
DỰ BÁO HOẠT ĐỘNG NGÂN HÀNG BẰNG THUẬT TOÁN RỪNG NGẪU NHIÊN

Đỗ Quang Hưng

Học viện Công nghệ Bưu chính Viễn thông

Email: quanghung.fcu@gmail.com; dqhung@ptit.edu.vn

Mã bài báo: JED-1426

Ngày nhận: 04/10/2023

Ngày nhận bản sửa: 02/01/2024

Ngày duyệt đăng: 11/01/2024

Mã DOI: 10.33301/JED.VI.1426

Tóm tắt:

Ngành ngân hàng có vai trò quan trọng trong việc ổn định vĩ mô của nền kinh tế. 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 kỹ thuật trí tuệ nhân tạo thuật toán rừng ngẫu nhiên (Random Forest - RF). Để chứng minh tính hiệu quả của mô hình dự báo dựa trên RF, các mô hình dự báo khác được dựa trên ba kỹ thuật trí tuệ nhân tạo khác là mạng nơ ron truyền thẳng nhiều lớp (ANN-MLP), mạng hàm cơ sở bán kính (RBF) và hồi quy tuyến tính (MLR) cũng được phát triển. 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 và tổng tiền gửi huy động. 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.

Từ khóa: Trí tuệ nhân tạo, dự báo hoạt động ngân hàng, hồi quy đa biến, mạng nơron, thuật toán rừng ngẫu nhiên, RBF.

Mã JEL: G21, C53.

Prediction of bank performance using random forest algorithm

Abstract:

The banking industry is regarded as the backbone of a country's modern economy. Globalization and technical innovation, on the other hand, have produced a highly competitive market in the banking and financial industry. The industry's performance is heavily dependent on the accuracy of managerial judgments. This research aims to predict bank performance using the Random Forest algorithm. To prove the effectiveness of the proposed model, other prediction models based on artificial intelligence techniques, including multi-layer feedforward neural network (ANN-MLP), RBF (Radial Basis Function) network, and multiple linear regression, are also developed. 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.

Keywords: Artificial intelligence, bank performance prediction, multiple linear regression, neural network, random forest, RBF.

JEL code: G21, C53.

1. Giới thiệu

Ngân hàng được coi là ngành quan trọng của nền kinh tế quốc dân. Ngân hàng có nhiệm vụ kiểm soát lạm phát, ổn định kinh tế vĩ mô; bảo đảm cung cấp nguồn vốn tín dụng và hệ thống thanh toán cho nền kinh tế; bảo đảm an ninh, an toàn tài chính, tiền tệ quốc gia, góp phần giữ vững ổn định chính trị, trật tự an toàn xã hội; bảo vệ quyền và lợi ích hợp pháp của người dân và doanh nghiệp. Trong bối cảnh toàn cầu hóa, các rào cản liên quan đến cạnh tranh đã giảm đi một phần rất lớn, điều này đã tạo ra các 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. Các công cụ toán học và thống kê có thể 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 (Artificial intelligence - AI) 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, 2009). Các kỹ thuật này thường được sử dụng trong các vấn đề dự báo phức tạp và phi tuyến. Trong các kỹ thuật trí tuệ nhân tạo, thuật toán rừng ngẫu nhiên (Random forest - RF) là một trong những kỹ thuật cho độ chính xác dự báo cao và tránh được hiện tượng quá khớp (overfitting). Thuật toán RF đã được sử dụng trong một số lĩnh vực kỹ thuật và công nghệ như dự báo khả năng chịu tải, phân loại gene.

Trong nghiên cứu này, thuật toán RF được sử dụng để dự báo hoạt động của các ngân hàng tại Việt Nam. Ngoài ra để chứng minh tính hiệu quả của thuật toán RF, một số kỹ thuật trí tuệ khác bao gồm mạng nơ ron truyền thẳng ANN-MLP, mạng RBF và kỹ thuật hồi quy MLR cũng được sử dụng trong nghiên cứu. Nghiên cứu này 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 dựa trên kỹ thuật trí tuệ nhân tạo được phát triển: mô hình dựa trên thuật toán RF có độ chính xác dự báo tốt nhất.

2. Tổng quan nghiên cứu

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 & 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 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 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 & cộng sự (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 DT, kỹ thuật Random Forest và mạng nơ ron, kết quả cho thấy mô hình cây quyết định 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 & cộng sự, 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 nghiên cứu được

thu thập từ 45 ngân hàng thương mại hoạt động tại Việt Nam trong giai đoạn 2002-2022.

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

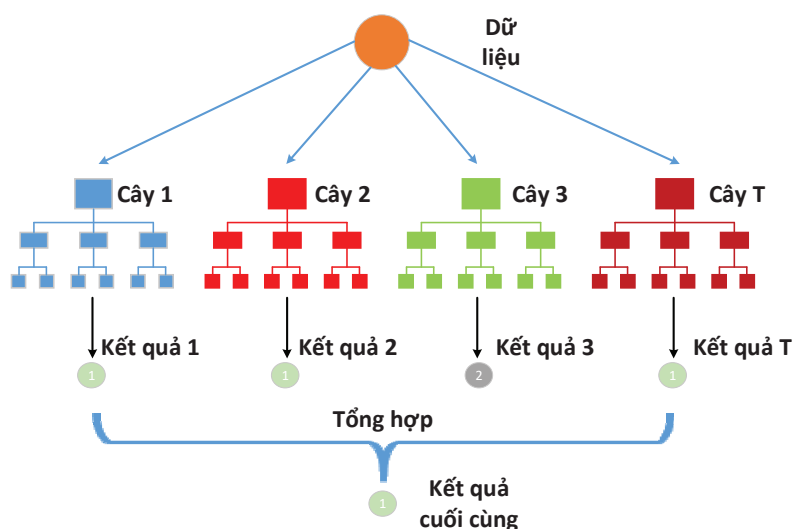
3.1. Thuật toán rừng ngẫu nhiên (Random forest - RF)

