ISSN 1859-3666
E-ISSN 2815-5726

MỤC LỤC

KINH TẾ VÀ QUẢN LÝ

QUẢN TRỊ KINH DOANH

ISSN 1859-3666
E-ISSN 2815-5726

ỨNG DỤNG CÁC THUẬT TOÁN MACHINE LEARNING TRONG THẨM ĐỊNH HỒ SƠ TÍN DỤNG TẠI NGÂN HÀNG

Phạm Thuỷ Tú
Trường Đại học Tài chính - Marketing
Email: pttu@ufm.edu.vn

Ngày nhận: *15/04/2024* ***Ngày nhận lại:*** *15/07/2024* ***Ngày duyệt đăng:*** *18/07/2024*

Ứng dụng các thuật toán Machine Learning thẩm định hồ sơ tín dụng được đánh giá mang lại nhiều thế mạnh trong xử lý dữ liệu tài chính. Nghiên cứu ứng dụng các thuật toán như Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, XGBoost mô phỏng khả năng phân loại hồ sơ tín dụng tại ngân hàng theo ba loại: tốt, xấu và đủ tiêu chuẩn. Kết quả thu được cho thấy Random Forest mang lại hiệu suất cao nhất với độ chính xác trên 92%; Naive Bayes, K-Nearest Neighbors, Decision Tree đạt hiệu suất dự đoán trên 80%; Logistic Regression và Support Vector Machine mang lại hiệu suất thấp (59% và 48%). Nhằm tăng cường tính phù hợp của dữ liệu đầu vào huấn luyện, nghiên cứu cũng sử dụng kết hợp một số kỹ thuật tiền xử lý dữ liệu như: tạo biến mới phù hợp với tiêu chí đánh giá từ bộ dữ liệu ban đầu, gán nhãn, xử lý giá trị ngoại lệ, phân tích lựa chọn đặc trưng tốt nhất, chuẩn hoá dữ liệu, cân bằng dữ liệu,... Kết quả cho thấy các kỹ thuật tiền xử lý dữ liệu cải thiện hiệu suất huấn luyện. Các kết quả thu được kỳ vọng có thể bổ sung thêm bằng chứng thực nghiệm đáng tin cậy cho các nghiên cứu khác có liên quan đến đề tài thẩm định hồ sơ tín dụng bằng các thuật toán machine learning.

Từ khóa*: Hồ sơ tín dụng, máy học (machine learning), ngân hàng, thẩm định tín dụng.*

JEL Classifications*: C53; E37; G17; G21; E27.*

DOI*: 10.54404/JTS.2024.192V.08*

1. Giới thiệu

Hoạt động thẩm định tín dụng đóng vai trò vô cùng quan trọng trong việc dự đoán và hỗ trợ đưa ra các quyết định hợp tác kinh doanh. Việc thẩm định hồ sơ tín dụng (HSTD) giúp đánh giá khả năng trả nợ của khách hàng là một trong những vấn đề quan trọng nhất đối với các tổ chức cho vay. Ngày nay, nhiều ngân hàng chú trọng việc phát triển nghiên cứu các giải pháp hỗ trợ thẩm định tín dụng theo định hướng ứng dụng công nghệ tài chính. Nhiều xu hướng công nghệ được ra đời từ nền tảng ứng dụng máy học, trí tuệ nhân tạo, dữ liệu lớn. Cùng với sự phát triển nhanh chóng của kỹ thuật học máy trong lĩnh vực khoa học máy tính, nhiều phương pháp khác nhau đã được đề xuất để tạo điều kiện thuận lợi cho việc thực hiện các phương pháp học máy để xác định đặc điểm hành vi trả nợ của khách hàng. Nhằm tăng tính chính xác, minh bạch trong việc thẩm định HSTD, nhiều nguồn thông tin được thu thập góp phần tăng cường hiệu quả cũng như tính minh bạch trong hoạt động thẩm định. Với một khối lượng lớn dữ liệu được thu thập từ nhiều nguồn khác nhau, việc xử lý tính toán theo phương trình

