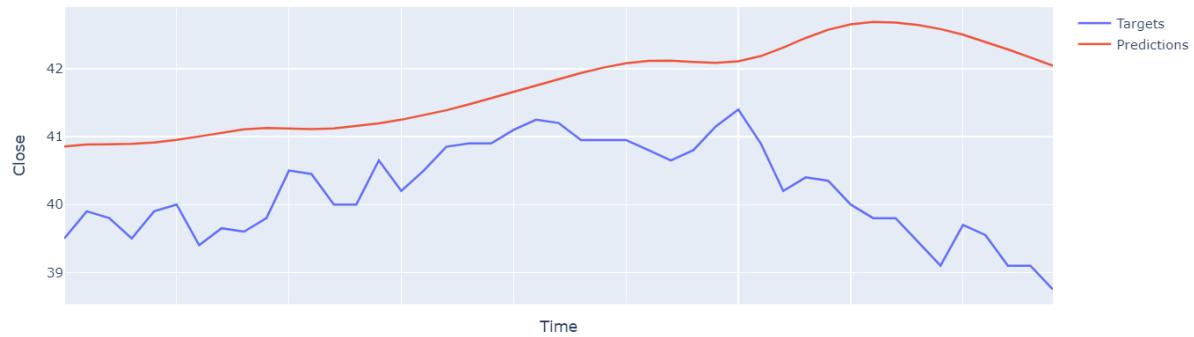
Hình 7.18: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình LSTM

Mô hình LSTM có kết quả dự đoán gần như là giống nhau với tất cả bản ghi trong tập test cho tất cả các ngân hàng Vietcombank, BIDV, và Techcombank. LSTM thường hoạt động tốt với bộ dữ liệu chuỗi thời gian dài hạn, trong trường hợp này mô hình có thể gặp khó khăn khi không có đủ dữ liệu.

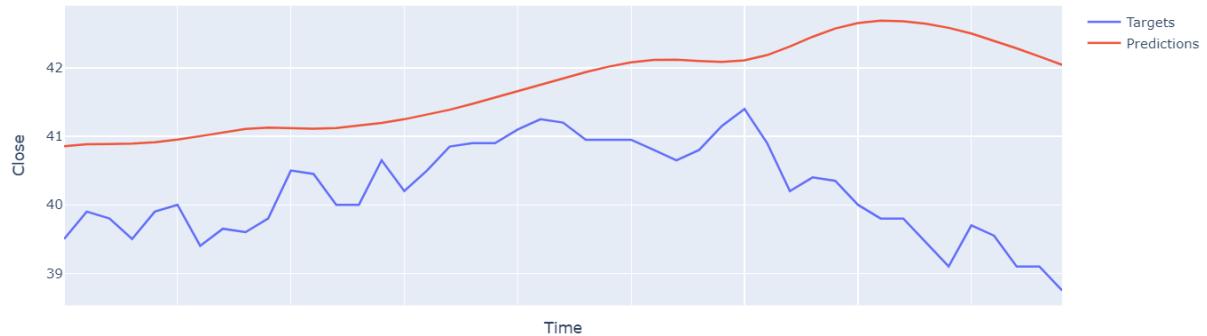
7.2.3.7 GRU



Hình 7.19: Kết quả dự đoán trên bộ dữ liệu Vietcombank của mô hình GRU



Hình 7.20: Kết quả dự đoán trên bộ dữ liệu BIDV của mô hình GRU



Hình 7.21: Kết quả dự đoán trên bộ dữ liệu Techcombank của mô hình GRU

Mô hình GRU đã có thể dự đoán được một chút xu hướng lên xuống trong giá trị dự đoán, tuy nhiên kết quả này không mang lại sự thay đổi rõ ràng. Mặc dù mô hình dự đoán có sự biến động, nhưng dự đoán không đủ chính xác để phản ánh xu hướng thực tế.

7.3 Phân tích kết quả

Đặt giả thuyết các chỉ số đánh giá của các mô hình bằng nhau:

- H_0 : Không có sự khác nhau giữa chỉ số đánh giá của các mô hình.

- H_1 : Có sự khác nhau giữa chỉ số đánh giá của các mô hình.

