

KẾT HỢP HỒI QUY LASSO, RIDGE, PRINCIPAL COMPONENT ANALYSIS VÀ CÁC MÔ HÌNH MACHINE LEARNING TRONG DỰ BÁO CHỈ SỐ ỔN ĐỊNH TÀI CHÍNH NGÂN HÀNG: NGHIÊN CỨU TẠI VIỆT NAM

Phạm Thủy Tú¹

Trường Đại học Tài chính - Marketing, TP. Hồ Chí Minh, Việt Nam

Phạm Anh Thủy

Trường Đại học Ngân hàng, TP. Hồ Chí Minh, Việt Nam

Ngày nhận: 19/09/2024; Ngày hoàn thành biên tập: 20/03/2025; Ngày duyệt đăng: 27/03/2025

DOI: <https://doi.org/10.38203/jiem.vi.092024.1186>

Tóm tắt: Nghiên cứu này đánh giá hiệu quả của hồi quy Lasso, Ridge và phân tích thành phần chính (PCA) khi kết hợp với các mô hình học máy trong dự báo chỉ số Z-score - thước đo mức độ ổn định tài chính ngân hàng. Dữ liệu nghiên cứu gồm 30 ngân hàng thương mại Việt Nam giai đoạn 2010-2023. Các mô hình học máy được sử dụng bao gồm Linear Regression, Gradient Boost, Decision Tree, Random Forest, KNN, SVR, AdaBoostRegressor và XGBRegressor. Hồi quy Lasso và Ridge hỗ trợ chọn lọc biến và giảm thiểu nhiễu trong ước lượng, trong khi PCA giúp giảm chiều dữ liệu. Kết quả cho thấy mô hình Gradient Boost kết hợp Lasso đạt hiệu suất dự báo cao nhất, trên 95%. Hơn nữa, sự kết hợp giữa Lasso, Ridge và học máy mang lại kết quả tốt hơn so với PCA. Các kỹ thuật tiền xử lý dữ liệu cũng đóng vai trò quan trọng trong việc nâng cao độ chính xác của mô hình. Nghiên cứu cung cấp bằng chứng thực nghiệm vững chắc về hiệu quả của việc tích hợp các phương pháp hồi quy và học máy trong đánh giá ổn định tài chính ngân hàng. Kết quả này không chỉ góp phần vào nền tảng học thuật về ứng dụng học máy trong tài chính mà còn hỗ trợ nhà quản lý trong việc ra quyết định chiến lược nhằm giảm thiểu rủi ro và tăng cường ổn định hệ thống ngân hàng.

Từ khóa: Lasso, Học máy, Ngân hàng, Phân tích thành phần chính (PCA), Ridge, Z-score

COMBINING LASSO, RIDGE, PRINCIPAL COMPONENT ANALYSIS, AND MACHINE LEARNING MODELS IN FORECASTING THE BANKING FINANCIAL STABILITY INDEX IN VIETNAM

Abstract: This study evaluates the effectiveness of Lasso and Ridge regression, along with Principal Component Analysis (PCA), when integrated with Machine Learning models in forecasting the Z-score—a key indicator of banking financial

¹ Tác giả liên hệ, Email: pttu@ufm.edu.vn

stability. The research utilizes secondary data from 30 Vietnamese commercial banks during the 2010-2023 period. The Machine Learning models employed include Linear Regression, Gradient Boost, Decision Tree, Random Forest, KNN, SVR, AdaBoostRegressor, and XGBRegressor. Lasso and Ridge regression aid in feature selection and noise reduction in estimation, while PCA assists in dimensionality reduction. The results indicate that the Gradient Boost model combined with Lasso achieves the highest forecasting performance, exceeding 95%. Furthermore, the integration of Lasso, Ridge, and Machine Learning models outperforms PCA. Data preprocessing techniques also play a crucial role in enhancing model accuracy. This study provides robust empirical evidence on the effectiveness of combining regression techniques with Machine Learning for assessing banking financial stability. The findings contribute not only to the academic understanding of Machine Learning applications in finance but also offer practical implications, assisting policymakers in making strategic decisions to mitigate risks and strengthen banking system stability.

Keywords: Lasso, Machine Learning, Banking, Principal Component Analysis, Ridge, Z-score

1. Giới thiệu

Ổn định tài chính được xem là điều kiện tiên quyết để tăng trưởng kinh tế bền vững, bởi hệ thống tài chính đóng vai trò trung gian cho hầu hết các giao dịch trong nền kinh tế. Khi hệ thống tài chính hoạt động ổn định, rủi ro được kiểm soát, tín dụng lưu thông hiệu quả và nền kinh tế có khả năng chống chịu với các cú sốc. Ngược lại, sự bất ổn tài chính có thể dẫn đến khủng hoảng ngân hàng, suy giảm lòng tin của nhà đầu tư và tăng trưởng kinh tế chậm lại. Do đó, đo lường và đánh giá mức độ ổn định tài chính của ngân hàng là một nhiệm vụ quan trọng đối với cả các nhà hoạch định chính sách và giới nghiên cứu học thuật. Chỉ số ổn định tài chính là một thước đo quan trọng để đánh giá sức khỏe và rủi ro tài chính của ngân hàng. Theo Altman (2013), có nhiều phương pháp được đề xuất để đo lường mức độ ổn định tài chính, từ các phương pháp phân tích tỷ lệ tài chính, phân tích đơn biến đến các mô hình kết hợp chỉ số, điển hình như mô hình của nhà kinh tế học Altman. Trong đó, phương pháp Z-score được sử dụng rộng rãi để đo lường mức độ ổn định tài chính của ngân hàng, nhờ khả năng phản ánh rủi ro vỡ nợ và mức độ an toàn vốn của hệ thống ngân hàng. Mô hình Z-score ngày càng cho thấy sự phù hợp với đặc điểm của hệ thống ngân hàng tại nhiều quốc gia (Goetz, 2018; Berger & cộng sự, 2017; Fernández & Garza-García, 2017; Mercieca & cộng sự, 2007; Phạm & cộng sự, 2021; Phạm & cộng sự, 2023; Võ & Dương, 2017; Võ & Mai, 2017).

