Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 34, Số 8 (2023), 108–122

www.jabes.ueh.edu.vn

JABES
Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
http://www.emeraldgrouppublishing.com/services/publishing/jabes/

Nghiên cứu mô hình học máy dự đoán xác suất vỡ nợ của khách hàng doanh nghiệp tại ngân hàng thương mại cổ phần ở Việt Nam

NGUYỄN QUỐC HÙNG [a,*], QUAN TOẠI MẪN [b], TRƯƠNG THỊ MINH LÝ [a]

[a] *Trường Đại học Kinh tế TP. Hồ Chí Minh*

[b] *Ngân hàng Thương mại cổ phần Á Châu*

THÔNG TIN

Ngày nhận: 01/06/2023
Ngày nhận lại: 06/08/2023
Duyệt đăng: 07/08/2023

Mã phân loại JEL:
C61; C63; C67.

Từ khóa:
Dự đoán vỡ nợ;
Decision Tree;
Random Forest;
XGBoost, CatBoost.

Keywords:
Loan Default Prediction;
Decision Tree;
Random Forest;
XGBoost;
CatBoost.

TÓM TẮT

Dự đoán khả năng vỡ nợ của khách hàng trong tương lai sử dụng các công nghệ hiện đại đang là xu hướng phát triển trong các tổ chức tài chính nói chung và ngân hàng nói riêng. Điều này là cần thiết để các tổ chức tài chính có các hướng xử lý kịp thời như giảm thiểu rủi ro tín dụng, phân tích quy trình tín dụng, và tối ưu hóa danh mục tín dụng. Bài báo sử dụng các dữ liệu liên quan đến thông tin tín dụng, tài chính và đặc điểm của khách hàng doanh nghiệp tại ngân hàng, thông qua phương pháp nghiên cứu định lượng, thu thập, xử lý dữ liệu để xây dựng mô hình máy học. Kết quả cho thấy mô hình XGBoost đạt có độ chính xác cao nhất với chỉ số F1-score đạt 0,84, đường cong ROC có diện tích AUC đạt 0,97 Qua đó, ngân hàng có thể áp dụng mô hình vào thực tế để hỗ trợ quyết định kinh doanh, giúp tăng cường khả năng dự báo rủi ro tín dụng, nâng cao hiệu suất hoạt động của ngân hàng và giảm thiểu các tổn thất không mong muốn.

Abstract

Predicting the probability of customer default in the future using modern technologies is a growing trend in the overall financial industry, specifically in the banking sector. It is necessary for financial institutions to have timely solutions such as credit risk reduction, credit process analysis, and credit portfolio optimization. The paper utilizes data related to credit information, financial indicators, and corporate customer characteristics from the bank, employing quantitative

* Tác giả liên hệ.
Email: *hungngq@ueh.edu.vn* (Nguyễn Quốc Hùng), *toaiman.quan@gmail.com* (Quan Toại Mẫn), *truongly@ueh.edu.vn* (Trương Thị Minh Lý).
Trích dẫn bài viết: Nguyễn Quốc Hùng, Quan Toại Mẫn & Trương Thị Minh Lý. (2023). Nghiên cứu mô hình học máy dự đoán xác suất vỡ nợ của khách hàng doanh nghiệp tại ngân hàng thương mại cổ phần ở Việt Nam. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 34*(8), 108–122.

research methods to collect and process data to build a Machine-Learning model. The result shows that the XGBoost model achieved the highest accuracy with an F1-score of 0.84, and the ROC curve had an AUC of 0.97. By utilizing these findings, the bank can implement the model into practice to facilitate business decision-making, enhance credit risk forecasting capabilities, improve operational efficiency, and mitigate undesirable losses.

1. Giới thiệu

Ngân hàng có nhiều hoạt động kinh doanh và đóng vai trò quan trọng trong nền kinh tế. Trong đó, hoạt động chủ yếu là cho vay khách hàng, cũng chính là hoạt động đem lại nguồn lợi nhuận lớn nhất cho ngân hàng. Đối với hầu hết các ngân hàng, các khoản vay chiếm một nửa hoặc nhiều hơn trong số tổng tài sản và khoảng một nửa đến hai phần ba doanh thu của họ. Hơn nữa, rủi ro trong hoạt động ngân hàng có xu hướng tập trung vào danh mục cho vay. Các khoản vay không thể thu hồi có thể gây ra các vấn đề tài chính nghiêm trọng cho ngân hàng và nền kinh tế (Rose & Hudgins, 2015). Do đó, việc nghiên cứu xây dựng mô hình để ước tính xác suất vỡ nợ của khách hàng là cần thiết để ngân hàng có thể nhận diện được các khách hàng có khả năng vỡ nợ trong tương lai nhằm hạn chế rủi ro phát sinh. Bên cạnh đó, ngân hàng có thể dựa trên xác suất vỡ nợ để chấm điểm tín dụng, đưa ra các quyết định về định hướng cấp tín dụng, chính sách lãi suất cho vay phù hợp, và dần hướng tới việc tự động hóa quy trình cho vay. Mô hình chấm điểm tín dụng là công cụ chấm điểm để đánh giá người xin cấp tín dụng. Điểm tín dụng có được từ quy trình đánh giá sử dụng hồ sơ cấp tín dụng để dự đoán xác suất mà người nộp đơn sẽ vỡ nợ hoặc khoản vay trở nên quá hạn. Hồ sơ cấp tín dụng được đại diện bởi một tập hợp các biến liên quan đến người xin cấp tín dụng đó, ví dụ như tuổi, các khoản thanh toán lịch sử, các khoản bảo đảm có thể được được sử dụng để xây dựng hồ sơ của người xin cấp tín dụng. Khi đó chấm điểm tín dụng là quá trình đánh giá hồ sơ của ứng viên và quyết định chấp nhận hay từ chối tín dụng của người nộp hồ sơ vay (Gonzales và cộng sự, 2004). Hiện nay, các ngân hàng Việt Nam đa phần vẫn áp dụng theo cách cho vay truyền thống, đánh giá dựa trên thu nhập, tài sản đảm bảo, lịch sử tín dụng của người đi vay với nhiều quy trình và nhiều nguồn lực từ các phòng ban, mà chưa ứng dụng các mô hình máy học để tự động hóa, hỗ trợ ra quyết định cấp tín dụng. Khi đó, ngân hàng tốn nhiều thời gian hơn để đánh giá hồ sơ vay, không đáp ứng kịp thời nhu cầu vay vốn của khách hàng, ảnh hưởng năng lực cạnh tranh với các ngân hàng khác. Bên cạnh đó, việc đánh giá chủ quan có thể gặp rủi ro sơ sót, đánh giá không đúng về người đi vay, dẫn đến việc mất khách hàng tiềm năng hoặc gây ra nợ xấu, ảnh hưởng đến tỷ lệ an toàn vốn và lợi nhuận của ngân hàng. Ngoài ra, việc ước tính được xác suất vỡ nợ của khách hàng cũng giúp đáp ứng được các quy định của Chuẩn mực báo cáo tài chính quốc tế (International Financial Reporting Standards – IFRS), trong đó yêu cầu các ngân hàng đánh giá và quản lý rủi ro tín dụng toàn diện hơn.

