

## Lựa chọn mô hình dự đoán xác suất vỡ nợ của khách hàng cá nhân vay tín chấp: Trường hợp Ngân hàng Thương mại Cổ phần Á Châu (ACB)

### Model selection for predicting the default probability of individual unsecured loans: The case of Asia Commercial Joint Stock Bank (ACB)

Vũ Hữu Thành<sup>1\*</sup>, Hoàng Thị Kim Diễm<sup>2</sup>

<sup>1</sup>Trường Đại học Mở Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam

<sup>2</sup>Ngân hàng Thương mại Cổ phần Á Châu, Thành phố Hồ Chí Minh, Việt Nam

\*Tác giả liên hệ, Email: thanh.vh@ou.edu.vn

#### THÔNG TIN

#### TÓM TẮT

DOI:10.46223/HCMCOUJS.  
econ.vi.20.3.3394.2025

Ngày nhận: 23/04/2024

Ngày nhận lại: 21/06/2024

Duyệt đăng: 22/11/2024

Mã phân loại JEL:

G21

Nghiên cứu này được thực hiện nhằm lựa chọn mô hình dự đoán xác suất vỡ nợ của hệ thống xếp hạng tín dụng nội bộ đối với khách hàng cá nhân. Từ đó, nghiên cứu cũng đề xuất tích hợp áp dụng mô hình vào quy trình tín dụng và đề xuất thu thập thông tin chính xác để tăng chất lượng dự đoán của mô hình. Trong nghiên cứu này, nhóm tác giả sử dụng bộ dữ liệu có thời gian lấy mẫu từ 01/01/2022 đến 31/12/2022 và thời gian quan sát từ 01/01/2023 đến 31/12/2023. Nhóm nghiên cứu đã sử dụng mô hình Logistic, 09 mô hình học máy và mô hình kết hợp Ensemble. Nhóm cũng sử dụng các biện pháp cân bằng dữ liệu và tiền xử lý dữ liệu trước khi đưa vào ước lượng mô hình. Kết quả, mô hình Logistic và mô hình kết hợp Ensemble là hai mô hình dự báo tốt nhất. Với ngưỡng xác suất phá sản tối ưu cho dự đoán độ nhạy, mô hình Logistic đã cho kết quả dự đoán trội hơn so với Ensemble. Ngoài ra, nghiên cứu cũng phát hiện các thông tin quan trọng để dự đoán xác suất vỡ nợ bao gồm: Trình độ học vấn, Loại hình tổ chức, Giới tính, Độ tuổi, Thời gian liên tục có thu nhập, Thời gian công tác, Thời hạn vay, Nhu cầu vay, Tổng thu nhập, Tổng chi phí, Nợ phải trả hàng tháng, Lịch sử tín dụng 06 tháng gần nhất.

#### ABSTRACT

This study was conducted to select a predictive model for the probability of default within an internal credit rating system for individual unsecured loans. Consequently, the study also suggests integrating the model into the credit process and proposes the collection of accurate information to enhance the model's predictive quality. In this research, the authors used a dataset sampled from January 01, 2022, to December 31, 2022, and observed from January 01, 2023, to December 31, 2023. We utilized nine Machine-Learning models and an Ensemble model. We also implemented data balancing and preprocessing features before model estimation. The Logistic and Ensemble models were the two best predictive models. With an optimal bankruptcy probability threshold for sensitivity prediction, the Logistic model performed better than the Ensemble. Additionally, the analysis revealed the importance of predictive variables, including Educational level, Organization type, Gender, Age, Continuous income duration, Employment duration, Loan term, Borrowing needs, Total income, Total expenses, Monthly debt obligations, and Credit history for the last six months.

*Từ khóa:*

dự báo; mô hình học máy;  
rủi ro tín dụng; xác suất  
vỡ nợ

*Keywords:*

prediction; machine learning  
models; credit risk; default  
probability

## 1. Giới thiệu

Hệ thống xếp hạng tín dụng nội bộ của Ngân hàng tại Việt Nam được quy định sử dụng trong (i) quá trình thẩm định và quyết định cho vay (Ngân hàng Nhà nước Việt Nam, 2018) và (ii) phân loại nợ và trích lập dự phòng rủi ro đối với rủi ro tín dụng (Ngân hàng Nhà nước Việt Nam, 2021). Hiện nay, Chính Phủ và Ngân hàng Nhà nước đang xây dựng hành lang pháp lý đối với lĩnh vực xếp hạng tín nhiệm nhằm nâng cao tính công khai minh bạch thông tin nhằm hỗ trợ cho các ngân hàng kiểm soát rủi ro tín dụng. Các Ngân Hàng Thương Mại (NHTM) cũng ngày càng nhận thấy tầm quan trọng của hệ thống này trong hoạt động tín dụng và quản trị rủi ro. Kết quả dự đoán xác suất vỡ nợ của hệ thống xếp hạng tín dụng nội bộ sẽ giúp ngân hàng có những ứng dụng trong công tác quản trị rủi ro như: hỗ trợ quá trình thẩm định và ra quyết định cho vay, xây dựng khung lãi suất nhằm tối đa hóa lợi nhuận của ngân hàng, đồng thời đảm bảo an toàn vốn của chủ sở hữu. Ngoài ra, các ngân hàng có thể giám sát danh mục tín dụng thông qua việc giám sát sự thay đổi phân hạng rủi ro của danh mục, từ đó giúp ngân hàng có những điều chỉnh kịp thời về chính sách, sản phẩm, định hướng tín dụng.

Trên thế giới và Việt Nam các nghiên cứu tiếp cận ở nhiều khía cạnh khác nhau về dự đoán xác suất vỡ nợ nhưng nhìn chung đều xoay quanh hai hướng: (i) giải thích/chứng minh các yếu tố có ảnh hưởng đến xác suất vỡ nợ của khách hàng hay (ii) tìm ra mô hình toán học tối ưu dự đoán xác suất vỡ nợ. Các nghiên cứu gần đây trên thế giới cho thấy xu hướng sử dụng Machine Learning để tìm ra mô hình tối ưu thay cho phương pháp hồi quy Logistic đã rất phổ biến. Chang và cộng sự (2022) chứng minh rằng Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), và Light Gradient Boosting Machine (LightGBM) có hiệu suất tốt hơn hồi quy Logistic (GLM). Nghiên cứu của Jepakoch (2020) thực hiện các phương pháp GLM, Linear Discriminant Analysis (LDA), Gradient Boosting (GB), DT, KNN, RF, SVM và XGBoost để dự đoán. XGBoost được xem là thuật toán tốt nhất cho bài toán dự đoán xác suất vỡ nợ.

Cho đến nay, các nghiên cứu được công bố ở Việt Nam vẫn chủ yếu sử dụng phương pháp Hồi quy Logistic cho mục tiêu giải thích các biến có ảnh hưởng đến xác suất vỡ nợ của khách hàng. Đối với các nghiên cứu dự đoán, nhóm tác giả nhận thấy chưa có nghiên cứu nào thực hiện dựa trên bộ dữ liệu khách hàng cá nhân mà phân biệt thời gian lấy mẫu và thời gian quan sát để áp dụng hiệu quả các mô hình học máy. Bên cạnh đó, các nghiên cứu này chưa quan tâm tới các hoạt động tiền xử lý dữ liệu như lựa chọn danh mục biến số dự đoán, biến đổi dữ liệu, và áp dụng một số mô hình dự đoán đang nổi trội gần đây như Ensemble. Ngoài ra, theo sự hiểu biết của nhóm tác giả, các nghiên cứu trước đây đã sử dụng ngưỡng xác suất vỡ nợ mặc định là 50% để so sánh tính chính xác của từng mô hình dự báo. Tuy nhiên, việc sử dụng ngưỡng mặc định là 50% là rất ít có ý nghĩa (Alam & ctg., 2020), đặc biệt là trong bối cảnh dự đoán xác suất vỡ nợ của ngân hàng. Các ngân hàng sẽ rất nhạy với dự đoán về vỡ nợ (tập trung vào chỉ số độ nhạy - Sensitivity) hơn là tập trung vào tính chính xác tổng thể (Accuracy). Bằng việc lựa chọn ngưỡng xác suất vỡ nợ mà từ đó đạt được Sensitivity cao nhất, nghiên cứu này sẽ giúp cho ngân hàng thiết lập được ngưỡng xác suất phá sản hay còn gọi là ngưỡng cảnh báo khi dự đoán cho từng hồ sơ tín dụng cá nhân. Kết quả là, nếu một hồ sơ vượt ngưỡng xác suất vỡ nợ, tín hiệu này sẽ được chuyển tới cấp có thẩm quyền để tiếp tục đưa ra những phân tích sâu hơn. Sau đó, cấp có thẩm quyền có thể đưa ra quyết định cấp vốn hoặc ngừng cấp vốn. Bên cạnh đóng góp vừa nêu, kết quả của nghiên cứu có thể giúp cho ngân hàng Á Châu nói riêng và các ngân hàng thương mại cổ phần nói chung một khung áp dụng mô hình dự đoán xác suất vỡ nợ đối với khách hàng cá nhân trong tiến trình phê duyệt.

