

DỰ BÁO HIỆU QUẢ HOẠT ĐỘNG CỦA NGÂN HÀNG BẰNG TRÍ TUỆ NHÂN TẠO

Đỗ Quang Hưng

Tóm tắt

Ngành ngân hàng có vai trò quan trọng, dẫn dắt nền kinh tế của một đất nước. Quá trình toàn cầu hóa và đổi mới sáng tạo trong lĩnh vực công nghệ đã tạo ra sự cạnh tranh trong lĩnh vực ngân hàng và tài chính. Hoạt động của các ngân hàng phụ thuộc rất nhiều vào độ chính xác của các quyết định quản lý. Mục tiêu của nghiên cứu này là dự báo hoạt động của ngân hàng dựa trên các kỹ thuật trí tuệ nhân tạo. Các mô hình dự báo được dựa trên các kỹ thuật trí tuệ nhân tạo như mạng nơ ron truyền thẳng nhiều lớp ANN-MLP, mạng RBF (Radial Basis Function), kỹ thuật Random Forest (RF) và hồi quy tuyến tính (MLR). Dữ liệu được sử dụng trong xây dựng mô hình gồm 405 mẫu được thu thập từ 45 ngân hàng hoạt động tại Việt Nam trong giai đoạn 2002-2022. Các chỉ số đầu ra dự báo bao gồm tổng các khoản vay (Total Loans) và tổng tiền gửi huy động (Total Deposits). Kết quả thực nghiệm và các chỉ số đánh giá mô hình xác định mô hình dự báo dựa trên kỹ thuật RF cho độ chính xác cao nhất, cụ thể như sau: RMSE; MAPE; MAE; R; Theil's U lần lượt là $6,2132 \times 10^7$; 1,2763; $2,8180 \times 10^7$; 0,9651; 0,1498 (đối với dự báo tổng các khoản vay) và $6,8408 \times 10^7$; 2,6092; $3,0893 \times 10^7$; 0,9741; 0,1469 (đối với dự báo tổng tiền gửi huy động).

Từ khóa: Hoạt động ngân hàng, trí tuệ nhân tạo, mạng nơ ron, random forest, RBF, hồi quy đa biến.

PREDICTION OF BANK PERFORMANCE USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Abstract

The banking industry is regarded as the backbone of a country's modern economy. Globalization and technological innovation have produced a highly competitive market in the banking and financial industry. Banks' performance is heavily dependent on the accuracy of managerial judgments. The objective of this research is to predict bank performance using several types of artificial intelligence techniques. Prediction models are based on artificial intelligence techniques such as Artificial Neural Networks and MultiLayer Perceptron (ANN-MLP), Radial Basis Function (RBF) network, Random Forest (RF) technique and multiple linear regression (MLR). The data used in developing models includes 405 samples collected from 45 banks in Vietnam during the period 2002-2022. Predicted outputs are total loans and total deposits. Experimental results and model evaluation criteria indicate that the prediction model based on RF technique provides the highest accuracy. Specifically, the values of RMSE ; MAPE ; MAE ; R ; Theil's U obtained by RF respectively are 6.2132×10^7 ; 1.2763; 2.8180×10^7 ; 0.9651; 0.1498 (for total loans prediction) and 6.8408×10^7 ; 2.6092; 3.0893×10^7 ; 0.9741; 0.1469 (for total deposit prediction).

Keywords: Bank performance, Artificial Intelligence, Neural Network, Random Forest, RBF, Multiple Linear Regression, Vietnam.

JEL classification: E5, G21, G24.

1. Mở đầu

Hiệu quả hoạt động của ngành tài chính và ngân hàng đóng một vai trò quan trọng trong sự ổn định kinh tế của bất kỳ quốc gia nào. Hơn nữa, bối cảnh toàn cầu hóa và tiến bộ công nghệ đã tạo ra một thị trường có tính cạnh tranh cao. Điều này ảnh hưởng đến tất cả các tổ chức trong tất cả các lĩnh vực và ngành nghề. Ngành ngân hàng Việt Nam

cũng không ngoại lệ. Các ngân hàng Việt Nam phải cạnh tranh không chỉ giữa các ngân hàng trong nước mà còn với các ngân hàng nước ngoài (Đoàn Việt Hùng, 2019). Điều này đòi hỏi các cấp quản lý trong ngành cần có khả năng đưa ra các quyết định điều hành chính xác. Việc dự báo sớm hiệu quả hoạt động giúp các ngân hàng có hướng xử lý kịp thời, giảm thiểu thời gian và quy trình phân

tích. Thông tin về dự báo hoạt động ngân hàng cũng cần thiết cho các nhà đầu tư và hoạch định chính sách. Trước đây, các công cụ toán học và thống kê thường được sử dụng để hỗ trợ người ra quyết định đưa ra những dự đoán về tình hình hoạt động và các thách thức trong tương lai. Ở các bài toán dự báo trong những nghiên cứu trước đây, các tác giả thường sử dụng dữ liệu lịch sử của chuỗi thời gian để cung cấp các ước lượng cho các giá trị trong tương lai. Tuy nhiên, trong những năm gần đây, các kỹ thuật trí tuệ nhân tạo (AI – artificial intelligence) đã chứng minh được khả năng khai thác dữ liệu cũng như dự báo hiệu quả hơn so với các phương pháp thống kê trong lĩnh vực tài chính. Kỹ thuật trí tuệ nhân tạo có thể tìm ra thông tin tiềm năng và quan trọng cần thiết từ dữ liệu (Lin và cộng sự, 2009). Các kỹ thuật này thường được sử dụng trong các vấn đề dự báo phức tạp và phi tuyến.

Trong nghiên cứu này, một số kỹ thuật trí tuệ bao gồm mạng nơron truyền thẳng ANN-MLP, mạng hàm cơ sở bán kính (Radial Basis Function - RBF), kỹ thuật rừng ngẫu nhiên (Random Forest – RF) và kỹ thuật hồi quy đa biến được sử dụng trong mô hình dự báo hiệu quả hoạt động của các ngân hàng thương mại Việt Nam. Đây là những kỹ thuật được coi là phổ biến và hiệu quả nhất trong dự báo (Ainan và Nur-E-Arefin, 2022). Nghiên cứu này cũng sẽ khám phá một ứng dụng của Học máy trong tài chính và khả năng áp dụng tại Việt Nam. Kết quả cho thấy, trong các mô hình được phát triển: mô hình dựa trên kỹ thuật Random Forest có độ chính xác dự báo tốt nhất.

2. Tổng quan

Giống như nhiều ngành và lĩnh vực khác, các ngân hàng đang ngày càng tìm cách tận dụng các ưu thế do công nghệ mang lại để cải thiện quy trình, năng suất và giảm chi phí. Do sự cạnh tranh ngày càng gay gắt nên các ngân hàng đang chạy đua trong việc áp dụng các công nghệ hiện đại để tự động hóa các quy trình vận hành và tăng năng lực phân tích bộ dữ liệu. Trong lĩnh vực dự báo hoạt động của ngân hàng, đã có nhiều công trình nổi bật trong đó sử dụng các kỹ thuật trí tuệ nhân tạo và học máy. Hao và Adsavakulchai, 2023 đã

sử dụng các kỹ thuật như cây quyết định, Naïve bayes và máy vector hỗ trợ để dự báo các khoản vay ngân hàng. Kết quả cho thấy cả ba kỹ thuật đều cho kết quả tương đối cao giống nhau, nhưng kỹ thuật cây quyết định dựa trên thuật toán J48 có hiệu quả tốt nhất với độ chính xác là 98.85%. Trong nghiên cứu của Assous (2022) đã sử dụng bốn kỹ thuật trí tuệ nhân tạo để phát triển mô hình dự báo hiệu quả hoạt động của các ngân hàng Saudi. Kỹ thuật cho độ chính xác cao nhất là kỹ thuật tự động tìm kiếm tương tác Chi-squared. Ledhem (2022) đã sử dụng các kỹ thuật học máy như hồi quy LASSO, random forest (RF), mạng nơron và kỹ thuật k láng giềng gần nhất (k -nearest neighbor - KNN) để dự báo hoạt động tài chính của các ngân hàng Indonesia. Kết quả cho thấy kỹ thuật RF cho kết quả tốt nhất. Appiahene và cộng sự (Appiahene, Missah và Najim, 2020) đã xây dựng mô hình dự báo hoạt động của các ngân hàng tại Ghana. Mô hình dự báo dựa trên các kỹ thuật cây quyết định (Decision Tree – DT), kỹ thuật RF và mạng nơron, kết quả cho thấy mô hình cây quyết định dựa trên thuật toán C5.0 cho kết quả dự báo tốt nhất.