Tổng hợp từ các nghiên cứu trước cho thấy có rất nhiều nhóm yếu tố có thể tác động đến chỉ số Z-score, bao gồm các chỉ tiêu về hiệu quả hoạt động ngân hàng, cấu trúc vốn, thanh khoản, quản trị rủi ro, điều kiện kinh tế vĩ mô và mức độ cạnh tranh trong ngành. Tuy nhiên, đối với các nghiên cứu sử dụng dữ liệu thứ cấp theo năm, vẫn tồn tại một số hạn chế như quy mô mẫu nhỏ, dữ liệu không đầy đủ hoặc không cập nhật thường xuyên. Do đó, việc lựa chọn các đặc trưng phù hợp để đưa

vào mô hình nghiên cứu là một thách thức lớn đối với giới nghiên cứu. Các phương pháp thống kê truyền thống như phân tích hồi quy tuyến tính hay phân tích phương sai gặp nhiều hạn chế trong việc điều chỉnh tham số mô hình dự báo cũng như xác định chính xác các ngưỡng biến động dữ liệu theo thời gian. Điều này đặt ra nhu cầu cấp thiết về việc ứng dụng các phương pháp tiên tiến hơn để cải thiện độ chính xác và hiệu suất của mô hình dự báo tài chính. Trong những năm gần đây, nhiều nghiên cứu đã phát triển các mô hình dự báo tài chính bằng cách kết hợp phương pháp thống kê với các thuật toán Machine Learning (ML) (Aidas & Aušrinė, 2021; Flavio & cộng sự, 2017; Le & Viviani, 2018; Chen & cộng sự, 2019).

Ứng dụng ML vào dự báo tài chính mang lại nhiều ưu điểm nổi bật so với các mô hình thực nghiệm truyền thống. Thứ nhất, ML có khả năng xử lý tập dữ liệu có số chiều cao mà không gặp phải vấn đề đa cộng tuyến. Thứ hai, các mô hình học máy có thể làm việc với tập dữ liệu không cân bằng, giữ lại toàn bộ thông tin có sẵn mà không cần giả định phân phối chuẩn. Thứ ba, ML có khả năng tự động tối ưu hóa các tham số mô hình mà không yêu cầu đầu vào chủ quan từ nhà nghiên cứu, giúp nâng cao độ chính xác dự báo (Alessi & Savona, 2021). Ngoài ra, hiệu suất dự báo của các mô hình học máy phụ thuộc trực tiếp vào tập hợp các yếu tố đầu vào. Sự phối hợp giữa các yếu tố này khi được xem xét trên nhiều góc độ khác nhau có thể tạo ra những cải tiến bất ngờ trong kết quả dự báo (Hellwig, 2021). Đặc biệt, các mô hình học máy có khả năng khắc phục bài toán kích thước tập dữ liệu nghiên cứu, một vấn đề phổ biến trong các nghiên cứu tài chính. Với sự phát triển của kiến trúc các mô hình học máy mới, vấn đề tập dữ liệu quá nhỏ hay quá lớn có thể được giải quyết, mở ra hướng đi tiềm năng trong dự báo kinh tế, không chỉ trong lĩnh vực tài chính mà còn trong các ứng dụng kinh tế vĩ mô và vi mô (Gogas & Papadimitriou, 2021).

Nghiên cứu này nhằm mục tiêu phân tích và tổng hợp các yếu tố tác động đến ổn định tài chính từ cơ sở lý thuyết và kết quả nghiên cứu trước. Đồng thời, nghiên cứu đề xuất thử nghiệm ba phương pháp giảm chiều dữ liệu nhằm tối ưu hóa tập biến đầu vào cho mô hình dự báo chỉ số Z-score. Phân tích thành phần chính (PCA) giúp giảm số chiều dữ liệu ban đầu, giảm độ phức tạp và tăng hiệu quả tính toán. Hồi quy Lasso giúp giữ lại các biến quan trọng nhất, loại bỏ các biến có mức độ tác động không đáng kể. Hồi quy Ridge giúp giảm thiểu hiện tượng phương sai lớn trong mô hình hồi quy, cải thiện tính ổn định của mô hình dự báo. Ngoài ra, nghiên cứu cũng áp dụng một số kỹ thuật tiền xử lý dữ liệu như chuẩn hóa dữ liệu (MinMaxScaler), xử lý ngoại lệ (outliers) và mã hóa dữ liệu biến phân loại (OnehotEncoder) để đảm bảo dữ liệu được xử lý chính xác trước khi đưa vào các mô hình học máy. Giá trị Z-score có dạng giá trị liên tục, vì vậy các thuật toán ML được đề xuất là Linear Regression (LR), Gradient Boost (GBR), Decision Tree (DT), Random Forest (RF), KNN (K-Nearest Neighbors), SVR (Support Vector Regression), AdaBoost, XGB (Extreme Gradient Boost).

Cấu trúc bài viết gồm 5 phần. Sau phần giới thiệu, phần 2 trình bày tổng quan các nghiên cứu trước và cơ sở lý thuyết về việc ứng dụng Lasso, Ridge, PCA và các thuật toán ML trong phát triển mô hình dự báo dự báo chỉ số tài chính. Phần 3 trình bày chi tiết về dữ liệu và phương pháp nghiên cứu thực nghiệm, bao gồm kịch bản ước lượng chỉ số ổn định tài chính và đánh giá hiệu suất mô hình trên dữ liệu thực tế của 30 ngân hàng thương mại Việt Nam giai đoạn 2010 đến 2023 bằng các thuật toán ML kết hợp lần lượt với ba kỹ thuật: Lasso, Ridge, PCA. Phần 4 trình bày chi tiết kết quả và các thảo luận liên quan đến kết quả thực nghiệm thu được. Cuối cùng, phần 5 đưa ra kết luận và một số đề xuất gợi ý dựa trên kết quả nghiên cứu thu được.

2. Tổng quan nghiên cứu

Việc đánh giá mức độ ổn định tài chính của ngân hàng là chủ đề nhận được nhiều sự quan tâm trong lĩnh vực tài chính - ngân hàng. Ổn định tài chính không chỉ phản ánh sức khỏe nội tại của một tổ chức tín dụng mà còn là chỉ báo quan trọng đối với toàn bộ hệ thống tài chính và nền kinh tế. Các nghiên cứu trước đây đề xuất nhiều phương pháp và chỉ số đo lường độ bền vững tài chính, trong đó chỉ số Z-score được đánh giá cao về tính ứng dụng và khả năng dự báo rủi ro vỡ nợ. Theo Altman & cộng sự (2017), mô hình Z-score tỏ ra hiệu quả trong việc dự báo sự ổn định tài chính của các ngân hàng với độ chính xác trung bình 0,75 và có thể nâng cao lên trên 0,90 khi kết hợp với các biến số bổ sung phù hợp với đặc thù quốc gia. Điều này cho thấy Z-score là một công cụ hữu ích để đánh giá mức độ an toàn của ngân hàng, giúp các nhà quản trị có thể đưa ra các biện pháp kịp thời nhằm giảm thiểu rủi ro tài chính.