Nghiên cứu này tập trung vào việc dự đoán khả năng vỡ nợ của khách hàng doanh nghiệp (KHDN), có ảnh hưởng lớn đến hoạt động kinh doanh và tài chính của các tổ chức tài chính, đặc biệt là ngân hàng. Trong bối cảnh phát triển của các công nghệ hiện đại, việc sử dụng học máy để dự đoán các sự kiện trong tương lai đang ngày càng trở nên phổ biến trong ngành tài chính. Do đó nghiên cứu này đặt mục tiêu xây dựng một mô hình học máy hiệu quả để dự đoán khả năng vỡ nợ của KHDN, từ đó

giúp ngân hàng nắm bắt và quản lý rủi ro tín dụng một cách hiệu quả. Bằng cách áp dụng phương pháp nghiên cứu định lượng và sử dụng công nghệ máy học, nghiên cứu đang xây dựng một mô hình dự đoán đáng tin cậy về khả năng vỡ nợ của KHDN. Kết quả từ nghiên cứu này hứa hẹn mang lại nhiều lợi ích trong việc quản lý rủi ro tín dụng cho ngân hàng. Qua đó, mô hình học máy được đề xuất trong nghiên cứu này có thể được triển khai rộng rãi trong ngành ngân hàng, giúp cải thiện hiệu quả hoạt động, giảm thiểu tổn thất không mong muốn và nâng cao khả năng dự báo, giảm thiểu rủi ro tín dụng.

2. Các nghiên cứu liên quan

Việc xem xét các nghiên cứu trước đây về chủ đề tương tự là quan trọng để làm rõ tình hình nghiên cứu và xác định vấn đề cần giải quyết. Thông qua việc xác định các kiến thức có sẵn và điểm mạnh, điểm yếu của những nghiên cứu trước sẽ giúp xây dựng cơ sở vững chắc và đưa ra đóng góp mới cho lĩnh vực nghiên cứu.

Nikolic và cộng sự (2013) đã sử dụng mô hình hồi quy Logistic để phát triển mô hình dự đoán chấm điểm tín dụng cho các đơn vị doanh nghiệp. Mô hình hóa dựa trên dữ liệu từ các báo cáo tài chính cuối năm trong 5 năm của các đơn vị doanh nghiệp Serbia với 350 biến tỷ lệ tài chính dữ liệu sự kiện dự báo vỡ nợ cho 7.590 doanh nghiệp, trong đó nhiều tỷ lệ tài chính được cho mới và chưa được đề cập trước đây. Phương pháp trọng số bằng chứng (Weight of Evidence – WOE) đã được áp dụng để chuyển đổi các tỷ lệ tài chính cho phù hợp với mô hình. Tiếp theo sử dụng phương pháp phân cụm để giảm danh sách các biến và loại bỏ các tỷ lệ tài chính có tương quan cao khỏi các bộ dữ liệu. Kết quả số lượng 350 biến giảm xuống một danh sách ngắn gồm 24 biến, được dùng để dự đoán sự kiện vỡ nợ của các đơn vị doanh nghiệp Serbia. Wang và cộng sự (2018) đã đề xuất mô hình chấm điểm hành vi (Ensemble Mixture Random Forest – EMRF) để ước tính xác suất vỡ nợ theo thời gian trong hoạt động cho vay ngang hàng Peer-to-Peer (P2P). Trong đó gồm hai mô hình, thứ nhất là mô hình Random Forest được sử dụng để xác định liệu người vay có vỡ nợ hay không, và thứ hai là mô hình Random Survival Forest xác định thời điểm mà người vay vỡ nợ. Kết quả phân tích thực nghiệm trên tập dữ liệu cho vay ngang hàng P2P tại Trung Quốc cho thấy mô hình EMRF có hiệu suất tốt trong việc dự đoán xác suất vỡ nợ biến động hàng tháng so với một số mô hình khác và cung cấp kết quả có ý nghĩa cho quản lý rủi ro sau cho vay đúng thời điểm. Guegan và cộng sự (2018) xây dựng các bộ phân loại nhị phân dựa trên mô hình máy học cơ bản và mô hình học sâu dựa trên mạng nơ-ron nhân tạo đa tầng nhằm dự đoán xác suất vỡ nợ cho vay trong tổ chức tín dụng. Dữ liệu được lựa chọn 10 đặc trưng quan trọng nhất từ các mô hình này, sử dụng chúng trong quá trình mô hình hóa để kiểm tra tính ổn định bằng cách so sánh hiệu suất của chúng trên các dữ liệu riêng biệt. Kết quả đánh giá cho thấy các bộ phân loại nhị phân ổn định hơn so với các mô hình dựa trên mạng nơ-ron nhân tạo đa tầng. Điều này mở ra một số hướng nghiên cứu mới liên quan đến việc sử dụng các hệ thống học sâu trong bài toán dự báo trong doanh nghiệp. Boughaci và cộng sự (2021) đã đề xuất một phương pháp kết hợp giữa kỹ thuật phân khúc và rừng ngẫu nhiên, với mục đích là thiết kế một mô hình xếp hạng tín dụng hiệu quả và nâng cao khả năng dự đoán phá sản tài chính. Phân khúc được sử dụng để chia dữ liệu thành các nhóm tương đối đồng nhất của các ứng viên vay vốn, phân loại các ứng viên đáng tin cậy và không đáng tin cậy. Trong giai đoạn phân khúc, tác giả sử dụng thuật toán phân cụm k-means (k-means clustering) tìm ra các cụm sao cho tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm của cụm gần nhất là nhỏ nhất để phân đoạn các ứng viên tương tự thành