Từ những lý do trên, nhóm tác giả sẽ thực hiện nghiên cứu để lựa chọn mô hình dự đoán xác suất vỡ nợ nhằm tích hợp vào hệ thống chấm điểm tín dụng nội bộ. Kết quả đó sẽ góp phần

nâng cao năng lực quản trị rủi ro tín dụng khách hàng cá nhân vay tín chấp tại Ngân hàng TMCP Á Châu. Bằng việc thu thập dữ liệu có phương pháp, cân bằng dữ liệu, lựa chọn biến số và lựa chọn các phương pháp mô hình học máy khác nhau để từ đó tìm ra các mô hình dự đoán tốt nhất và các biến số đóng vai trò quan trọng nhất trong dự đoán. Từ đó, nhóm tác giả cho rằng nghiên cứu này là cần thiết và có những đóng góp về khoa học và thực tiễn cho Ngân hàng Á Châu. Bên cạnh đó, kết quả phân tích này cũng là nguồn tham khảo cho các bên liên quan như Ngân hàng thương mại khác và Ngân hàng nhà nước.

## **2. Cơ sở lý thuyết và các nghiên cứu trước**

### **2.1. Tín dụng cá nhân, rủi ro tín dụng, và xác suất vỡ nợ**

Theo Điều 20 Luật các tổ chức tín dụng 2010 quy định: Cấp tín dụng là việc ngân hàng “thỏa thuận để tổ chức, cá nhân sử dụng một khoản tiền hoặc cam kết cho phép sử dụng một khoản tiền theo nguyên tắc có hoàn trả bằng nghiệp vụ cho vay, chiết khấu, cho thuê tài chính, bao thanh toán, bảo lãnh ngân hàng và các nghiệp vụ cấp tín dụng khác” (Quốc hội nước Cộng hòa Xã hội Chủ nghĩa Việt Nam, 2010, tr. 12). Trên cơ sở định nghĩa “tín dụng ngân hàng”, tín dụng khách hàng cá nhân có thể được hiểu là hình thức tín dụng mà ở đó ngân hàng đóng vai trò là người chuyển nhượng quyền sử dụng vốn của mình cho KHCN sử dụng trong một thời gian nhất định và phải hoàn trả cả gốc lẫn lãi.

Rủi ro tín dụng sẽ phát sinh khi các cá nhân được phê duyệt các khoản vay. Theo BCBS (2000) rủi ro tín dụng (credit risk) được định nghĩa là người vay có khả năng không đáp ứng các nghĩa vụ của mình theo các điều khoản đã thỏa thuận. Theo khoản 24 Điều 2 Thông tư 41/2016/TT-NHNN giải thích chi tiết hơn: “Rủi ro tín dụng là rủi ro do khách hàng không thực hiện hoặc không có khả năng thực hiện một phần hoặc toàn bộ nghĩa vụ trả nợ theo hợp đồng hoặc thỏa thuận với ngân hàng, chi nhánh ngân hàng nước ngoài” (Ngân hàng Nhà nước Việt Nam, 2016, tr. 02).

Rủi ro tín dụng một khi được lượng hóa sẽ hình thành nên xác suất vỡ nợ (Probability of Default - PD). Xác suất vỡ nợ là một thành phần quan trọng được áp dụng trong nhiều phân tích rủi ro tín dụng và hoạt động quản lý rủi ro. Xác suất vỡ nợ có thể được ước tính từ hai nhóm phương pháp chính: (i) từ hoạt động chấm điểm tín dụng trước khi vay nợ thông qua các mô hình chấm điểm tín dụng và (ii) từ các mô hình xác suất và học máy. Hoạt động chấm điểm tín dụng là một phương pháp truyền thống được các ngân hàng áp dụng thông qua các mô hình chấm điểm tín dụng khác nhau để hỗ trợ ngân hàng ra quyết định trước khi cho vay (Walusala & ctg., 2017). Tuy nhiên, phương pháp này tồn tại một số nhược điểm như không đánh giá đầy đủ khả năng trả nợ hoặc không cải thiện được khả năng học hỏi từ dữ liệu thực tế (Teles & ctg., 2020). Chính vì vậy, việc tích hợp các mô hình xác suất và học máy là cần thiết để cải thiện quá trình lượng hóa xác suất vỡ nợ (Teles & ctg., 2020).

### **2.2. Lược khảo các nghiên cứu trước**

Việc lược khảo các nghiên cứu trước sẽ giúp ích cho nhóm tác giả khi phát hiện ra những mô hình dự đoán được sử dụng phổ biến và hiệu quả. Đồng thời, các nghiên cứu trước cũng giúp ích trong quá trình xây dựng các yếu tố đầu vào của mô hình dự đoán. Khi dẫn chiếu các nghiên cứu trước đây, nhóm tác giả sẽ tập trung vào các nghiên cứu áp dụng kỹ thuật mới, được trích dẫn trong nhiều nghiên cứu và có đầy đủ thông tin trong quá trình thực hiện xây dựng mô hình.

Một trong các nghiên cứu đáng chú ý là của Chang và cộng sự (2022). Các tác giả sử dụng dữ liệu từ 282,763 khoản vay, chọn ra 16 biến quan trọng, và so sánh các mô hình LR, SVM, DT, RF, XGBoost, và LightGBM, trong đó XGBoost có hiệu suất tốt nhất với độ chính xác khoảng 88%. Giang (2021) sử dụng bộ dữ liệu Kaggle để đánh giá xác suất vỡ nợ qua các mô hình KNN, CART, NB, SVM, DNN, CNN, RNN, và LSTM, trong đó LSTM và RNN cho

kết quả tốt nhất. Jepkoech (2020) áp dụng các phương pháp LDA, GB, DT, KNN, Random Forest, SVB, và XGB trên dữ liệu tín dụng của Rwanda, kết quả cho thấy XGBoost vượt trội với Recall = 0.9939, F1 = 0.9892 và độ chính xác 99.56%.

Qua quá trình lược khảo, chúng tôi nhận thấy mô hình được sử dụng nhiều nhất trong các nghiên cứu để đánh giá xác suất vỡ nợ của KHCCN bao gồm: *Decision Tree (DT - C5.0)*, *Support Vector Machines (SVM)*, *Random Forest (RF)*, *K-Nearest Neighbors (KNN)*, *Logistic Regression (GLM)*. Ngoài ra có các phương pháp ít phổ biến hơn là: *Naive Bayes (NB)*, *Neural Network (NN)*, *Linear Discriminant Analysis (LDA)*, *Gradient Boosting (GB)* và *Deep Learning (DL)*.

### 3. Phương pháp nghiên cứu

#### 3.1. Xác định biến số phụ thuộc (biến số được dự đoán - Predicted variable)

Trong phần viết này nhóm nghiên cứu sẽ mô tả hai nội dung: xác định cách thức đo lường và thời gian quan sát biến phụ thuộc.