Kế thừa các nghiên cứu trước đây, bài viết này lựa chọn Z-score làm thước đo ổn định tài chính của ngân hàng do lợi thế dễ tính toán và phù hợp với đặc điểm hoạt động của ngân hàng thương mại Việt Nam (Goetz, 2018; Fernández & Garza-García, 2017; Berger & cộng sự, 2017; Pham & cộng sự, 2021; Phạm & cộng sự, 2023; Võ & Dương, 2017; Võ & Mai, 2017). Giá trị Z-score càng cao thể hiện mức độ ổn định tài chính càng lớn, nghĩa là ngân hàng có khả năng chống chịu cao hơn trước các cú sốc kinh tế và tài chính. Do đó, việc dự báo sớm giá trị Z-score có thể giúp các nhà quản trị ngân hàng chủ động điều chỉnh chiến lược kinh doanh và quản trị rủi ro, qua đó nâng cao tính bền vững của hệ thống ngân hàng.

Sự phát triển nhanh chóng của công nghệ tài chính (Fintech) và trí tuệ nhân tạo đã mở ra nhiều hướng tiếp cận mới trong việc dự báo tài chính. Trong đó, học máy (ML) được xem là công cụ hiệu quả nhờ khả năng xử lý dữ liệu phức tạp, phát hiện các mối quan hệ phi tuyến và tự động tối ưu hóa mô hình dự báo. ML không chỉ giúp nâng cao độ chính xác mà còn giảm thiểu những hạn chế của các mô hình thống kê truyền thống. Flavio & cộng sự (2017) đã chỉ ra rằng ML có thể cải thiện đáng kể độ chính xác dự báo khả năng khủng hoảng tài chính và sự thất bại của doanh nghiệp. Nghiên cứu này đã so sánh hiệu suất của các thuật toán như Random

Forest (RF), Neural Networks và Support Vector Machine (SVM), cho thấy ML có ưu thế vượt trội so với các phương pháp thống kê truyền thống trong việc nhận diện sớm các dấu hiệu rủi ro. Tương tự, Barboza & cộng sự (2017) cũng khẳng định rằng việc áp dụng các mô hình thông minh giúp tăng cường khả năng dự báo nguy cơ vỡ nợ và cải thiện chất lượng ra quyết định tài chính. Bên cạnh đó, một số nghiên cứu như Aidas & Aušrinė (2021) đã ứng dụng ML vào dự báo sức khỏe tài chính cho các doanh nghiệp nhỏ và vừa. Kết quả cho thấy thuật toán RF mang lại độ chính xác cao nhất, đồng thời chỉ ra rằng việc điều chỉnh tham số rủi ro theo từng giai đoạn là cần thiết để tối ưu hóa hiệu suất mô hình. Những nghiên cứu này không chỉ khẳng định vai trò quan trọng của ML trong tài chính mà còn nhấn mạnh sự cần thiết của việc tích hợp các thuật toán hiện đại để tăng cường độ chính xác dự báo.

Chất lượng dự báo của mô hình phụ thuộc lớn vào việc lựa chọn biến đầu vào. Các nghiên cứu trước đây đã chỉ ra nhiều yếu tố có tác động đáng kể đến Z-score, bao gồm quy mô ngân hàng, GDP, lạm phát, tỷ lệ dự phòng rủi ro tín dụng, tốc độ tăng trưởng tổng tài sản và tỷ lệ dư nợ trên tổng tiền gửi (Delis & Kouretas, 2011; Laeven & Levine, 2009; Demircuc-Kunt & Huizinga, 2013; Fu & cộng sự, 2014; Võ & Mai, 2020; Phạm & cộng sự, 2023). Ngoài ra, sự hiện diện của ngân hàng nước ngoài cũng ảnh hưởng đến mức độ ổn định tài chính (Chen & cộng sự, 2019; Claessens & Neeltje, 2014; Fetene & Soyoun, 2021; Wu & cộng sự, 2017; Phạm & cộng sự, 2023) và năng lực cạnh tranh (Fu & cộng sự, 2014; Phạm & cộng sự, 2023; Võ & Dương, 2017). Do đó, việc lựa chọn các biến số phù hợp là rất quan trọng để đảm bảo tính chính xác của mô hình dự báo. Một trong những thách thức chính trong nghiên cứu tài chính là số lượng biến độc lập lớn có thể dẫn đến hiện tượng đa cộng tuyến, nội sinh và tự tương quan, làm sai lệch kết quả dự báo. Để khắc phục vấn đề này, các phương pháp như Lasso, Ridge và PCA đã được áp dụng rộng rãi trong nghiên cứu tài chính (Ben & cộng sự, 2023). Lasso giúp loại bỏ các biến không cần thiết, giảm hiện tượng đa cộng tuyến và quá khớp dữ liệu, trong khi Ridge cải thiện tính ổn định của mô hình bằng cách điều chỉnh trọng số của các biến có quan hệ chặt chẽ. PCA giúp tạo ra các biến mới từ tổ hợp các biến gốc, giúp giảm chiều dữ liệu mà vẫn giữ được phần lớn thông tin quan trọng.

Tại Việt Nam, ứng dụng các mô hình học máy trong xây dựng mô hình dự báo tài chính vẫn còn sơ khai, với số lượng nghiên cứu hạn chế và thiếu sự kết hợp giữa các phương pháp tối ưu hóa mô hình. Đặc biệt, chưa có nghiên cứu nào áp dụng đồng thời Lasso, Ridge, PCA cùng các thuật toán ML tiên tiến để phát triển mô hình dự báo Z-score—một chỉ số quan trọng đo lường mức độ ổn định tài chính và rủi ro vỡ nợ của ngân hàng. Trong bối cảnh ngành ngân hàng Việt Nam đang đối mặt với nhiều biến động kinh tế vĩ mô, rủi ro tín dụng gia tăng và sức ép từ hội nhập tài chính quốc tế, việc phát triển một mô hình dự báo chính xác, linh hoạt và có tính tổng quát cao là hết sức cần thiết. Do đó, nghiên cứu này kết hợp Lasso, Ridge, PCA với các mô hình học máy tiên tiến nhằm tận dụng ưu điểm của từng phương pháp từ

chọn lọc biến, giảm đa cộng tuyến, cải thiện hiệu suất mô hình đến tối ưu hóa khả năng dự báo. Sự kết hợp này không chỉ giúp mô hình hoạt động hiệu quả hơn trên tập dữ liệu thực tế mà còn cung cấp một khung phương pháp luận có thể mở rộng và áp dụng cho các nghiên cứu tài chính khác trong tương lai.

3. Dữ liệu và phương pháp nghiên cứu

3.1 Dữ liệu nghiên cứu

Nghiên cứu thực hiện trên bộ dữ liệu thu thập các chỉ số tài chính của 30 ngân hàng thương mại Việt Nam trong khoảng thời gian từ 2010 - 2023. Trong đó, các biến kiểm soát đặc trưng cho ngân hàng thương mại Việt Nam sử dụng nguồn dữ liệu thứ cấp được thu thập từ báo cáo tài chính, yếu tố đo lường mức độ gia nhập từ các ngân hàng nước ngoài được thu thập từ Ngân hàng Nhà nước, các yếu tố bên ngoài thuộc môi trường vĩ mô được thu thập từ nguồn công bố của Quỹ tiền tệ quốc tế (IMF). Dữ liệu được làm sạch và chuẩn hoá trước khi đưa vào huấn luyện bởi các mô hình học máy đề xuất.

