DOI: https://doi.org/10.62985/j.huit_ojs.vol25.no3.203

XÂY DỰNG MÔ HÌNH DỰ BÁO THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM DƯỚI TÁC ĐỘNG CỦA CÁC YẾU TỐ VĨ MÔ

Phạm Thị Hà An

Khoa Tài chính ngân hàng, Trường Đại học Văn Lang

69/68 Đường Đặng Thùy Trâm, P.13, Quận Bình Thạnh, Thành phố Hồ Chí Minh

Email: *an.pth@vlu.edu.vn*

Ngày nhận bài: 20/5/2025; Ngày chấp nhận đăng: 26/6/2025

TÓM TẮT

Nghiên cứu này nhằm mục đích ứng dụng các mô hình học máy để đánh giá và so sánh khả năng dự báo xu hướng chỉ số VNIndex, HNX Index, Upcom Index và lợi nhuận của các ngành dưới tác động của các nhân tố vĩ mô. Nghiên cứu tập trung đánh giá và so sánh hiệu suất dự báo của các mô hình học máy bao gồm: Linear Regression, K-nearest Neighbors, Random Forest, Lasso Regression và Ridge Regression với bộ dữ liệu được sử dụng bao gồm dữ liệu lịch sử của các nhân tố vĩ mô: cung tiền M2, lãi suất huy động, lãi suất cho vay liên ngân hàng, chỉ số giá tiêu dùng (CPI), tỷ giá USD/VND; các chỉ số chứng khoán (Dow Jones, Nikkei 225, S&P 500, VN-Index, VN30-Index, Upcom-Index); các chỉ số khác (giá dầu, giá vàng) và dữ liệu tỷ suất sinh lợi của 5 nhóm ngành (Ngân hàng, Chứng khoán, Bất động sản, Thép, Bán lẻ). Nghiên cứu sử dụng các thông số như R2, MSE, RMSE, MAE để đánh giá hiệu suất dự báo của các mô hình. Kết quả nghiên cứu cho thấy các mô hình Linear Regression, Random Forest, Lasso Regression có hiệu suất dự báo tốt hơn các mô hình còn lại khi biến mục tiêu là tỷ suất sinh lời của các chỉ số giá chứng khoán. Trong khi đó, mô hình Ridge Regression lại chiếm ưu thế trong việc dự báo tỷ suất sinh lời của các nhóm ngành.

Từ khóa: Học máy, vĩ mô, dự báo, Linear, KNN, Random Forest, Lasso, Ridge.

1. GIỚI THIỆU NGHIÊN CỨU

Thị trường chứng khoán là một phần quan trọng của hệ thống tài chính, là nơi mà các loại tài sản tài chính được phát hành, giao dịch và định giá dựa trên yếu tố cung cầu; cũng là kênh huy động vốn dài hạn cho doanh nghiệp thông qua việc phát hành cổ phiếu, trái phiếu mang lại nguồn tài trợ trọng yếu để mở rộng hoạt động kinh doanh, sản xuất và phát triển.

Trải qua hơn hai thập kỷ hình thành và phát triển, thị trường chứng khoán Việt Nam đã từng bước khẳng định vai trò là kênh dẫn vốn trung và dài hạn quan trọng cho nền kinh tế. Từ khi chính thức đi vào hoạt động vào năm 2000, thị trường đã trải qua nhiều giai đoạn thăng trầm gắn liền với các biến động trong nước và quốc tế, từ khủng hoảng tài chính toàn cầu năm 2008 đến đại dịch COVID-19 và những biến động địa chính trị gần đây. Mặc dù vậy, thị trường vẫn đạt được nhiều cột mốc đáng ghi nhận như sự ra đời của các chỉ số ngành, sự tham gia ngày càng nhiều của nhà đầu tư cá nhân và tổ chức trong và ngoài nước, cũng như sự phát triển của các sản phẩm tài chính phái sinh.

Trong giai đoạn hiện nay, khi Việt Nam đẩy mạnh quá trình hội nhập kinh tế quốc tế, Chính phủ đã và đang thể hiện quyết tâm rõ rệt trong việc nâng hạng thị trường chứng khoán từ "thị trường cận biên" lên "thị trường mới nổi" vào năm 2025. Đây không chỉ là một mục tiêu kỹ thuật mà còn là bước đi chiến lược nhằm tăng cường uy tín và sức hấp dẫn của thị trường tài chính Việt Nam trên trường quốc tế. Việc được nâng hạng sẽ mở ra cơ hội thu hút dòng vốn ngoại lớn hơn, đặc biệt là từ các quỹ đầu tư chỉ số toàn cầu, từ đó cải thiện tính thanh khoản, minh bạch và hiệu quả vận hành của thị trường.