##### 3.1.1. Xác định đo lường

Biến số phụ thuộc có thể được xác định theo một trong ba cách:

- *Cách tiếp cận 1*: Tham chiếu BCBS (2006), Basel có đưa ra hướng dẫn rằng một khách hàng được coi là vỡ nợ nếu khách hàng bị quá hạn trên 90 ngày.
- *Cách tiếp cận 2*: Khoản 8 Điều 3 Thông tư 11/2021/TT-NHNN ban hành ngày 30/07/2021 “Nợ xấu là nợ thuộc các nhóm 3, 4 và 5 theo phân loại nợ” (Ngân hàng Nhà nước Việt Nam, 2021, tr. 05).
- *Cách tiếp cận 3*: Siddiqi (2006) hướng dẫn phân tích Roll Rate.

Do mẫu dữ liệu không đủ lớn để thực hiện phân tích Roll Rate nên trong nghiên cứu này nhóm tác giả sẽ xây dựng định nghĩa Good/Bad theo cách tiếp cận 1 và cách tiếp cận 2 được nêu ở trên, cụ thể như sau:

- Định nghĩa quan sát Bad (vỡ nợ): (i) Phát sinh trễ hạn từ 90 ngày trở lên hoặc (ii) Phát sinh nợ cơ cấu gia hạn và/hoặc dời kỳ.
- Định nghĩa quan sát Good (không vỡ nợ): là những quan sát không Bad.

##### 3.1.2. Xác định thời gian quan sát

Thời gian quan sát là khoảng thời gian quan sát hành vi của khách hàng để xác định Vỡ nợ/Không vỡ nợ. Có một số cách tiếp cận để xác định thời gian quan sát như:

*Cách tiếp cận 1*: Tham chiếu BCBS (2006), mô hình xác suất vỡ nợ có kỳ đánh giá là 12 tháng.

*Cách tiếp cận 2*: Theo Siddiqi (2006), xác định thời gian quan sát cần thông qua phân tích Vintage.

Trong nghiên cứu này do những giới hạn về dữ liệu nên tác giả sẽ sử dụng cách tiếp cận 1 với thời gian quan sát biến phụ thuộc là 12 tháng.

#### 3.2. Thu thập dữ liệu

Nhóm tác giả sử dụng bộ dữ liệu có thời gian lấy mẫu từ 01/01/2022 đến 31/12/2022 và thời gian quan sát từ 01/01/2023 đến 31/12/2023. Bộ dữ liệu cuối cùng sử dụng xây dựng và kiểm định mô hình bao gồm 2,917 quan sát, trong đó có 2,766 quan sát tốt, chiếm tỷ lệ 94.82% và 151 quan sát xấu, chiếm tỷ lệ 5.18%.

**Bảng 1***Bộ Dữ Liệu Xây Dựng và Kiểm Định Mô Hình*

<b>Định nghĩa</b>	<b>Số lượng quan sát</b>	<b>Tỷ lệ %</b>
Tốt	2,766	94.82%
Xấu	151	5.18%
<b>Tổng</b>	<b>2,917</b>	<b>100%</b>

*Nguồn:* Tổng hợp của nhóm tác giả

Mẫu dữ liệu tác giả thu thập được như Bảng 1 bị tình trạng không cân bằng (imbalance) khi chỉ có 5.18% số quan sát xấu trên tổng số dữ liệu. Để cân bằng dữ liệu, nhóm tác giả sẽ sử dụng phương pháp SMOTE và BOTH để cân bằng dữ liệu và so sánh kết quả mô hình với tập dữ liệu gốc (xem Bảng 2).

**Bảng 2***Sampling Mẫu Xây Dựng*

<b>Phương pháp</b>	<b>Tỷ lệ Xấu/Tốt</b>	<b>Tốt</b>	<b>Xấu</b>
Phương pháp SMOTE	1:1	2,005	1,998
Phương pháp Under	Không thực hiện được do quan sát xấu quá ít		
Phương pháp BOTH	1:1	1,031	1,011

*Nguồn:* Tổng hợp của nhóm tác giả

### **3.3. Xác định tập dữ liệu huấn luyện và tập dữ liệu kiểm định**

Với mục tiêu kiểm định mô hình cho kết quả khách quan nhất, nhóm tác giả sẽ tiến hành chia tập dữ liệu ra làm tập huấn luyện (Train) chiếm 70% và tập kiểm định (Test) chiếm 30% (xem Bảng 3). Dữ liệu sẽ được chuẩn hóa trước khi phân tích.

**Bảng 3***Mẫu Dữ Liệu Xây Dựng và Kiểm Định Mô Hình*

<b>Bộ dữ liệu</b>	<b>Tập dữ liệu</b>	<b>Good</b>	<b>Bad</b>
Bộ dữ liệu gốc	Tập huấn luyện (Train)	1,936	105
	Tập kiểm định (Test)	830	46
	<b>Tổng</b>	<b>2,766</b>	<b>151</b>
Bộ dữ liệu SMOTE	Tập huấn luyện (Train)	1,403	1,399
	Tập kiểm định (Test)	602	599
	<b>Tổng</b>	<b>2,005</b>	<b>1,998</b>
Bộ dữ liệu BOTH	Tập huấn luyện (Train)	722	708
	Tập kiểm định (Test)	309	303
	<b>Tổng</b>	<b>1,031</b>	<b>1,011</b>

*Nguồn:* Tổng hợp của nhóm tác giả

### 3.4. Xác định biến số dự đoán (Predictor variables)

Tác giả tiến hành xây dựng tập hợp biến dự đoán thông qua ba nguồn thông tin như sau:

- Nguồn 1: Nghiên cứu định tính được thực hiện thông qua các bước sau: Đầu tiên, nhóm nghiên cứu thực hiện phỏng vấn 14 chuyên gia phê duyệt tại Hội sở và 04 chuyên gia phân tích tín dụng có nhiều kinh nghiệm trong phân tích hồ sơ tín chấp. Tất cả các thông tin để đánh giá khách hàng theo tờ trình tín dụng theo chuyên gia sẽ là thông tin biến đầu vào cho mô hình.

- Nguồn 2: Tổng hợp từ các nghiên cứu có liên quan trước đây.

- Nguồn 3: Nhóm tác giả tạo ra các biến phái sinh là các biến tỷ lệ có thể chia tỷ lệ được từ nguồn 1 và 2.

Kết quả, có 68 biến số được đưa vào danh mục. Để lựa chọn biến dự đoán, nhóm tác giả thực hiện tính toán bằng chứng có trọng số (Weight Of Evidence - WOE) và phân tích thông tin giá trị của biến số (Information Value - IV) theo đề xuất của Siddiqi (2006). Theo Siddiqi (2006), IV của một biến số cần lớn hơn 0.02. Sau khi phân tích, có 28 biến số thỏa mãn điều kiện vừa nêu và được trình bày tại Bảng 4.

**Bảng 4**

*Information Value (IV) của Biến Số*

STT	Biến dự đoán	Ý nghĩa	IV
1	gioitinh	Giới tính	0.40
2	NCTN	Nhu cầu/Thu nhập	0.32
3	Nganhnghekinhdoanh	Ngành nghề kinh doanh	0.32
4	Nghenghiiep	Nghề nghiệp	0.26
5	Loaihinhthochuc	Loại hình công ty nơi khách hàng công tác	0.24
6	NPTTgltctn	Nợ phải trả hàng tháng/Thời gian liên tục có thu nhập	0.24
7	Chuevu	Chức vụ	0.23
8	NCthoihan	Nhu cầu/Thời hạn vay	0.22
9	Tuoi	Tuổi	0.21
10	Tongthunhap	Tổng thu nhập	0.20
11	Tongchiph	Tổng chi phí	0.20
12	Thoihan	Thời hạn vay	0.19
13	Trinhdohocvan	Trình độ học vấn	0.18
14	Nhucaucaptindung	Số tiền vay	0.18
15	Nophaitrahangthang	Nợ phải trả hàng tháng	0.16
16	cptn	Chi phí/Thu nhập	0.16
17	TNTglientuc	Thu nhập/Thời gian liên tục có thu nhập	0.15
18	Tglientuccothunhap	Thu nhập/Thời gian liên tục có thu nhập	0.15
19	NCTuoi	Nhu cầu/Tuổi	0.14