3.2 Phương pháp nghiên cứu

Nghiên cứu thực hiện kết hợp đồng thời cả hai phương pháp nghiên cứu định tính và nghiên cứu định lượng. Trước tiên, bài viết lược khảo cơ sở lý thuyết và kết quả từ các nghiên cứu trước. Tiếp đó, dữ liệu được thu thập, tổng hợp và xử lý tính toán các yếu tố cho bộ dữ liệu thực nghiệm. Kế thừa từ các nghiên cứu của Allen & cộng sự (2009), Chen & cộng sự (2019), Claessens & Neeltje (2014), Fetene & Soyung (2021), Delis & Kouretas (2011), Goetz (2018), Laeven & Levine (2009), Demircuc-Kunt & Huizinga (2013), Fernández & Garza-García (2017), Fu & cộng sự (2014), Rajhi & Hassairi (2013), Pham & cộng sự (2021), Phạm & cộng sự (2023), Võ & Dương (2017), Võ & Mai (2017) và Wu & cộng sự (2017), các yếu tố được xác định như sau:

Chỉ số $Z\text{-score}_{it}$ đo lường mức độ ổn định tài chính của ngân hàng, được tính bằng công thức:

$$Z\text{-score}_{i,t} = \frac{ETA_{i,t} + ROAA_{i,t}}{\sigma(ROAA)}$$

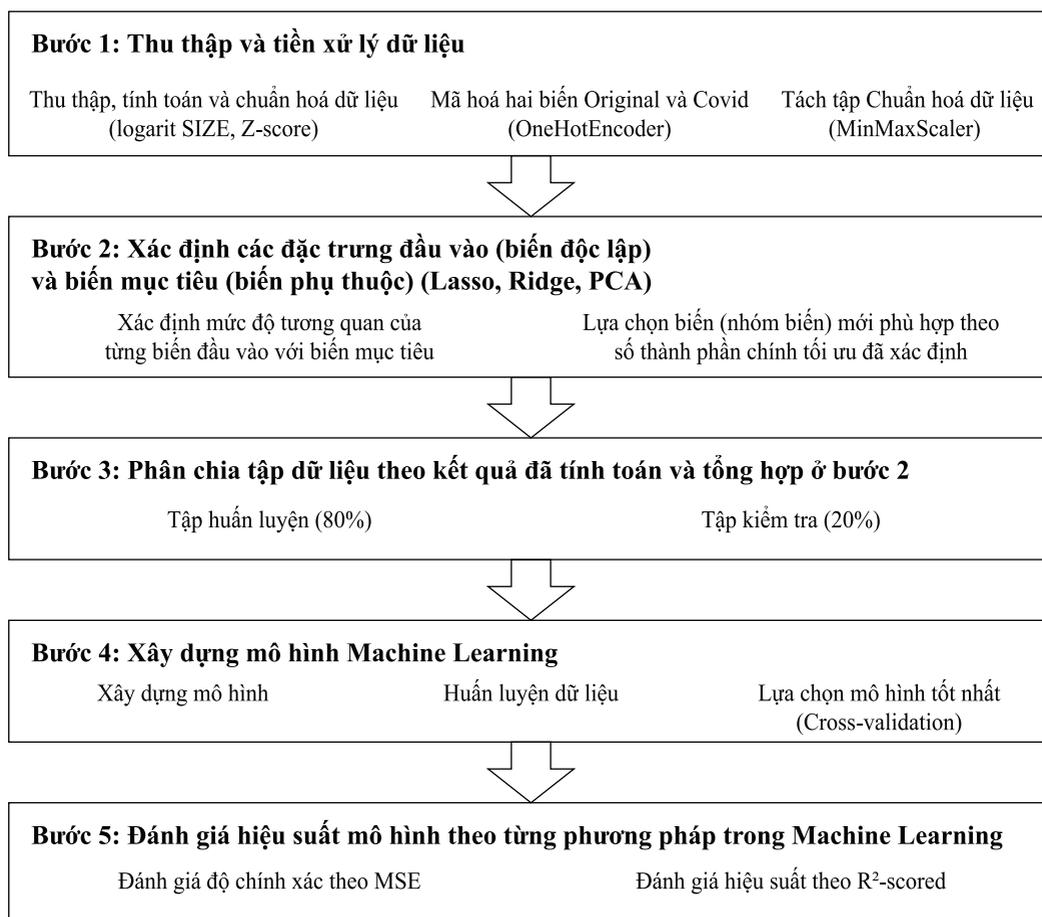
trong đó, i là biến đại diện ngân hàng, t là thời gian; ETA là tỷ lệ vốn chủ sở hữu (VCSH) trên tổng tài sản; $ROAA$ là tỷ suất sinh lời trên tổng tài sản bình quân của ngân hàng. Các yếu tố tác động đến chỉ số $Z\text{-score}$ được trình bày trong Bảng 1.

Bảng 1. Tổng hợp các yếu tố tác động đến Z-score từ các nghiên cứu trước

Biến	Ý nghĩa	Công thức tính
Z-score	Chỉ số ổn định tài chính ngân hàng	$Z\text{-score}_{it} = (ETA_{it} + ROAA_{it}) / \text{độ lệch chuẩn } (ROAA_{it})$
LERNER	Năng lực cạnh tranh	$Lerner_{it} = (P_{it} - MC_{it})/P_{it}$ trong đó, P là giá đầu ra được tính bằng tổng doanh thu trên tổng tài sản, MC là chi phí biên của ngân hàng.
FS1	Tỷ lệ số lượng NHNNg	= Tổng số lượng chi nhánh NHNNg / Tổng số lượng NH Việt Nam
FS2	Tỷ trọng tài sản NHNNg	= TTS NHNNg / TTS hệ thống tổ chức tín dụng Việt Nam
ETA	Quy mô VCSH trên TTS	= VCSH / TTS
SIZE	Quy mô ngân hàng	= ln (TTS)
LTA	Tỷ lệ cho vay	= Tổng cho vay / TTS
HDV	Tỷ lệ huy động vốn	= Tổng huy động vốn / TTS
LDR	Hiệu quả sử dụng vốn huy động	= Tổng huy động vốn / Tổng cho vay
ROE	Lợi nhuận trên vốn chủ sở hữu	= Lợi nhuận sau thuế / VCSH
CTI	Hiệu quả sử dụng chi phí	= Tổng chi phí hoạt động / Tổng thu nhập
LLP	Tỷ lệ dự phòng rủi ro tín dụng	= Tổng chi phí dự phòng rủi ro tín dụng / Tổng dư nợ cho vay
HHI	Khả năng đa dạng hoá thu nhập	$= 1 - [(NON/NI)^2 + (NET/NI)^2]$ trong đó, NON: thu nhập ngoài lãi; NET: thu nhập lãi thuần, NI = NON + NET.
MS	Sức mạnh thị trường	= TTS NH / TTS toàn hệ thống tín dụng
GROTA	Tốc độ tăng trưởng tổng tài sản (TTS)	$= (TTS_t - TTS_{t-1})/TTS_{t-1}$
Original	Hình thức sở hữu	= 1 :Sở hữu Nhà nước = 0 :Sở hữu tư nhân
GDP	Tốc độ tăng trưởng GDP	$= (GDP_t - GDP_{t-1})/GDP_{t-1}$
INF	Tỷ lệ lạm phát	Nguồn thu thập IMF

Nguồn: Tính toán và tổng hợp của nhóm tác giả

Toàn bộ quy trình thực nghiệm được tóm lược như Hình 1.