truyền thống có thể phức tạp, gây tốn nhiều chi phí thời gian và nhân lực. Trong bối cảnh đẩy mạnh công nghệ số và số hóa, việc thẩm định HSTD đặc biệt các gói dịch vụ hỗ trợ vay online cần nhanh chóng và tính chính xác cao. Hoạt động này có thể tiết kiệm rất nhiều chi phí, đồng thời thu hút được rất nhiều khách hàng tiềm năng thông qua các gói dịch vụ phù hợp với đa thành phần. Nhận thức được tầm quan trọng của các giá trị thông tin tiềm năng từ kho dữ liệu lớn (bigdata), các thuật toán trong máy học (Machine Learning, viết tắt là ML) được vận dụng vào quá trình phát triển công nghệ tài chính nằm tăng cường hiệu suất hoạt động trong quá trình thẩm định HSTD.

Trước đây, phần lớn các ngân hàng xây dựng mô hình thẩm định HSTD bằng hai phương pháp phổ biến: phương pháp chuyên gia truyền thống và phương pháp mô hình thống kê tần suất. Ngày nay, cùng với sự phát triển mạnh mẽ của công nghệ tài chính, các mô hình thẩm định tín dụng hiện đại ứng dụng các thuật toán ML đang được đánh giá rất cao và ứng dụng trong nhiều ngân hàng bởi những ưu điểm nổi bật như: *Thứ nhất,* kết quả thẩm định nhanh chóng do thời gian thực hiện của máy nhanh hơn con người; *Thứ hai*, năng suất thẩm định vượt trội so với phương pháp chuyên gia truyền thống. Trong cùng một khoảng thời gian, việc xử lý thẩm định HSTD bằng phương pháp ứng dụng máy học sẽ tiết kiệm được rất nhiều thời gian và công sức; *Thứ ba,* tiết kiệm thời gian, chi phí lao động rất nhiều trong các khâu của quy trình thẩm định bằng phương pháp chuyên gia truyền thống; *Thứ tư,* kết quả thẩm định là nhất quán dựa trên số điểm tín nhiệm được chấm bởi mô hình ứng dụng máy học. Đồng thời, thời gian trả kết quả cũng nhanh do các khâu hoàn toàn tự động thực hiện bằng máy, không gặp phải vấn đề bất đồng bộ khi có sự tham gia của nhiều đối tượng chuyên gia khác nhau; *Thứ năm,* mô hình được xây dựng có thể xem xét toàn diện các dữ liệu đầu vào và linh hoạt trong việc điều chỉnh tăng hay giảm các yếu tố đầu vào mà không cần phải lo lắng về vấn đề thời gian thực hiện dự báo; *Thứ sáu,* mô hình có ứng dụng các kỹ thuật máy học có thể linh hoạt trong việc thử nghiệm các kịch bản đánh giá khác nhau. Đồng thời có thể linh hoạt tích hợp thêm các nguồn thông tin đầu vào khác từ bigdata để tăng cường hiệu quả và tính chính xác cho mô hình dự báo so với phương pháp truyền thống. Ngoài ra, việc thẩm định bằng việc thu thập thông tin từ rất nhiều nguồn sẽ giúp gia tăng năng suất và hiệu quả cho công tác thẩm định rất nhiều. Mô hình công nghệ có thể hoạt động với công suất thay thế cho khối lượng công việc của hàng trăm chuyên gia.

Bài viết "Ứng dụng các thuật toán machine learning trong thẩm định hồ sơ tín dụng tại ngân hàng" của tác giả được thực hiện với kỳ vọng các kết quả có thể bổ sung thêm các bằng chứng thực nghiệm về hiệu quả của các thuật toán ML trong phát triển mô hình thẩm định HSTD. Trong phần tiếp theo, bài viết sẽ trình bày về cơ sở lý thuyết và một số kết quả nghiên cứu trước về việc ứng dụng các thuật toán ML trong xây dựng mô hình thẩm định HSTD. Sau đó, tác giả mô tả quá trình xây dựng mô hình và đánh giá kết quả thông qua chỉ số hiệu suất mô hình trên dữ liệu nghiên cứu. Cuối cùng, bài viết tổng kết và thảo luận các kết quả thực nghiệm thu được. Từ đó, đề xuất một số giải pháp phát triển tiếp theo để cải thiện hiệu suất mô hình huấn luyện bằng các thuật toán ML.