các nhóm. Sau đó, tác giả áp dụng kỹ thuật học rừng ngẫu nhiên trên dữ liệu đã phân khúc. Các nghiên cứu thực nghiệm được thực hiện trên sáu tập dữ liệu tài chính nổi tiếng của các kích thước khác nhau. Kết quả nghiên cứ khá khả quan và cho thấy tính hiệu quả của kỹ thuật đề xuất cho phân khúc ứng viên. Khi phân khúc được sử dụng kết hợp với phân loại, phương pháp kết quả đã cải thiện đáng kể hiệu suất phân loại. Ngoài ra, mô hình dự đoán khả năng vỡ nợ hai giai đoạn, kết hợp phân cụm k-means và mô hình mô tả miền véc tơ hỗ trợ (Support Vector Domain Description – SVDD) để dự đoán khả năng vỡ nợ cũng đã được đề xuất trong nghiên cứu của Yuan và cộng sự (2021). Mô hình này sẽ sử dụng dữ liệu thuộc tính của doanh nghiệp ở thời điểm trước đó và tình trạng vỡ nợ tại thời điểm hiện tại để đào tạo mô hình. Kết quả nghiên cứu cho thấy độ chính xác dự đoán của mô hình hai giai đoạn này tốt hơn so với các mô hình một giai đoạn chỉ sử dụng phân cụm k-means hoặc SVDD. Nghiên cứu cho thấy mô hình có khả năng dự đoán vỡ nợ trong năm năm tiếp theo (AUC> 0,85) và cũng chỉ ra rằng "lợi nhuận giữ lại/tổng tài sản", "chi phí tài chính/doanh thu gộp" và "loại ý kiến kiểm toán" là ba đặc điểm chính trong dự báo khả năng vỡ nợ cho các doanh nghiệp niêm yết tại Trung Quốc. Nghiên cứu này đóng góp cho lĩnh vực nghiên cứu dự đoán khả năng vỡ nợ đa giai đoạn bằng cách chứng minh rằng sự kết hợp của các phương pháp khác nhau là đáng xem xét để cải thiện hiệu suất của các mô hình dự đoán vỡ nợ.

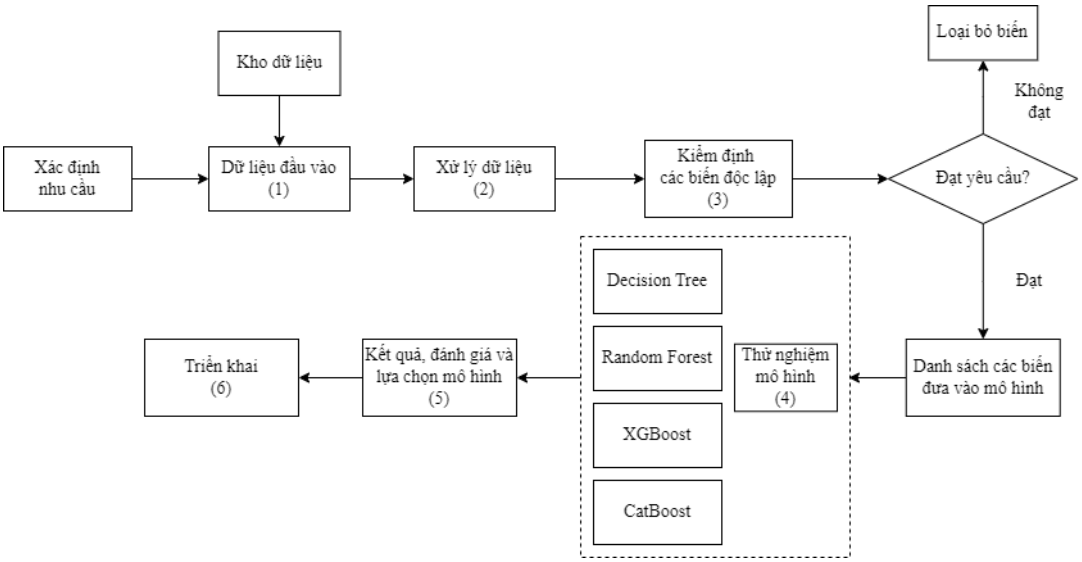
Tại Việt Nam đã một số nghiên cứu về chủ đề này, cụ thể Đặng Thị Thu Hằng (2019) đã sử dụng phương pháp ước lượng xác xuất vỡ nợ (Possibility of Default – PD) thông qua mô hình Logistic để đánh giá và dự báo rủi ro vỡ nợ của KHDN. Nghiên cứu này dựa trên các thông tin công khai từ báo cáo tài chính của các doanh nghiệp đã niêm yết trên thị trường chứng khoán Việt Nam. Từ những kết quả thu được, tác giả đã tiến hành phân tích và đưa ra những khuyến nghị nhằm tăng cường khả năng áp dụng các tiêu chuẩn quản trị rủi ro tín dụng quốc tế cho các ngân hàng tại Việt Nam. Phi Hồng Hạnh (2015) đã sử dụng mô hình Logit trong nghiên cứu về phân tích các chỉ tiêu tài chính để đánh giá khả năng trả nợ của doanh nghiệp nhằm hỗ trợ các ngân hàng trong việc xếp hạng tín dụng và ra quyết định về việc cho vay hoặc phân loại nợ. Ngoài ra, mô hình này cũng giúp cho các doanh nghiệp có thể tiếp cận các nguồn tài chính dựa trên thông tin tín dụng có sẵn một cách dễ dàng và tin cậy hơn. Việc ứng dụng mô hình định lượng về chấm điểm tín dụng và xếp hạng khách hàng để nâng cao khả năng quản trị rủi ro và hiệu quả kinh doanh cũng đã được thực hiện trong nghiên cứu Nguyễn Quốc Chiến và Nguyễn Thị Thúy Quỳnh (2022) tại chi nhánh huyện Mường Ảng, Điện Biên của Ngân hàng Nông nghiệp và Phát triển Nông thôn. Nghiên cứu đã dựa trên tập dữ liệu thu thập bao gồm thông tin của 1.892 khách hàng cá nhân để phân tích, xây dựng và lựa chọn các biến quan trọng, sau đó sử dụng mô hình hồi quy Logistic chấm điểm tín dụng để xếp hạng khách hàng khi nộp hồ sơ vay. Kết quả đánh giá chính xác khả năng trả nợ của khách hàng so với phương pháp thẩm định chuyên gia mà ngân hàng đang áp dụng. Nghiên cứu Hoàng Thanh Hải và cộng sự (2018) đã sử dụng mô hình hồi quy logistic để đánh giá ảnh hưởng của các biến giải thích đến xác suất vỡ nợ của khách hàng tín dụng cá nhân và xây dựng một mô hình dự báo xác suất vỡ nợ. Kết quả từ mô hình dự báo cho thấy tỷ lệ dự đoán định lượng trên 80% độ chính xác và chỉ số đánh giá định tính (Area Under the Receiver – AUC) đo lường diện tích nằm dưới đường cong (Receiver Operating Characteristic Curve – ROC) chiếm gần 0,7% diện tích trên cả dữ liệu huấn luyện và dữ liệu kiểm định. Ngoài ra, trong một nghiên cứu khác của nhóm tác giả Đỗ Năng Thắng và Nguyễn Văn Huân (2019) đã tiến hành khảo sát và đề xuất một bộ các yếu tố ảnh hưởng đến khả năng trả nợ của khách hàng. Dữ liệu thu thập 210 mẫu quan sát để làm bộ dữ liệu và làm sạch bằng phần mềm SPSS. Mô hình dựa trên hồi quy logistic nhị phân của Maddala (1983) đã được áp dụng để tìm hiểu cách mà từng yếu tố đơn

lẻ của khách hàng ảnh hưởng đến khả năng trả nợ của họ. Kết quả nghiên cứu cho thấy mức độ ảnh hưởng của từng yếu tố đối với khả năng trả nợ của khách hàng được xác định rõ ràng. Điều này giúp các nhà quản lý ngân hàng có cái nhìn trực quan và sâu sắc hơn để đưa ra quyết định cho vay chính xác và hạn chế rủi ro.