Cùng với sự phát triển của Công nghiệp 4.0, Machine learning ngày càng được sử dụng rộng rãi trong hầu hết các lĩnh vực, trong đó Tài chính – Ngân hàng là một trong những ngành tiêu biểu ứng dụng học máy vào việc dự báo. Đối với tệp dữ liệu lịch sử lớn và phức tạp, các mô hình học máy có khả

năng xử lý, phân tích và dự đoán xu hướng, mẫu và thông tin tiềm ẩn. Điều này giúp tối ưu hóa quy trình xử lý dữ liệu và đưa ra những dự đoán có độ tin cậy cao hơn về các xu hướng trong tương lai, từ đó hỗ trợ quá trình ra quyết định kinh doanh và chiến lược.

Trong các nghiên cứu trước đây, dữ liệu chứng khoán thường được nghiên cứu theo 2 hướng chính là dự báo và nghiên cứu ảnh hưởng. Đối với hướng nghiên cứu ảnh hưởng, hàng loạt các nghiên cứu trên thế giới và Việt Nam về ảnh hưởng của các nhân tố vĩ mô đến thị trường chứng khoán, cụ thể: nghiên cứu của Nguyễn Thị Như Quỳnh và cộng sự [1] đo lường tác động của 6 nhân tố kinh tế vĩ mô đến chỉ số VNIndex giai đoạn 2008-2018 bằng mô hình VECM; nghiên cứu của Nguyễn An Phú [2] về tác động của các nhân tố vĩ mô đến thị trường Chứng khoán Việt Nam; Nghiên cứu của Celebi và Hönig [3] về tác động của các yếu tố kinh tế vĩ mô đến thị trường chứng khoán Đức: Bằng chứng cho giai đoạn khủng hoảng, trước và sau khủng hoảng; Nghiên cứu của Muhammad Kamran Khan và Jian-Zhou Teng [4] về phản ứng của thị trường chứng khoán đối với các biến số kinh tế vĩ mô: Đánh giá bằng mô phỏng độ trễ phân phối tự hồi quy động.

Bên cạnh đó, hướng nghiên cứu dự báo thường tập trung vào tối ưu hóa các mô hình để mang lại hiệu quả dự báo tốt nhất cho việc dự báo chỉ số hoặc giá chứng khoán. Một số nghiên cứu theo hướng dự báo: Nghiên cứu của Harahap và cộng sự [5] sử dụng các biến số đầu vào là các nhân tố vĩ mô để dự báo chỉ số Nikkei 225 (N225) và Nikkei 400 (N400); Polamuri và cộng sự [6] đã thực hiện nghiên cứu về việc dự báo thị trường Chứng khoán bằng cách sử dụng nhiều mô hình học máy như Linear Regression, Multivariate Regression, Random Forest, và Extra Tree Regressor; Dự báo biến động cho thị trường chứng khoán kết hợp các biến kinh tế vĩ mô dựa trên GARCH-MIDAS và mô hình học sâu của Wu và cộng sự [7]. Ngoài ra, cũng có nghiên cứu sử dụng cả hai hướng nghiên cứu ảnh hưởng và dự báo như nghiên cứu của Lê Hoàng Anh và Nguyễn Lê Thanh Thy [8] về ứng dụng học máy vào nghiên cứu thị trường chứng khoán Việt Nam dưới ảnh hưởng của các nhân tố vĩ mô. Mặc dù mỗi hướng nghiên cứu có mục tiêu riêng nhưng có thể thấy được việc sử dụng kết quả của hướng nghiên cứu về ảnh hưởng để xác định các biến đầu vào cho hướng nghiên cứu về dự báo, kết hợp với các phương pháp học máy để xây dựng một mô hình dự báo sẽ mang lại một kết quả mới hơn. Trong nghiên cứu này, tác giả sẽ thực hiện dự báo các chỉ số chứng khoán (VNIndex, VN30-Index, Upcom-Index) và tỷ suất sinh lợi của một số nhóm ngành chính trên thị trường (Ngân hàng, Chứng khoán, Thép, Bất động sản và Bản lẻ) dưới tác động của các yếu tố vĩ mô (Lãi suất, lạm phát, tỷ giá, các nền kinh tế lớn, giá dầu, giá vàng,…) với dạng dữ liệu theo tháng. Với mục đích chính là xác định và nâng cao hiệu quả của việc sử dụng các mô hình học máy để phân tích và dự báo các chỉ số thị trường chứng khoán Việt Nam.