STT	Biến dự đoán	Ý nghĩa	IV
20	Loaihoppdong	Loại hợp đồng lao động	0.13
21	tglttgct	Thời gian liên tục có thu nhập/Thời gian công tác	0.13
22	NPTTN	Nợ phải trả hàng tháng/Thu nhập	0.12
23	Tgct	Thời gian công tác	0.09
24	Tinhtrangsohuu	Tình trạng sở hữu	0.07
25	NPTcp	Nợ phải trả hàng tháng/Chi phí	0.06
26	Lichsutindung6thang	Lịch sử tín dụng 06 tháng	0.03
27	Tinhtranghonnhan	Tình trạng hôn nhân	0.03
28	Hinhthucnhanthunhap	Hình thức nhận thu nhập	0.02

Nguồn: Tổng hợp của nhóm tác giả

### 3.5. Các mô hình học máy được áp dụng

Dựa trên tổng hợp các nghiên cứu trước đây như đã nêu tại Bảng 1, tác giả sẽ tiến hành thực hiện chạy mô hình phổ biến nhất trong xây dựng hệ thống xếp hạng là Logistic Regression (GLM), đồng thời tác giả sẽ sử dụng thêm 08 phương pháp Machine Learning thường dùng khác để so sánh khả năng dự báo của mô hình. Sau đó, nghiên cứu sẽ áp dụng mô hình kết hợp “Ensemble”. Đây là một kỹ thuật trong học máy dùng để cải thiện độ chính xác của dự đoán bằng cách kết hợp nhiều mô hình học máy đơn lẻ thành một mô hình tổng hợp. Các phương pháp học máy bao gồm: Adaboost, Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest, Neural Network, XGBoost, và Decision Tree.

### 3.6. Phương pháp đánh giá chất lượng mô hình dự đoán

Các tiêu chí để đánh giá mô hình phân loại thường bao gồm AUC/GINI, KS, và ma trận nhầm lẫn (Confusion Matrix).

#### Chỉ số AUC và đường cong ROC

ROC Curve (The Receiver Operating Characteristic Curve) là một đường cong biểu diễn hiệu suất phân loại của một mô hình phân loại tại các ngưỡng (threshold). Về cơ bản, nó hiển thị True Positive Rate (TPR) so với False Positive Rate (FPR) đối với các giá trị ngưỡng khác nhau.

AUC (Area Under the Curve) là chỉ số được tính toán dựa trên đường cong ROC nhằm đánh giá khả năng phân loại của mô hình tốt như thế nào. Phần diện tích nằm dưới đường cong ROC và trên trục hoành chính là AUC, có giá trị nằm trong khoảng [0, 1]. Theo Hosmer và cộng sự (2013) và Siddiqi (2006), mô hình có AUC  $\geq 70\%$  là đạt yêu cầu. Mối quan hệ của AUROC và AR được thể hiện qua công thức  $GINI = 2 * AUC - 1$ .

#### Chỉ số KS (Kolmogorov-Smirnov test)

KS đo lường mức độ chênh lệch lớn nhất giữa tỷ lệ % lũy kế của khách hàng tốt và khách hàng xấu. Càng nhiều khách hàng xấu phân bố ở các dải điểm thấp và càng nhiều khách hàng tốt phân bố ở các dải điểm cao thì càng cho thấy mô hình có khả năng phân biệt khách hàng tốt.

#### Ma trận nhầm lẫn (Confusion Matrix)

Confusion Matrix đưa ra thống kê số lượng quan sát vỡ nợ/không vỡ nợ được dự báo đúng/sai, từ đó tính toán độ chính xác của dự báo. Mỗi ngưỡng cắt (threshold) của xác suất vỡ nợ

sẽ được xác định để chia khách hàng thành hai nhóm vỡ nợ/không vỡ nợ. Theo mặc định, ma trận nhầm lẫn sẽ được thực hiện khi ngưỡng cắt (threshold) là 0.5. Nghĩa là, một quan sát được dự đoán là Bad (vỡ nợ) nếu xác suất vỡ nợ lớn hơn 50% và ngược lại. Tuy nhiên, ngưỡng 0.5 không phải là ngưỡng cắt tốt nhất trong các trường hợp đánh giá chất lượng của mô hình dự đoán (Alam & ctg., 2020). Một trong các phương pháp ưa dùng để tìm ra ngưỡng cắt đó là áp dụng chỉ số Youden Index (J) do Youden (1950) đề xuất và được áp dụng rộng rãi trong y khoa. Mục tiêu của chỉ số này là giúp nhà nghiên cứu tìm ra độ nhạy (Sensitive) và độ chuyên (Specificity) cao nhất. Tuy nhiên, ưu tiên của ngân hàng là dự đoán Bad nên ngưỡng cắt nào đem lại tỷ lệ dự đoán Bad tốt nhất trong tổng số Bad thực tế (chỉ số Sensitivity). Alam và cộng sự (2020) đưa ra khuyến nghị chọn ngưỡng cắt là điểm gần nhất với đỉnh bên trái của đường ROC. Tuy vậy, Zou và cộng sự (2016) đã khuyến nghị rằng việc sử dụng ROC là chưa đủ và có nhiều giới hạn. Chúng tôi sẽ thực hiện theo hướng dẫn của Zou và cộng sự (2016) và Boehmke và Greenwell (2019) để tìm ra ngưỡng cắt tốt nhất trong việc dự đoán Bad.

Một khi ngưỡng cắt được thiết lập, ma trận sẽ được hình thành để đo lường khả năng dự báo như Hình 1.

## Hình 1

*Ma Trận Nhầm Lẫn*

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

*Nguồn:* Tổng hợp của nhóm tác giả

Trong đó:

- True Positive - TP: Số quan sát dự báo là vỡ nợ và thực tế vỡ nợ.
- True Negative - TN: Số quan sát dự báo là không vỡ nợ và thực tế không vỡ nợ.
- False Negative - FN: Số quan sát dự báo là không vỡ nợ và thực tế vỡ nợ.
- False Positive - FP: Số quan sát dự báo là vỡ nợ và thực tế không vỡ nợ.

Các chỉ số được tính toán từ Confusion Matrix để đánh giá mức độ hiệu quả của mô hình như sau: Accuracy, Recall hay Sensitivity, Precision, F1, và Specificity.

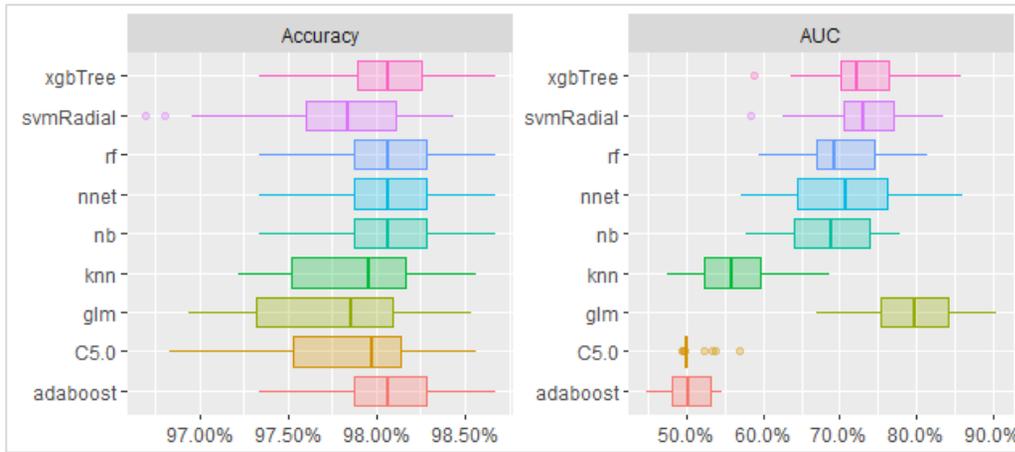
## 4. Kết quả nghiên cứu

### 4.1. Kết quả các mô hình dự đoán

Tác giả sử dụng ngôn ngữ R (giao diện Rstudio) với gói package Caret, do Kuhn và cộng sự (2022) phát triển, thực hiện chạy các mô hình Machine Learning với (i) tập dữ liệu gốc, (ii) tập dữ liệu sampling bằng phương pháp SMOTE, (iii) tập dữ liệu sampling bằng phương pháp BOTH. Kết quả các chỉ số mô hình được biểu diễn tại Hình 2, 3, và 4.