Ở Việt Nam, trí tuệ nhân tạo đã và đang được quan tâm và phát triển ở một số lĩnh vực. Trong lĩnh vực ngân hàng và tài chính có một số nghiên cứu tiêu biểu như phát hiện gian lận thẻ tín dụng (Nguyễn Thị Liên, 2018), chấm điểm tín dụng (Giang Thị Thu Huyền, 2021), dự báo khách hàng rời bỏ dịch vụ ngân hàng (Đào Công Ân, 2018). Tuy nhiên, ứng dụng kỹ thuật trí tuệ nhân tạo trong dự báo hoạt động của ngân hàng hiện vẫn chưa có. Trong nghiên cứu này, tác giả đã sử dụng bốn kỹ thuật trí tuệ nhân tạo để xây dựng mô hình dự báo hiệu quả hoạt động của các ngân hàng Việt Nam. Dữ liệu sử dụng trong mô hình được thu thập từ 45 ngân hàng thương mại hoạt động tại Việt Nam trong giai đoạn 2002-2022.

3. Các kỹ thuật trí tuệ nhân tạo sử dụng trong dự báo

3.1. Mạng ANN-MLP

Là công cụ tính toán phổ biến trong lĩnh vực trí tuệ nhân tạo, có cấu trúc gồm một tập các đơn vị tính toán và được chia thành nhiều lớp như

Hình 1. Mức độ liên kết giữa các đơn vị được xác định bởi một tập giá trị trọng số. Tham số thiên lệch được sử dụng để tăng độ thích nghi của mạng với bài toán đặt ra. Số lớp và các đơn vị trong mỗi lớp phụ thuộc vào từng bài toán và được xác định bằng thử nghiệm. Số lượng đơn vị của lớp ra bằng số biến của vectơ lời giải.

Mạng nơron nhân tạo gồm có một nhóm các nơron nhân tạo (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút. Trong đó mạng nhận thức nhiều lớp (Multilayer Perceptron –MLP), hay còn gọi là mạng truyền thẳng nhiều lớp, mở rộng của mô hình mạng nhận thức, là mạng nơron nhân tạo được sử dụng phổ biến nhất, đặc biệt là mạng MLP có một lớp ẩn. Các nghiên cứu cho thấy rằng một mạng nơron truyền thẳng nhiều lớp với một lớp ẩn có thể xấp xỉ hóa tất cả các hàm số liên tục (Cuomo và cộng sự, 2022; Raviv, Lupyan, và Green 2022), do đó được ứng dụng trong rất nhiều lĩnh vực (Masini, Medeiros, và Mendes 2023). Hình 1 là một mạng nơron truyền thẳng nhiều lớp

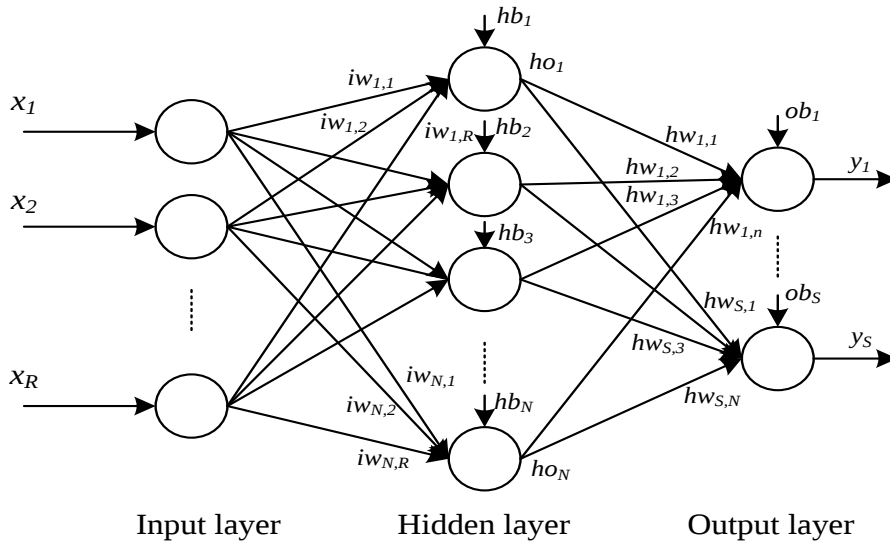
gồm 3 lớp. Với R , N , và S là số lượng nút vào, nút ẩn và nút ra; iw và hw là các trọng số của nút vào và nút ẩn; hb và ob là các vectơ độ lệch bias của lớp ẩn và lớp ra; x là vectơ các đầu vào; ho là các vectơ đầu ra của lớp ẩn; và y là vectơ đầu ra. Mạng nơron trong Hình 1 được trình bày thông qua công thức sau:

$$h_{oi} = f\left(\sum_{j=1}^R iw_{i,j} \cdot x_j + hb_i\right), \quad \text{với } i = 1, \dots, N \quad (1)$$

$$y_i = f\left(\sum_{k=1}^N hw_{i,k} \cdot ho_k + ob_i\right), \quad \text{với } i = 1, \dots, S \quad (2)$$

Trong đó, f là hàm kích hoạt (hàm chuyển)

Khi xây dựng một mô hình mạng nơron, cần phải xác định số lớp và số nút trong mỗi lớp. Một mạng có nhiều lớp và nút thì mạng sẽ phức tạp. Khi độ phức tạp của mô hình quá cao sẽ có hiện tượng quá khớp (quá chính xác trên tập dữ liệu huấn luyện), có thể dẫn đến việc dự đoán nhầm lẫn, và chất lượng mô hình không còn tốt trên dữ liệu kiểm tra (Caruana, Lawrence và Giles 2001).



Hình 1: Mạng nơron truyền thẳng MLP ba lớp

Chức năng của một mạng nơron được quyết định bởi cấu trúc mạng (số lớp, số nút trên mỗi lớp, liên kết giữa các lớp), các trọng số của các liên kết. Cấu trúc mạng thường cố định, và các trọng số được quyết định bởi các thuật toán huấn luyện. Tiến trình điều chỉnh các trọng số để mạng

“nhận biết” được quan hệ giữa đầu vào và đích mong muốn được gọi là học (learning) hay huấn luyện (training). Nhiều thuật toán đã được áp dụng để tìm ra tập trọng số tối ưu làm giải pháp cho các bài toán, chia làm hai nhóm chính: học có giám

sát (supervised learning) và học không có giám sát (unsupervised learning).

Học có giám sát là mạng được huấn luyện bằng cách cung cấp cho nó các cặp mẫu đầu vào và các đầu ra mong muốn (target values). Sự khác biệt giữa các đầu ra thực tế so với các đầu ra mong muốn được thuật toán sử dụng để thích ứng các trọng số trong mạng. Điều này thường được đưa ra như một bài toán xấp xỉ hàm số: cho dữ liệu huấn luyện bao gồm các cặp mẫu đầu vào x , và một đích tương ứng t , mục đích là tìm ra hàm $f(x)$ thỏa mãn tất cả các mẫu học đầu vào.

Để huấn luyện một mạng và xét xem nó thực hiện tốt đến đâu, ta cần xây dựng một hàm mục tiêu (cost function) để cung cấp cách thức đánh giá khả năng mô hình. Có một số hàm cơ bản được sử dụng như tổng bình phương lỗi (sum of squared error - SSE) và trung bình bình phương lỗi (mean squared error - MSE). Trong quá trình huấn luyện, sẽ đạt được phương án tối ưu hoặc gần tối ưu tương ứng với các vector trọng số và độ lệch.

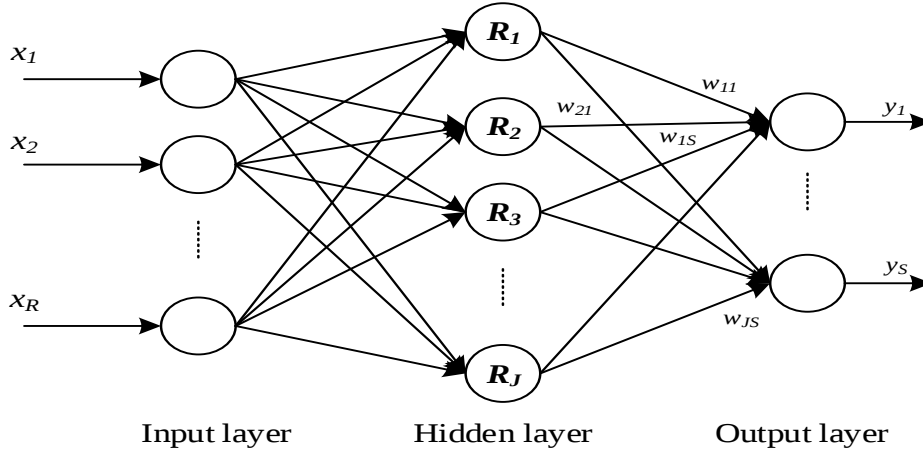
Giả thiết là có m cặp đầu vào và đầu ra mong muốn, x_k, t_k , với $k = 1, 2, \dots, m$. Trong quá trình huấn luyện, các giá trị iw , hw , hb , và ob sẽ được thay đổi để tối thiểu hóa hàm mục tiêu E , giả thiết E sử dụng hàm MSE sẽ được biểu diễn như sau:

$$MSE = \frac{1}{m} \sum_{k=1}^m e_k^2 = \frac{1}{m} \sum_{k=1}^m (t_k - y_k)^2 \quad (3)$$

Với y_k là đầu ra thực tế và t_k là đầu ra mong muốn.