Thuật toán rừng ngẫu nhiên (Random forest - RF) là phương pháp phân lớp thuộc tính được phát triển bởi Breiman (2001). Dựa trên ý tưởng kết hợp nhiều mô hình phân loại/hồi quy thành tập hợp các mô hình phân loại/hồi quy để cho tính chính xác cao hơn so với chỉ một mô hình phân loại. RF được xây dựng dựa trên nền tảng thuật toán phân lớp Classification and Regression Trees (CART) sử dụng kỹ thuật có tên gọi là bagging (Hart & cộng sự, 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 1). 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 (m_{try}). Giá trị mặc định của tham số này là căn bậc hai của n với n 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 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.

Mỗi cây được tạo ra dựa trên một tập mẫu huấn luyện ngẫu nhiên được lấy ra từ tập mẫu huấn luyện ban đầu với cùng độ lớn theo nguyên tắc lấy mẫu có hoàn lại (phương pháp Bootstrap) có nghĩa là một mẫu có thể được lấy nhiều lần. Các cây ra quyết định dựa trên tập mẫu vừa tạo ra với nguyên tắc chỉ sử dụng một số lượng biến đầu vào tại mỗi nút phân chia. Kết quả cuối cùng là giá trị trung bình kết quả thu được từ tất cả các cây ra quyết định. Bằng cách sử dụng thật nhiều cây ra quyết định, sai số dự báo của mô hình sẽ được giảm. Cụ thể các bước như sau:

- Xác định m là số lượng cây sẽ được xây dựng.
- Xác định m là số lượng thuộc tính được sử dụng để phân chia tại mỗi nút của cây. m là tổng số các thuộc tính (n). được giữ không đổi trong suốt quá trình xây dựng cây.
- Xây dựng cây quyết định. Trong đó mỗi cây quyết định được xây dựng như sau: (1) xây dựng tập mẫu khởi động (bootstrap) với m mẫu từ việc hoán vị tập các mẫu ban đầu. Mỗi cây sẽ được dựng từ tập mẫu khởi động này; (2) Khi xây dựng cây quyết định, tại mỗi nút sẽ chọn ra m thuộc tính và m thuộc tính này được sử dụng để tìm ra cách phân chia tốt nhất; (3) Mỗi cây quyết định được phát triển lớn nhất có thể và không bị cắt xén.

Hình 1: Thuật toán RF



- Sau khi xây dựng được rừng ngẫu nhiên, để phân lớp cho đối tượng, thu thập kết quả phân lớp đối tượng này trên tất cả các cây quyết định và xác định kết quả cuối cùng (bằng phương pháp lấy bình quân giá trị dự báo hoặc “bỏ phiếu”) của các cây quyết định để làm kết quả cuối cùng của thuật toán. Tỷ lệ lỗi của cây phụ thuộc vào độ mạnh của từng cây quyết định thành phần và mối quan hệ giữa các cây đó.

Các thông số ảnh hưởng đến hiệu suất của thuật toán là: số lượng cây quyết định được xây dựng; kỹ thuật lấy mẫu (có sử dụng hay không sử dụng kỹ thuật bootstrap); số lượng biến được sử dụng tại mỗi nút; và đặc điểm của tập dữ liệu bao gồm thông số đầu vào và thông số đầu ra.

3.2. Mạng nơ-ron truyền thẳng nhiều lớp (Multi-layer Perceptron 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 ví dụ Hình 2. Mức độ liên kết giữa các đơn vị được xác định bởi một tập giá trị trọng số. Tham số bias (thiên vị) 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 vector lời giải.

Mạng nơ-ron nhân tạo gồm có một nhóm các nơ-ron 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 perceptron nhiều lớp (Multilayer perceptron - MLP), hay còn gọi là mạng truyền thẳng nhiều lớp, mở rộng của mô hình mạng perceptron, là mạng nơ-ron 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 nơ-ron truyền thẳng nhiều lớp với một lớp ẩn có thể xấp xỉ hóa tất cả các hàm số liên tục (Cuomo, 2022; Raviv & cộng sự, 2022), do đó được ứng dụng trong rất nhiều lĩnh vực (Masini & cộng sự, 2023). Hình 2 là một mạng nơ-ron truyền thẳng nhiều lớp gồm 3 lớp. Với x , h , và y là số lượng nút vào, nút ẩn và nút ra; w và b là các trọng số của nút vào và nút ẩn; w và b là các véc tơ độ lệch bias của lớp ẩn và lớp ra; x là véc tơ các đầu vào; h là các véc tơ đầu ra của lớp ẩn; và y là véc tơ đầu ra. Mạng nơ-ron trong Hình 2 được trình bày thông qua công thức sau:

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

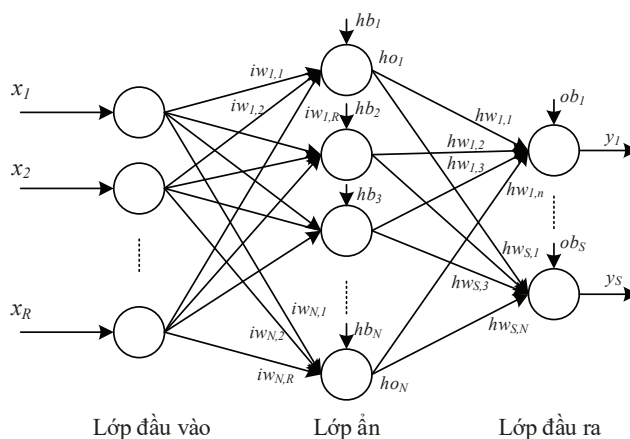
$$y_i = f\left(\sum_{k=1}^N hw_{i,k}.ho_k + ob_i\right), \text{ với } i = 1, \dots, S \quad (2)$$

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

Khi xây dựng một mô hình mạng nơ-ron, 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 (overfitting), có thể dẫn đến việc dự đoán nhầm lẫn, và chất lượng mô hình không còn tốt trên dữ liệu kiểm tra (Caruana & cộng sự, 2001).

Chức năng của một mạng nơ-ron 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

Hình 2: Mạng nơ-ron truyền thẳng MLP ba lớp



bởi các thuật toán huấn luyện. Quá trình điều chỉnh các trọng số để mạng “nhận biết” được quan hệ giữa đầu vào và đích mong muốn được gọi là học hay huấn luyện. 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 và học không có giám sát.

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. 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 hay hàm chi phí (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 véc-tơ trọng số và độ lệch. Giả thiết là có m cặp đầu vào và đầu ra mong muốn, x_k, t_k với $k=1,2,...,m$. 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:

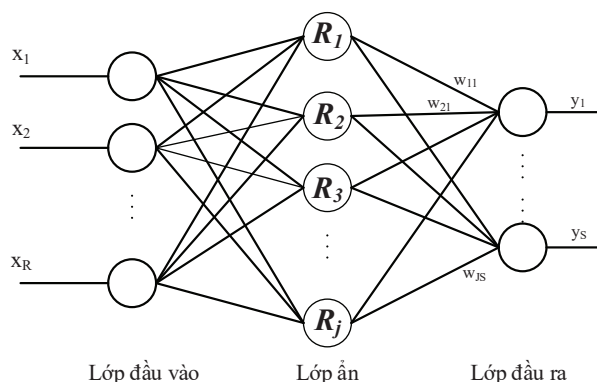
$$MSE = \frac{1}{m} \sum_{k=1}^m e_k^2 = \frac{1}{m} \sum_{k=1}^m (t_k - y_k)^2 \quad (3)$$

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

3.3. Mạng hàm cơ sở bán kính RBF (Radial Basis Function)

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 3. Mặc dù kiến trúc của mạng (RBF) tương đối đơn giản, nhưng có khả năng tổng hợp cao (Jiang & cộng sự, 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 (Batoool & cộng sự, 2013; Guan & cộng sự, 2016).

Hình 3: 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 y_s là số lượng đầu ra, R_j là số neural trong lớp ẩn và w_{js} là trọng số liên kết giữa nốt 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 véc tơ 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 Euclidean 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.

3.4. Hồi quy tuyến tính đa biến (Multiple Linear Regression - MLR)

Hồi quy tuyến tính đa biến (Multiple linear regression - MLR) là một trong 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 hiệu quả đối với phần lớn các bài toán. Hồi quy tuyến tính là một mô hình tuyến tính, ví dụ: một mô hình trong đó giả định mối quan hệ tuyến tính giữa các biến đầu vào () và biến đầu ra duy nhất (). Nói cách khác, có thể được tính toán từ sự kết hợp tuyến tính của các biến đầu vào (). Khi có một biến đầu vào duy nhất (), 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 đa 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 & Abdulazeez, 2020).

4. Phương pháp nghiên cứu

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 (Kosmidou & Zopounidis, 2008; Bod'a & Piklová, 2018; Appiahene & cộng sự, 2020; Nguyễn Minh Kiều & Nguyễn Ngọc Thùy Trang, 2020; Wei & cộng sự, 2021). 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 & Lindley, 1977). Một trong những sự khác biệt trong các cách tiếp cận là bởi biến tiền gửi huy động có cả đặc điểm của biến đầu vào và đầu ra. Về cơ bản, có hai cách tiếp cận chính như sau:

Coi ngân hàng là một tổ chức trung gian tài chính 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, tài sản cố định, tiền gửi huy động. Đầu ra là lợi nhuận.

Coi ngân hàng là một tổ chức cung cấp các dịch vụ và sản phẩm tài chính 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, số lao động, số chi nhánh. Đầu ra là những chỉ số liên quan đến dịch vụ cung cấp cho khách hàng là tổng tiền gửi huy động và tổng các khoản vay. 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.