Nguồn: Đề xuất của nhóm tác giả

Để kiểm tra và kiểm soát khả năng quá khớp dữ liệu (overfitting) của mô hình, nghiên cứu đã áp dụng phương pháp kiểm định chéo (cross-validation) nhằm đánh giá mức độ tổng quát và khả năng khái quát hóa trên tập dữ liệu kiểm tra. Cụ thể, nghiên cứu sử dụng k-fold cross-validation, trong đó dữ liệu được chia thành k tập con (folds), với mỗi lần huấn luyện, mô hình được đào tạo trên k-1 tập và kiểm tra trên tập còn lại. Quá trình này được lặp lại k lần, đảm bảo rằng tất cả dữ liệu đều được sử dụng cả trong huấn luyện và kiểm tra, giúp giảm thiểu sai lệch do phân bố dữ liệu. Phương pháp này cho phép đánh giá hiệu suất thực sự của mô hình, hạn chế rủi ro của việc mô hình hoạt động tốt trên tập huấn luyện nhưng kém trên dữ liệu mới. Bên cạnh đó, nghiên cứu cũng so sánh kết quả giữa các tập huấn luyện và kiểm tra thông qua các thước đo hệ số xác định (R^2), sai số dự báo (MSE) để phát hiện dấu hiệu quá khớp dữ liệu (nếu có). Các kỹ thuật điều chỉnh giảm số lượng đặc trưng như Lasso, Ridge hoặc PCA được kết hợp để cải thiện tính ổn định của mô hình.

4. Kết quả và thảo luận

4.1 Phân tích khám phá dữ liệu từ các yếu tố trong mô hình nghiên cứu

Dữ liệu các yếu tố dự kiến đưa vào nghiên cứu được thực hiện thống kê mô tả trong Bảng 2.

Bảng 2. Thống kê mô tả

Các yếu tố đặc trưng	Số quan sát	Trung bình	Độ lệch chuẩn	Giá trị nhỏ nhất	Phân bố tứ phân vị			Giá trị lớn nhất
					25%	50%	75%	
Z-score	415	16,0262	9,2589	0,4930	9,2518	13,8265	20,7883	61,6600
FS1	415	0,2141	0,0255	0,1837	0,1875	0,2391	0,2391	0,2391
FS2	415	0,1019	0,0057	0,0930	0,0972	0,1006	0,1078	0,1130
SIZE	415	18,7460	1,2366	15,9227	17,8665	18,7512	19,6166	21,5566
ETA	415	0,0899	0,0396	0,0262	0,0624	0,0788	0,1046	0,2554
LTA	415	0,5648	0,1242	0,1448	0,4782	0,5828	0,6658	0,7881
HDV	415	0,6500	0,1263	0,2508	0,5609	0,6589	0,7383	0,9122
LDR	415	0,8811	0,1842	0,3633	0,7653	0,8766	0,9968	1,7893
CTI	415	0,8782	0,0945	0,5378	0,8269	0,8914	0,9466	1,5160
LLP	415	0,0270	0,0593	0,0001	0,0068	0,0112	0,0194	0,6766
HHI	415	0,3346	0,2088	-2,7370	0,2811	0,3761	0,4406	0,4995
GROTA	415	0,1876	0,2100	-0,3924	0,0848	0,1450	0,2270	1,4701
MS	415	0,0274	0,0323	0,0015	0,0059	0,0134	0,0316	0,1261
Original	415	0,1687	0,3749	0,0000	0,0000	0,0000	0,0000	1,0000
Covid	415	0,3614	0,4810	0,0000	0,0000	0,0000	1,0000	1,0000
LERNER	415	0,1605	0,0840	0,0006	0,1030	0,1555	0,2097	0,4604
ROE	415	0,1034	0,0841	-0,5633	0,0413	0,0926	0,1568	0,3033
GDP	415	0,0600	0,0153	0,0259	0,0550	0,0642	0,0699	0,0812
INF	415	0,0694	0,1081	-0,0172	0,0186	0,0341	0,0436	0,4230

Nguồn: Tính toán của nhóm tác giả

Dựa trên kết quả thống kê mô tả có thể thấy giá trị Z-score có độ lệch chuẩn tương đối lớn, xem xét trong phạm vi tứ phân vị sẽ có nhiều giá trị ngoại lai. Vấn đề này có thể dễ dàng được giải thích bởi tính khác biệt về quy mô cũng như chiến lược kinh doanh của từng ngân hàng trong từng năm, Original chỉ bao gồm 2 giá trị “0, 1” và số lượng ngân hàng nghiên cứu đa số mang giá trị 0.

Dữ liệu nghiên cứu có sự phân hóa và chênh lệch lớn ở các yếu tố như FS1, Original, Covid, cùng độ nhọn cao (CTI, LLP, HHI, GROTA, ROE) và phân bố lệch (GDP lệch phải, MS và INF lệch trái), ảnh hưởng đến độ chính xác dự báo. Do đó,