3.2. Mạng RBF

Kiến trúc của mạng hàm cơ sở bán kính RBF (Radial Basis Function) là một loại mạng nơron nhân tạo truyền thẳng bao gồm 3 lớp: lớp đầu vào, lớp ẩn và lớp đầu ra như trong Hình 2. Mặc dù kiến trúc của mạng (RBF) tương đối đơn giản, nhưng có khả năng tổng hợp cao (Jiang, Cao và Chen 2016). Mạng RBF được sử dụng trong các bài toán phân loại và hồi quy trong nhiều lĩnh vực khác nhau như nhận dạng mẫu và xấp xỉ hàm (Batool và cộng sự, 2013; Guan, Zhu và Song 2016).



Hình 2: Mạng RBF

Như Hình 2, đầu ra được tính theo công thức sau:

$$y_s = \sum_{j=1}^J w_{js} R_j(x), s = 1, 2, \dots, S \quad (4)$$

Trong đó S là số lượng đầu ra, J là số nơron trong lớp ẩn và w_{js} là trọng số liên kết giữa nôt thứ j trong lớp ẩn với node thứ s ở lớp đầu ra. Có một số hàm cơ sở hướng tâm, tuy nhiên hàm được sử dụng phổ biến nhất là:

$$R_j(x) = \exp\left(-\frac{\|x - c_j\|^2}{2\sigma_j^2}\right), j = 1, 2, \dots, J \quad (5)$$

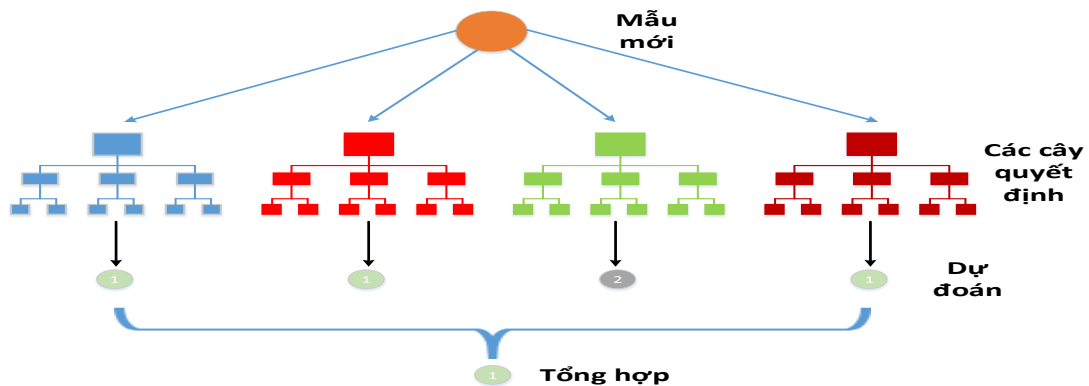
Trong đó x là vector đầu vào, mỗi vector đầu vào được thể hiện bằng một vector N -chiều; c_j và σ_j là tâm và độ lệch (độ rộng) của RBF; $\|x - c_j\|$ là chuẩn dựa trên khoảng cách Euclid của x và c_j , hay còn gọi là khoảng cách giữa hai vector x và c_j .

Thông qua huấn luyện mạng RBF, mối quan hệ giữa đầu vào và đầu ra sẽ được thiết lập. Việc huấn luyện mạng RBF thông qua việc xác định các tham số tâm, độ rộng của hàm cơ sở và các trọng số kết nối.

3.3. Random Forest (RF)

Thuật toán Random Forest (rừng ngẫu nhiên) là phương pháp phân lớp thuộc tính được phát triển bởi Leo Breiman (Leo Breiman, 2001). Random forest được xây dựng dựa trên nền tảng thuật toán cây phân loại và hồi quy (Classification and Regression Trees - CART) sử dụng kỹ thuật có tên gọi là học kết hợp (Hart, Stork và Duda 2000). Kỹ thuật này cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây để phân chia cho mức tiếp theo của cây phân lớp (Hình 3). Bằng cách chia nhỏ không gian tìm kiếm thành các cây nhỏ hơn như vậy cho phép thuật toán có thể phân loại một cách rất nhanh chóng cho dù

không gian thuộc tính rất lớn. Các tham số đầu vào của thuật toán khá đơn giản bao gồm số các thuộc tính được chọn trong mỗi lần phân chia (mtry). Giá trị mặc định của tham số này là căn bậc hai của p với p là số lượng các thuộc tính. Tương tự như thuật toán CART, RF vẫn sử dụng công thức Gini là công thức tính toán việc phân chia cây. số lượng cây được tạo ra là không hạn chế và cũng không sử dụng bất kỳ kỹ thuật để hạn chế mở rộng cây. Cần phải lựa chọn tham số cho biết số lượng cây (ntree) sẽ được sinh ra sao cho đảm bảo rằng sẽ mỗi một thuộc tính sẽ được kiểm tra một vài lần. Sau khi một số lượng lớn các cây được tạo ra các cây này “bỏ phiếu” cho lớp phổ biến nhất.



Hình 3: Thuật toán Random Forest

3.4. Hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression - LR) là một trong những những thuật toán cơ bản nhất của học máy, thuộc nhóm học có giám sát. Hồi quy tuyến tính là một phương pháp đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các bài toán. Hồi quy tuyến tính là một mô hình tuyến tính, ví dụ: một mô hình trong đó giá định mỗi quan hệ tuyến tính giữa các biến đầu vào (x) và biến đầu ra duy nhất (y). Nói cách khác, y có thể được tính toán từ sự kết hợp tuyến tính của các biến đầu vào (x). Khi có một biến đầu vào duy nhất (x), phương pháp này được gọi là hồi quy tuyến tính đơn giản (simple linear regression). Khi có nhiều biến đầu vào, ta có phương pháp là hồi quy tuyến tính nhiều biến (multiple linear regression). Các kỹ thuật khác nhau có thể được sử dụng để chuẩn bị hoặc huấn luyện phương trình hồi quy tuyến tính từ dữ liệu, trong đó phương pháp phổ biến nhất được gọi là bình phương nhỏ nhất thông thường (Ordinary least squares). Mô hình hồi quy này được gọi là

Ordinary least squares linear regression, hay còn gọi ngắn gọn là Least squares regression (Maulud và Abdulazeez 2020).

4. Xây dựng mô hình dự báo

4.1. Lựa chọn đầu vào và đầu ra

Có rất nhiều nghiên cứu khác nhau trong và ngoài nước thảo luận nhằm xác định đầu ra và đầu vào của ngân hàng (Wei, Ye và Zhang, 2021; Appiahene, Missah và Najim, 2020; Kosmidou và Zopounidis, 2008; Nguyễn Minh Kiều và Nguyễn Ngọc Thùy Trang, 2020; Boďa và Piklová 2018). Tuy nhiên, không có cách tiếp cận nào là hoàn hảo trong việc xác định đầu ra và đầu vào của ngân hàng vì không có cách tiếp cận nào có thể phản ánh được tất cả các hoạt động, vai trò của ngân hàng với tư cách là chủ thể cấp các dịch vụ trung gian tài chính (Sealey Jr và Lindley 1977). Sự khác biệt trong cách tiếp cận là bởi biến tiền gửi huy động (Total Deposits) có cả đặc điểm của biến đầu vào và đầu ra. Có hai cách tiếp cận chính như sau:

Coi ngân hàng là một tổ chức trung gian tài chính (Intermediation Approach) kết nối khu vực

tiết kiệm và khu vực đầu tư của nền kinh tế, để phân tích và đánh giá hiệu quả hoạt động của ngân hàng. Với cách tiếp cận này, các đầu vào được sử dụng bao gồm: chi phí nhân viên (Personnel Expenses), tài sản cố định (Total Fixed Assets), tiền gửi huy động (Total Deposits). Đầu ra là lợi nhuận (Profits Before Tax).