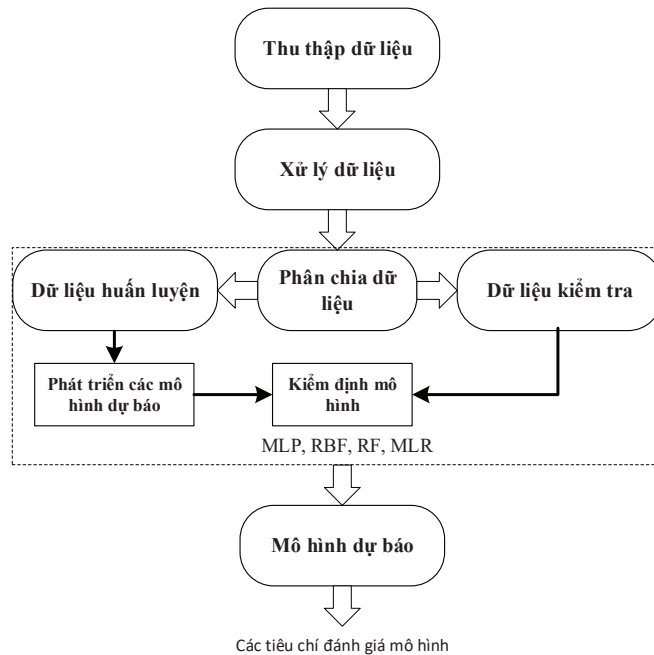
4.2. Dữ liệu

Dữ liệu sử dụng trong nghiên cứu được thu thập từ 45 ngân hàng Việt Nam trong giai đoạn 2002-2022 (Le & cộng sự, 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	Giá trị nhỏ nhất	Giá trị lớn nhất	Giá trị trung bình	Độ lệch chuẩn
Tài sản cố định (Đơn vị: tỷ đồng)	769,00	11.114.537,00	1.742.134,40	2.503.932,12
Số lao động	45,00	39.950,00	6.906,97	8.593,60
Số chi nhánh	2,00	2.400,00	268,83	457,02
Tổng tiền gửi huy động (Đơn vị: tỷ đồng)	7.314,00	1.404.875.777,00	134.050.816,11	233.066.086,72
Tổng các khoản vay (Đơn vị: tỷ đồng)	50.721,00	1.345.632.643,00	122.985.141,52	218.774.472,25

4.3. Xây dựng các mô hình dự báo

Hình 4: Các bước xây dựng mô hình

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, mã hóa các biến nhóm, chuẩn hóa dữ liệu,...

Bước 2: Phân chia dữ liệu

Bước này chuẩn bị dữ liệu để xây dựng mô hình. Dữ liệu được chia thành hai phần: dữ liệu dùng để phục vụ cho huấn luyện, phát triển mô hình; dữ liệu dùng để kiểm tra mô hình. Trong nghiên cứu này, tác giả sử dụng 2/3 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 và gán nhãn cho dữ liệu, thường được gọi là học hay huấn luyện. Trong đó: là các dữ liệu đầu vào, là đầu ra của dự báo. Các kỹ thuật học có giám sát ANN-MLP, RBF, RF, MLR đã được sử dụng trong nghiên cứu.

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 tiêu đánh giá trên tập dữ liệu kiểm tra. Nếu không đạt được kết quả mong muốn thì các tham số 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 sẽ chọn ra được mô hình dự báo tốt nhất.

4.3.1. 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 3 units, hàm kích hoạt là gradient liên hiệp được sử dụng để tăng tốc độ tính toán. Hàm chi phí là sai số toàn phương trung bình (MSE- Mean squared error). Đối với mô hình RBF, số lượng hàm Gaussian là 2, hàm kích hoạt là gradient liên hiệp. Đối với mô hình dựa trên thuật toán RF, số lượng cây xây dựng là 100 và không giới hạn độ sâu của cây. Đối với dự báo tổng các khoản vay, mô hình MLR có công thức như sau:

$$\text{Tổng các khoản vay} = 2.899,9008 \times \text{Số lao động} - 21.723,9927 \times \text{Số chi nhánh} + 2,0568 \times \text{Tài sản cố định} + 0,8487 \times \text{Tổng tiền gửi huy động} - 8.561.646,342$$

Đối với dự báo tổng tiền gửi huy động, mô hình MLR có công thức như sau:

$$\text{Tổng tiền gửi huy động} = -2.193,5462 \times \text{Số lao động} + 21.709,9094 \times \text{Số chi nhánh} + 4,161 \times \text{Tài sản cố định} + 1,058 \times \text{Tổng các khoản vay} + 6.000.922,381$$

4.3.2. 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 $\varepsilon_i = t_k - y_i$ đều nhau thì nên chọn MSE để đánh giá. Ngược lại, nếu giá trị sai số ε_i 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:

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

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 và thảo luận

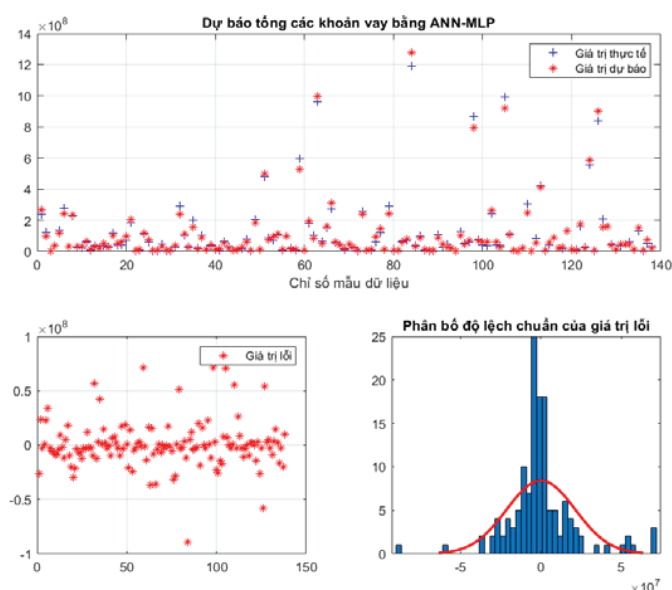
Các chỉ số đánh giá mô hình được trình bày trong Bảng 2. Các giá trị $RMSE$, $MAPE$, MAE , R và $Theil's U$ của mô hình dựa trên kỹ thuật RF tương ứng là $2,0276 \times 10^7$; 1,4467; $1,4044 \times 10^7$; 0,9948 và 0,0447 đối với dự báo tổng các khoản vay; $2,9291 \times 10^7$; 4,2898; $1,9158 \times 10^7$; 0,9932 và 0,0585 đối dự báo tổng tiền gửi huy động. 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). 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
Tổng các khoản vay	ANN-MLP	2,1110	1,6432	1,4185	0,9943	0,0464
	RF	2,0276	1,4467	1,4044	0,9948	0,0447
	RBF	2,3602	1,5406	1,4713	0,9931	0,0517
	MLR	2,7020	1,6402	1,6856	0,9947	0,0574
Tổng tiền gửi huy động	ANN-MLP	4,2323	4,5217	1,9736	0,9871	0,0862
	RF	2,9291	4,2898	1,9158	0,9932	0,0585
	RBF	3,9472	5,2293	2,3105	0,9897	0,0804
	MLR	3,3857	5,1542	1,9081	0,9931	0,0686