Với $\alpha = 0.02$, ta có:

7.3.1 Chỉ số đánh giá

7.3.1.1 Dự đoán 1 bước

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	119.617825	19.936304	5.042028	0.007061
Residual	13.0	51.402322	3.954025	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-6.4086	0.0212	-12.0159	-0.8013	True
1	3	GRU	Ridge	-6.4064	0.0213	-12.0137	-0.7991	True
2	5	GRU	SVM	-6.4086	0.0212	-12.0159	-0.8013	True

Hình 7.22: ANOVA và Tukey's HSD với MSE của các chỉ số đánh giá với các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.007061 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	18.423909	3.070652	16.270061	0.000023
Residual	13.0	2.453492	0.188730	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower
0	1	GRU	Linear Regression	-2.3213	0.0003	-3.5464
1	2	GRU	Random Forest	-1.5519	0.0101	-2.7769
2	3	GRU	Ridge	-2.3099	0.0003	-3.5349
3	5	GRU	SVM	-2.3213	0.0003	-3.5464
4	6	LSTM	Linear Regression	-2.0344	0.0010	-3.2594
5	7	LSTM	Random Forest	-1.2649	0.0412	-2.4900
6	8	LSTM	Ridge	-2.0229	0.0011	-3.2480
7	10	LSTM	SVM	-2.0344	0.0010	-3.2594
8	13	Linear Regression	SARIMAX	1.6921	0.0121	0.3225
9	18	Ridge	SARIMAX	1.6807	0.0128	0.3110
10	20	SARIMAX	SVM	-1.6921	0.0121	-3.0618

Hình 7.23: ANOVA và Tukey's HSD với RMSE của các chỉ số đánh giá với các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000023 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	15.287611	2.547935	14.374886	0.000046
Residual	13.0	2.304238	0.177249	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-2.1830	0.0004	-3.3702	-0.9958	True
1	2	GRU	Random Forest	-1.7465	0.0030	-2.9337	-0.5593	True
2	3	GRU	Ridge	-2.1734	0.0004	-3.3606	-0.9862	True
3	5	GRU	SVM	-2.1830	0.0004	-3.3702	-0.9958	True
4	6	LSTM	Linear Regression	-1.8150	0.0021	-3.0022	-0.6278	True
5	7	LSTM	Random Forest	-1.3785	0.0190	-2.5657	-0.1913	True
6	8	LSTM	Ridge	-1.8054	0.0023	-2.9926	-0.6182	True
7	10	LSTM	SVM	-1.8150	0.0021	-3.0022	-0.6278	True

Hình 7.24: ANOVA và Tukey's HSD với MAE của các chỉ số đánh giá với các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.00046 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	0.012002	0.002000	3.548797	0.026374
Residual	13.0	0.007328	0.000564	NaN	NaN

Hình 7.25: ANOVA và Tukey's HSD với MAPE của các chỉ số đánh giá với các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.026374 > 0.02$) nên chấp nhận giả thuyết H_0 , không có sự khác nhau giữa chỉ số đánh giá của các mô hình.

7.3.1.2 Dự đoán nhiều bước

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	71.980024	14.396005	12.983416	0.000171
Residual	12.0	13.305594	1.108799	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-4.2200	0.0038	-7.1079	-1.3321	True
1	3	GRU	Ridge	-4.3623	0.0029	-7.2502	-1.4744	True
2	4	GRU	SVM	-4.3604	0.0029	-7.2483	-1.4725	True
3	5	LSTM	Linear Regression	-4.5158	0.0022	-7.4037	-1.6279	True
4	6	LSTM	Random Forest	-3.0549	0.0361	-5.9428	-0.1670	True
5	7	LSTM	Ridge	-4.6580	0.0017	-7.5459	-1.7701	True
6	8	LSTM	SVM	-4.6561	0.0017	-7.5440	-1.7682	True

Hình 7.26: ANOVA và Tukey's HSD với MSE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000171 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	8.537477	1.707495	15.206072	0.000078
Residual	12.0	1.347484	0.112290	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-1.3751	0.0031	-2.2942	-0.4561	True
1	3	GRU	Ridge	-1.5377	0.0012	-2.4567	-0.6186	True
2	4	GRU	SVM	-1.5431	0.0012	-2.4622	-0.6241	True
3	5	LSTM	Linear Regression	-1.4704	0.0018	-2.3895	-0.5514	True
4	7	LSTM	Ridge	-1.6329	0.0007	-2.5520	-0.7139	True
5	8	LSTM	SVM	-1.6384	0.0007	-2.5574	-0.7194	True