**Hình 2**

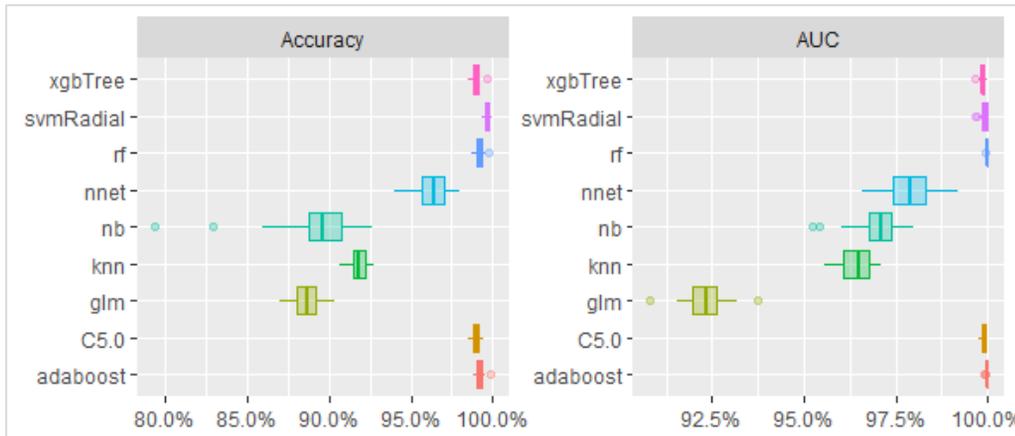
*Kết Quả Machine Learning Tập Dữ Liệu Gốc*



Nguồn: Tổng hợp của nhóm tác giả

**Hình 3**

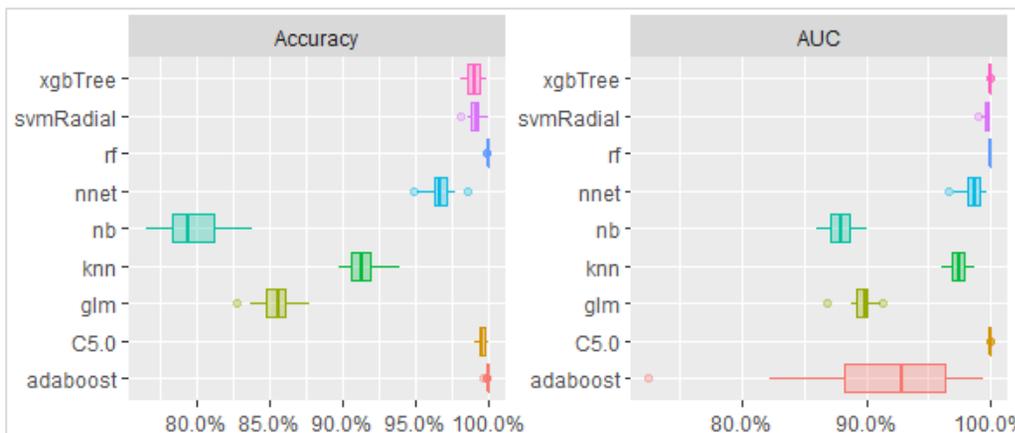
*Kết Quả Machine Learning Tập Dữ Liệu SMOTE*



Nguồn: Tổng hợp của nhóm tác giả

**Hình 4**

*Kết Quả Machine Learning Tập Dữ Liệu BOTH*



Nguồn: Tổng hợp của nhóm tác giả

Nhóm tác giả tiến hành thực hiện kiểm định lại mô hình trên tập Test. Với các tập dữ liệu xây dựng khác nhau kết quả AUC, KS của các phương pháp trên tập Test cũng thay đổi khác nhau như Bảng 5, cụ thể:

- Mô hình GLM trên dữ liệu gốc có chỉ số AUC cao nhất. GLM, NB, RF, XGBTree có kết quả dự báo trên tập Test ổn định (AUC  $\geq$  70%) trên cả bộ dữ liệu gốc, SMOTE, và BOTH.
- Dữ liệu gốc: GLM và NB dự báo tốt. Các mô hình ADABOOST, KNN, NN, SVM, C5.0 (DT) dự báo kém (AUC  $<$  70%).
- SMOTE: các mô hình tốt và đồng đều nhất, các mô hình đều đạt yêu cầu trừ C5.0 (DT). Điều này cũng tương đồng với các kết quả của các nghiên cứu trước đây.
- BOTH: NB có kết quả dự báo tốt nhất, các mô hình ADABOOST và KNN dự báo kém (AUC  $<$  70%).

## Bảng 5

Kết Quả AUC, KS Machine Learning trên Tập Test

Mô hình	AUC			KS		
	Dữ liệu gốc	SMOTE	BOTH	Dữ liệu gốc	SMOTE	BOTH
ADABOOST	52%	73%	47%	5%	42%	6%
GLM	85%	80%	77%	61%	55%	48%
KNN	59%	74%	61%	19%	52%	26%
NB	80%	77%	83%	62%	58%	63%
NNET	56%	72%	72%	17%	39%	37%
RF	76%	82%	79%	41%	61%	50%
SVM	66%	74%	76%	34%	49%	55%
C5.0 (DT)	50%	60%	71%	0%	23%	40%
XGBTREE	73%	70%	72%	42%	45%	42%

Nguồn: Tổng hợp tính toán của nhóm tác giả

### 4.2. Kết quả mô hình kết hợp - Ensemble

Từ kết quả các mô hình Machine Learning ở trên cho thấy mô hình xây dựng trên chính bộ dữ liệu gốc cho kết quả dự báo mô hình tốt hơn các phương pháp sampling dữ liệu (SMOTE và BOTH). Các mô hình GLM, NB, RF, và XGBTREE ổn định trên tập Test với cả 03 bộ dữ liệu. Do đó, nhóm tác giả sẽ tiến hành phân tích định hướng xây dựng mô hình kết hợp “Ensemble” từ kết quả xác suất vỡ nợ của 04 mô hình Logistic Regression (GLM), Naive Bayes (NB), Random Forest (RF), XGBoost (XGBTREE) trên tập dữ liệu gốc.

#### Phân tích tương quan và tầm quan trọng của các mô hình Machine Learning

Khi kết hợp kết quả đầu ra của các mô hình Machine Learning với nhau sử dụng kỹ thuật “Ensemble”, cần kiểm chứng mức độ tương quan của các mô hình này và tầm quan trọng của từng mô hình. Các nội dung vừa nêu được mô tả tại Bảng 6.

Kết quả cho thấy hệ số tương quan giữa NB-RF, XGBTREE-NB, và XGBTREE-RF đang rất cao. Hệ số tương quan cao nghĩa là các mô hình cho kết quả dự báo tương tự nhau, điều này sẽ làm giảm hiệu quả khi kết hợp các mô hình này với nhau.

**Bảng 6***Hệ Số Tương Quan và Tầm Quan Trọng của Các Mô Hình trong Mô Hình Kết Hợp “Ensemble”*

	GLM	RF	NB	XGBTREE	Tầm quan trọng
GLM	1				<b>100</b>
RF	0.658	1			<b>35</b>
NB	0.769	0.998	1		<b>16</b>
XGBTREE	0.757	0.993	0.994	1	<b>3</b>

Nguồn: Tổng hợp tính toán của nhóm tác giả

Xét ở mức độ quan trọng, trong cấu phần tạo nên mô hình kết hợp “Ensemble” thì mô hình GLM đóng vai trò quan trọng nhất, tiếp đó là mô hình RF, NB. Mô hình Xgbtree gần như không tham gia trong mô hình kết hợp.

Từ các phân tích tương quan và mức độ quan trọng của các mô hình tác giả sẽ sử dụng mô hình GLM và RF để xây dựng mô hình kết hợp “Ensemble”.