Dựa trên các bài nghiên cứu trong và ngoài nước có liên quan đến việc ứng dụng mô hình máy học để dự đoán rủi ro vỡ nợ của khách hàng, tác giả nhận thấy các nghiên cứu đã cho thấy nhiều mô hình có thể được ứng dụng để giải quyết bài toán dự đoán khả năng vỡ nợ của khách hàng, và một số phương pháp mới về việc xây dựng các biến độc lập, áp dụng WOE để chuyển đổi biến trước khi đưa vào mô hình như trong nghiên cứu của Nikolic và cộng sự (2013). Tuy nhiên, các nghiên cứu ở Việt Nam hầu hết đều sử dụng mô hình hồi quy Logistic, chưa ứng dụng các mô hình học máy khác để so sánh hiệu quả của các mô hình trên cùng một bộ dữ liệu để đánh giá. Ngoài ra, các nghiên cứu này chưa đưa ra ngưỡng xác định khách hàng vỡ nợ để dự đoán sự kiện vỡ nợ xảy ra trong tương lai mà chỉ dự đoán khách hàng vỡ nợ tại một thời điểm. Đây là những khoảng trống giúp cho nhóm tác giả đề xuất mô hình dự báo xác định khách hàng vỡ nợ trong tương lai.

3. Mô hình nghiên cứu

Trong nghiên cứu này, nhóm tác giả đề xuất xây dựng mô hình dự báo dựa trên hành vi của quá khứ và hiện tại nhằm xác định khách hàng vỡ nợ trong tương lai trên khoảng thời gian quan sát được xác định trước là 12 tháng, chi tiết ở các bước thực hiện như Hình 1:



Hình 1. Quy trình xây dựng mô hình học máy và ứng dụng mô hình với dự đoán khách hàng vỡ nợ

- Bước 1: Dựa trên mục tiêu dự đoán khách hàng có khả năng vỡ nợ, tác giả đã tiến hành nghiên cứu và xác định các tiêu chí cần thiết để đưa vào dữ liệu đầu vào phù hợp. Dựa trên một số nghiên cứu của Zekic-Susac và cộng sự (2004), Luo và cộng sự (2020), Li và cộng sự (2021) thì các tiêu chí chủ yếu được sử dụng để đưa vào dữ liệu đầu vào cho việc dự đoán vỡ nợ bao gồm:

+ *Thông tin tín dụng,* như: giá trị khoản vay, lãi suất, kỳ hạn vay.

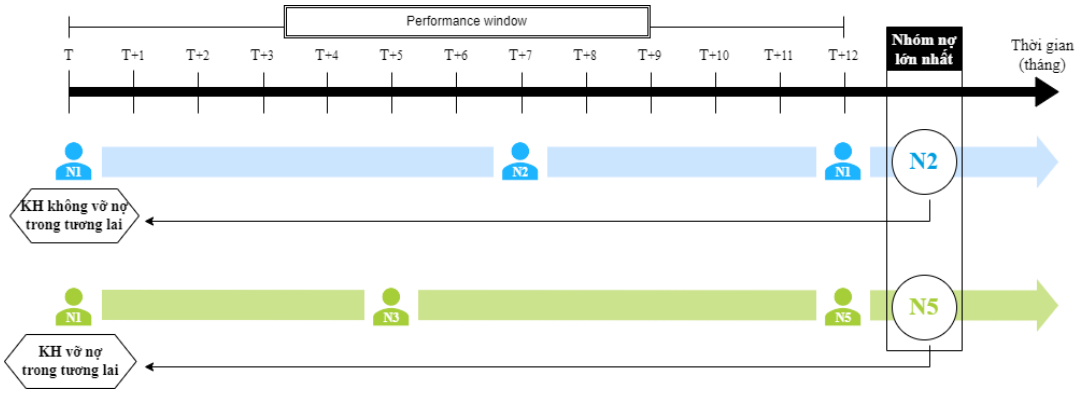
+ *Thông tin tài chính,* như: tình hình lợi nhuận, nợ phải trả, tài sản, vốn chủ sở hữu, doanh số gửi rút, phí dịch vụ.

+ *Đặc điểm doanh nghiệp,* như: ngành nghề, khu vực, số lượng nhân viên, số năm kinh doanh.

Sau đó, tác giả tiến hành thu thập, trích lọc các trường dữ liệu có liên quan từ kho dữ liệu trong hệ thống của ngân hàng. Kho dữ liệu ngân hàng là một hệ thống lưu trữ và quản lý các thông tin và dữ liệu liên quan đến hoạt động và giao dịch của ngân hàng. Do các dữ liệu thu thập được theo các tiêu chí trên đang theo cấp độ tài khoản nên nhóm tác giả thực hiện chuyển đổi dữ liệu theo cấp độ khách hàng để phục vụ cho mục đích nghiên cứu, ví dụ như các trường kỳ hạn, lãi suất của các tài khoản sẽ được thay đổi bằng cách tính bình quân kỳ hạn, và bình quân lãi suất theo từng khách hàng thay vì theo từng tài khoản.

- Bước 2: Tác giả thực hiện tiền xử lý dữ liệu thông qua việc kiểm tra tính hợp lệ của dữ liệu như đối với các trường thông tin về số dư, số năm kinh doanh không được là số âm, trường thông tin về số lượng nhân viên phải là số nguyên. Đối với các trường dữ liệu bị thiếu sẽ được thay thế bằng giá trị trung bình hoặc loại bỏ trường nếu không có ảnh hưởng lớn, loại bỏ dữ liệu ngoại lai do các dữ liệu ngoại lai chỉ chiếm tỷ trọng nhỏ trong tập dữ liệu nghiên cứu. Từ bộ dữ liệu sẵn có, tác giả tiến hành tạo các tiêu chí mới và biến mục tiêu cho mô hình máy học. Việc dự đoán xác suất vỡ nợ của khách hàng trong tương lai dựa trên thông tin hành vi của khách hàng; do đó, một số tiêu chí của khách hàng có thể được tạo thêm thông qua cấu phần Tần suất (Frequency) và Giá trị (Monetary) trong phương pháp RFM (Recency – Frequency - Monetary), ví dụ như số lần hoặc doanh số gửi rút trong vòng 1 tháng hoặc 3 tháng gần nhất. Ngoài ra, tác giả cũng tạo thêm một số tiêu chí như tỷ lệ tăng trưởng so với tháng trước của khách hàng về số dư, doanh số gửi rút...