Hình 7.27: ANOVA và Tukey's HSD với RMSE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000078 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	7.784222	1.556844	21.745043	0.000012
Residual	12.0	0.859144	0.071595	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-1.3661	0.0005	-2.1000	-0.6323	True
1	2	GRU	Random Forest	-1.0762	0.0036	-1.8100	-0.3424	True
2	3	GRU	Ridge	-1.5211	0.0002	-2.2550	-0.7873	True
3	4	GRU	SVM	-1.5259	0.0002	-2.2598	-0.7921	True
4	5	LSTM	Linear Regression	-1.3385	0.0006	-2.0724	-0.6047	True
5	6	LSTM	Random Forest	-1.0486	0.0045	-1.7824	-0.3148	True
6	7	LSTM	Ridge	-1.4935	0.0002	-2.2274	-0.7597	True
7	8	LSTM	SVM	-1.4983	0.0002	-2.2322	-0.7645	True

Hình 7.28: ANOVA và Tukey's HSD với MAE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000012 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có thể thấy có sự khác nhau giữa chỉ số đánh giá của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	0.005536	0.001107	3.373126	0.039178
Residual	12.0	0.003939	0.000328	NaN	NaN

Hình 7.29: ANOVA và Tukey's HSD với MAPE của các chỉ số đánh giá với các mô hình dự đoán nhiều bước

Riêng với giá trị kiểm định MAPE, $P > \alpha$ ($0.39178 > 0.02$) nên chấp nhận giả thuyết H_0 , kết luận không có sự khác nhau giữa giá trị kiểm định của các mô hình.

Kết luận chung: Sử dụng Tukey's HSD để kiểm định, ta thấy rõ các nhóm đang có sai số chênh lệch với nhau. Xét giá trị MSE, RMSE và MAE của các mô hình deep learning lớn hơn nhiều so với các mô hình machine learning. Cho thấy các mô hình machine learning như Linear Regression, Ridge Regression, Random Forest và SVM cho kết quả dự đoán tốt hơn so với GRU và LSTM.

7.3.2 Hiệu suất

Với $\alpha = 0.02$, ta có:

7.3.2.1 Thời gian huấn luyện

Đặt giả thuyết:

- H_0 : Không có sự khác nhau giữa thời gian huấn luyện của các mô hình.
- H_1 : Có sự khác nhau giữa thời gian huấn luyện của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	35.823165	5.970528	3.93723	0.016235
Residual	14.0	21.230000	1.516429	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	13	Linear Regression	SARIMAX	3.6776	0.0323	0.2443	7.1108	True
1	18	Ridge	SARIMAX	3.6535	0.0338	0.2202	7.0867	True

Hình 7.30: ANOVA và Tukey's HSD với thời gian huấn luyện của các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.016235 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có sự khác nhau giữa thời gian huấn luyện của các mô hình.

Sử dụng Tukey's HSD để kiểm định, ta thấy rõ các nhóm đang có sai số chênh lệch với nhau. Thời gian huấn luyện của mô hình SARIMAX lớn hơn thời gian huấn luyện của các mô hình machine learning Linear Regression, Ridge Regression. Cho thấy hiệu suất của các mô hình machine learning tối ưu hơn.

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	174.769200	34.953840	23.486733	0.000008
Residual	12.0	17.858852	1.488238	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	\
0	2	GRU	Random Forest	7.4150	0.0001	4.0693	
1	6	LSTM	Random Forest	7.8334	0.0001	4.4877	
2	9	Linear Regression	Random Forest	9.1478	0.0000	5.8020	
3	12	Random Forest	Ridge	-9.0612	0.0000	-12.4070	
4	13	Random Forest	SVM	-6.7594	0.0002	-10.1051	
		upper	reject				
0	10.7607	True					
1	11.1792	True					
2	12.4935	True					
3	-5.7155	True					
4	-3.4136	True					

Hình 7.31: ANOVA và Tukey's HSD với thời gian huấn luyện của các mô hình dự đoán nhiều bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000008 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có sự khác nhau giữa thời gian huấn luyện của các mô hình.

Sử dụng Tukey's HSD để kiểm định, ta thấy rõ các nhóm đang có sai số chênh lệch với nhau. Kết quả cho thấy thời gian huấn luyện của mô hình Random Forest lớn hơn nhiều so với các mô hình còn lại.