Coi ngân hàng là một tổ chức cung cấp các dịch vụ và sản phẩm tài chính (Production Approach) cho khách hàng, bao gồm lưu trữ tiền, cho vay tiền, chuyển khoản tiền, đầu tư và tư vấn tài chính. Đầu vào là tài sản cố định (Fixed assets), Số lao động (Number of Employees), Số chi nhánh (Number of Branches). Đầu ra là những chỉ số liên quan đến dịch vụ cung cấp cho khách hàng là tổng

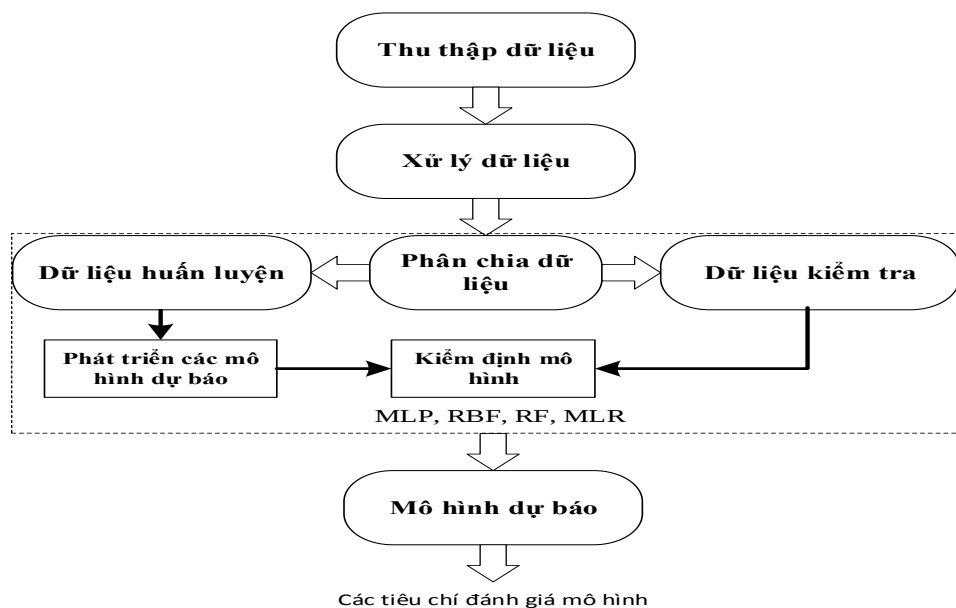
tiền gửi huy động (Total Deposits) và tổng các khoản vay (Total Loans). Trong nghiên cứu này, tác giả sử dụng cách tiếp cận coi ngân hàng là một tổ chức cung cấp dịch vụ và sản phẩm tài chính.

4.2. Dữ liệu

Dữ liệu sử dụng trong nghiên cứu được thu thập từ 45 ngân hàng Việt Nam trong giai đoạn 2002-2022 (Le Duong Quang Tu, Ho Huu Tin, Ngo Thanh, Nguyen Tien Dat và Tran Hong Son, 2022). Dữ liệu gồm 644 mẫu được tổng hợp từ các báo cáo và báo cáo tài chính thường niên của các ngân hàng. Sau khi loại bỏ những mẫu thiếu dữ liệu hoặc dữ liệu không phù hợp, có 405 mẫu được sử dụng trong nghiên cứu. Bảng 1 cung cấp một số thống kê mô tả của dữ liệu.

Bảng 1: Thống kê mô tả

Biến	Min	Max	Trung bình	Độ lệch chuẩn
Tài sản cố định (Fixed assets) (Đơn vị: tỷ đồng)	769,00	11.114.537,00	1.742.134,40	2.503.932,12
Số lao động (Number of Employees)	45,00	39.950,00	6.906,97	8.593,60
Số chi nhánh (Number of Branches)	2,00	2.400,00	268,83	457,02
Tổng tiền gửi huy động (Total Deposits) (Đơn vị: tỷ đồng)	7.314,00	1.404.875.777,00	134.050.816,11	233.066.086,72
Tổng các khoản vay (Total Loans) (Đơn vị: tỷ đồng)	50.721,00	1.345.632.643,00	122.985.141,52	218.774.472,25



Hình 4: Các bước xây dựng mô hình

4.3. Xây dựng các mô hình dự báo

Các bước xây dựng mô hình được trình bày trong Hình 4, cụ thể như sau:

Bước 1: Chuẩn bị dữ liệu

Bước này sẽ thực hiện thu thập dữ liệu và tiền xử lý dữ liệu. Dữ liệu sau khi được thu thập, cần phải được xử lý, làm sạch và biến đổi trước khi một kỹ thuật học máy có thể được huấn luyện trên những bộ dữ liệu này. Các kỹ thuật này bao gồm: xử lý dữ liệu bị khuyết thiếu, mã hóa các biến nhóm (encoding categorical variables), chuẩn hóa dữ liệu, ...

Bước 2: Phân chia dữ liệu

Bước này chuẩn bị dữ liệu để xây dựng mô hình. Kiểm tra chéo (cross-validation) thường được sử dụng để chia tập dữ liệu thành hai phần, phục vụ cho huấn luyện (training datasets) và kiểm tra mô hình (testing dataset). Trong nghiên cứu này, tác giả sử dụng 2/3 tập dữ liệu cho mục đích huấn luyện, phần còn lại (1/3) dữ liệu cho mục đích kiểm tra.

Bước 3: Xây dựng mô hình

Mục đích của bước này là tìm ra hàm $f(x)$ và gán nhãn cho dữ liệu, thường được gọi là học hay training. Trong đó: x là các dữ liệu đầu vào, y là đầu ra của dự báo. Các kỹ thuật học giám sát (supervised learning) là MLP, RBF, RF, MLR sẽ được sử dụng.

Bước 4: Kiểm tra

Các dữ liệu mới sẽ được đưa vào để kiểm tra, đánh giá.

Bước 5: Đánh giá và chọn ra mô hình tốt nhất.

Việc đánh giá được thực hiện thông qua các chỉ số lỗi tìm được trên tập dữ liệu. Nếu không đạt được kết quả mong muốn thì các tham số (turning parameter) của các thuật toán phải được thay đổi để tìm ra các mô hình tốt hơn và thực hiện kiểm tra, đánh giá lại. Cuối cùng chọn ra mô hình dự báo tốt nhất.

Tham số của các mô hình

Các mô hình được thực hiện trên phần mềm Matlab R2022b và Weka 3.9. Tham số chính của các mô hình cụ thể như sau: Đối với mô hình ANN-MLP, mạng nơron truyền thẳng nhiều lớp

với một lớp ẩn được sử dụng. Lớp ẩn có 2 nút mạng, hàm kích hoạt là gradient liên hợp được sử dụng để tăng tốc độ tính toán. Hàm chi phí là MSE. Đối với mô hình RBF, số lượng hàm Gauss cơ sở là 2, hàm kích hoạt là gradient liên hợp. Đối với mô hình Random Forest, số lượng cây xây dựng là 100 và không giới hạn độ sâu của cây. Đối với dự báo Tổng các khoản vay (Total Loans), mô hình hồi quy MLR có công thức như sau:

Tổng các khoản vay = $10.172,5429 \cdot \text{Số lao động} - 32.315,5208 \cdot \text{Số chi nhánh} + 54,7589 \cdot \text{Tài sản cố định} - 33.986.096,7684$.

Đối với dự báo Tổng tiền gửi huy động (Total Deposits), mô hình hồi quy MLR có công thức như sau:

Tổng tiền gửi huy động = $7.882,3132 \cdot \text{Số lao động} + 62,7792 \cdot \text{Tài sản cố định} - 29.761.876,9481$.

Các chỉ số đánh giá mô hình

Sai số dự báo là chênh lệch giữa giá trị thực và giá trị dự báo nhằm đánh giá chất lượng hay sự phù hợp của mô hình dự báo tại cùng một thời điểm. Sai số dự báo cũng là căn cứ để thực hiện việc điều chỉnh mô hình dự báo.

Căn của sai số bình phương trung bình (Root Mean Squared Error- RMSE):

$$RMSE = \sqrt{\frac{1}{m} \sum_{k=1}^m (t_k - y_k)^2} \quad (6)$$

Với t_k là giá trị mong muốn, y_k là giá trị dự báo của mô hình, m là tổng số mẫu.

Sai số tương đối trung bình (Mean Absolute Percent Error- MAPE).

$$MAPE = \frac{1}{m} \sum_{k=1}^m \left| \frac{t_k - y_k}{t_k} \right| \quad (7)$$

Sai số tuyệt đối trung bình MAE (Mean Absolute Error).

$$MAE = \frac{1}{m} \sum_{k=1}^m |t_k - y_k| \quad (8)$$

Các chỉ số MAE và MSE và RMSE có đặc tính, công năng như nhau và thường cho cùng một kết quả khi đánh giá. Tuy nhiên, nếu giá trị sai số $\varepsilon_t = t_t - y_t$ đều nhau thì nên chọn MSE để đánh giá. Ngược lại, nếu giá trị sai số ε_t quá khác biệt, MAE nên được lựa chọn. Tiêu chí RMSE là căn bậc hai của tiêu chí MSE nên hai tiêu chí về

bản chất là một; điều khác biệt là giá trị của tiêu chí RMSE bé hơn.

Tiêu chí MAPE giúp đánh giá sai số một cách tương đối, do đó thường được áp dụng khi đánh giá sai số dự báo với các bộ số liệu khác nhau. Ngược lại, với cùng một bộ số liệu nhưng áp dụng nhiều phương pháp dự báo khác nhau thì không nên áp dụng tiêu chí MAPE vì tính phức tạp trong tính toán.