### **Kết quả mô hình kết hợp “Ensemble”**

Trên tập Train của dữ liệu gốc mô hình kết hợp “Ensemble” có chỉ số AUC là 0.6534, thấp hơn mô hình GLM (0.7929), RF (0.7036). Đối với chỉ số Accuracy, mô hình kết hợp “Ensemble” đạt 0.9802, thấp hơn mô hình RF (0.9804) và cao hơn mô hình GLM (0.9775). Tất cả các kết quả vừa nêu được mô tả tại Bảng 7.

**Bảng 7***Kết Quả AUC và Accuracy của Mô Hình GLM, RF, và Ensemble*

	AUC	Accuracy
GLM	0.7929	0.9775
RF	0.7036	0.9804
Ensemble	0.6534	0.9802

Nguồn: Tổng hợp tính toán của nhóm tác giả

### **Lựa chọn mô hình dự đoán**

Khi phân tích trên tập Test (Bảng 8), GLM và Ensemble cho kết quả AUC và KS tốt nhất. Vì vậy nhóm nghiên cứu lựa chọn GLM và Ensemble để thực hiện các phân tích tiếp theo.

**Bảng 8***Kết Quả AUC và KS của Mô Hình GLM, RF, Ensemble trên Tập Test*

Mô hình	AUC	KS
GLM	85%	61%
ENSEMBLE	84%	63%
RF	76%	41%

Nguồn: Tổng hợp tính toán của nhóm tác giả

Tiếp theo, nhóm tác giả tiến hành so sánh Confusion Matrix của 02 mô hình vì mục tiêu của Ngân hàng là phát hiện được các trường hợp Bad với ít sai lầm nhất. Tuy nhiên, nhóm

nghiên cứu sẽ không sử dụng ngưỡng cắt (threshold) mặc định là 0.5 mà tìm ra ngưỡng tối ưu cho từng mô hình theo hướng dẫn của Zou và cộng sự (2016) và Boehmke và Greenwell (2019). Kết quả cho thấy, ngưỡng cắt tối ưu của GLM và Ensemble lần lượt là 0.12 và 0.15.

Với ngưỡng cắt vừa nêu, mô hình GLM đã có kết quả vượt hơn so với Ensemble ở toàn bộ các chỉ số như: Accuracy, Sensitivity, Precision, và F1 (xem Bảng 9).

### Bảng 9

*Confusion Matrix của GLM và Ensemble trên Tập Test*

	GLM (Ngưỡng cắt PD = 12%)			Ensemble (Ngưỡng cắt PD = 15%)		
	Bad	Good	Dự đoán	Bad	Good	Dự đoán
Bad	30	19	39	20	67	87
Good	16	811	837	26	763	789
<b>Thực tế</b>	<b>46</b>	<b>830</b>	<b>876</b>	<b>46</b>	<b>830</b>	<b>876</b>
Accuracy	96%			89%		
Sensitivity	65%			43%		
Precision	77%			23%		
F1	70%			30%		

*Nguồn:* Tổng hợp tính toán của nhóm tác giả

Cụ thể hơn, với thực tế là có 46 trường hợp không có khả năng trả nợ, GLM dự đoán tốt nhất được 30 trường hợp (Sensitivity = 65%) trong khi đó Ensemble dự đoán được thấp hơn với 20 trường hợp đúng (Sensitivity = 43%).

#### 4.3. Vai trò của biến số dự đoán

Mặc dù sử dụng cùng một tập dữ liệu, hiệu quả dự đoán của các mô hình khá tương đồng nhau. Tuy nhiên, mức độ quan trọng của biến là khác nhau trong mỗi mô hình. Để phân tích mức độ quan trọng, nhóm tác giả sử dụng tính năng Variable Importance trong thư viện Caret. Sau đó, nhóm nghiên cứu sẽ xếp hạng theo thứ tự từ 01 đến 28 đối với từng mô hình và được tổng hợp tại Bảng 10.

### Bảng 10

*Vai Trò của Biến Số Dự Đoán*

STT	Biến	Ý nghĩa	Mô hình GLM	Mô hình Ensemble
1	Trinhdohocvan	Trình độ học vấn	1	19
2	Loaihinhthochuc	Loại hình tổ chức	2	2
3	gioitinh	Giới tính	3	7
4	tglttgct	Thời gian liên tục có thu nhập/Thời gian công tác	4	13
5	NPTTgltctn	Nợ phải trả hàng tháng/Thời gian liên tục có thu nhập	5	1

STT	Biến	Ý nghĩa	Mô hình GLM	Mô hình Ensemble
6	Thoihan	Thời hạn vay	6	5
7	NCTN	Nhu cầu/Thu nhập	7	4
8	Tuoi	Tuổi	8	11
9	Lichsutindung6thang	Lịch sử tín dụng 06 tháng	9	24
10	Tglientuccothunhap	Thời gian liên tục có thu nhập	10	3
11	NCthoihan	Nhu cầu/Thời hạn vay	11	6
12	TNTglientuc	Thu nhập/Thời gian liên tục có thu nhập	12	15
13	NPTTN	Nợ phải trả hàng tháng/Thu nhập	13	22
14	Chucvu	Chức vụ	14	17
15	Nghenghiiep	Nghề nghiệp	15	12
16	Tongthunhap	Tổng thu nhập	16	9
17	Tinhtranghonhan	Tình trạng hôn nhân	17	25
18	Tgct	Thời gian công tác	18	16
19	Hinhthucnhanthunhap	Hình thức nhận thu nhập	19	26
20	Loaihoptdong	Loại hợp đồng lao động	20	27
21	Tongchiphii	Tổng chi phí	21	10
22	Nhucaucaptindung	Số tiền vay	22	18
23	Tinhtrangsohuu	Tình trạng sở hữu	23	23
24	cptn	Chi phí/Thu nhập	24	20
25	NCTuoi	Nhu cầu/Tuổi	25	21
26	NPTcp	Nợ phải trả hàng tháng/Chi phí	26	28
27	Nganhnghekinhdoanh	Ngành nghề kinh doanh	27	14
28	Nophaitrahangthang	Nợ phải trả hàng tháng	28	8

*Nguồn:* Tổng hợp tính toán của nhóm tác giả

Từ những kết quả tại Bảng 10, các thông tin quan trọng cần thu thập khi thực hiện thu thập và phân tích hồ sơ tín dụng bao gồm: *Trình độ học vấn, Loại hình tổ chức, Giới tính, Tuổi, Thời gian liên tục có thu nhập, Thời gian công tác, Thời hạn vay, Nhu cầu vay, Tổng thu nhập, Tổng chi phí, Nợ phải trả hàng tháng, Lịch sử tín dụng 06 tháng gần nhất.*

#### 4.4. Thảo luận

Thông qua một tiến trình lựa chọn dữ liệu, biến đổi và làm sạch dữ liệu, và lựa chọn mô hình, có thể coi GLM và Ensemble là hai mô hình có chất lượng dự đoán tốt nhất. Mặc dù kết quả của nghiên cứu này có khác biệt so với một số nghiên cứu có áp dụng mô hình Ensemble như ElMasry và cộng sự (2019) hay Alam và cộng sự (2020) khi các tác giả phát hiện ra Ensemble có kết quả dự đoán tốt hơn so với GLM. Tuy vậy, sự khác biệt đó là bình thường vì thực tế là không có mô hình nào vượt trội trong tất cả các bối cảnh. Trong tình huống của ACB, mô hình trội hơn chính là GLM. Chi tiết hơn, khi sử dụng ngưỡng cắt xác suất vỡ nợ để có chỉ số Sensitivity tốt nhất thì GLM đã cho thấy kết quả vượt trội hơn ở chỉ số Sensitivity. Nghĩa là, GLM dự đoán đúng