Dựa trên thông tin các khoản vay của khách hàng tại các thời điểm quan sát khác nhau, và khung thời gian hiệu suất (Performace window) để làm thời gian quan sát hiệu quả của các khoản vay này là 12 tháng để thực hiện đánh giá biến mục tiêu. Nếu trong khoảng thời gian này có bất cứ khoản vay nào của khách hàng có khoản vay từ nhóm 3 trở lên thì khách hàng này được xem là vỡ nợ và được gắn cờ vỡ nợ cho biến mục tiêu và ngược lại (Hình 2). Kết quả thu được bộ dữ liệu với 119 biến độc lập và biến mục tiêu là khách hàng vỡ nợ hoặc không vỡ nợ.



Hình 2. Minh họa xác định KHDN có khả năng vỡ nợ trong tương lai

Nhiều bài toán phân loại trong thực tế có phân phối không cân bằng giữa các lớp, ví dụ như phát hiện gian lận, phát hiện vỡ nợ, phát hiện thư rác và dự đoán khách hàng rời đi. Dữ liệu mất cân bằng được phân thành hai mức độ như sau trong nghiên cứu Brownlee (2020):

+ Mất cân bằng nhẹ (Slight Imbalance) là khi có sự phân bố không đồng đều một lượng nhỏ quan sát trong tập dữ liệu đào tạo, ví dụ 4:6.

+ Mất cân bằng nghiêm trọng (Severe Imbalance) là khi có sự phân bố không đồng đều một lượng lớn quan sát trong tập dữ liệu đào tạo, ví dụ 1:100 hoặc nhiều hơn.

Krawczyk (2016) chỉ ra rằng hầu hết các nghiên cứu hiện đại về mất cân bằng lớp dữ liệu tập trung vào tỷ lệ mất cân bằng từ 1:4 đến 1:100. Đối với bộ dữ liệu hiện tại của ngân hàng lấy mẫu có tỷ lệ nợ xấu thấp, nên việc dữ liệu bị mất cân bằng xảy ra đối với nhóm khách hàng vỡ nợ và không vỡ nợ là đúng với thực tế. Với dữ liệu gốc ban đầu, tỷ lệ mất cân bằng là 1:62 tương ứng với 62 khách hàng vay thì có 1 khách hàng vỡ nợ (chiếm 2%). Do đó, để xử lý việc mất cân bằng dữ liệu, tác giả lựa chọn đưa tỷ lệ khách hàng vỡ nợ và không vỡ nợ về 1:3, và lựa chọn phương pháp giảm mẫu dữ liệu (Under-sampling) với phân tổ (stratify) theo từng tháng nhằm phản ánh đúng dữ liệu thật trong bộ dữ liệu đã lấy mẫu và đảm bảo tính đại diện. Ngoài ra, việc giảm số lượng mẫu trong nhóm đa số còn giúp giảm độ phức tạp tính toán và thời gian huấn luyện mô hình. Kết quả thu được bộ dữ liệu có 53.125 quan sát, trong đó có 40.129 quan sát tốt (KHDN không vỡ nợ trong 12 tháng tới) và 12.996 quan sát xấu (KHDN vỡ nợ trong 12 tháng tới).

- Bước 3: Tác giả tiến hành lựa chọn biến độc lập có ảnh hưởng đến biến mục tiêu bằng chỉ số IV lớn hơn 0,02, và kiểm định T-Test có p-value thấp dưới ngưỡng ý nghĩa 0,05.

Chuyển đổi biến định tính thành định lượng bằng trọng số bằng chứng WOE là một phương pháp phổ biến, thể hiện sức mạnh dự báo của mỗi biến đối với biến mục tiêu. WOE là một số dương hoặc âm đại diện cho sức mạnh của một biến định tính đối với biến mục tiêu trong mô hình. WOE được tính bằng cách so sánh tỷ lệ giữa tỷ lệ khả năng xảy ra vỡ nợ và tỷ lệ khả năng không xảy ra vỡ nợ giữa các nhóm của biến định tính. Việc chuyển đổi biến định tính sang WOE để lượng hóa, giúp cho việc đánh giá sức mạnh của biến đó đối với biến mục tiêu trở nên dễ dàng và có thể so sánh được giữa các biến với nhau. WOE cũng có thể được sử dụng để tạo ra các biến tương tự như các biến giả, mà không cần phải tạo ra nhiều biến. Cụ thể, WOE của một nhóm được tính bằng công thức sau:

$$WOE_i = ln\left(\frac{Distr\ Good_i}{Distr\ Bad_i}\right) \quad (1)$$

Trong đó:

Distr $Good_i$ là tỷ trọng của số quan sát tốt (không vỡ nợ) trong nhóm thứ i, được tính bằng số quan sát tốt của nhóm thứ i chia cho tổng số quan sát tốt.

Distr Bad_i là tỷ trọng của số quan sát xấu (vỡ nợ) trong nhóm thứ i, được tính bằng số quan sát xấu của nhóm thứ i chia cho tổng số quan sát xấu.

Ngoài ra, ta có IV (InformationValue) là một chỉ số dùng để đánh giá sức mạnh của một biến định tính trong mô hình hồi quy, bằng cách tính tổng WOE của tất cả các nhóm của biến đó. Cụ thể, công thức tính IV như sau:

$$IV = \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr } Bad_i) * WOE \quad (2)$$

Trong đó:

Distr $Good_i$ là tỷ trọng của số quan sát tốt (không vỡ nợ) trong nhóm thứ i, được tính bằng số quan sát tốt của nhóm thứ i chia cho tổng số quan sát tốt.

Distr Bad_i là tỷ trọng của số quan sát xấu (vỡ nợ) trong nhóm thứ i, được tính bằng số quan sát xấu của nhóm thứ i chia cho tổng số quan sát xấu.

WOE là Weight of Evidence, được tính từ công thức (1)

Theo quy ước trong nghiên cứu của Siddiqi (2016), khả năng dự báo của biến dựa trên giá trị của IV có thể được phân loại như sau:

IV < 0,02: không có khả năng dự báo

0,02 ≤ IV < 0,1: khả năng dự báo yếu

0,1 ≤ IV < 0,3: khả năng dự báo trung bình

0,3 ≤ IV: khả năng dự báo mạnh

Lưu ý trong trường hợp biến có giá trị IV lớn hơn 0,5 thì nên được kiểm tra xem có gây ra hiện tượng dự đoán quá cao hay không. Chúng có thể được loại bỏ khỏi quá trình mô hình hóa hoặc sử dụng một cách có kiểm soát.