Hệ số tương quan R : Có giá trị từ -1 đến 1, được dùng để đo lường mức độ phụ thuộc tuyến tính giữa giá trị thực tế và giá trị dự báo. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là không có liên hệ giữa hai biến số; ngược lại nếu bằng -1 hay 1 có nghĩa là giữa giá trị thực tế và giá trị dự báo có một mối liên hệ tuyệt đối. Nếu $R < 0$ có nghĩa là khi t tăng cao thì y giảm và ngược lại; nếu $R > 0$ có nghĩa là khi t tăng cao thì y cũng tăng, và khi t giảm cao thì y cũng giảm theo.

$$R = \frac{\sum_{k=1}^m (t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^m (t_k - \bar{t})^2 \cdot \sum_{k=1}^m (y_k - \bar{y})^2}} \quad (9)$$

$$\text{Với } \bar{t} = \frac{1}{m} \sum_{k=1}^m t_k \text{ và } \bar{y} = \frac{1}{m} \sum_{k=1}^m y_k.$$

Theil's U: Hệ số này được sử dụng để so sánh các mô hình dự báo, công thức như sau:

$$U = \frac{\sqrt{\sum_{k=1}^m (t_k - y_k)^2}}{\sum_{k=1}^m t_k^2 + \sum_{k=1}^m y_k^2} \quad (10)$$

Giá trị U nằm trong khoảng từ 0 đến 1, U càng tiến về 0 thì mô hình dự báo càng chính xác.

5. Kết quả

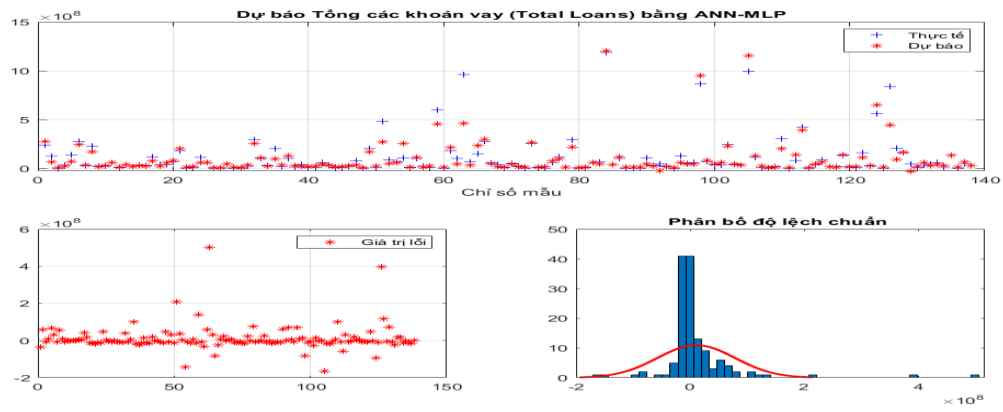
Các chỉ số đánh giá mô hình được trình bày trong Bảng 2. Các giá trị $RMSE$, $MAPE$, MAE , R và $Theil's U$ của mô hình dựa trên kỹ thuật RF tương ứng là $6,2132 \times 10^7$; 1,2763; $2,8180 \times 10^7$; 0,9651 và 0,1498 đối với dự báo tổng các khoản vay (Total Loans); $6,8408 \times 10^7$; 2,6092; $3,0893 \times 10^7$; 0,9741 và 0,1469 đối dự báo tổng tiền gửi huy động (Total Deposits). Mô hình được đánh giá là tốt khi các giá trị $RMSE$, $MAPE$, và MAE nhỏ, R gần giá trị 1 và $Theil's U$ gần giá trị 0. Trong Bảng 2, các giá trị tốt nhất đối với mỗi chỉ số được in đậm và nghiêng. Dễ dàng nhận thấy mô hình dự báo dựa trên kỹ thuật RF là mô hình tốt nhất (9 tiêu chí tốt nhất trong 10 tiêu chí xét). Hầu hết các mô hình dựa trên kỹ thuật trí tuệ nhân tạo đều cho kết quả chấp nhận được.

Bảng 2: Các chỉ số đánh giá của các mô hình dự báo

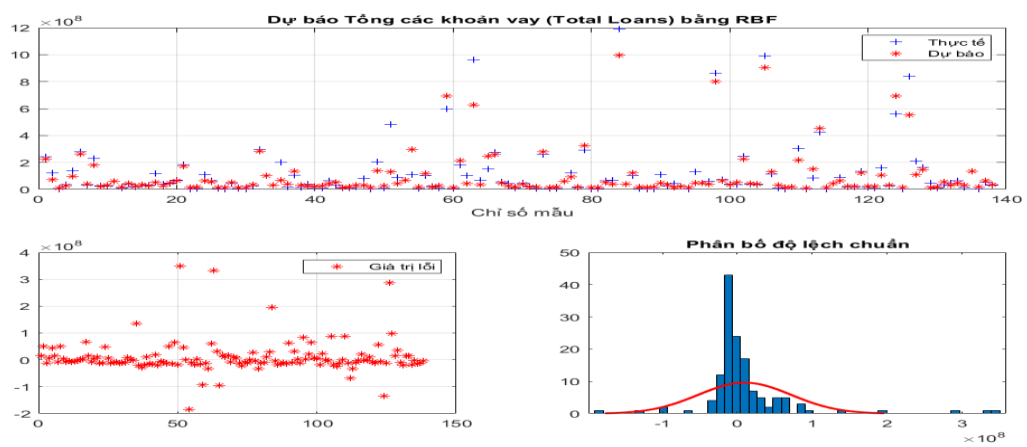
Đầu ra dự báo	Mô hình	RMSE ($\times 10^7$)	MAPE	MAE ($\times 10^7$)	R	Theil's U
	ANN-MLP	6,8574	1,7920	3,0543	0,9397	0,1570
Tổng các khoản vay (Total Loans)	Random Forest	6,2132	1,2763	2,8180	0,9651	0,1498
	RBF	6,2541	2,4085	3,0821	0,9542	0,1467
	Hồi quy tuyến tính	7,2158	6,2813	4,4039	0,9331	0,1673
Tổng tiền gửi huy động (Total Deposits)	ANN-MLP	7,4915	4,5206	3,8875	0,9478	0,1549
	Random Forest	6,8408	2,6092	3,0893	0,9741	0,1469
	RBF	8,3028	5,5054	4,2925	0,9376	0,1747
	Hồi quy tuyến tính	9,4856	10,9306	5,1364	0,9138	0,2006

Các Hình 5 đến Hình 12 thể hiện mức độ phù hợp giữa giá trị thực tế và giá trị dự báo bởi các mô hình dựa trên kỹ thuật trí tuệ nhân tạo đã phát triển cho việc dự báo tổng các khoản vay

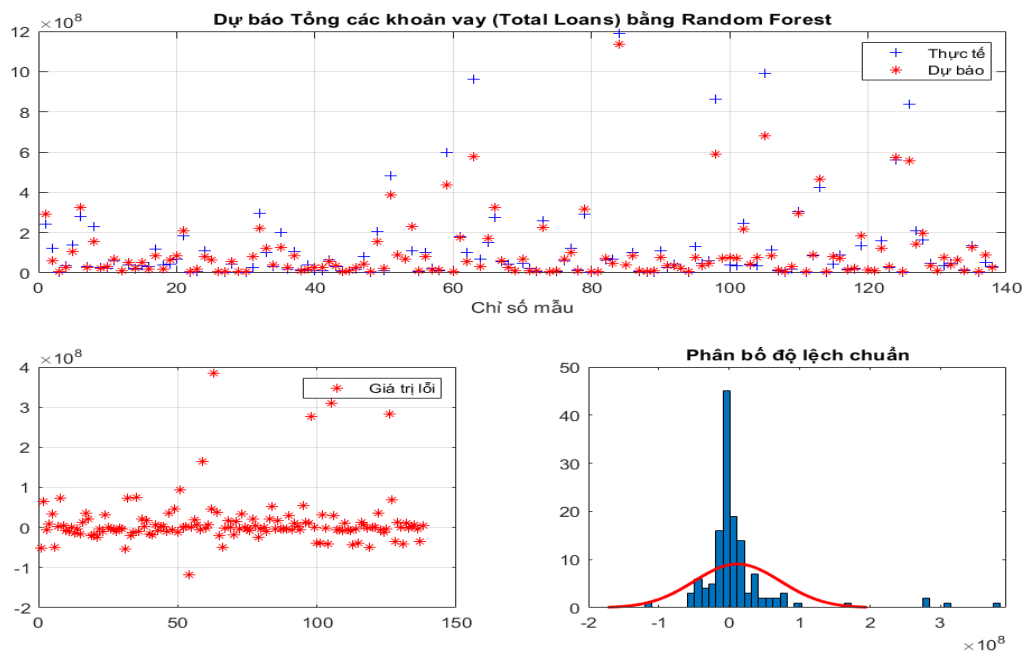
(Total Loans) và tổng tiền gửi huy động (Total Deposits). Trong mỗi hình, thể hiện giá trị thực tế và giá trị dự báo được, giá trị lỗi và phân bố độ lệch chuẩn của lỗi.