số trường hợp vỡ nợ cao hơn 22% (= 65% - 23%). Theo sự hiểu biết của nhóm tác giả, các nghiên cứu trước đây liên quan tới cùng chủ đề nghiên cứu đã sử dụng ngưỡng xác suất vỡ nợ mặc định là 50% để so sánh tính chính xác của từng mô hình dự báo. Tuy nhiên, việc sử dụng ngưỡng mặc định là 50% là rất ít có ý nghĩa, đặc biệt là trong bối cảnh dự đoán xác suất vỡ nợ của ngân hàng. Các ngân hàng sẽ rất nhạy với dự đoán về Bad (tập trung vào chỉ số Sensitivity) hơn là tập trung vào tính chính xác tổng thể (Accuracy). Bằng việc lựa chọn ngưỡng xác suất phá vỡ nợ mà từ đó đạt được Sensitivity cao nhất, nghiên cứu này sẽ giúp cho ngân hàng thiết lập được ngưỡng xác suất phá sản khi dự đoán cho từng hồ sơ tín dụng cá nhân. Đây chính là ngưỡng cảnh báo mà ngân hàng cần có những phân tích sâu hơn để có thể đưa ra các quyết định có tiếp tục xử lý hồ sơ ở giai đoạn tiếp theo hay không. Cụ thể, khi ngân hàng đưa các thông tin của từng cá nhân vào mô hình GLM (đã được thiết lập), giá trị dự đoán xác suất phá sản sẽ được tính toán từ kết quả của mô hình. Nếu giá trị này thấp hơn 12%, ngân hàng sẽ thực hiện các thủ tục tiếp theo. Ngược lại, nếu giá trị vượt ngưỡng 12%, ngân hàng sẽ quyết định ngừng hồ sơ hoặc thu thập thêm thông tin để ra các quyết định tiếp theo. Ứng dụng vừa nêu cũng chính là một trong các đóng góp thực tiễn của đề tài mà các nghiên cứu thực nghiệm trước chưa đề cập tới.

Bên cạnh việc tìm ra mô hình dự đoán tốt nhất, nghiên cứu còn đánh giá tầm quan trọng của các biến số dự báo. Các biến số quan trọng được xác định bao gồm: *Trình độ học vấn, Loại hình tổ chức, Giới tính, Tuổi, Thời gian liên tục có thu nhập, Thời gian công tác, Thời hạn vay, Nhu cầu vay, Tổng thu nhập, Tổng chi phí, Nợ phải trả hàng tháng, Lịch sử tín dụng 06 tháng gần nhất*. Đây chính là các biến số mà ngân hàng nên cải thiện chất lượng thông tin được thu thập để giúp cho chất lượng mô hình dự đoán được cải thiện.

## **5. Kết luận, hàm ý quản trị, và hạn chế của nghiên cứu**

### **5.1. Kết luận**

Nghiên cứu này được thực hiện với mục đích là lựa chọn mô hình dự đoán xác suất vỡ nợ nhằm nâng cao năng lực quản trị rủi ro tín dụng khách hàng cá nhân vay tín chấp tại Ngân hàng TMCP Á Châu. Nhóm tác giả đã sử dụng 09 mô hình máy học bao gồm Logistic Regression, Adaboost, Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest, Neural Network, XGBoost, Decision Tree và 01 mô hình kết hợp Ensemble bằng cách sử dụng xác suất đầu ra của các mô hình máy học hiệu quả. Nhóm tác giả đã thực hiện tổng cộng 27 mô hình Machine Learning (09 phương pháp Machine Learning với 03 tập dữ liệu khác nhau) và 01 mô hình kết hợp Ensemble. Từ kết quả các mô hình Machine Learning ở trên cho thấy mô hình xây dựng trên chính bộ dữ liệu gốc cho kết quả dự báo mô hình tốt hơn các phương pháp sampling dữ liệu (SMOTE và BOTH). Các mô hình GLM, NB, RF, và XGBTREE ổn định trên tập Test với cả 03 bộ dữ liệu. Do đó, nhóm tác giả đã tiến hành phân tích định hướng xây dựng mô hình kết hợp “Ensemble” từ kết quả xác suất vỡ nợ của 04 mô hình GLM, NB, RF, và XGBTREE trên tập dữ liệu gốc. Thông qua các phân tích tương quan và mức độ quan trọng của các mô hình, mô hình kết hợp “Ensemble” đã được xây dựng từ mô hình GLM và RF. Như vậy, GLM và Ensemble đã được lựa chọn để thực hiện các phân tích tiếp theo. Từ các kết quả (i) phân tích AUC và Accuracy trên tập huấn luyện và (ii) AUC và KS trên tập kiểm tra, hai mô hình được lựa chọn cho phân tích dự đoán xác suất vỡ nợ là GLM và Ensemble.

Tiếp theo, nhóm tác giả tiến hành so sánh Confusion Matrix của 02 mô hình với ngưỡng cắt xác suất vỡ nợ tối ưu của GLM và Ensemble lần lượt là 0.12 và 0.15. Với ngưỡng cắt vừa nêu, mô hình GLM đã có kết quả vượt hơn so với Ensemble ở toàn bộ các chỉ số như: Accuracy, Sensitivity, Precision, và F1. Cụ thể, xét về dự đoán vỡ nợ, với thực tế là có 46 trường hợp không có khả năng trả nợ, GLM dự đoán tốt nhất được 30 trường hợp (Sensitivity = 65%) trong khi đó Ensemble dự đoán được thấp hơn với 20 trường hợp đúng (Sensitivity = 43%).

## 5.2. Hàm ý quản trị

Căn cứ vào kết quả dự báo, nghiên cứu đề xuất một số nội dung tới ngân hàng ACB và Ngân hàng nhà nước như sau:

### (i) Đối với Ngân hàng TMCP Á Châu

Nghiên cứu này đề xuất hai giải pháp như sau đối với ngân hàng:

#### **Giải pháp 1:** Ứng dụng mô hình vào quy trình tín dụng KHCN tín chấp

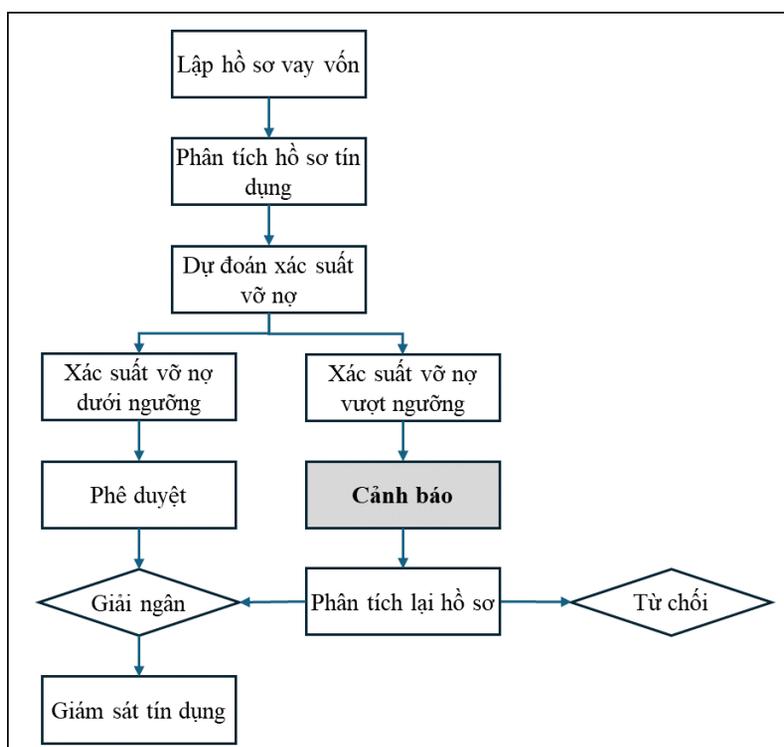
Xây dựng quy trình tín dụng có ứng dụng mô hình như Hình 5. Cụ thể: sau khi hồ sơ tín dụng đã được kiểm soát thông tin ở bước phân tích tín dụng, các thông tin này được sử dụng để mô hình (Mô hình Logistic và mô hình Ensemble) tính toán xác suất vỡ nợ. Các hồ sơ có xác suất vỡ nợ của vượt ngưỡng của một trong 02 mô hình (GLM và Ensemble lần lượt là 0.12 và 0.15), hệ thống sẽ cảnh báo đến cấp phê duyệt cần xem xét kỹ lại một lần nữa các hồ sơ này. Sau đó, cấp phê duyệt sẽ quyết định “từ chối” hoặc “giải ngân”.