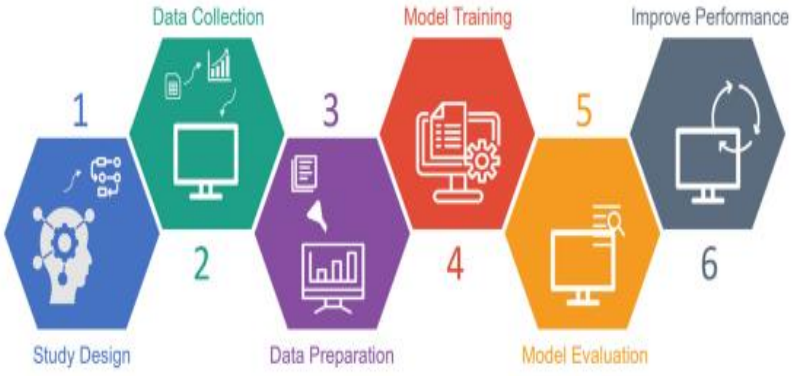
2. PHƯƠNG PHÁP NGHIÊN CỨU

Bài nghiên cứu sử dụng phương pháp định lượng, hồi quy với dữ liệu chuỗi thời gian. Đầu tiên, bộ dữ liệu bao gồm dữ liệu lịch sử của các nhân tố vĩ mô như: cung tiền M2, lãi suất huy động, lãi suất cho vay liên ngân hàng, chỉ số giá tiêu dùng (CPI), tỷ giá USD/VND; các chỉ số chứng khoán (Dow Jones, Nikkei 225, S&P 500, VN-Index, VN30-Index, Upcom-Index) và các chỉ số khác (giá dầu, giá vàng) được thu thập theo tháng từ ngày 01/01/2010 đến 31/12/2023 từ hai nguồn chính là FiinPro-X và Bloomberg.com. Ngoài ra, dữ liệu tỷ suất sinh lợi của 5 nhóm ngành tiêu biểu (Ngân hàng, Chứng khoán, Bất động sản, Thép, Bán lẻ) trên thị trường chứng khoán được thu thập theo cách tính nội bộ của Công ty Cổ phần Chứng khoán Rồng Việt.

Để cung cấp các giải pháp mới và thông minh cho các câu hỏi nghiên cứu ngày càng phức tạp, các mô hình học máy cần phải học hỏi và phát triển từ dữ liệu hiện có, cũng như liên tục cải thiện chúng trong môi trường luôn thay đổi. Các bước cơ bản trong việc xây dựng mô hình ML bao gồm: thiết kế nghiên cứu, thu thập dữ liệu, chuẩn bị dữ liệu, huấn luyện mô hình, đánh giá mô hình và cải thiện hiệu suất.

Bộ dữ liệu sau đó được tiền xử lý, kiểm định tính dừng, tính toán TSSL, độ trễ và chuẩn hóa dữ liệu thành bộ dữ liệu hoàn chỉnh trước khi đưa vào các mô hình học máy. Quy trình nghiên cứu được tiến hành theo Hình 1. Cụ thể, Bước đầu tiên là xác định rõ mục tiêu nghiên cứu, câu hỏi nghiên cứu, biến đầu vào và biến đầu ra cần dự báo hoặc phân loại. Đây là nền tảng giúp định hướng toàn bộ quy trình sau đó. Sau khi xác định được mục tiêu, tác giả tiến hành thu thập dữ liệu từ các nguồn khác nhau như dữ liệu kinh tế vĩ mô, tài chính, thị trường, v.v. Dữ liệu cần đảm bảo tính đầy đủ, độ tin cậy và có liên quan đến bài toán nghiên cứu. Dữ liệu sau khi thu thập sẽ được làm sạch, chuẩn hóa, xử lý các giá trị thiếu, biến đổi định dạng và chia tập dữ liệu thành tập huấn luyện và kiểm định. Đây là bước rất quan trọng để tăng độ chính xác của mô hình. Tác giả sử dụng các mô hình học máy như: Linear Regression,

KNN, Random Forest, Lasso Regression, Ridge Regression cho bài nghiên cứu. Sau đó, tác giả thêm vào các thư viện cần thiết như: pandas, numpy, matplotlib, seaborn os, statsmodels, sklearn và matplotlib.pyplot,… và bắt đầu xây dựng các mô hình bằng cách chia tách bộ dữ liệu hoàn chỉnh thành tập huấn luyện và tập kiểm tra, với tỷ lệ 80:20. Sau khi chạy mô hình, tác giả xem xét và đánh giá về kết quả dự báo qua thông số như R2, MSE, RMSE, MAE để đánh giá hiệu suất dự báo của từng mô hình.