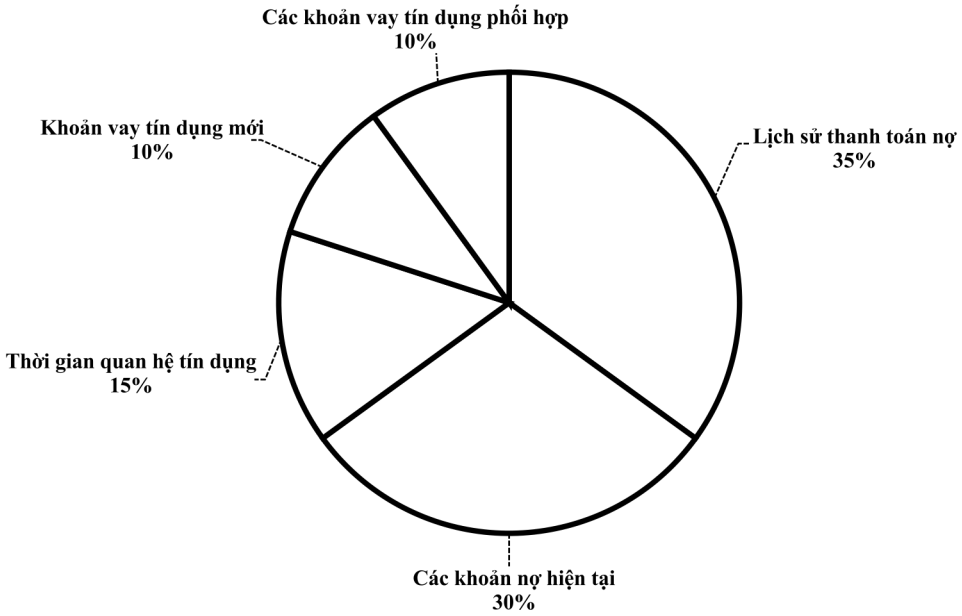
2. Tổng quan nghiên cứu và khuôn khổ lý thuyết

Anderson (2007) cho rằng thẩm định HSTD có thể được định nghĩa là "việc sử dụng các mô hình thống kê để chuyển đổi dữ liệu liên quan thành các thước đo số nhằm hướng dẫn các quyết định tín dụng. Đó là sự công nghiệp hóa của niềm tin; sự phát triển hợp lý trong tương lai của xếp hạng tín dụng chủ quan". Trong thực tế, mỗi tổ chức tài chính sẽ có một phương pháp hoặc tỷ

trọng chấm điểm, từ đó phân loại các HSTD. Theo Trung tâm thông tin tín dụng quốc gia Việt Nam (gọi tắt là Tổ chức CIC), điểm tín dụng của khách hàng sẽ được chấm dựa trên năm yếu tố chính, bao gồm: Lịch sử thanh toán nợ, các khoản nợ hiện tại, thời gian quan hệ tín dụng, khoản vay tín dụng mới và các khoản vay tín dụng. Để thẩm định một HSTD, kết quả thẩm định cho vay hay từ chối được xem xét trên bộ tiêu chí như sau:

- *Thời gian quan hệ tín dụng:* Là chỉ tiêu phản ánh thời gian tài khoản tín dụng được mở. Thời gian này càng dài thì càng được đánh giá cao, vì ngân hàng hay tổ chức đánh giá hành vi tài chính của khách hàng tổng thể và toàn diện hơn.

- *Tín dụng mới (10%):* Việc mở thêm các khoản tín dụng mới thường không được ưa chuộng, nhất là trong một thời gian ngắn. Các khoản tín dụng của khách hàng được mở càng lâu



(Nguồn: Tham khảo từ CIC.org (2022))

Biểu đồ 1: *Các yếu tố thành phần xét duyệt tín dụng*

Trong đó:

- *Lịch sử thanh toán nợ (35%):* Chỉ tiêu này phản ánh việc khách hàng trả tiền đúng hạn, có trả hết nợ hay trả nợ không đúng hạn hay không. Điểm tín dụng sẽ được tính dựa trên lịch sử thanh toán nợ, khách hàng trả nợ đúng thời hạn và nghiêm túc thì điểm tín dụng sẽ cao.