Bảng 3. Ma trận tương quan

	Z-score	FS1	FS2	SIZE	ETA	LTA	HDV	LDR	CTI	LLP	HHI	GROTA	MS	LERNER	ROE	GDP	INF
Z-score	1,00																
FS1	-0,47	1,00															
FS2	0,22	-0,42	1,00														
SIZE	-0,46	0,39	-0,23	1,00													
ETA	0,72	-0,25	0,13	-0,57	1,00												
LTA	-0,09	0,39	-0,25	0,42	-0,09	1,00											
HDV	-0,04	0,21	-0,12	0,35	-0,34	0,58	1,00										
LDR	-0,06	0,21	-0,14	0,11	0,26	0,54	-0,34	1,00									
CTI	-0,04	-0,23	0,06	-0,30	-0,25	-0,20	0,20	-0,43	1,00								
LLP	0,01	-0,15	0,06	-0,06	-0,12	-0,12	0,14	-0,25	0,36	1,00							
HHI	-0,05	0,27	-0,07	0,28	-0,06	0,18	0,18	0,05	-0,35	-0,05	1,00						
GROTA	-0,08	-0,17	0,23	-0,16	0,07	-0,29	-0,40	0,07	-0,22	-0,08	-0,02	1,00					
MS	-0,28	0,03	-0,03	0,80	-0,40	0,41	0,31	0,13	-0,14	-0,03	0,17	-0,08	1,00				
LERNER	0,04	0,08	-0,01	0,16	0,35	0,10	-0,23	0,33	-0,74	-0,32	-0,05	0,19	0,09	1,00			
ROE	-0,19	0,26	-0,09	0,46	-0,05	0,28	-0,09	0,40	-0,86	-0,33	0,38	0,17	0,30	0,56	1,00		
GDP	0,10	-0,19	-0,23	-0,10	0,00	-0,06	-0,04	-0,03	0,09	0,06	-0,07	0,05	0,00	-0,08	-0,07	1,00	
INF	-0,15	-0,36	0,30	-0,28	0,21	-0,29	-0,47	0,15	-0,13	-0,09	-0,11	0,51	-0,03	0,27	0,12	0,10	1,00

Nguồn: Tổng hợp của nhóm tác giả

nghiên cứu áp dụng mã hóa và chuẩn hóa dữ liệu trước khi hồi quy, sử dụng MinMaxScaler để quy đổi giá trị về khoảng [0;1], trong đó “0” thể hiện tình trạng bất ổn tài chính và “1” phản ánh trạng thái tốt nhất. Bộ dữ liệu huấn luyện cũng tồn tại sự chênh lệch lớn về Z-score giữa các ngân hàng do khác biệt về quy mô, cơ cấu tài sản và chiến lược kinh doanh. Vì vậy, các yếu tố như SIZE, Z-score được chuẩn hóa bằng logarit tự nhiên. Đồng thời, phân tích tứ phân vị cho thấy nhiều giá trị ngoại lai, ví dụ ETA - 18 giá trị ngoại lai, LLP - 43 giá trị ngoại lai, GDP - 60 giá trị ngoại lai. Để hạn chế mất mát dữ liệu và nâng cao độ chính xác, các giá trị ngoại lai được thay thế bằng trung vị của từng biến số.

Kết quả ma trận tương quan (Bảng 3) cho thấy một số cặp biến có mức độ tương quan cao như MS - SIZE, CTI - ROE, ETA - Z-score hay CTI - LERNER, làm gia tăng nguy cơ đa cộng tuyến, có thể dẫn đến sai lệch hồi quy và quá khớp mô hình dự báo. Để xử lý vấn đề này, nghiên cứu xem xét ba phương pháp loại bỏ biến, tạo biến mới hoặc sử dụng kỹ thuật chọn lọc biến tối ưu nhằm giảm mức độ tương quan mà vẫn bảo toàn thông tin quan trọng. Tuy nhiên, việc thêm hoặc bớt biến đều cần cân nhắc kỹ lưỡng để đảm bảo độ chính xác và tính tin cậy của mô hình. Do đó, nghiên cứu tiến hành thử nghiệm với ba phương pháp Lasso, Ridge và PCA nhằm đánh giá hiệu suất xử lý dữ liệu đầu vào, xác định phương pháp tối ưu nhất. Tiêu chí đánh giá bao gồm sai số bình phương trung bình (MSE) và hệ số xác định (R^2), phù hợp với biến mục tiêu có giá trị liên tục. Các kết quả phân tích sẽ được thảo luận để so sánh mức độ hiệu quả của từng phương pháp xử lý đặc trưng đầu vào.

4.2 Kết quả nghiên cứu

Sau khi dữ liệu được xử lý ngoại lai, mã hóa biến gốc (OneHotEncoder) và chuẩn hóa (MinMaxScaler), 8 mô hình học máy bao gồm Linear Regression, Gradient Boost, Decision Tree, Random Forest, KNN, SVR, AdaBoostRegressor và XGBRegressor được sử dụng để tính toán. Kết quả thực nghiệm trên bộ dữ liệu nghiên cứu được trình bày chi tiết tại Bảng 4.

Bảng 4. Hiệu suất dự báo ứng dụng các mô hình Machine Learning

Các thuật toán Machine Learning		Hiệu suất mô hình	Lasso	Ridge	PCA	Không dùng Lasso, Ridge hay PCA
Linear Regression	Tập huấn luyện	MSE	0,0039	0,0031	0,0039	0,0524
		R^2 -scored	0,8242	0,8601	0,8244	0,8759
	Tập kiểm tra	MSE	0,0036	0,0030	0,0043	0,0384
		R^2 -scored	0,8498	0,8743	0,8208	0,8402
Gradient Boost	Tập huấn luyện	MSE	0,0000	~0,0000	0,0002	0,0035
		R^2 -scored	1,0000	~1,0000	0,9897	0,9891
	Tập kiểm tra	MSE	0,0000	0,0002	0,0029	0,0327
		R^2 -scored	0,9987	0,9927	0,8791	0,7757

Bảng 4. Hiệu suất dự báo ứng dụng các mô hình Machine Learning (tiếp theo)

Các thuật toán Machine Learning		Hiệu suất mô hình	Lasso	Ridge	PCA	Không dùng Lasso, Ridge hay PCA
Decision Tree	Tập huấn luyện	MSE	0,0000	0,0000	0,0000	0,0000
		R^2 -scored	1,0000	1,0000	1,0000	1,0000
	Tập kiểm tra	MSE	0,0000	0,0002	0,0070	0,0735
		R^2 -scored	0,9983	0,9922	0,7056	0,4988
Random Forest	Tập huấn luyện	MSE	0,0000	0,0000	0,0006	0,0096
		R^2 -scored	0,9982	0,9990	0,9742	0,9546
	Tập kiểm tra	MSE	0,0000	0,0002	0,0036	0,0296
		R^2 -scored	0,9987	0,9922	0,8497	0,7838
KNN	Tập huấn luyện	MSE	0,0004	0,0059	0,0064	0,1103
		R^2 -scored	0,9812	0,7339	0,7147	0,7065
	Tập kiểm tra	MSE	0,0010	0,0117	0,0113	0,1436
		R^2 -scored	0,9580	0,5079	0,5229	0,6155
SVR	Tập huấn luyện	MSE	0,0040	0,0042	0,0042	0,0046
		R^2 -scored	0,8224	0,8134	0,8125	0,8547
	Tập kiểm tra	MSE	0,0040	0,0054	0,0060	0,0323
		R^2 -scored	0,8336	0,7713	0,7486	0,8803
AdaBoostRegressor	Tập huấn luyện	MSE	0,0002	0,0003	0,0039	0,0230
		R^2 -scored	0,9923	0,9883	0,8253	0,9277
	Tập kiểm tra	MSE	0,0002	0,0004	0,0067	0,0497
		R^2 -scored	0,9932	0,9840	0,7185	0,8160
XGBRegressor	Tập huấn luyện	MSE	0,0000	0,0000	0,0000	6,2236
		R^2 -scored	0,9999	0,9999	0,9999	0,9999
	Tập kiểm tra	MSE	0,0009	0,0006	0,0052	0,0390
		R^2 -scored	0,9614	0,9734	0,7826	0,6842

Chú thích: Giá trị biểu diễn được tính xấp xỉ, làm tròn 4 số thập phân.