Sau khi biến đổi biến định tính thành định lượng WOE, cùng với các biến định lượng khác trong dữ liệu, tác giả tiến hành kiểm định sự khác biệt trung bình của các biến định lượng so với biến mục tiêu là khách hàng vỡ nợ hay không vỡ nợ để thực hiện lọc các biến không có sự khác biệt giữa hai nhóm khách hàng trước khi đưa vào mô hình dự báo. Cụ thể, để thực hiện kiểm định T-Test, tác giả xác định giả thuyết H_0 là không có sự khác biệt trung bình giữa các biến định lượng của khách hàng vỡ nợ và khách hàng không vỡ nợ, giả thuyết H_1 là có sự khác biệt trung bình giữa các biến định lượng của khách hàng vỡ nợ và khách hàng không vỡ nợ. Giá trị t (t-score) được tính toán bằng cách lấy hiệu của trung bình của hai nhóm (nhóm khách hàng vỡ nợ và không vỡ nợ) chia cho sự khác biệt tiêu chuẩn giữa hai nhóm. Giá trị t cao hơn cho thấy sự khác biệt lớn hơn giữa hai nhóm, công thức như sau:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} \quad (3)$$

Trong đó:

$\overline{X_1}$, $\overline{X_2}$: trung bình của biến định lượng cho khách hàng vỡ nợ và khách hàng không vỡ nợ.

s_1, s_2: độ lệch chuẩn của biến định lượng cho khách hàng vỡ nợ và khách hàng không vỡ nợ.

n_1, n_2: kích thước mẫu của khách hàng vỡ nợ và khách hàng không vỡ nợ.

Mức ý nghĩa (alpha) là mức giá trị tối đa của xác suất chấp nhận giả thuyết sai (tức là bác bỏ giả thuyết đúng). Mức ý nghĩa thường được chọn là 0,05, nghĩa là xác suất chấp nhận giả thuyết sai là 5%. Giá trị p (p-value) là xác suất được tính toán từ phân phối t-student với mức ý nghĩa được xác định trước. Giá trị p-value thể hiện xác suất để t-score lớn hơn hoặc bằng giá trị quy định nếu không có sự khác biệt thực sự giữa hai nhóm. Nếu p-value thấp (dưới ngưỡng ý nghĩa 5%), ta có thể bác bỏ giả thuyết không khác biệt giữa hai nhóm và kết luận rằng có sự khác biệt đáng kể giữa các nhóm

khách hàng vỡ nợ và không vỡ nợ cho từng biến định lượng. Kết quả thu được 48 biến độc lập thỏa mãn điều kiện (Bảng 1).

Bảng 1.

Danh sách các biến độc lập sử dụng cho mô hình dự đoán khả năng vỡ nợ của KHDN

STT	Tên biến	Mô tả
1	A1	Tổng dư nợ hiện tại
2	A2	Số lượng lao động
3	A3	Số lượng lao động bảo hiểm xã hội bình quân
4	A4	Số năm kinh doanh
5	A5	Tỷ số khả năng trả nợ
6	A6	Giá trị tài sản
7	A7	Lãi suất
8	A8	Kỳ hạn
9	A9	Doanh số gửi
10	A10	Số lần gửi
11	A11	Số lần rút
12	A12	Số lần gửi tiền mặt
13	A13	Số lần rút tiền mặt
14	A14	Phí thanh toán quốc tế
15	A15	Phí dịch vụ
16	A16	Phí bảo lãnh
17	A17	Phí khác
18	A18	Tổng doanh thu
19	A19	Tổng nguồn vốn
20	A20	Tổng dư nợ trong 3 tháng gần nhất
21	A21	Số dư nợ trung bình trong 3 tháng gần nhất
22	A22	Số dư nợ tối đa trong 3 tháng gần nhất
23	A23	Tổng doanh số gửi trong 3 tháng gần nhất
24	A24	Doanh số gửi trung bình trong 3 tháng gần nhất
25	A25	Doanh số gửi tối đa trong 3 tháng gần nhất
26	A26	Tổng số lần gửi trong 3 tháng gần nhất
27	A27	Số lần gửi trung bình trong 3 tháng gần nhất
28	A28	Số lần gửi tối đa trong 3 tháng gần nhất

STT	Tên biến	Mô tả
29	A29	Tổng doanh số rút trong 3 tháng gần nhất
30	A30	Doanh số rút trung bình trong 3 tháng gần nhất
31	A31	Doanh số rút tối đa trong 3 tháng gần nhất
32	A32	Tổng số lần rút trong 3 tháng gần nhất
33	A33	Số lần rút trung bình trong 3 tháng gần nhất
34	A34	Số lần rút tối đa trong 3 tháng gần nhất
35	A35	Tổng số lần gửi tiền mặt trong 3 tháng gần nhất
36	A36	Số lần gửi tiền mặt trung bình trong 3 tháng gần nhất
37	A37	Số lần gửi tiền mặt tối đa trong 3 tháng gần nhất
38	A38	Tổng số lần rút tiền mặt trong 3 tháng gần nhất
39	A39	Số lần rút tiền mặt trung bình trong 3 tháng gần nhất
40	A40	Số lần rút tiền mặt tối đa trong 3 tháng gần nhất
41	A41	Tổng phí thanh toán quốc tế trong 3 tháng gần nhất
42	A42	Phí thanh toán quốc tế trung bình trong 3 tháng gần nhất
43	A43	Phí thanh toán quốc tế tối đa trong 3 tháng gần nhất
44	A44	Tổng phí khác trong 3 tháng gần nhất
45	A45	Phí khác trung bình trong 3 tháng gần nhất
46	A46	Phí khác tối đa trong 3 tháng gần nhất
47	A47	WOE từng nhóm ngành
48	A48	WOE từng nhóm khu vực

Bước 4: Bộ dữ liệu sau khi qua các bước trong quy trình chuẩn bị dữ liệu để tiến tới mô hình hóa bao gồm các biến độc lập được nêu trong Bảng 1.

Từ kết quả của bộ dữ liệu sau khi xử lý, tác giả thực hiện chia dữ liệu thành hai phần là tập dữ liệu huấn luyện (70%) tương ứng với 37.187 quan sát, và tập dữ liệu kiểm tra (30%) tương ứng với 15.938 quan sát (Bảng 2).

Bảng 2.

Tập dữ liệu sử dụng cho mô hình dự đoán khả năng vỡ nợ của KHDN

Tập dữ liệu	Tỷ lệ	Số quan sát	Số quan sát không vỡ nợ	Số quan sát vỡ nợ
Huấn luyện	70%	37.187	28.090	9.097
Kiểm tra	30%	15.938	12.039	3.899
Tổng	100%	53.125	40.129	12.996

Tiếp đến, tác giả tiến hành xây dựng mô hình học máy dựa trên tập dữ liệu huấn luyện với các thuật toán Decision Tree, Random Forest, XGBoost, CatBoost và thực hiện đánh giá kết quả trên tập dữ liệu kiểm tra.