Hình 1. Quy trình nghiên cứu

Để tìm ra mô hình có hiệu suất dự báo cao nhất đối với từng biến mục tiêu, tác giả tiến hành cải thiện mức độ hiệu quả cho từng mô hình. Các phương pháp cải thiện tác giả đã sử dụng bao gồm: chuẩn hóa dữ liệu bằng StandardScaler, tìm test_size tối ưu bằng cách sử dụng cross-validation, điều chỉnh tham số tối ưu bằng GridSearchCV,… tùy thuộc vào đặc điểm của từng mô hình.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả huấn luyện và hiệu suất dự báo của các mô hình

Đối với biến mục tiêu là r_vnindex, kết quả hiệu suất dự báo của các mô hình như sau:

Bảng 1. Hiệu suất dự báo của các mô hình với biến mục tiêu r_vnindex

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.613	0.517	20.167	4.491	3.953
Random Forest	0.851	-0.173	42.191	6.495	5.265
KNN	0.141	-0.153	41.465	6.439	5.114
Lasso	0.853	0.008	32.501	5.701	4.559
Ridge	0.222	-0.048	37.684	6.139	4.944

Mô hình Linear Regression có giá trị R2_test là 51.6%, cao nhất trong số các mô hình so sánh. Giá trị R2_test cao cho thấy mô hình có khả năng dự đoán tốt trên dữ liệu kiểm tra. Đồng thời, mô hình Linear Regression có RMSE và MAE lần lượt là 4.491 và 3.953, thấp nhất trong số các mô hình so sánh.

Bảng 2. Hiệu suất dự báo của các mô hình với biến mục tiêu r_vn30

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.196	0.011	33.826	5.816	4.575
Random Forest	0.842	-0.030	5.763	33.214	4.636
KNN	0.085	-0.183	42.549	6.523	5.317
Lasso	0.847	-0.001	32.290	5.682	4.529
Ridge	0.230	-0.066	38.323	6.191	5.024

Mặc dù tất cả các mô hình đều có giá trị R2_test thấp hoặc âm, mô hình Lasso Regression có giá trị R2_test gần bằng 0 nhất (-0.001), điều này cho thấy mô hình này có khả năng dự đoán tốt hơn các mô hình khác. Ngoài ra, mô hình Lasso Regression có MSE, RMSE và MAE thấp nhất trong số các mô hình so sánh, lần lượt là 32.290, 5.682 và 4.529. Điều này cho thấy mô hình này có sai số dự đoán nhỏ hơn, dự báo chính xác hơn.

Bảng 3. Hiệu suất dự báo của các mô hình với biến mục tiêu r_upcom

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.248	-0.110	39.926	6.319	5.178
Random Forest	0.848	0.071	11.416	130.332	5.368
KNN	0.095	-0.149	41.324	6.428	5.267
Lasso	0.856	0.047	133.635	11.560	5.531
Ridge	0.248	-0.108	39.837	6.312	5.170

Mô hình Random Forest Regression cho thấy hiệu suất tốt nhất với chỉ số R2_test cao nhất (0.071) và MSE thấp nhất (11.416) trong số các mô hình được so sánh. Mặc dù RMSE của nó cao hơn so với các mô hình khác, điều này có thể do các dự đoán cực trị làm tăng RMSE. Tuy nhiên, MSE thấp cho thấy mô hình này vẫn có khả năng dự đoán chính xác hơn.

Bảng 4. Hiệu suất dự báo của các mô hình với biến mục tiêu Bán lẻ (banle)

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.501	0.364	22.852	4.780	4.019
Random Forest	0.832	-0.148	0.086	0.007	0.067
KNN	0.091	-0.435	0.006	0.079	0.068
Lasso	0.837	-0.090	0.007	0.084	0.066
Ridge	0.423	0.179	29.536	5.435	4.508

Mô hình Linear Regression cho thấy hiệu suất tốt nhất với chỉ số R2_test cao nhất (36.44%) và MSE, RMSE hợp lý. Mặc dù Random Forest và Lasso có MSE và RMSE thấp hơn, giá trị R2_test âm cho thấy chúng không phù hợp cho biến mục tiêu này. Mô hình Linear có MAE (4.019) khá tốt, cho thấy sai số trung bình tuyệt đối của mô hình này hợp lý và thấp hơn so với Ridge.