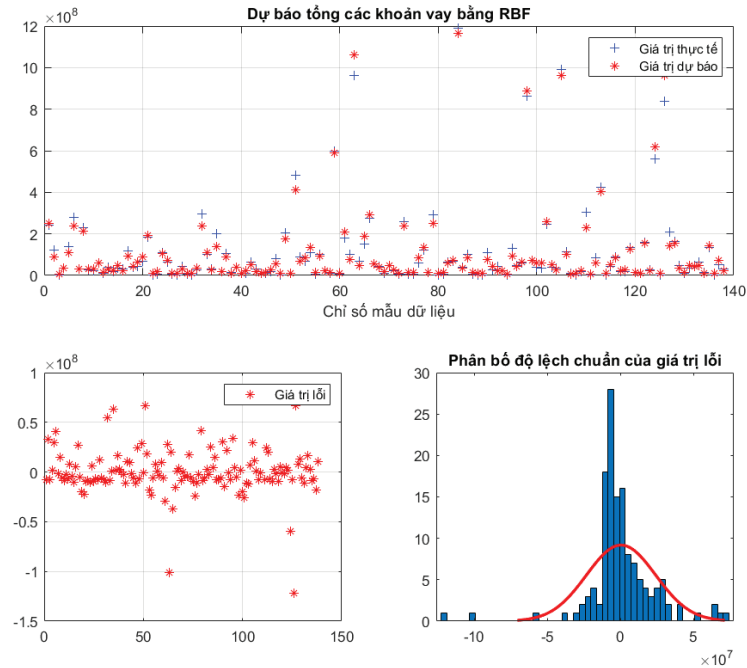
Các Hình 5 đến Hình 12 thể hiện mức độ phù hợp giữa giá trị thực tế và giá trị dự báo bởi các mô hình dựa trên kỹ thuật trí tuệ nhân tạo đã phát triển cho việc dự báo tổng các khoản vay và tổng tiền gửi huy động. Trong mỗi hình, thể hiện giá trị thực tế và giá trị dự báo được, giá trị lỗi và phân bố độ lệch chuẩn của giá trị lỗi.

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

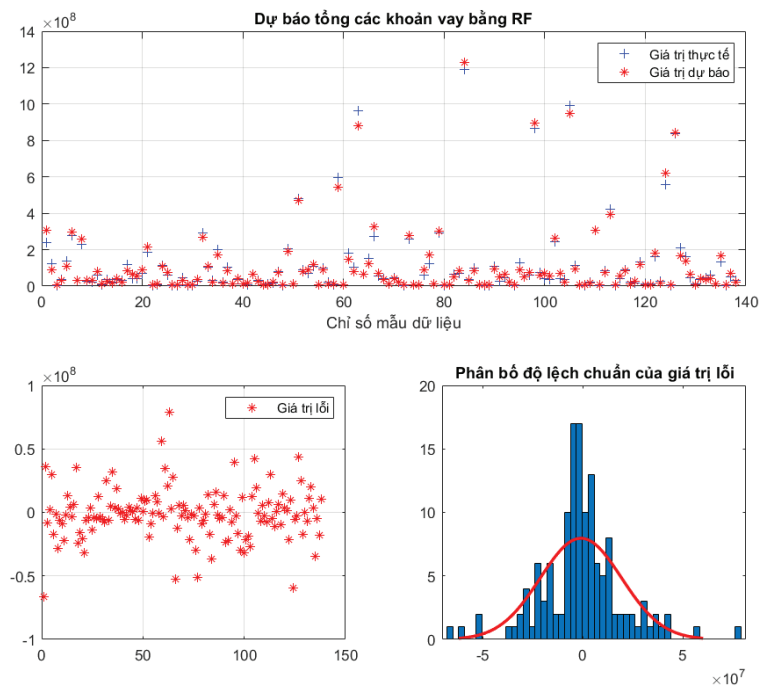


Việc so sánh các giá trị thực tế và giá trị dự báo của các mô hình cũng được thể hiện qua biểu đồ phân tán trong Hình 13 và Hình 14. Biểu đồ phân tán thể hiện mối tương quan giữa hai tập giá trị thực tế và giá trị dự báo bởi các mô hình. Trong hình vẽ, đường thẳng 1:1 thể hiện giá trị thực tế và giá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