- *Các khoản nợ tín dụng (30%):* Phản ánh tất cả các món nợ, tỷ lệ nợ tín dụng được tạo nên từ tổng số các khoản vay mà ngân hàng cấp. Theo các chuyên gia, người có điểm số lý tưởng có xu hướng duy trì tỷ lệ nợ tín dụng ở mức trung bình khoảng 7%.

và có hoạt động trong ít nhất 6 tháng sẽ càng tăng điểm tín dụng và giúp họ xây dựng được một lịch sử tín dụng lâu dài, vững chắc.

- *Loại tín dụng (10%):* Tất cả các loại tín dụng khách hàng có như: thẻ tín dụng, các khoản vay (vay học phí, vay mua nhà, vay mua xe…). Các chuyên gia cho rằng việc từng sử dụng nhiều đòn bẩy tài chính và trả nợ đúng hạn cho thấy người đi vay có khả năng xử lý tốt các loại nợ tín dụng.

Sau khi xét duyệt hồ sơ và kiểm định các thông tin khách hàng cung cấp, các tổ chức tài chính sẽ căn cứ vào trọng số quy đổi thành điểm tín dụng của khách hàng và phân loại HSTD. Từ

đó, kết quả thẩm định HSTD hỗ trợ người cho vay đưa ra quyết định trong việc cấp tín dụng tiêu dùng. Những kỹ thuật này quyết định ai sẽ nhận được tín dụng, họ sẽ nhận được bao nhiêu tín dụng và chiến lược hoạt động nào sẽ nâng cao khả năng sinh lời của người đi vay đối với người cho vay. Các mô hình thẩm định HSTD đã mang lại nhiều thành công đặc biệt trong lĩnh vực tài chính và ngân hàng. Theo (Benton và cộng sự, 2005), phát triển công cụ thẩm định HSTD là việc sử dụng các mô hình thống kê để xác định khả năng người đi vay tiềm năng sẽ không trả được nợ. Một công cụ đánh giá rủi ro được vận hành theo phương thức tự động hoá với độ chính xác thì các đơn vị cho vay (tổ chức tín dụng) mới mở rộng khoản vay của họ một cách có hiệu quả (Thomas và cộng sự, 2002). Các nhà nghiên cứu và thực hành đã cố gắng phát triển các phương pháp đánh giá rủi ro tín dụng nhằm dự đoán khả năng đáp ứng nghĩa vụ của các doanh nghiệp và cá nhân. Những phương pháp này bao gồm phương pháp dựa trên chuyên gia, phương pháp phân tích thống kê tần suất và gần đây hơn là ứng dụng các phương pháp học máy.

Thẩm định tín dụng là một trong những lĩnh vực đầu tiên ứng dụng kỹ thuật học máy trong kinh tế. Một số nghiên cứu thực hiện trên các thuật toán trong học máy như: hồi quy Logistic - Logistic Regression (Crook và cộng sự, 2007; Finlay, 2011; Kruppa và cộng sự, 2013), cây quyết định - Decision tree và các biến thể cải tiến của nó (Coffman, 1986; Finlay, 2011; Vieira và cộng sự, 2019; Hué và cộng sự, 2018); K láng giềng - K-Nearest Neighbors (KNN) (Henley và Hand, 1997; Finlay, 2011; Kruppa và cộng sự, 2013; Mukid và cộng sự, 2018); Máy vectơ hỗ trợ - Support Vector Machine (SVM) (Baesens và cộng sự, 2003; Bellotti và Crook, 2009; Harris, 2015; Goh và Lee, 2019); XGBoost (Wang và cộng sự, 2022) hay Rừng ngẫu nhiên - Random Forest (Kruppa và cộng sự, 2013; Zhou và cộng sự, 2023).