7.3.2.2 Thời gian tính toán

Đặt giả thuyết:

- H_0 : Không có sự khác nhau giữa thời gian tính toán của các mô hình.
- H_1 : Có sự khác nhau giữa thời gian tính toán của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	6.0	0.025636	0.004273	22.871101	0.000002
Residual	14.0	0.002615	0.000187	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-0.0688	0.0004	-0.1069	-0.0307	True
1	2	GRU	Random Forest	-0.0577	0.0021	-0.0958	-0.0196	True
2	3	GRU	Ridge	-0.0694	0.0003	-0.1075	-0.0313	True
3	4	GRU	SARIMAX	-0.0652	0.0007	-0.1033	-0.0271	True
4	5	GRU	SVM	-0.0692	0.0004	-0.1073	-0.0311	True
5	6	LSTM	Linear Regression	-0.0887	0.0000	-0.1268	-0.0505	True
6	7	LSTM	Random Forest	-0.0775	0.0001	-0.1157	-0.0394	True
7	8	LSTM	Ridge	-0.0892	0.0000	-0.1273	-0.0511	True
8	9	LSTM	SARIMAX	-0.0850	0.0000	-0.1231	-0.0469	True
9	10	LSTM	SVM	-0.0891	0.0000	-0.1272	-0.0509	True

Hình 7.32: ANOVA và Tukey's HSD với thời gian tính toán của các mô hình dự đoán một bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000002 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có sự khác nhau giữa thời gian huấn luyện của các mô hình.

	df	sum_sq	mean_sq	F	PR(>F)
Model	5.0	0.026205	0.005241	15.690252	0.000067
Residual	12.0	0.004008	0.000334	NaN	NaN

	index	group1	group2	meandiff	p-adj	lower	upper	reject
0	1	GRU	Linear Regression	-0.0936	0.0005	-0.1437	-0.0435	True
1	2	GRU	Random Forest	-0.0559	0.0261	-0.1060	-0.0058	True
2	3	GRU	Ridge	-0.0929	0.0005	-0.1431	-0.0428	True
3	4	GRU	SVM	-0.0909	0.0006	-0.1410	-0.0407	True
4	5	LSTM	Linear Regression	-0.0780	0.0022	-0.1281	-0.0279	True
5	7	LSTM	Ridge	-0.0774	0.0024	-0.1275	-0.0272	True
6	8	LSTM	SVM	-0.0753	0.0030	-0.1254	-0.0252	True

Hình 7.33: ANOVA và Tukey's HSD với thời gian tính toán của các mô hình dự đoán nhiều bước

Xét ANOVA, ta thấy $P < \alpha$ ($0.000067 < 0.02$) nên bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_1 , có sự khác nhau giữa thời gian huấn luyện của các mô hình.

Sử dụng Tukey's HSD để kiểm định thời gian tính toán nhiều bước, ta thấy rõ các nhóm đang có sai số chênh lệch với nhau. Kết quả cho thấy thời gian tính toán của các mô hình deep learning LSTM, GRU lâu hơn các mô hình machine learning.

7.4 Đánh giá chung

Dựa trên các chỉ số đánh giá và kiểm định, có thể thấy các mô hình machine learning như Linear Regression, Ridge Regression, Random Forest và SVM cho kết quả dự đoán tốt hơn so với GRU và LSTM. Nguyên nhân có thể đến từ việc các mô hình deep learning cần nhiều dữ liệu để học được các mẫu. Bộ dữ liệu nhóm đang dùng để thực

hiện đề tài có lượng dữ liệu là khá ít cho deep learning hoạt động hiệu quả. Trong khi đó, các mô hình machine learning truyền thống, tuy đơn giản nhưng phù hợp để dự đoán với dữ liệu nhỏ nên cho kết quả chính xác cao hơn. Bên cạnh đó, mô hình SAMIMAX cũng không được ưu tiên sử dụng trong bài toán này vì bị giới hạn khả năng dự đoán. SARIMAX chỉ hoạt động tốt trong dự đoán 1 bước tiếp theo. Khi muốn dự đoán nhiều bước (multi-step), mô hình phải lại dùng dữ liệu bước 1 làm đầu vào. Điều này gây ra sai số trong quá trình dự đoán và dẫn đến việc kết quả không chính xác.