Hình 6: Dự báo tổng các khoản vay bằng RBF



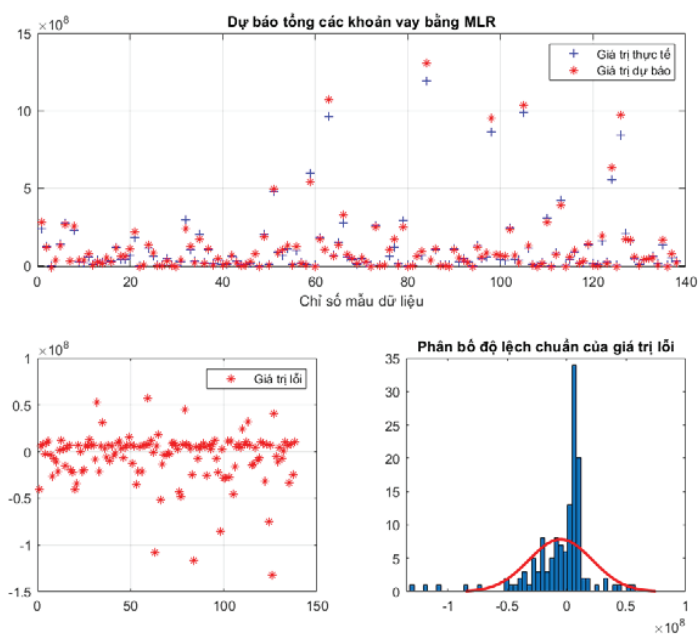
Hình 7: Dự báo tổng các khoản vay bằng RF



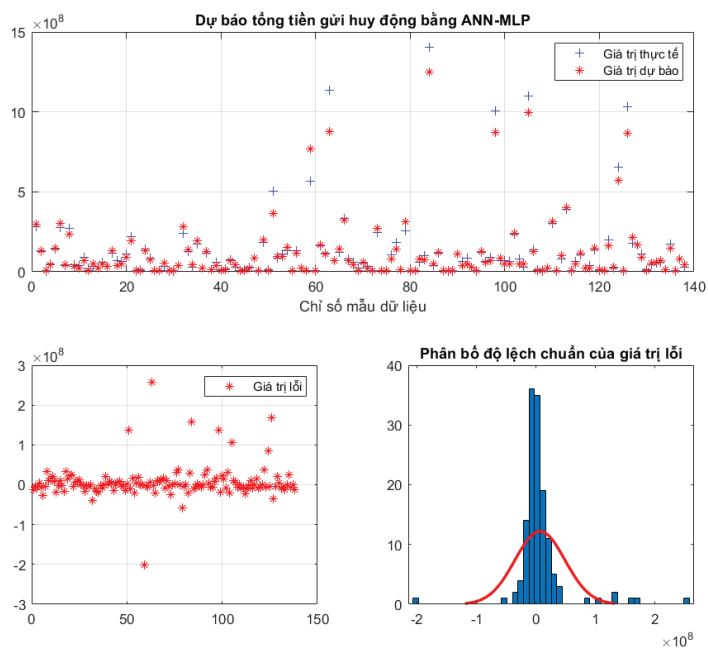
thực tế. Quan sát Hình 13 và Hình 14 nhận thấy giá trị dự báo bởi mô hình RF và giá trị thực tế tương đối khớp nhau do các điểm tập trung gần nhau và có thể vẽ được một đường thẳng đi qua các điểm này. Điều này thể hiện mối tương quan là rất mạnh.

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

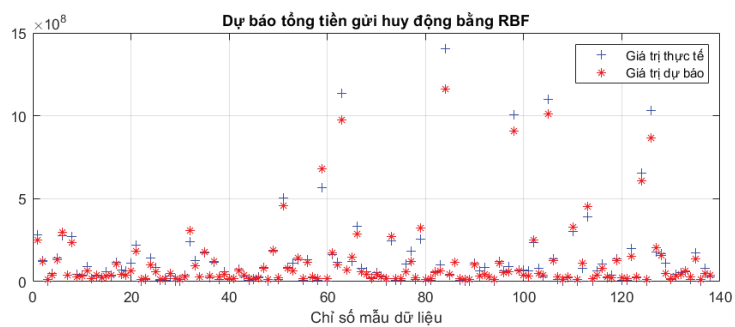
Hình 8: Dự báo tổng các khoản vay bằng MLR