Trong nghiên cứu về lượng hoá hiệu suất của các kỹ thuật trong máy học, Lessmann và các cộng sự (2015) đã so sánh 41 thuật toán với nhiều tiêu chí đánh giá khác nhau và một số bộ dữ liệu chấm điểm tín dụng. Kết quả nghiên cứu của họ chỉ ra rằng phương pháp rừng ngẫu nhiên mang lại hiệu suất phân loại vượt trội hơn so với hồi quy logistic và dần trở thành một trong những mô hình tiêu chuẩn trong ngành tính điểm tín dụng. Trong thập kỷ qua, các kỹ thuật học máy ngày càng được các ngân hàng và công ty phát triển công nghệ tài chính (fintech) sử dụng như các mô hình thách thức (ACPR, 2020) hoặc thường được kết hợp với dữ liệu "mới" (mạng xã hội hoặc truyền thông, dấu chân kỹ thuật số,…) từ nguồn "dữ liệu lớn" (Gambacorta và cộng sự, 2020; Kumar và cộng sự, 2021). Hiệu suất của các mô hình dựa trên máy học đã được cải thiện đáng kể từ khi áp dụng các phương pháp tổng hợp (tổng hợp), đặc biệt là các phương pháp đóng gói và tăng cường (Finlay, 2011; Lessmann và cộng sự, 2015). Nhiều nghiên cứu cho thấy các phương pháp tổng hợp luôn tốt hơn phương pháp hồi quy logistic vì phương pháp sau không phù hợp với các hiệu ứng phi tuyến tính này (Gambacorta và cộng sự, 2019). Đặc biệt, trong những năm gần đây, các nghiên cứu về ứng dụng phối hợp các kỹ thuật trong máy học vào xây dựng mô hình chấm điểm tín dụng được thực hiện đã cho các kết quả ngày càng dự đoán tốt hơn, cho kết quả nhanh chóng và chính xác hơn (Vieira và cộng sự, 2019; Teles và cộng sự, 2020; Assef và cộng sự, 2020; Bono và cộng sự, 2021; Machado và Karray, 2022).

Ngoài ra, để xây dựng các mô hình học tốt, các yếu tố dự đoán đầu vào quan trọng phải được chọn và bộ dữ liệu cần được tiền xử lý trước khi đưa vào huấn luyện với mô hình cụ thể. Trong quá trình huấn luyện dữ liệu với các mô hình, các nghiên cứu trước cũng cho thấy việc dữ liệu đầu vào đã được tiền xử lý bằng các kỹ thuật phù hợp sẽ góp phần đáng kể trong việc cải thiện hiệu suất

mô hình. Một số kỹ thuật thường được sử dụng như: làm sạch dữ liệu, mã hoá dữ liệu (LabelEncoder, StandardEncoder,...), cân bằng dữ liệu (SMOTE), chuẩn hoá dữ liệu (MinMaxScaler),... Có nhiều kỹ thuật khác nhau được sử dụng tuy nhiên việc cải thiện hiệu suất hay không còn tuỳ thuộc vào khả năng vận dụng linh hoạt các kỹ thuật phù hợp với từng đặc trưng của bộ dữ liệu nghiên cứu (Huang và cộng sự, 2015; Obaid và cộng sự, 2019; Shrawan, 2020; Gómez và cộng sự, 2024).

Nhìn chung, kết quả từ nhiều nghiên cứu trước cho thấy rằng nhiều kỹ thuật và phương pháp học máy được đưa vào ứng dụng trong quá trình xây dựng mô hình thẩm định HSTD. Tuy nhiên, mỗi phương pháp hay kỹ thuật sẽ phù hợp với đặc thù ngân hàng khác nhau. Vì vậy, việc xây dựng mô hình thẩm định HSTD dựa vào bộ tiêu chuẩn xét duyệt phù hợp với chiến lược phát triển của ngân hàng vô cùng quan trọng. Điều này có thể đóng vai trò quyết định trong việc giữ vững hay phá vỡ ổn định tài chính của ngân hàng. Chính vì vậy, trong nghiên cứu này chúng tôi sẽ lựa chọn các chỉ tiêu (dữ liệu đầu vào huấn luyện) dựa trên bộ tiêu chí xét duyệt tín dụng tại các ngân hàng thương mại hiện nay đang sử dụng.