Nguồn: Tính toán và tổng hợp của nhóm tác giả

Sau khi tiền xử lý, việc ứng dụng trực tiếp các mô hình học máy trên dữ liệu gốc cho thấy hiệu suất dự báo kém, thể hiện qua chỉ số R^2 và MSE. MSE cao phản ánh sai số lớn, trong khi R^2 biến động giữa mẫu huấn luyện và kiểm tra, làm giảm độ tin cậy của dự báo. Tuy nhiên, hiệu suất mô hình được cải thiện đáng kể khi áp dụng các phương pháp Lasso, Ridge và PCA. Kết quả thu được trên hai tập dữ liệu huấn luyện và kiểm tra cho thấy:

Kết quả đo lường hiệu suất dự báo chỉ số Z-score với dữ liệu tập huấn luyện

Phương pháp Ridge cải thiện hiệu suất có độ tương đồng cao so với Lasso. Tuy theo mô hình học máy mà từng phương pháp lựa chọn đặc trưng mang lại hiệu suất dự đoán chênh lệch nhau. Ví dụ, huấn luyện trên LinearRegression sử dụng Ridge hiệu quả hơn còn trong 7 thuật toán còn lại Lasso chiếm ưu thế hơn và vượt trội so

với PCA. Trong ba phương pháp thì PCA mang lại hiệu suất thấp nhất. Theo đó, so với PCA thì hiệu suất mô hình dự báo cải thiện rõ rệt khi sử dụng phương pháp Lasso và Ridge. Kết quả này có thể giải thích bởi PCA chủ yếu được sử dụng để giảm chiều dữ liệu bằng cách tạo ra các thành phần chính mới, thay vì tập trung vào tối ưu hóa dự báo. Trong khi đó, Lasso và Ridge không chỉ thực hiện chọn lọc và điều chỉnh trọng số của biến mà còn tối ưu hóa mô hình dự báo.

Kết quả đo lường hiệu suất dự báo chỉ số Z-score với tập dữ liệu kiểm tra

Phương pháp Lasso chiếm ưu thế hơn so với các phương pháp còn lại. Ví dụ, kết quả huấn luyện bằng mô hình LinearRegression và XGBRegressor hiệu quả hơn khi kết hợp với phương pháp Ridge. Trong khi 6 các mô hình học máy còn lại khi kết hợp với phương pháp Lasso thì kết quả dự báo đạt hiệu suất tốt hơn. Kết quả này nhấn mạnh việc sử dụng quá nhiều yếu tố đầu vào có thể gây nhiễu thông tin ảnh hưởng đến kết quả dự báo. Hiệu suất huấn luyện mô hình kết hợp kỹ thuật Lasso và Ridge vượt trội so với kỹ thuật PCA. Kết quả Bảng 4 cho thấy phương pháp Lasso kết hợp với Gradient Boost cho kết quả tốt nhất với hệ số R^2 -scored lên đến 99,83%, MSE xấp xỉ 0,0000. Kết quả thực nghiệm cũng cho thấy hệ số MSE được cải thiện đáng kể khi có ứng dụng các phương pháp Lasso, Ridge hay PCA.

5. Hàm ý chính sách và kết luận

So với việc ứng dụng trực tiếp các mô hình học máy trên bộ dữ liệu sau khi tiền xử lý cơ bản thì việc kết hợp với 3 phương pháp Lasso, Ridge, PCA sẽ cải thiện đáng kể - chất lượng dự báo. Kết quả cho thấy sau khi áp dụng Lasso, Ridge, PCA đều có MSE rất nhỏ và R^2 được cải thiện rõ rệt. Xét về tổng thể, phương pháp Lasso mang lại kết quả dự báo với độ chính xác tốt nhất. Kỹ thuật Lasso có thể loại bỏ đi các đặc trưng cho là không liên quan, trong tình huống này, Lasso chỉ giữ lại 3 đặc trưng (Z -score_1, LERNER, MS). Phương pháp Ridge thực hiện theo nguyên tắc chỉ điều chỉnh hệ số của từng yếu tố chứ không loại bỏ biến. Phương pháp PCA tạo ra các thành phần chính mới được cho là mang lại hiệu suất dự báo tốt nhất (trong nghiên cứu này là 12 thành phần). Tùy theo mục tiêu nghiên cứu hay quan điểm của nhà quản trị, việc quyết định loại bỏ bớt biến tránh dư thừa dữ liệu hay giữ lại các biến tránh mất mát thông tin cần được cân nhắc thận trọng để lựa chọn phương pháp phát triển mô hình dự báo hiệu quả nhất.

Bộ dữ liệu nghiên cứu có kích cỡ mẫu tương đối nhỏ nhưng số đặc trưng đầu vào lại khá lớn (18 đặc trưng). Do đó, áp dụng các phương pháp như Lasso, Ridge, PCA giúp giảm nguy cơ quá khớp dữ liệu. Trong nghiên cứu tài chính, vấn đề đa cộng tuyến và tương quan cao giữa các đặc trưng luôn là thách thức đáng quan tâm. Việc tổng hợp cơ sở lý thuyết và lược khảo nghiên cứu trước đây cho thấy số lượng đặc trưng đầu vào khá nhiều. Loại bỏ quá nhiều biến có thể gây mất mát thông tin, trong khi giữ lại quá nhiều đặc trưng dễ dẫn đến nhiễu dữ liệu và sai lệch dự báo.

Triển khai các kỹ thuật tiền xử lý trước khi ứng dụng với các mô hình học máy là giải pháp quan trọng nhằm giảm thiểu rủi ro này. Không có thuật toán tối ưu tuyệt đối

mà chỉ có phương pháp phù hợp nhất với từng bối cảnh và đặc trưng dữ liệu, do đó cần đánh giá cẩn trọng trước khi lựa chọn các mô hình học máy phù hợp. Bên cạnh đó, quá trình xây dựng bộ dữ liệu đầu vào phải đảm bảo tuân thủ các quy định và chính sách ngân hàng hiện hành, đồng thời cần có sự giám sát từ chuyên gia tài chính để đảm bảo tính chính xác và đáng tin cậy của dữ liệu huấn luyện. Vấn đề về kích cỡ dữ liệu cũng ảnh hưởng trực tiếp đến kết quả dự báo của mô hình. Đây cũng là một trong những hạn chế của nghiên cứu này. Vì vậy, định hướng nghiên cứu kế tiếp là mở rộng kích thước bộ mẫu trong khoảng thời gian dài hơn. Đồng thời, xem xét trong nhiều bối cảnh khác nhau, tiếp cận thêm các phương pháp khác như Artificial Neural Network (ANN), Convolutional Neural Network (CNN) ... để tăng độ tin cậy cũng như bổ sung thêm các bằng chứng thực nghiệm đáng tin cậy cho các nghiên cứu liên quan.