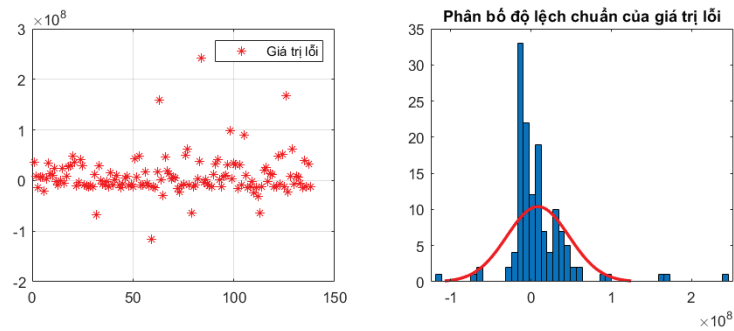


Hình 9: Dự báo tổng tiền gửi huy động bằng ANN-MLP

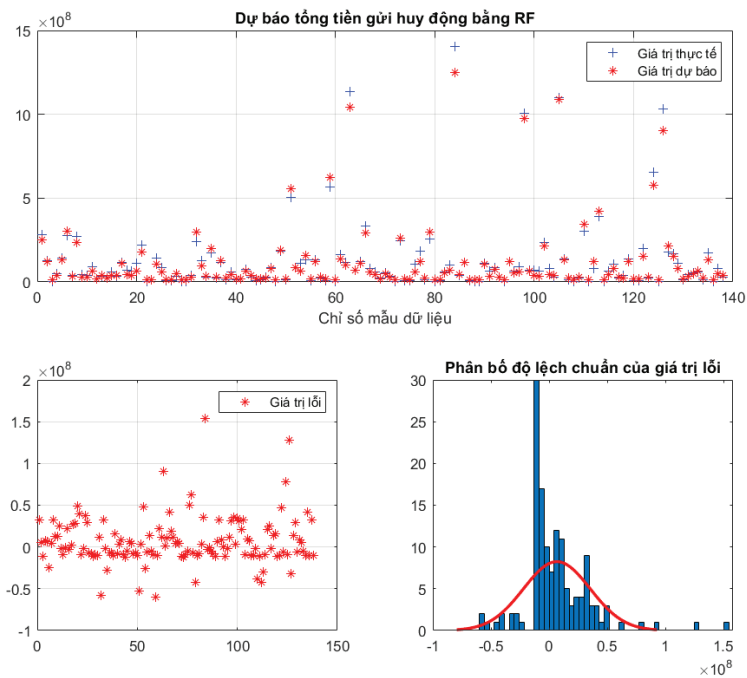


Hình 10: Dự báo tổng tiền gửi huy động bằng RBF

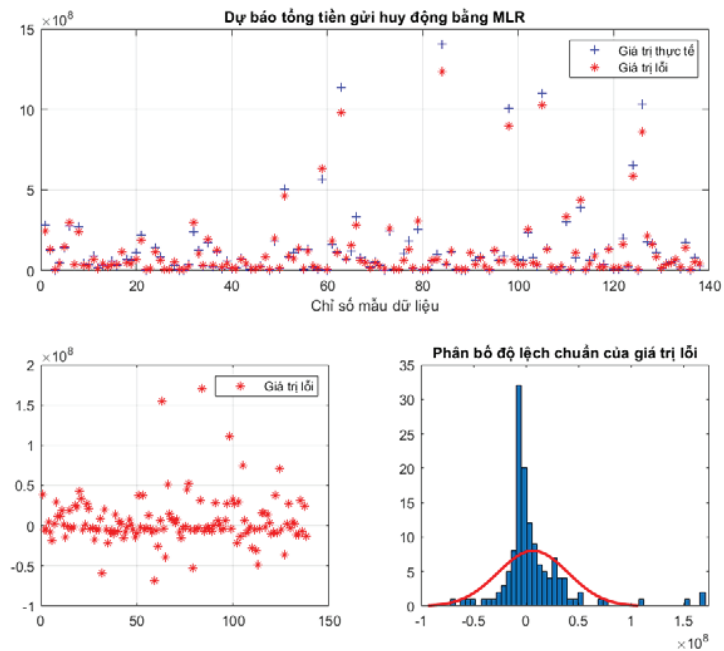




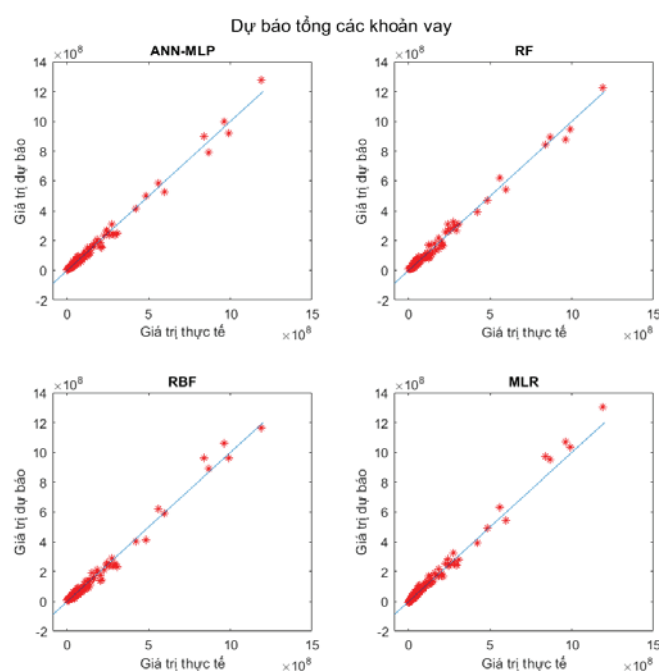
Hình 11: Dự báo tổng tiền gửi huy động bằng RF