8 Tổng kết

Trong đề tài này, nhóm đã thực hiện tìm hiểu, phân tích và xây dựng mô hình dự đoán giá cổ phiếu của ba Ngân hàng Vietcombank, BIDV và Techcombank. Với bộ dữ liệu thu thập nhóm đã tiến hành các phương pháp làm sạch dữ liệu, phân tích và khai phá dữ liệu (EDA). Sau đó, nhóm tiến hành chọn lọc các thuộc tính phục vụ cho bài toán. Cuối cùng, nhóm thực hiện xây dựng mô hình dự đoán bằng cách chạy thực nghiệm ba bộ dữ liệu trên 7 mô hình khác nhau. Kết quả dự đoán được đánh giá bằng các chỉ số MSE, RMSE, MAE và MAPE, sau đó thực hiện phân tích ANOVA và kiểm định lại bằng Tukey's HSD với mức ý nghĩa 0.02. Kết quả cho thấy các mô hình machine learning phù hợp để sử dụng cho bài toán này nhất.

Tuy nhiên, đồ án của nhóm cũng còn tồn tại vài điểm hạn chế và còn nhiều phần cần được bổ sung và cải thiện. Bộ dữ liệu nhóm thu thập được còn chưa nhiều về số lượng, dẫn đến kết quả của các mô hình deep learning chưa tốt. Điều này khiến chưa đánh giá khách quan được hiệu suất của hai mô hình deep learning trong đề tài. Bên cạnh đó, các thành viên của nhóm cũng chưa tiếp xúc với những đề tài liên quan đến cổ phiếu nói riêng hay đề tài kinh tế nói chung nên chưa có kinh nghiệm trong quá trình làm việc với các chỉ số kinh tế. Nhóm cũng chưa tiếp cận được nhiều các nguồn dữ liệu về kinh tế, tài chính nên còn hạn hẹp về vấn đề nguồn tham khảo.

References

- [1] VnExpress. “Chứng khoán biến động mạnh nhất một tháng.” [Truy cập lần cuối: 12/03/2024]. (Mar. 2024), [Online]. Available: <https://www.dnse.com.vn/senses/tin-tuc/co-phieu-ngan-hang-duy-nhat-tang-gia-tuan-qua-35023350>.
- [2] “Vnstock3 - giải pháp phân tích chứng khoán mở cho người việt.” [Truy cập lần cuối: 12/03/2024]. (), [Online]. Available: <https://pypi.org/project/vnstock/>.
- [3] “Vietstockfinance.” (), [Online]. Available: <https://finance.vietstock.vn/>.
- [4] “Tổng sản phẩm nội địa.” (), [Online]. Available: https://vi.wikipedia.org/wiki/T%E1%BB%95ng_s%E1%BA%A3n_ph%E1%BA%A9m_n%E1%BB%99i_%C4%91%E1%BB%8Ba.
- [5] “Chỉ số giá tiêu dùng.” [Truy cập lần cuối: 12/03/2024]. (), [Online]. Available: https://vi.wikipedia.org/wiki/Ch%E1%BB%89_s%E1%BB%91_gi%C3%A1_t%C3%A1i%C3%A9u_d%C3%B9ng.
- [6] F. Alharbi and D. Csala, “A seasonal autoregressive integrated moving average with exogenous factors (sarimax) forecasting model-based time series approach,” *Inventions*, vol. 7, p. 94, Oct. 2022. DOI: [10.3390/inventions7040094](https://doi.org/10.3390/inventions7040094).
- [7] A. Hoerl and R. Kennard, “Ridge regression,” *Encyclopedia of Statistical Sciences*, vol. 8, pp. 129–136, Aug. 2006. DOI: [10.1002/0471667196.ess2280.pub2](https://doi.org/10.1002/0471667196.ess2280.pub2).
- [8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324). [Online]. Available: <https://doi.org/10.1023/A:1010933404324>.
- [9] M. Awad and R. Khanna, “Support vector machines for classification,” in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress, 2015, pp. 39–66, ISBN: 978-1-4302-5990-9. DOI: [10.1007/978-1-4302-5990-9_3](https://doi.org/10.1007/978-1-4302-5990-9_3). [Online]. Available: https://doi.org/10.1007/978-1-4302-5990-9_3.
- [10] R. M. Schmidt, *Recurrent neural networks (rnns): A gentle introduction and overview*, 2019. arXiv: [1912.05911 \[cs.LG\]](https://arxiv.org/abs/1912.05911). [Online]. Available: <https://arxiv.org/abs/1912.05911>.
- [11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 1997. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1915014>.