4. Kết quả đánh giá

Bảng 3 thống kê các hệ số để đánh giá mô hình phân loại, theo đó, để đạt được mục tiêu xây dựng mô hình để dự đoán khách hàng vỡ nợ trong 12 tháng tới thì mô hình XGBoost đứng đầu cả 4 chỉ số Recall, F_1-score, Accuracy, AUC. Riêng đối với chỉ số Precision, mô hình XGBoost thấp hơn mô hình Random Forest.

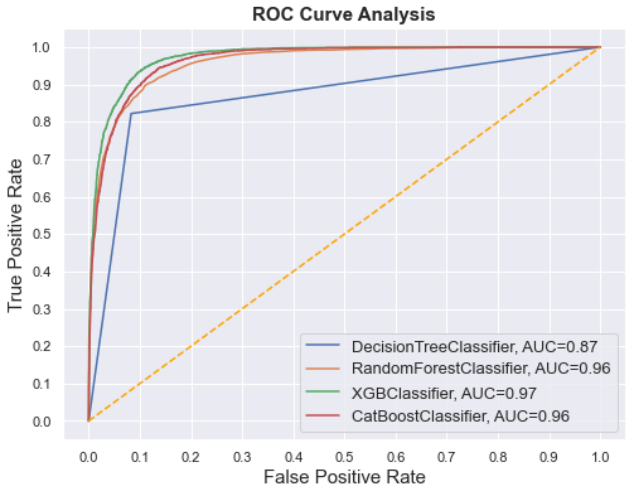
Bảng 3.

Tổng hợp các hệ số đánh giá mô hình

	Mô hình	Precision	Recall	F_1-score	Accuracy	AUC
KHDN vỡ nợ trong 12 tháng tới	Decision Tree	0,76	0,82	0,79	0,89	0,87
	Random Forest	**0,88**	0,71	0,79	0,91	0,96
	XGBoost	0,86	**0,82**	**0,84**	**0,92**	**0,97**
	CatBoost	0,84	0,77	0,80	0,91	0,96

Ghi chú: - Precision là tỷ lệ dự đoán chính xác trong các trường hợp dự đoán là vỡ nợ;

- Recall là tỷ lệ dự dự đoán chính xác trong các trường hợp thực tế là vỡ nợ;
- F1-score là tỷ lệ trung bình điều hòa của Precision và Recall;
- Accuracy là tỷ lệ dự đoán chính xác trong tất cả trường hợp;
- AUC là diện tích dưới đường cong ROC (Receiver Operating Characteristic).
- Các giá trị in đậm thể hiện mô hình tốt nhất theo từng chỉ số.



Hình 3. Tổng hợp đường cong ROC đánh giá mô hình

Hình 3 cho thấy tổng hợp các đường cong ROC để đánh giá mô hình phân loại, theo đó mô hình XGBoost cho thấy khả năng dự đoán mạnh nhất so với các mô hình còn lại với AUC là 0,97, tiếp đến là mô hình CatBoost với AUC là 0,96, mô hình Random Forest với AUC là 0,96 và cuối cùng là Decision Tree với AUC là 0,87.

5. Thảo luận

Sau khi huấn luyện tập dữ liệu với các mô hình Decision Tree, Random Forest, XGBoost, CatBoost, nhóm tác giả đánh giá dựa trên kết quả có được với tập dữ liệu kiểm tra với các chỉ số Accuracy, Precision, Recall, F_1-score, và AUC. Kết quả cho thấy mô hình XGBoost có khả năng dự đoán tốt nhất so với các mô hình khác với các chỉ số Accuracy (0,92), Recall (0,82), F_1-score (0,84), và AUC (0,97). Mô hình XGBoost riêng chỉ có chỉ số Precision (0,86) thấp hơn so với mô hình Random Forest (0,88). Tuy nhiên, trong trường hợp đánh giá mô hình phân loại khách hàng có khả năng vỡ nợ, chúng ta nên chú ý đến sai lầm loại 2 hơn là sai lầm loại 1. Sai lầm loại 2 là khi mô hình dự đoán khách hàng sẽ không vỡ nợ nhưng thực tế lại vỡ nợ. Khi xảy ra sai lầm loại 2, ngân hàng sẽ cho cho vay cho những khách hàng có khả năng vỡ nợ, dẫn đến thất thoát tài chính và tiềm ẩn nguy cơ rủi ro cho ngân hàng. Trong khi đó, sai lầm loại 1 là khi mô hình dự đoán khách hàng sẽ vỡ nợ nhưng thực tế lại không vỡ nợ. Trong trường hợp này, ngân hàng sẽ từ chối cho vay đối với những khách hàng có khả năng thanh toán đầy đủ, gây mất cơ hội kinh doanh cho ngân hàng, nhưng sẽ ít nguy hiểm hơn so với sai lầm loại 2 đã đề cập ở trên. Dựa trên mục tiêu giảm thiểu rủi ro tín dụng cho ngân hàng thì tác giả ưu tiên chọn mô hình giải quyết tốt sai lầm loại 2, nghĩa là khách hàng thực tế là vỡ nợ nhưng dự đoán là không vỡ nợ, được thể hiện qua chỉ số Recall. Do đó, tác giả chọn mô hình XGBoost để xây dựng mô hình dự đoán khách hàng vỡ nợ trong 12 tháng tới tại ngân hàng.

So với các nghiên cứu trước đó, kết quả của nghiên cứu này cho thấy mô hình XGBoost có F_1-score đạt 0,84 và AUC trên dữ liệu kiểm định đạt 0,97. Điều này chứng tỏ mô hình XGBoost có khả năng dự đoán khả năng vỡ nợ của KHDN một cách hiệu quả và đáng tin cậy. So với các phương pháp truyền thống trong việc đánh giá tín dụng, việc sử dụng mô hình học máy đã mang lại những cải tiến đáng kể về độ chính xác và hiệu suất. Tuy nhiên, nghiên cứu cũng gặp một số hạn chế như dữ liệu sử dụng trong nghiên cứu này chỉ bao gồm thông tin từ khách hàng của một ngân hàng cụ thể, điều này có thể ảnh hưởng đến khả năng ứng dụng của mô hình trong các ngân hàng khác. Ngoài ra, mô hình chỉ xây dựng dựa trên các yếu tố nội tại của khách hàng mà chưa xem xét các yếu tố bên ngoài như biến động kinh tế, chính sách tài chính và tình hình thị trường. Những yếu tố này cũng có thể ảnh hưởng đến khả năng vỡ nợ của khách hàng mà chưa được mô hình hóa trong nghiên cứu này. Do đó, có thể cần tiếp tục nghiên cứu và tích hợp những yếu tố bên ngoài này vào mô hình để cải thiện độ chính xác và đáng tin cậy của dự đoán. Ngoài ra, nghiên cứu cũng có thể khám phá thêm các biến đổi và tiền xử lý dữ liệu phù hợp để tăng tính ổn định và chính xác của mô hình. Điều này đòi hỏi một quá trình nghiên cứu và phát triển tiếp theo để đảm bảo tính tin cậy và ứng dụng hiệu quả của mô hình trong thực tế.