2. Phương pháp và dữ liệu nghiên cứu

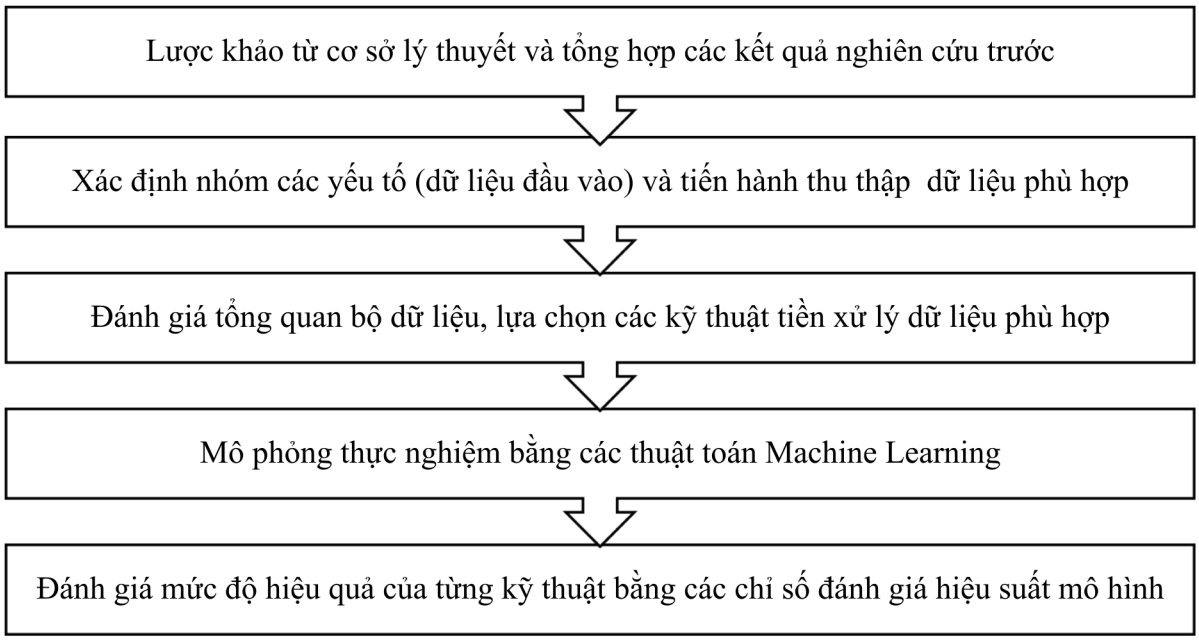
2.1. Phương pháp nghiên cứu

Dựa vào mục tiêu, đối tượng và phạm vi nghiên cứu đã đề xuất, tác giả sử dụng phương pháp định tính kết hợp định lượng. Theo đó, trình tự thực hiện nghiên cứu tiến hành như sau:

Ngôn ngữ lập trình Python và các thư viện hỗ trợ khai phá dữ liệu thống kê được tác giả chọn để mô phỏng trong xuyên suốt toàn bộ nghiên cứu.