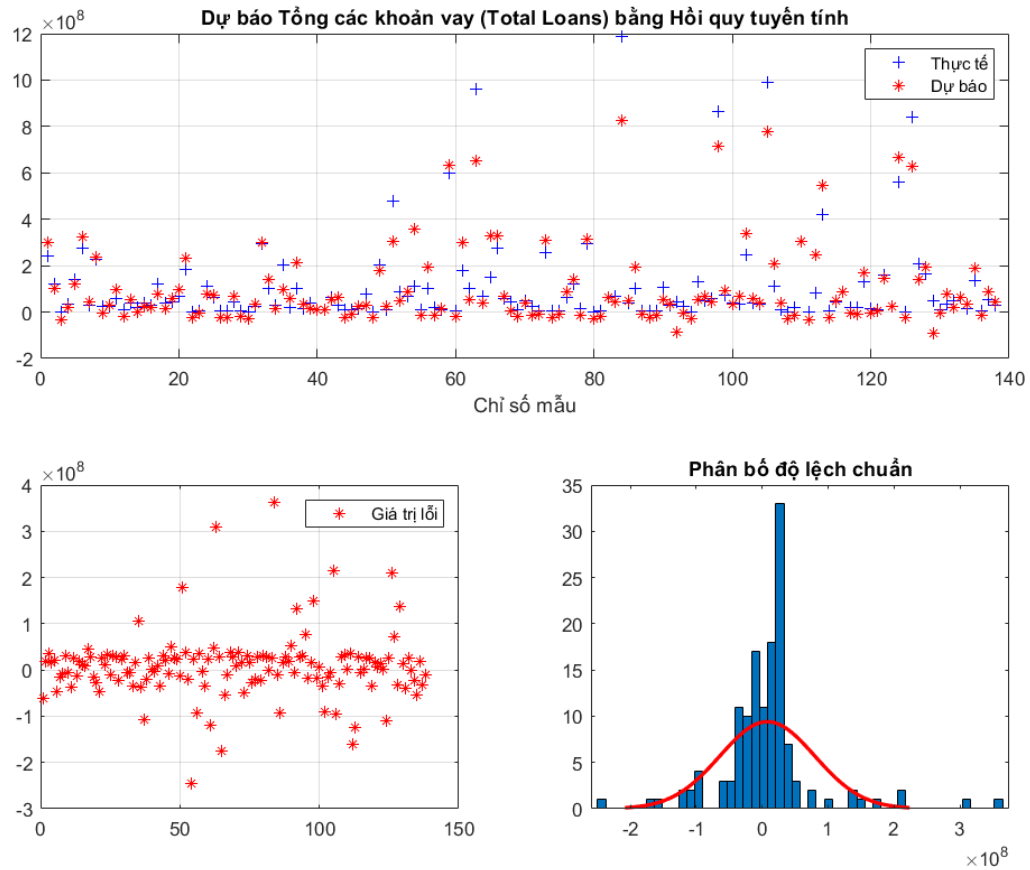
Hình 5: Dự báo tổng các khoản vay bằng ANN-MLP



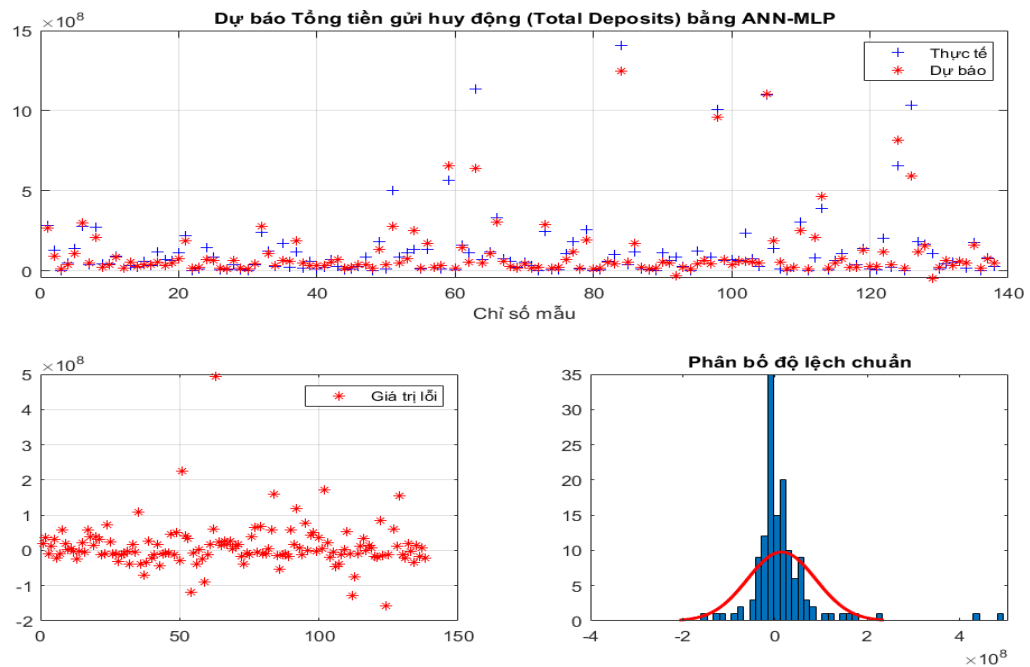
Hình 6: Dự báo tổng các khoản vay bằng RBF



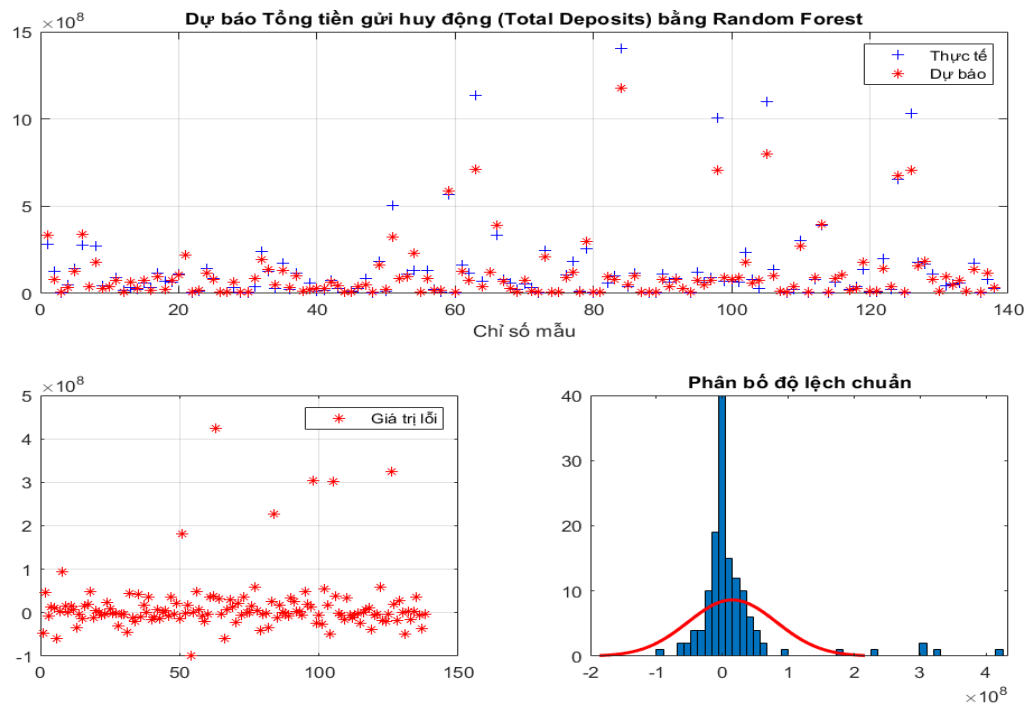
Hình 7: Dự báo tổng các khoản vay bằng Random Forest



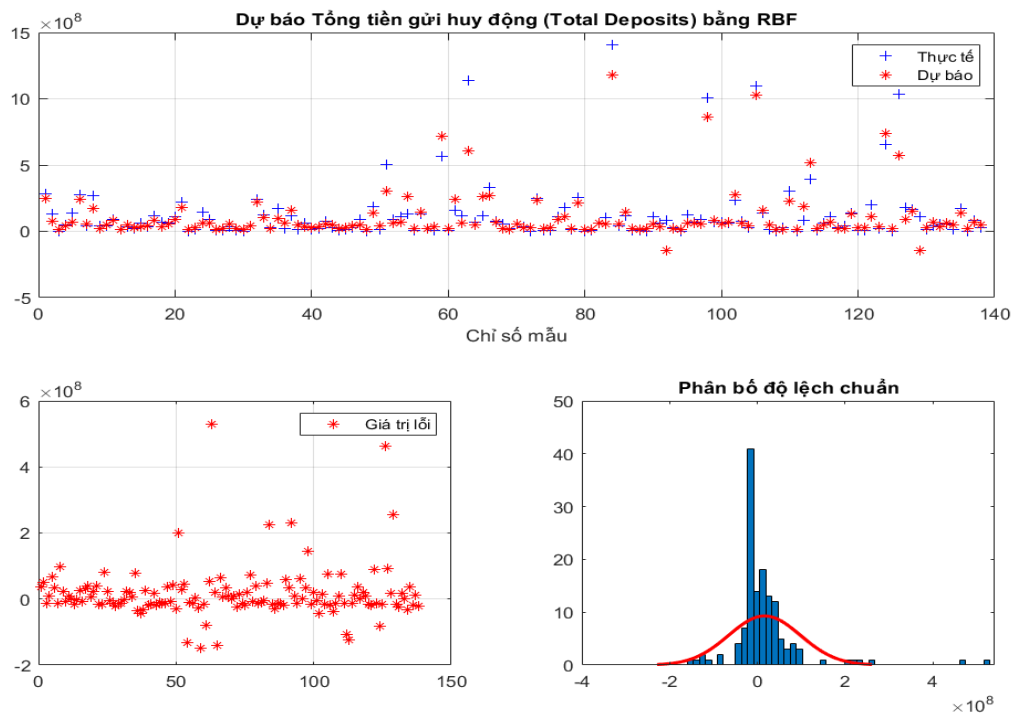
Hình 8: Dự báo tổng các khoản vay bằng MLR