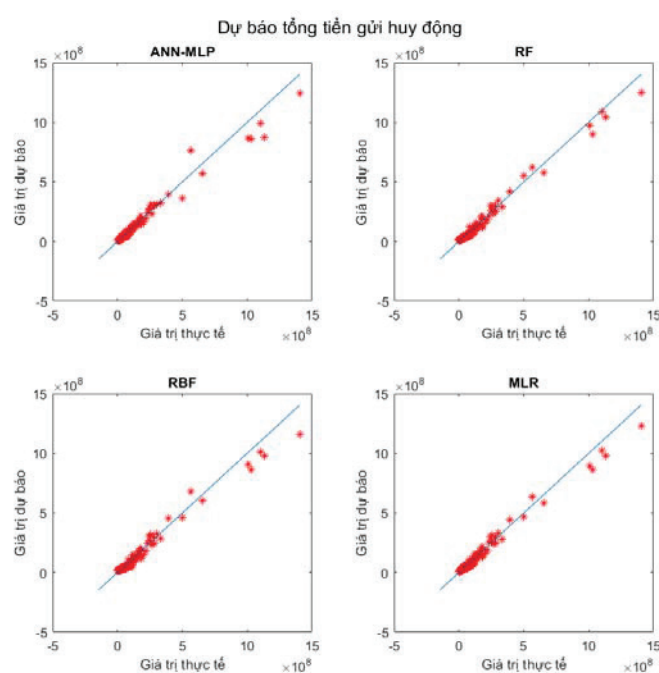
Hình 12: Dự báo tổng tiền gửi huy động bằng MLR



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



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



6. Kết luận

Dự báo hoạt động 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 và tổng tiền gửi huy động chính dựa trên các thuật toán rừng ngẫu nhiên RF. Một số kỹ thuật trí tuệ nhân tạo khác bao gồm ANN-MLP, RBF và hồi quy tuyến tính MLR cũng được sử dụng. Kết quả thực nghiệm dựa trên các chỉ tiêu đánh giá đã khẳng định rằng mô hình dự báo dựa trên thuật toán RF 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, tư vấn 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 điều chỉnh và lựa chọn các tham số cũng cần được xem xét.

Tài liệu tham khảo

- Appiahene, P., Missah, Y.M. & Najim, U. (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.
- Assous, H.F. (2022), 'Prediction of banks efficiency using feature selection method: comparison between selected machine learning models', *Complexity*, 2022, 1-15.
- Batool, F. (2013), 'Gamma radiations induced improvement in dyeing properties and colorfastness of cotton fabrics dyed with chicken gizzard leaves extracts', *Radiation Physics and Chemistry*, 89, 33-37.
- Bod'a, M. & Piklová, Z. (2018), 'The production or intermediation approach?: It matters', in *Contemporary Trends and Challenges in Finance: Proceedings from the 3rd Wroclaw International Conference in Finance*, Springer, 111-120.
- Breiman, L. (2001), 'Random forests', *Machine learning*, 45, 5-32.
- Caruana, R., Lawrence, S. & Giles, L. (2001), 'Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping', in *Advances in Neural Information Processing Systems*, Denver, CO, USA.
- Cuomo, S. (2022), 'Scientific machine learning through physics-informed neural networks: Where we are and what's next', *Journal of Scientific Computing*, 92(3), p.88.
- Đào Công Ân (2018), 'Mô hình dự báo churn cho khách hàng bằng phương pháp học máy suy diễn phương sai', luận văn thạc sĩ, Trường Đại học Bách khoa Hà Nội.
- 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.
- 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.
- Guan, X., Zhu, Y. & Song, W. (2016), 'Application of RBF neural network improved by peak density function in intelligent color matching of wood dyeing', *Chaos, Solitons and Fractals*, 89, 485-490.
- Hao, C. & Adsavakulchai, S. (2023), 'The use of machine learning algorithms for bank loan prediction', *European Economic Letters (EEL)*, 13(3), 735-741.
- Hart, P.E., Stork, D.G. & Duda, R.O. (2000), *Pattern classification*, Wiley Hoboken.
- Jiang, J., Cao, D. & Chen, H. (2016), 'Boundary value problems for fractional differential equation with causal operators', *Applied Mathematics and Nonlinear Sciences*, 1(1), 11-22.
- Kosmidou, K. & Zopounidis, C. (2008), 'Measurement of bank performance in Greece', *South-Eastern Europe Journal of Economics*, 1(1), 79-95.
- Le, T.D.Q., Tin, H.H., Ngo, T., Nguyen, D.T. & Tran, S.H. (2022), 'A dataset for the Vietnamese banking system (2002-2021)', *Data*, 7(9), p.120.
- Ledhem, M.A. (2022), 'Data mining techniques for predicting the financial performance of Islamic banking in

-
- Indonesia', *Journal of Modelling in Management*, 17(3), 896-915.
- Lin, S.W. (2009), 'Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks', *Expert Systems with Applications*, 36(9), 1543-11551.
- Masini, R.P., Medeiros, M.C. & Mendes, E.F. (2023), 'Machine learning advances for time series forecasting', *Journal of economic surveys*, 37(1), 76-111.
- Maulud, D. & Abdulazeez, A.M. (2020), 'A review on linear regression comprehensive in machine learning', *Journal of Applied Science and Technology Trends*, 1(4), 140-147.
- Nguyễn Minh Kiều & 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.
- Nguyễn Thị Liên, Nguyễn Thị Thu Trang & Nguyễn Chiến Thắng (2018), 'Phương pháp học máy trong phát hiện gian lận thẻ tín dụng-một nghiên cứu thực nghiệm', *Tạp Chí Kinh Tế & Phát Triển*, 256(II), 118-126.
- Raviv, L., Lupyan, G. & Green, S.C. (2022), 'How variability shapes learning and generalization', *Trends in cognitive sciences*, 26(6), 462-483.
- Sealey Jr, C.W. & Lindley, J.T. (1977), 'Inputs, outputs, and a theory of production and cost at depository financial institutions', *The journal of finance*, 32(4), 1251-1266.
- Wei, J., Ye, T. & Zhang, Z. (2021), 'A machine learning approach to evaluate the performance of rural bank', *Complexity*, 2021, 1-10.