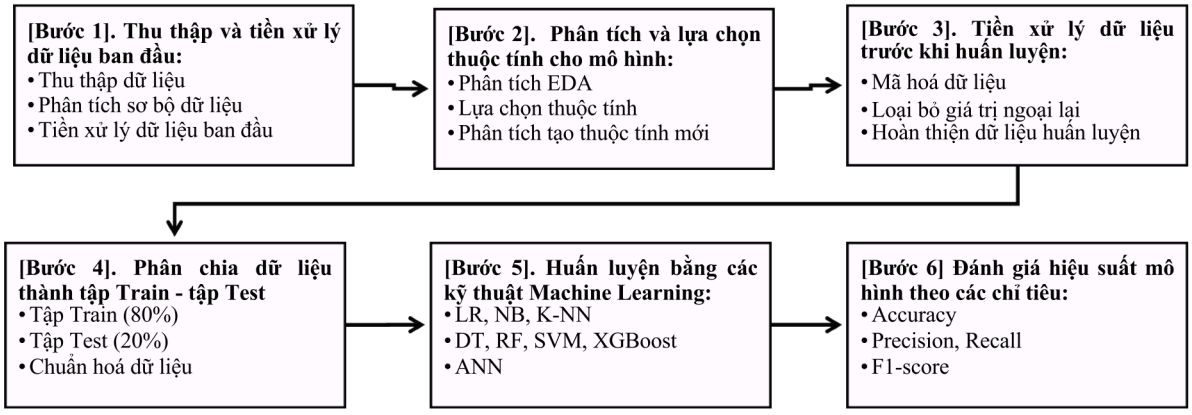
2.2. Dữ liệu nghiên cứu

Nghiên cứu sử dụng bộ dữ liệu từ UCI. Dữ liệu tín dụng từ UCI là một bộ dữ liệu tốt được sử dụng cho các cuộc thi và đã được các nhà nghiên cứu trên thế giới sử dụng rộng rãi trong các công trình thực nghiệm về thẩm định HSTD. Bộ dữ liệu này bao gồm thông tin của 150000 người vay vốn.



(Nguồn: Đề xuất bởi tác giả)

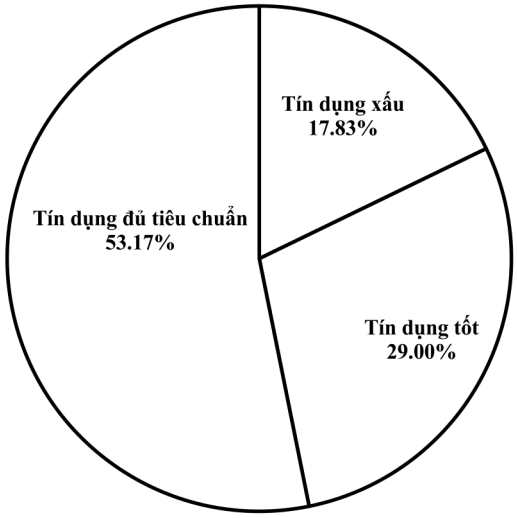
Hình 1: *Quy trình thực hiện nghiên cứu*



(Nguồn: Đề xuất bởi tác giả)

Hình 2: *Quy trình mô phỏng thực nghiệm huấn luyện dữ liệu*

Các bản ghi được gán nhãn tương ứng với ba phân lớp "tín dụng xấu" (nhãn 0.0), "tín dụng tốt" (nhãn 1.0) và "tín dụng đủ tiêu chuẩn" (nhãn 2.0, nhóm trung tính).



(Nguồn: Tông hợp bởi tác giả)

Biểu đồ 2: *Tỷ lệ phân loại hồ sơ tín dụng trên bộ dữ liệu ban đầu*

3. Kết quả nghiên cứu và các thảo luận

3.1. Kết quả phân tích khám phá các nhân tố của dữ liệu đầu vào trong mô hình huấn luyện

Như kết quả đã đề cập trước đó (Biểu đồ 2), bộ dữ liệu ban đầu có số lượng dòng giữa các phân lớp không đồng đều nhau, điều này có thể ảnh hưởng đến độ chính xác của mô hình dự đoán. Hai kỹ thuật cân bằng dữ liệu thường được sử dụng là UnderSampler và OverSampler. Trong nghiên cứu này, vì biến mục tiêu phân loại thành 3 phân lớn, có lớp trung tính vì vậy khi sử dụng UnderSampler thường dữ liệu thuộc lớp trung tính sẽ bị xoá bỏ. Để hạn chế rủi ro có thể mất mát thông tin trong các bộ hồ sơ bị xoá bỏ, tác giả ứng dụng kỹ thuật OverSampler tiến hành cân bằng dữ liệu để đảm bảo tính chính xác cao hơn trong quá trình huấn luyện mô hình (Biểu đồ 3). Hiện tượng

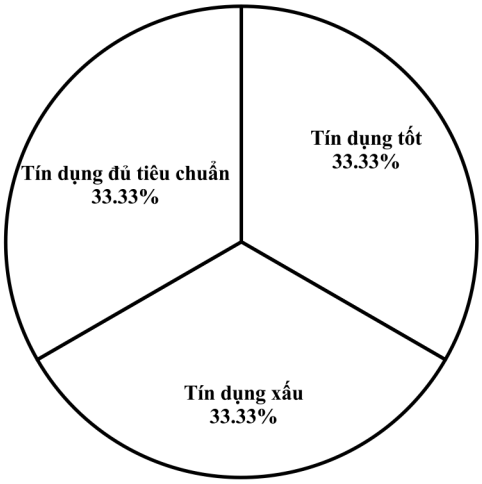
Bảng 1: *Bảng phân loại và mô tả thuộc tính của bộ dữ liệu nghiên cứu*