Bảng 5. Hiệu suất dự báo của các mô hình với biến mục tiêu Bất động sản (bds)

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.111	-0.099	0.007	0.085	0.071
Random Forest	0.816	-0.066	0.106	0.011	0.072
KNN	0.102	-0.232	0.007	0.085	0.071
Lasso	0.847	-0.001	0.011	0.103	0.070
Ridge	0.431	0.182	29.399	5.422	4.271

Mô hình Ridge Regression cho thấy hiệu suất tốt nhất với chỉ số R2_test cao nhất (0.182), cho thấy mô hình này có khả năng giải thích biến thiên của biến mục tiêu tốt hơn so với các mô hình khác. Mô hình Ridge có MAE (4.271) thấp nhất, cho thấy sai số trung bình tuyệt đối của mô hình này tốt hơn so với các mô hình khác.

Bảng 6. Hiệu suất dự báo của các mô hình với biến mục tiêu Chứng khoán (ck)

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.149	-0.065	0.014	0.120	0.096
Random Forest	0.824	-0.018	0.118	0.014	0.099
KNN	0.110	-0.287	0.016	0.126	0.109
Lasso	0.848	0.051	0.013	0.113	0.096
Ridge	0.513	0.348	23.438	4.841	4.096

Dựa trên bảng kết quả, mô hình Ridge Regression được chọn làm mô hình dự báo tối ưu với giá trị R2_test cao nhất (34.8%) và MAE nhỏ nhất (4.09). Mặc dù MSE (23.438) và RMSE (4.841) của Ridge Regression cao hơn, giá trị R2_test cao khẳng định hiệu suất dự báo vượt trội của mô hình này.

Bảng 7. Hiệu suất dự báo của các mô hình với biến mục tiêu Ngân hàng (nh)

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.213	-0.504	0.009	0.093	0.075
Random Forest	0.832	-0.053	0.072	0.005	0.058
KNN	0.152	-0.186	0.007	0.084	0.067
Lasso	0.847	0.031	0.005	0.069	0.055
Ridge	0.503	0.272	26.162	5.115	4.073

Mô hình Ridge Regression là lựa chọn tốt nhất cho biến mục tiêu nh dựa trên chỉ số R2_test cao nhất (27.23%). Dù MSE và RMSE cao, R2_test cao vượt trội cho thấy mô hình này có khả năng dự báo chính xác và ổn định nhất so với các mô hình còn lại.

Bảng 8. Hiệu suất dự báo của các mô hình với biến mục tiêu Thép (thep)

Mô hình	R2_train	R2_test	MSE	RMSE	MAE
Linear	0.155	-0.346	0.012	0.109	0.089
Random Forest	0.822	-0.008	0.083	0.007	0.066
KNN	0.097	-0.281	0.010	0.099	0.078
Lasso	0.842	0.026	0.007	0.081	0.061
Ridge	0.461	0.295	25.353	5.035	4.099

Mô hình Ridge Regression có giá trị R2_test cao nhất (29.8%), cho thấy mô hình này có khả năng giải thích biến thiên của biến mục tiêu tốt nhất so với các mô hình khác.

4. KẾT LUẬN

Nhìn chung, kết quả sau cải thiện của các mô hình Linear Regression, KNN và Lasso cho thấy hiệu suất dự báo chỉ cải thiện trên một số biến mục tiêu cụ thể. Chẳng hạn Linear Regression sau cải thiện làm tăng hiệu suất dự báo của biến mục tiêu r_vnindex, với R2_test tăng từ 47.2% lên 51.6%. Tuy nhiên, đối với hiệu suất dự báo cho các biến còn lại có hệ số R2 thấp hoặc âm, cho thấy cho thấy mô hình Linear Regression sau cải thiện chỉ phù hợp để dự báo TSSL của toàn bộ thị trường (r_vnindex). Đối với mô hình Random Forest và Ridge Regression, kết quả sau cải thiện không khả quan với mục đích cải thiện hiệu quả dự báo. Do đó, tiếp tục sử dụng mô hình trước khi cải thiện để mang lại hiệu

suất dự báo tốt nhất đối với các biến mục tiêu. Các kết quả đạt được cho thấy các mô hình Linear Regression, Random Forest, Lasso Regression có hiệu suất dự báo tốt hơn các mô hình còn lại khi biến mục tiêu là TSSL của các chỉ số giá chứng khoán. Trong khi đó, mô hình Ridge Regression lại chiếm ưu thế trong việc dự báo TSSL của phần lớn các ngành trong biến mục tiêu.