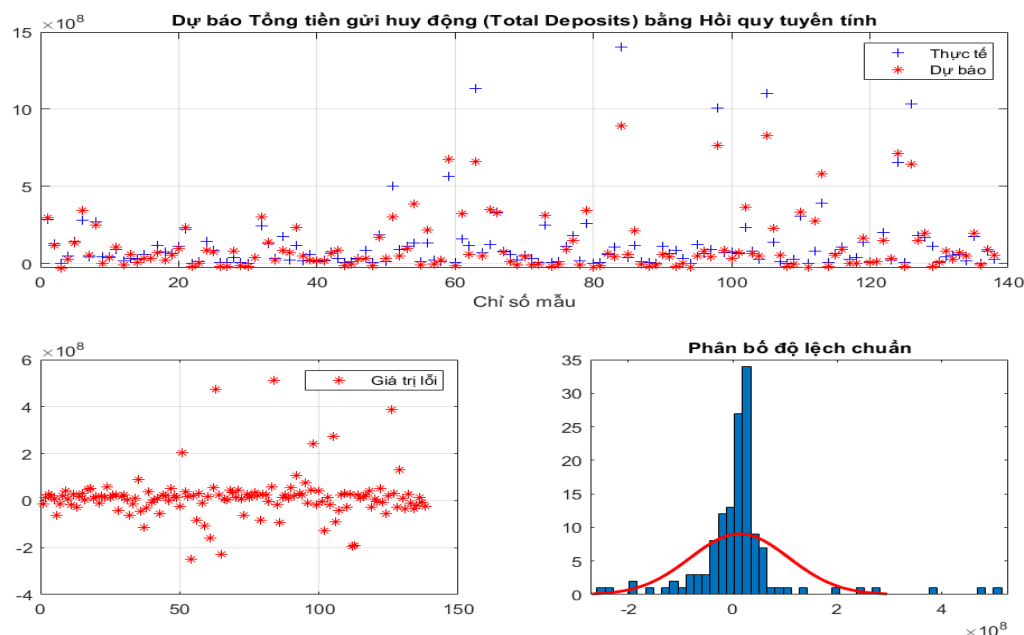
Hình 9: Dự báo tổng tiền gửi huy động bằng ANN-MLP



Hình 10: Dự báo tổng tiền gửi huy động bằng RBF



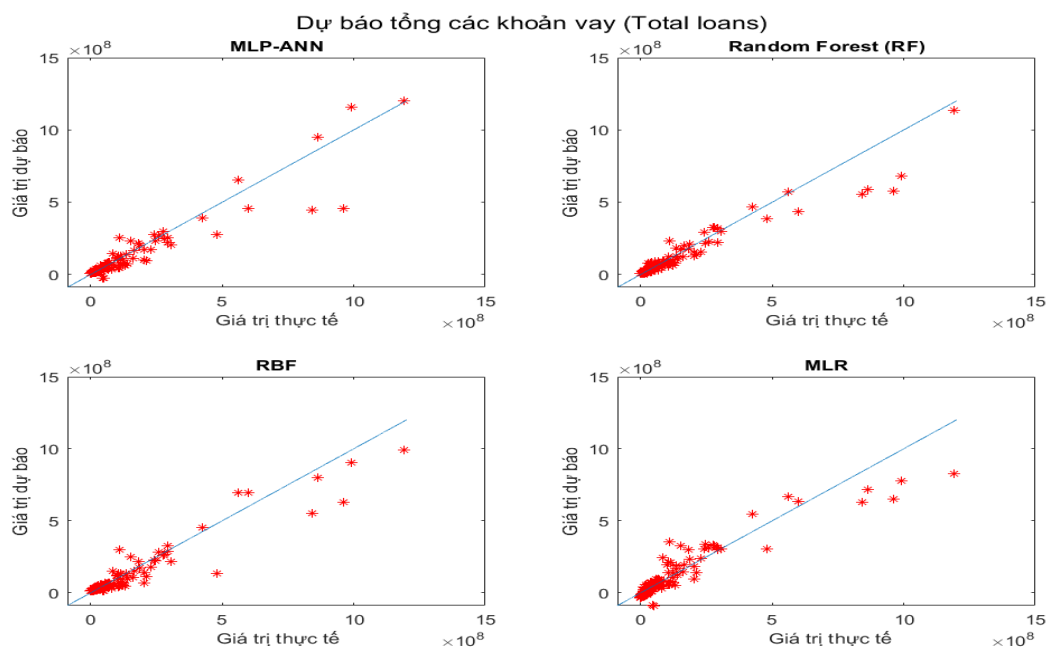
Hình 11: Dự báo tổng tiền gửi huy động bằng RF



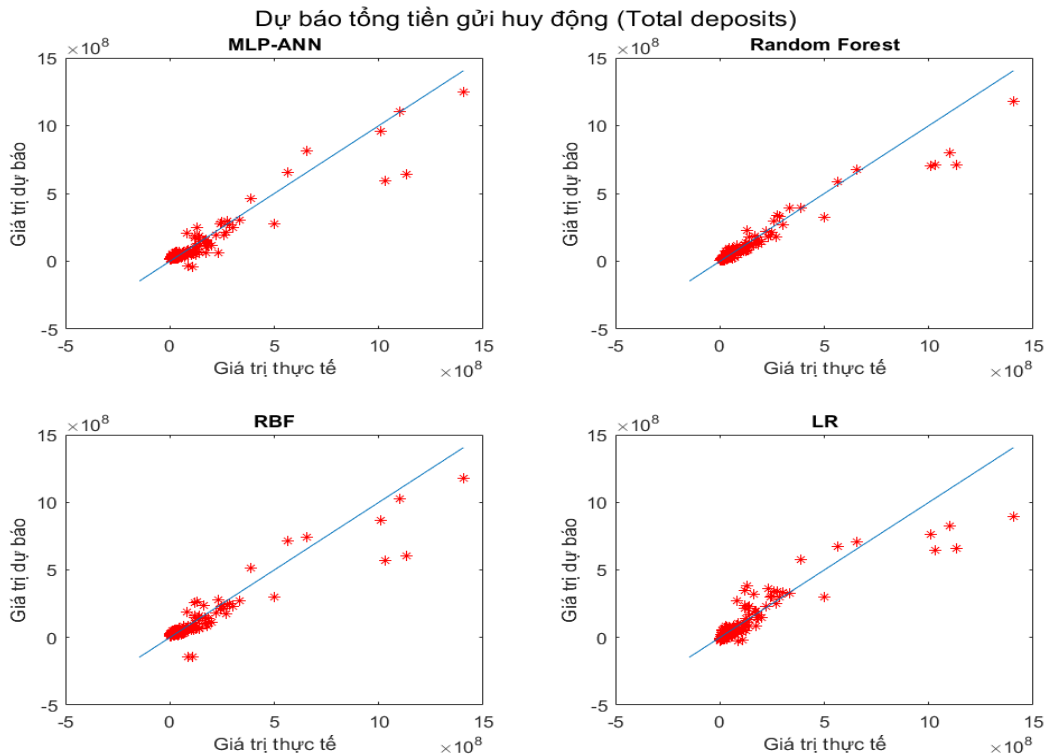
Hình 12: Dự báo tổng tiền gửi huy động bằng MLR

Việc so sánh các giá trị thực tế và giá trị dự báo của các mô hình cũng được thể hiện qua biểu đồ phân tán trong Hình 13 và Hình 14. Biểu đồ phân tán thể hiện mối tương quan giữa hai tập giá trị thực tế và giá trị dự báo bởi các mô hình. Trong hình vẽ, đường thẳng 1:1 thể hiện giá trị thực tế và giá trị dự báo trùng nhau. Nếu tập các giá trị thực

tế và giá trị dự báo tập trung quanh đường thẳng 1:1 thì giá trị dự báo gần với giá trị thực tế. Quan sát Hình 13 và Hình 14 nhận thấy giá trị dự báo bởi mô hình RF và giá trị thực tế tương đối khớp nhau do các điểm tập trung gần nhau và có thể vẽ được một đường thẳng đi qua các điểm này. Điều này thể hiện mối tương quan là rất mạnh.