Giải pháp này có thể đem lại một số lợi ích cụ thể như sau:

- Kết quả xác suất vỡ nợ của mô hình là một công cụ hỗ trợ để cảnh báo cho cấp phê duyệt tín dụng (cơ chế phê duyệt con người) các hồ sơ rủi ro cao theo phương pháp thống kê để xem xét hồ sơ kỹ hơn mà không làm ảnh hưởng đến quy trình tín dụng hiện tại.
- Ứng dụng này sẽ giúp các mô hình đã xây dựng được vận hành thực tế, có dữ liệu để đánh giá mô hình, so sánh với phê duyệt của con người, làm cơ sở cho việc áp dụng chính thức mô hình với các ứng dụng thực tế như ra quyết định phê duyệt tự động (thay cơ chế con người), áp dụng lãi suất cho vay, đánh giá rủi ro tín dụng thông qua mô hình theo phương pháp IRB của Basel 2 và Ngân hàng Nhà nước.

### Hình 5

Quy Trình Cấp Tín Dụng Tích Hợp Mô Hình Dự Đoán Xác Suất Vỡ Nợ



Nguồn: Đề xuất của nhóm tác giả

Tuy nhiên, để giải pháp trên có thể áp dụng được, ngân hàng cần chuẩn bị các công việc như sau:

- Xây dựng quy định, quy trình tín dụng và chương trình tin học ứng dụng mô hình đo lường xác suất vỡ nợ.
- Xây dựng quy định, quy trình và chương trình tin học xây dựng, vận hành, kiểm định mô hình để kiểm soát rủi ro mô hình.
- Thu thập dữ liệu các hồ sơ bị từ chối cấp tín dụng giúp tăng số lượng quan sát vỡ nợ khi xây dựng mô hình và kiểm định ngoại mẫu mô hình vì dữ liệu của ngân hàng luôn thiếu tính cân bằng giữa số quan sát vỡ nợ - không vỡ nợ theo tiêu chuẩn thống kê.
- Xem xét mua các sản phẩm chấm điểm tín dụng của bên thứ ba uy tín và có độ bao quát về dữ liệu nhiều hơn như Scoring hay eKYC của Trusting Scocial, CIC, PCB, ... để so sánh với kết quả của mô hình Ngân hàng tự xây dựng.
- Phải nâng cao khả năng bảo mật thông tin khách hàng và các thông tin tín dụng nội bộ khi ứng dụng mô hình.

### **Giải pháp 2:** Các thông tin của hồ sơ tín dụng cần thu thập chính xác

Ngân hàng cần xây dựng cơ chế thu thập chính xác các thông tin được tổng hợp từ các biến số quan trọng của mô hình như tác giả đã phân tích tại Mục 4.3, bao gồm các thông tin: *Tuổi, Trình độ học vấn, Loại hình tổ chức, Giới tính, Thời gian liên tục có thu nhập, Thời gian công tác, Thời hạn vay, Nhu cầu vay, Nợ phải trả hàng tháng, Tổng thu nhập, Tổng chi phí, Lịch sử tín dụng 06 tháng gần nhất.*

#### **(ii) Đối với Ngân hàng Nhà nước**

Thứ nhất, Ngân hàng Nhà nước cần xây dựng khuôn khổ pháp lý, quy định cụ thể về việc xây dựng, kiểm định hệ thống xếp hạng tín dụng nội bộ, đặc biệt là các mô hình Machine Learning. Đồng thời có quy định cụ thể về việc ứng dụng mô hình trong công tác thẩm định, phê duyệt cấp tín dụng để các ngân hàng ứng dụng đồng bộ và có kiểm soát.

Thứ hai, khi hệ thống xếp hạng tín dụng được ứng dụng trong hệ thống ngân hàng đòi hỏi Ngân hàng Nhà nước phải có đội ngũ am hiểu, có kiến thức về mô hình để thanh tra, kiểm tra tổ chức tín dụng.

### **Tài liệu tham khảo**

- Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *Ieee Access*, 8(1), 201173-201198.
- BCBS. (2000). *Principles for the management of credit risk*. <https://www.bis.org/publ/bcbs75.htm>
- BCBS. (2006). *International convergence of capital measurement and capital standards*. <https://www.bis.org/publ/bcbs128.htm>
- Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Chapman and Hall/CRC.
- Chang, A. H., Yang, L. K., Tsaih, R. H., & Lin, S. K. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data. *Journal*, 6(2), 303-325.

- Elmasry, W., Akbulut, A., & Zaim, A. H. (2019). Empirical study on multiclass classification-based network intrusion detection. *Computational Intelligence*, 35(4), 919-954.
- Giang, H. T. T. (2021). Một số kỹ thuật học máy cho chấm điểm tín dụng [Some machine learning techniques for credit scoring]. *Tạp chí Khoa học & Đào tạo Ngân hàng*, 227, 34-40.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Jepkoech, B. G. (2020). *Hybrid approach of machine learning algorithms in automated credit scoring prediction in banking sector: Case of a Rwandan commercial bank* [Doctoral dissertation, University of Rwanda]. <https://dr.ur.ac.rw/bitstream/handle/123456789/1461/Biwott%20Gloria%20Jepkoech.pdf?sequence=1&isAllowed=y>
- Kuhn, M., Wing, J., Weston, S., & Williams, A. (2022). *Package 'caret' - Classification and regression training*. <https://github.com/topepo/caret/>
- Ngân hàng Nhà nước Việt Nam. (2016). *Thông tư 41/2016/TT-NHNN ngày 30 tháng 12 năm 2016 quy định tỷ lệ an toàn vốn đối với ngân hàng, chi nhánh ngân hàng nước ngoài* [Circular No. 41/2016/TT-NHNN dated December 30, 2016 stipulating the capital adequacy ratio for banks and foreign bank branches]. <https://www.sbv.gov.vn>
- Ngân hàng Nhà nước Việt Nam. (2018). *Thông tư số 13/2018/TT-NHNN ngày 18 tháng 05 năm 2018 về hệ thống kiểm soát nội bộ của tổ chức tín dụng, chi nhánh ngân hàng nước ngoài* [Circular No. 13/2018/TT-NHNN dated May 18, 2018 on the internal control system of credit institutions and foreign bank branches]. <https://www.sbv.gov.vn>
- Ngân hàng Nhà nước Việt Nam. (2021). *Thông tư 11/2021/TT-NHNN ngày 30 tháng 07 năm 2021 Quy định về phân loại tài sản có, mức trích, phương pháp trích lập dự phòng rủi ro và việc sử dụng dự phòng để xử lý rủi ro trong hoạt động của tổ chức tín dụng, chi nhánh ngân hàng nước ngoài* [Circular No. 11/2021/TT-NHNN dated July 30, 2021, stipulating the classification of assets, the provisioning rates and methods for risk provisions, and the use of provisions to handle risks in the operations of credit institutions and foreign bank branches]. <https://chinhphu.vn/default.aspx?pageid=27160&docid=203811>
- Quốc hội nước Cộng hòa Xã hội Chủ nghĩa Việt Nam. (2010). *Luật các tổ chức tín dụng số 47/2010/QH12 ngày 16 tháng 06 năm 2010* [Law on Credit Institutions No. 47/2010/QH12 dated June 16, 2010]. <https://vanban.chinhphu.vn>
- Siddiqi, N. (2006). *Credit risk scorecards developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons.
- Teles, G., Rodrigues, J. J., Saleem, K., Kozlov, S., & Rabêlo, R. A. (2020). Machine learning and decision support system on credit scoring. *Neural Computing and Applications*, 32(1), 9809-9826.
- Walusala, S., Rimiru, R., & Otieno, C. (2017). A hybrid machine learning approach for credit scoring using PCA and logistic regression. *International Journal of Computer (IJC)*, 27(1), 84-102.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32-35.
- Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5(1), 2-8.