6. Kết luận

Trong bài báo này, nhóm tác giả đã đề xuất mô hình học máy dự đoán khả năng vỡ nợ của KHDN, giúp ngân hàng giảm thiểu rủi ro tín dụng, tối ưu hóa danh mục cho vay, và cải thiện hiệu quả quy trình phân tích tín dụng. Bằng việc áp dụng mô hình dự đoán, ngân hàng có thể đưa ra quyết định kinh doanh chính xác hơn, từ đó giảm thiểu các tổn thất không mong muốn, cho vay các khách hàng có ít rủi ro hơn và tối ưu hóa lợi nhuận. Trong đó sử dụng một số mô hình máy học cơ bản và thu thập bộ dữ liệu mẫu từ một ngân hàng thương mại tại Việt Nam từ năm 2016 đến năm 2021. Kết quả của nghiên cứu cho thấy mô hình XGBoost là mô hình phù hợp nhất trong việc dự đoán khả năng vỡ nợ, với chỉ số F1-score đạt 0,84, đường cong ROC có diện tích AUC đạt 0,97. Ngoài ra, mô hình XGBoost còn giải quyết tốt sai lầm loại 2 về việc dự đoán khả năng vỡ nợ của KHDN. Do đó, việc áp dụng mô hình này đã đáp ứng mục tiêu của nghiên cứu đặt ra về việc chủ động giám sát, đánh giá khách hàng kịp thời, qua đó hỗ trợ kiểm soát nợ xấu. Ngoài ra, nghiên cứu này còn đóng góp quan trọng vào việc quản lý rủi ro trong hoạt động kinh doanh của Ngân hàng. Nghiên cứu này cũng đóng góp vào việc tự động hóa quy trình tín dụng, nhờ mô hình máy học thì quy trình xét duyệt tín dụng có thể được thực hiện tự động và hiệu quả, giúp giảm thiểu sai sót và tăng tốc độ xử lý hồ sơ vay. Điều này không chỉ giúp tiết kiệm thời gian và chi phí mà còn tăng cường sự linh hoạt và đáng tin cậy trong quy trình tín dụng của Ngân hàng. Bên cạnh đó, mô hình dự đoán khả năng vỡ nợ cũng hỗ trợ ngân hàng trong việc tối ưu hóa danh mục tín dụng, hạn chế rủi ro và đảm bảo lợi nhuận cho ngân hàng.

Lời cảm ơn

Bài báo này được tài trợ bởi Trường Đại học Kinh tế TP. Hồ Chí Minh theo Đề tài Nghiên cứu khoa học cấp trường, Mã số: CTD-2022-14, Đề tài: "*Xây dựng mô hình kết hợp giữa tập mờ hình ảnh Picture Fuzzy Set và Fuzzy TOPSIS cho bài toán ra quyết định tín dụng Ngân hàng", Chủ nhiệm:* TS. Nguyễn Quốc Hùng.

Tài liệu tham khảo

Boughaci, D., Alkhawaldeh, A. A. K., Jaber, J. J., & Hamadneh, N. (2021). Classification with segmentation for credit scoring and bankruptcy prediction. *Empirical Economics, 61*(3), 1281–1309. doi:10.1007/s00181-020-01901-8

Brownlee, J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*: Machine Learning Mastery.

Phạm Quốc Chiến, & Nguyễn Thị Thúy Quỳnh. (2022). Nâng cao chất lượng tín dụng bằng chấm điểm khách hàng tại Ngân hàng Nông nghiệp và Phát triển Nông thôn chi nhánh huyện Mường Ảng, Điện Biên. *Tạp chí Khoa học (Trường Đại học Tây Bắc)*, *25*, 31–36.

Addo, P.M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Computational Methods for Risk Management in Economics and Finance, 6*(2), 38. doi:10.3390/risks6020038

Guegan, D., Addo, P., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Computational Methods for Risk Management in Economics and Finance, 6*(2), 38. doi:10.3390/risks6020038

Hoàng Thanh Hải, Trần Đình Chúc, & Nguyễn Quỳnh Hoa. (2018). Mô hình hồi quy Logistic trong đo lường xác suất vỡ nợ khách hàng tín dụng cá nhân. *Tạp chí Kinh tế & Quản trị Kinh doan*h, *7*, 92–100.

Đặng Thị Thu Hằng. (2019). Ứng dụng mô hình logistic trong quản trị rủi ro tín dụng. *Tạp chí Thị trường Tài chính Tiền tệ, 11,* 30–34.

Phi Hồng Hạnh. (2015). Đánh giá khả năng trả nợ của doanh nghệp bằng ứng dụng Mô hình Logit. *Tạp chí Khoa học & Đào tạo Ngân hàng, 155,* 45–51.

Gonzales, F., Haas, F., Persson, M., Toledo, L., Violi, R., Wieland, M., & Zins, C. (2004). Market dynamics associated with credit ratings: A literature review. *European Central Bank Occasional Paper Series, 16*.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232. doi:10.1007/s13748-016-0094-0

Li, B., Xiao, B., & Yang, Y. (2021). Strengthen credit scoring system of small and micro businesses with soft information: Analysis and comparison based on neural network models. *Journal of Intelligent & Fuzzy Systems, 40*, 1–18. doi:10.3233/JIFS-200866

Luo, Z., Hsu, P., & Xu, N. (2020). SME default prediction framework with the effective use of external public credit data. *Sustainability, 12*, 7575. doi:10.3390/su12187575

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts, London, England: MIT press.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. *Expert Systems with Applications, 40*(15), 5932–5944. doi:10.1016/j.eswa.2013.05.022

Rose, & Hudgins. (2015). Bank Management and Financial Services. New York: *McGraw-Hill*.

Siddiqi, N. (2016). *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, Second Edition*. Wiley.

Đỗ Năng Thắng, & Nguyễn Văn Huân (2019). Đề xuất cảnh báo rủi ro trong cho vay khách hàng doanh nghiệp của ngân hàng thương mại ở Việt Nam. *Tạp Chí Khoa học thương mại, 131*, 55–63.

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A Novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications, 27*, 74–82. doi:10.1016/j.elerap.2017.12.006

Yuan, K., Chi, G., Zhou, Y., & Yin, H. (2021). A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. *Research in International Business and Finance, 59*, 101536. doi:10.1016/j.ribaf.2021.101536

Zekic-Susac, M., Sarlija, N., & Benšić, M. (2004). Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models. *In 26th International Conference on Information Technology Interfaces*, 265–270.