TÀI LIỆU THAM KHẢO

1. Nguyễn Thị Như Quỳnh, Võ Thị Hương Linh - Tác động của một số yếu tố kinh tế vĩ mô đến chỉ số giá chứng khoán tại Việt Nam, Tạp chí Khoa học Đại học Mở Thành phố Hồ Chí Minh - Kinh tế và Quản trị Kinh doanh **14** (3) (2019)47–63 https://doi.org/10.46223/HCMCOUJS.econ.vi.14.3.477.2019.
2. Nguyễn An Phú - Tác động của các yếu tố kinh tế vĩ mô đến thị trường chứng khoán Việt Nam, Tạp chí Tài chính, Kỳ 2 tháng 7/2024.
3. Celebi K., Hönig M. - The impact of macroeconomic factors on the german stock market: evidence for the crisis, pre- and post-crisis Periods, International Journal of Financial Studies **7** (2) (2019). https://doi.org/10.3390/ijfs7020018
4. Megaravalli A.V., Sampagnaro G. - Macroeconomic indicators and their impact on stock markets in ASIAN 3: A pooled mean group approach, Cogent Economics & Finance **6** (1) (2018) 1432450–1432450. https://doi.org/10.1080/23322039.2018.1432450
5. Harahap L.A., Lipikorn R., and Kitamoto A. - Nikkei stock market price index prediction using machine learning, J. Phys.: Conf. Ser. **1566** (1) (2020) 012043. https://doi.org/10.1088/1742-6596/1566/1/012043
6. Polamuri D., Srinivas K., and Mohan A. - Stock market prices prediction using random forest and extra tree regression, International Journal of Recent Technology and Engineering **8** (2019) 1224-1228. https://doi.org/10.35940/ijrte.C4314.098319
7. Wu X., Yin X., and Mei X. - Forecasting the volatility of european union allowance futures with climate policy uncertainty using the EGARCH-MIDAS model, Sustainability **14** (7) (2022). https://doi.org/10.3390/su14074306.
8. Lê Hoàng Anh, Nguyễn Lê Thanh Thy - Ứng dụng phương pháp học máy trong giao dịch chứng khoán theo chỉ báo bằng ngôn ngữ lập trình Python, Tạp chí Khoa học và công nghệ Trường Đại học Bình Dương **7** (1) (2024). https://doi.org/10.56097/binhduonguniversityjournalofscienceandtechnology.v7i1.212.

ABSTRACT

BUILDING A FORECAST MODEL OF VIETNAM'S STOCK MARKET UNDER THE IMPACT OF MACRO FACTORS

Pham Thi Ha An

Faculty of Finance and Banking, Van Lang University

69/68 Dang Thuy Tram Street, Ward 13, Binh Thanh District, Ho Chi Minh City

Email: *an.pth@vlu.edu.vn*

The purpose of this research is to apply machine learning models to evaluate and compare the forecasting capabilities of the VNIndex, HNX Index, Upcom Index trends, and the profitability of the industries under the influence of macroeconomic factors. The study focuses on assessing and comparing the forecasting performance of various machine learning models, including Linear Regression, K-nearest Neighbors, Random Forest, Lasso Regression, and Ridge Regression. The dataset utilized comprises historical data of macroeconomic factors such as money supply M2, deposit interest rates, overnight lending rates, consumer price index (CPI), USD/VND exchange rate; stock indices (Dow

Jones, Nikkei 225, S&P 500, VN-Index, VN30-Index, Upcom-Index); other indices (oil prices, gold prices); and profitability data of five industry groups (Banking, Securities, Real Estate, Steel, Retail). The research employs metrics such as R2, MSE, RMSE, and MAE to evaluate the forecasting performance of the models. The results reveal that the Linear Regression, Random Forest, and Lasso Regression models exhibit superior forecasting performance compared to the other models when the target variable is the profitability of stock indices. On the other hand, the Ridge Regression model demonstrates higher performance in forecasting the profitability of industries.

Keywords: machine learning, macroeconomic factors, forecasting, Linear Regression, KNN, Random Forest, Lasso, Ridge.