Cuối cùng, với sự phát triển ngày càng lớn mạnh và tính ưu việt thấy rõ của việc ứng dụng các mô hình học máy trong lĩnh vực tài chính, chiến lược xây dựng các mô hình dự báo chỉ số tài chính cần được quan tâm đầu tư phát triển thích hợp. Điều này phụ thuộc rất lớn bởi quan điểm, tầm nhìn và chiến lược điều hành của người lãnh đạo các ngân hàng.

Tài liệu tham khảo

- Aidas, M. & Aušrinė, L. (2021), "Financial distress prediction for small and medium enterprises using machine learning techniques", *Engineering Economics*, Vol. 32 No. 1, pp. 4-14.
- Alessi, L. & Savona, R. (2021), "Machine learning for financial stability", *Data Science for Economics and Finance*, Springer, pp. 65-87.
- Altman, E.I. (2013), "Predicting financial distress of companies: revisiting the Z-Score and ZETA® models", *Handbook of Research Methods and Applications in Empirical Finance*, Massachusetts: Edward Elgar, pp. 428-456.
- Altman, E., Iwanicz-Drozdowska, M. & Laitinen, E. &. (2017), "Financial distress prediction in an international context: a review and empirical analysis of Altman's Z-Score model", *Journal of International Financial Management & Accounting*, Vol. 28 No. 2, pp. 131-171.
- Barboza, F., Kimura, H. & Altman, E. (2017), "Machine learning models and bankruptcy prediction", *Expert Systems with Applications*, Vol. 83, pp. 405-417.
- Ben, S.J., Stef, N. & Carmona, P. (2023), "Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering", *Computational Economics*, Vol. 61, pp. 715-741.
- Berger, A.N., Klapper, L.F. & Turk-Ariss, R. (2017), "Bank competition and financial stability", *Handbook of Competition in Banking and Finance*, Edward Elgar Publishing, pp. 185-204.
- Chen, M.A., Wu, Q. & Yang, B. (2019), "How valuable is fintech innovation?", *The Review of Financial Studies*, Vol. 32 No. 5, pp. 2062-2106.
- Chen, S., Imran, M., Hashmi, S.H. & Shaikh, R. (2019), "Bank competition, foreign bank entry, and risk-taking behavior: cross country evidence", *Journal of Risk and Financial Management*, Vol. 12 No. 3, pp. 1-26.

- Claessens, S. & Neeltje, H.V. (2014), "Foreign banks: trends and impact", *Journal of Money, Credit and Banking*, Vol. 46 No. 1, pp. 295-326.
- Delis, M. & Kouretas, G. (2011), "Interest rates and bank risk-taking", *Journal of Banking & Finance*, Vol. 35 No. 4, pp. 840-855.
- Demirguc-Kunt, A. & Huizinga, H. (2013), "Are banks too big to fail or too big to save? International evidence from equity prices and CDS spreads", *Journal of Banking & Finance*, Vol. 37 No. 3, pp. 875-894.
- Fernández, R.O. & Garza-García, J.G. (2017), "The relationship between bank competition and financial stability: a case study of the Mexican banking industry", *Ensayos Revista de Economía, Universidad Autónoma de Nuevo Leon*, Facultad de Economía, No. 1, pp. 103-120.
- Fetene, H.B. & Soyoun, K. (2021), "A dynamic analysis on foreign bank entry Nexus economic growth in Sub-Sahara African countries", *Cogent Economics & Finance*, Vol. 9 No. 1, pp. 1-21.
- Flavio, B., Herbert, K. & Edward, A. (2017), "Machine learning models and bankruptcy prediction", *Expert Systems with Applications*, Vol. 83, pp. 405-417.
- Fu, X.M., Lina, Y.R. & Mol, P. (2014), "Bank competition and financial stability in Asia Pacific", *Journal of Banking and Finance*, Vol. 38, pp. 64-77.
- Goetz, M.R. (2018), "Competition and bank stability", *Journal of Financial Intermediation*, Vol. 35, pp. 57-69.
- Gogas, P. & Papadimitriou, T. (2021), "Machine learning in economics and finance", *Computational Economics*, Vol. 57, pp. 1-4.
- Hellwig, K.P. (2021), "Predicting fiscal crises: a machine learning approach", IMF Working Paper, pp. 1-65.
- Laeven, L. & Levine, R. (2009), "Bank governance, regulation and risk taking", *Journal of Financial Economics*, Vol. 93 No. 2, pp. 259-275.
- Le, H.H. & Viviani, J.L. (2018), "Predicting bank failure: an improvement implementing a machine-learning approach to classical financial ratios", *Research in International Business and Finance*, Vol. 44, pp. 16-25.
- Mercieca, S., Schaeck, K. & Wolfe, S. (2007), "Small European banks: benefits from diversification?", *Journal of Banking & Finance*, Vol. 31 No. 7, pp. 1975-1998.
- Phạm, T.T., Đào, L.K.O. & Dương, N.T.T. (2023), "Tác động từ sự gia nhập của ngân hàng ngoại đến ổn định tài chính các ngân hàng thương mại Việt Nam trong bối cảnh hội nhập", *Tạp chí Khoa học và Thương mại*, pp. 47-58.
- Pham, T.T., Dao, L.K. & Nguyen, V.C. (2021), "The determinants of bank's stability: a system GMM panel analysis", *Cogent Business & Management*, Vol. 8 No. 1, pp. 1-18.
- Võ, V.X. & Dương, A.T. (2017), "Các yếu tố ảnh hưởng đến sức cạnh tranh của các ngân hàng thương mại Việt Nam giai đoạn 2005-2014", *Tạp chí Kinh tế và Kinh doanh*, Vol. 33 No. 1, pp. 12-22.
- Võ, V. & Mai, X.Đ. (2017), "Sở hữu nước ngoài và rủi ro thanh khoản của các ngân hàng thương mại Việt Nam", *Tạp chí Khoa học ĐHQGHN: Kinh tế và Kinh doanh*, pp. 1-19.
- Wu, J., Chen, M., Jeon, B.N. & Wang, R. (2017), "Does foreign bank penetration affect the risk of domestic banks? Evidence from emerging economies", *Journal of Financial Stability*, Vol. 31, pp. 45-61.