Hình 13: Biểu đồ phân tán của các mô hình khi dự báo tổng các khoản vay (Total loans)



Hình 14: Biểu đồ phân tán của các mô hình khi dự báo tổng tiền gửi huy động (Total deposits)

Dựa vào các kết quả thu được, có thể kết luận rằng mô hình dựa trên kỹ thuật RF cho kết quả tin cậy và chính xác cao nhất. Do đó, với bộ dữ liệu đã thu thập được, mô hình dự báo dựa trên kỹ thuật RF có thể được ứng dụng trong việc dự báo hoạt động của các ngân hàng.

6. Kết luận và hướng phát triển

Dự báo hoạt động của ngân hàng là một lĩnh vực nghiên cứu thu hút sự quan tâm của nhà quản lý, nhà điều hành và nhà phân tích dữ liệu vì tính ứng dụng cao của kết quả dự báo. Trong nghiên cứu này, tác giả đã phát triển các mô hình dự báo hoạt động của ngân hàng thông qua dự báo hai chỉ số đầu ra tổng các khoản vay (Total loans) và tổng tiền gửi huy động (Total deposits) chính dựa trên các kỹ thuật trí tuệ nhân tạo bao gồm ANN-MLP, RBF, RF và hồi quy tuyến tính. Kết quả thực nghiệm dựa trên các chỉ tiêu đánh giá đã khẳng

định rằng mô hình dự báo dựa trên kỹ thuật RF (Random Forest) cho độ chính xác dự báo cao nhất và tin cậy nhất.

Do hạn chế về mặt số liệu nên các yếu tố khác ảnh hưởng đến hoạt động của ngân hàng như chỉ số ổn định chính trị và tình hình kinh tế thế giới chưa đưa được vào mô hình dự báo. Hướng nghiên cứu tiếp theo sẽ là: nâng cấp các mô hình đã được xây dựng trong bài báo thành một hệ hỗ trợ ra quyết định hoàn chỉnh phục vụ cho dự báo giá của một số mặt hàng, bao gồm các thành phần: hệ thống máy tính, cơ sở dữ liệu, quản lý mô hình, quản lý cơ sở tri thức, giao tiếp với người dùng. Đồng thời cần tiếp tục bổ sung thêm các đầu vào khác có thể ảnh hưởng đến sự hoạt động của các ngân hàng. Để tăng độ chính xác dự báo của các mô hình, việc tinh chỉnh các tham số cũng cần được xem xét.

REFERENCE

- [1]. Ainan, Ummey Hany, and Md Nur-E-Arefin. (2022). Prediction of bank performance using machine learning and genetic algorithm hybrid models. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, 1–6. IEEE.
- [2]. Appiahene, Peter, Yaw Marfo Missah, and Ussiph Najim. (2020). Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks. *Advances in Fuzzy Systems* 2020: 1–12.

- [3]. Assous, Hamzeh F. (2022). Prediction of Banks Efficiency Using Feature Selection Method: Comparison between Selected Machine Learning Models. *Complexity* 2022.
- [4]. Batool, Fatima, Shahid Adeel, Muhammad Azeem, Ali Ahmad Khan, Ijaz Ahmad Bhatti, Abdul Ghaffar, and Naeem Iqbal. (2013). Gamma Radiations Induced Improvement in Dyeing Properties and Colorfastness of Cotton Fabrics Dyed with Chicken Gizzard Leaves Extracts. *Radiation Physics and Chemistry*. <https://doi.org/10.1016/j.radphyschem.2013.03.045>.
- [5]. Bod'a, Martin, and Zuzana Piklová. (2018). The Production or Intermediation Approach?: It Matters. In *Contemporary Trends and Challenges in Finance: Proceedings from the 3rd Wroclaw International Conference in Finance*, 111–20. Springer.
- [6]. Breiman, Leo. (2001). Random Forests. *Machine Learning* 45: 5–32.
- [7]. Caruana, Rich, Steve Lawrence, and Lee Giles. (2001). Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems*.
- [8]. Cuomo, Salvatore, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. (2022). Scientific Machine Learning through Physics-Informed Neural Networks: Where We Are and What's next. *Journal of Scientific Computing* 92 (3): 88.
- [9]. Đào Công Ân. (2018). *Mô hình dự báo churn cho khách hàng bằng phương pháp học máy suy diễn phương sai*. Trường Đại học Bách khoa Hà Nội.
- [10]. Guan, Xuemei, Yuren Zhu, and Wenlong Song. (2016). Application of RBF Neural Network Improved by Peak Density Function in Intelligent Color Matching of Wood Dyeing. *Chaos, Solitons and Fractals*. <https://doi.org/10.1016/j.chaos.2016.02.015>.
- [11]. Hao, Chen, and Suwannee Adsavakulchai. (2023). The Use of Machine Learning Algorithms for Bank Loan Prediction. *European Economic Letters (EEL)* 13 (3): 735–41.
- [12]. Hart, Peter E, David G Stork, and Richard O Duda. (2000). *Pattern Classification*. Wiley Hoboken.
- [13]. Đoàn Việt Hùng. (2019). Cạnh tranh và hiệu quả hoạt động của các ngân hàng thương mại tại Việt Nam: tiếp cận bằng phương pháp lasso. *Tạp Chí Khoa Học Lạc Hồng* 8: 8–13.
- [14]. Giang Thị Thu Huyền. (2021). Một số kỹ thuật học máy cho chấm điểm tín dụng. *Tạp Chí Khoa Học & Đào Tạo Ngân Hàng* 227: 34–40.
- [15]. Jiang, Jingfei, Dengqing Cao, and Huatao Chen. (2016). Boundary Value Problems for Fractional Differential Equation with Causal Operators. *Applied Mathematics and Nonlinear Sciences*. <https://doi.org/10.21042/amns.2016.1.00002>.
- [16]. Nguyễn Minh Kiều và Nguyễn Ngọc Thùy Trang. (2020). Phân tích hiệu quả kỹ thuật của các ngân hàng thương mại cổ phần tại Việt Nam. *Tạp Chí Khoa Học Đại Học Mở Thành Phố Hồ Chí Minh-Kinh Tế và Quản Trị Kinh Doanh* 15 (3): 22–40.
- [17]. Kosmidou, Kyriaki, and Constantin Zopounidis. (2008). Measurement of Bank Performance in Greece. *South-Eastern Europe Journal of Economics* 1 (1): 79–95.
- [18]. Le Duong Quang Tu, Ho Huu Tin, Ngo Thanh, Nguyen Tien Dat, and Tran Hong Son. (2022). A Dataset for the Vietnamese Banking System (2002–2021). *Data* 7 (9): 120.
- [19]. Ledhem, Mohammed Ayoub. (2020). Data Mining Techniques for Predicting the Financial Performance of Islamic Banking in Indonesia. *Journal of Modelling in Management* 17 (3): 896–915.
- [20]. Lin, Shih-Wei, Yeou-Ren Shiue, Shih-Chi Chen, and Hui-Miao Cheng. (2009). Applying Enhanced Data Mining Approaches in Predicting Bank Performance: A Case of Taiwanese Commercial Banks. *Expert Systems with Applications* 36 (9): 11543–51.
- [21]. Masini, Ricardo P, Marcelo C Medeiros, and Eduardo F Mendes. (2023). Machine Learning Advances for Time Series Forecasting. *Journal of Economic Surveys* 37 (1): 76–111.

- [22]. Maulud, Dastan, and Adnan M Abdulazeez. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends* 1 (4): 140–47.
- [23]. Nguyễn Thị Liên. (2018). Phương pháp học máy trong phát hiện gian lận thẻ tín dụng-một nghiên cứu thực nghiệm. *Tạp Chí Kinh Tế & Phát Triển*.
- [24]. Raviv, Limor, Gary Lupyan, and Shawn C Green. (2022). How Variability Shapes Learning and Generalization. *Trends in Cognitive Sciences* 26 (6): 462–83.
- [25]. Sealey Jr, Calvin W, and James T Lindley. 1977. “Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions.” *The Journal of Finance* 32 (4): 1251–66.
- [26]. Wei, Jun, Tao Ye, and Zhe Zhang. 2021. “A Machine Learning Approach to Evaluate the Performance of Rural Bank.” *Complexity* 2021: 1–10.

Thông tin tác giả:

Đỗ Quang Hưng

- Đơn vị công tác: Học viện Công nghệ Bưu chính viễn thông
- Địa chỉ email: dqhung@ptit.edu.vn; quanghung.fcu@gmail.com

Ngày nhận bài: 5/10/2023

Ngày nhận bản sửa: 9/11/2023

Ngày duyệt đăng: 25/12/2023