- [12] R. Dey and F. M. Salem, *Gate-variants of gated recurrent unit (gru) neural networks*, 2017. arXiv: [1701.05923 \[cs.NE\]](https://arxiv.org/abs/1701.05923). [Online]. Available: <https://arxiv.org/abs/1701.05923>.
- [13] R. Henson, “Analysis of variance (anova),” in Dec. 2015, vol. 1, pp. 477–481, ISBN: 9780123973160. DOI: [10.1016/B978-0-12-397025-1.00319-5](https://doi.org/10.1016/B978-0-12-397025-1.00319-5).
- [14] A. Barnette, J. Barnette, and J. McLean, “Document resume ed 427 043 tm 029 421; the tukey honestly significant difference procedure and it's control of type i error rate,” Nov. 1998.

Đánh giá thành viên

Bảng đánh giá thành viên			
STT	Thành viên	MSSV	Mức độ đóng góp
1	Lưu Bảo Uyên	22521640	35%
2	Lê Vy	22521703	20%
3	Trần Lương Vân Nhi	22521044	20%
4	Trương Nhật Quang	22521207	20%
5	Nguyễn Anh Hải Ngọc	22520955	5%

Bảng phân công chi tiết				
STT	Nhiệm vụ	Phân công	Hoàn thành	Mức độ hoàn thành
1	Tìm đề tài, tìm nguồn lấy dữ liệu	Cả nhóm	Cả nhóm: Hoàn thành tốt nhiệm vụ	100%
2	Tìm hiểu về các biến cần thu thập	Bảo Uyên	Cả nhóm: Hoàn thành tốt nhiệm vụ	100%
3	Tìm hiểu về phương pháp phân tích dữ liệu nghiên cứu	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
4	Viết bản phác thảo sơ bộ	Vân Nhi	Hoàn thành tốt nhiệm vụ	100%
5	Tìm thêm biến ngoại vi	Cả nhóm	Nhật Quang, Bảo Uyên, Lê Vy: Hoàn thành tốt nhiệm vụ, Vân Nhi, Hải Ngọc: Không hoàn thành nhiệm vụ	Nhật Quang, Bảo Uyên, Lê Vy: 100%, Vân Nhi, Hải Ngọc: 0%
6	Phân tích định tính	Lê Vy, Nhật Quang	Hoàn thành tốt nhiệm vụ	100%
7	Lấy các chỉ số vĩ mô	Vân Nhi	Hoàn thành tốt nhiệm vụ	100%
Source code theo Python				
8	Cào dữ liệu tỷ giá	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
9	Làm sạch dữ liệu	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
10	Chuẩn bị dữ liệu	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
11	Phân tích các chỉ số về công ty	Hải Ngọc	Hoàn thành nhiệm vụ	80%
12	Phân tích các chỉ số về giá cổ phiếu + biến ngoại vi	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
13	Tìm hiểu về Random Forest, PCA để chọn đặc trưng	Lê Vy, Nhật Quang	Hoàn thành tốt nhiệm vụ	100%
14	Chạy SARIMAX	Hải Ngọc	Không hoàn thành nhiệm vụ	0%
15	Chạy Machine Learning	Nhật Quang	Hoàn thành tốt nhiệm vụ	97%
16	Chạy Deep Learning	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%

17	Đánh giá mô hình (hiệu suất, chỉ số)	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
18	Phân tích (nhận xét biểu đồ) các chỉ số của doanh nghiệp	Hải Ngọc	Hoàn thành nhiệm vụ	80%
19	Phân tích (nhận xét biểu đồ) các chỉ số vĩ mô	Vân Nhi	Hoàn thành tốt nhiệm vụ	100%
20	Phân tích (nhận xét biểu đồ) các chỉ số về giá cổ phiếu + biến ngoại vi	Vân Nhi	Hoàn thành tốt nhiệm vụ	100%
Source code theo R				
21	Phân tích	Lê Vy	Hoàn thành tốt nhiệm vụ	100%
22	Chuẩn bị dữ liệu	Lê Vy	Hoàn thành tốt nhiệm vụ	100%
23	SARIMAX	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
24	Machine learning	Nhật Quang	Hoàn thành tốt nhiệm vụ	100%
25	Deep learning	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
26	Đánh giá mô hình (hiệu suất, chỉ số)	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
27	Tổng hợp và hiệu chỉnh source code	Bảo Uyên	Hoàn thành tốt nhiệm vụ	100%
28	Viết báo cáo	Vân Nhi	Hoàn thành tốt nhiệm vụ	100%
29	Làm slide	Bảo Uyên, Lê Vy	Hoàn thành tốt nhiệm vụ	100%