STT	Tiêu chí	Thuộc tính	Mô tả
1	Lịch sử thanh toán nợ	*Delay_from_due_date*	Số ngày bị trì hoãn kể từ ngày đến hạn thanh toán
		Num_of_Delayed_Payment	Số lần thanh toán bị trì hoãn
		Payment_of_Min_Amount	Cho biết số tiền thanh toán tối thiểu có được đáp ứng hay không
		Outstanding_Debt	Số tiền nợ tồn đọng
		Payment_Behaviour	Hành vi thanh toán của khách hàng
2	Các khoản nợ hiện tại	*Num_of_Loan*	Số khoản vay khách hàng có
		Type_of_Loan	Loại khoản vay mà khách hàng thực hiện
		Changed_Credit_Limit	Cho biết giới hạn tín dụng có bị thay đổi hay không
		Interest_Rate	Lãi suất áp dụng cho khoản vay
		Num_Credit_Inquiries	Số lượng yêu cầu tín dụng thực hiện
		Credit_Utilization_Ratio	Tỷ lệ tín dụng được sử dụng trên tín dụng hiện có
		Total_EMI_per_month	Tổng số tiền trả góp hàng tháng mà khách hàng thanh toán
3	Thời gian quan hệ tín dụng - Khoảng vay tín dụng mới	*Credit_History_Age*	Tuổi của lịch sử tín dụng
4	Các khoản vay tín dụng phối hợp	*Num_Bank_Accounts*	Số tài khoản ngân hàng khách hàng có
		Num_Credit_Card	Số thẻ tín dụng khách hàng có
		Credit_Mix	Kết hợp các loại tài khoản tín dụng khác nhau do khách hàng nắm giữ
5	Thông tin bổ sung của khách hàng	*Amount_invested_monthly*	Số tiền khách hàng đầu tư hàng tháng
		Monthly_Balance	Số dư hàng tháng trong tài khoản
		Monthly_Inhand_Salary	Lương hàng tháng sau khi trừ các khoản
		Num_Bank_Accounts	Số tài khoản ngân hàng có
		Customer_ID	Mã định danh mỗi khách hàng
		Name	Tên của khách hàng
		Age	Tuổi của khách hàng
		Occupation	Nghề nghiệp của khách hàng
		Annual_Income	Thu nhập hàng năm của khách hàng

(Nguồn: Tổng hợp và phân loại bởi tác giả)

dữ liệu bị nhiễu hoặc quá khớp có thể được kiểm soát thông qua kết quả độ phân bố hiệu suất dự báo (căn cứ vào kết quả accuracy, precision, recall, f1-score).

Khi quan sát và tiến hành phân tích bộ dữ liệu thu thập, tác giả phát hiện còn tồn tại một số vấn đề như: còn nhiều giá trị rỗng (chứa giá trị nan, null), có chứa giá trị ngoại lệ (outliers), cột không cần thiết trong quá trình huấn luyện (ID, Custommer_ID, Name, SNN), dữ liệu dạng chuỗi, lỗi chính tả và lỗi sai định dạng trong nhập liệu,... Vì vậy, trước khi đưa vào huấn luyện mô hình, tác giả tiến hành thực hiện các kỹ thuật làm sạch và tiền xử lý dữ liệu. Bên cạnh đó, để tăng



(Nguồn: Tổng hợp bởi tác giả)

Biểu đồ 3: *Tỷ lệ phân loại hồ sơ tín dụng sau khi áp dụng kỹ thuật cân bằng dữ liệu*

thông tin hữu ích cho việc đánh giá phân loại hồ sơ, một số chỉ tiêu mới được tính toán bổ sung dựa trên bộ dữ liệu gốc ban đầu. Các kỹ thuật xử lý dữ liệu được thực hiện cụ thể như sau (Bảng 2):

Bảng 2: *Các phương án đề xuất tiền xử lý dữ liệu*

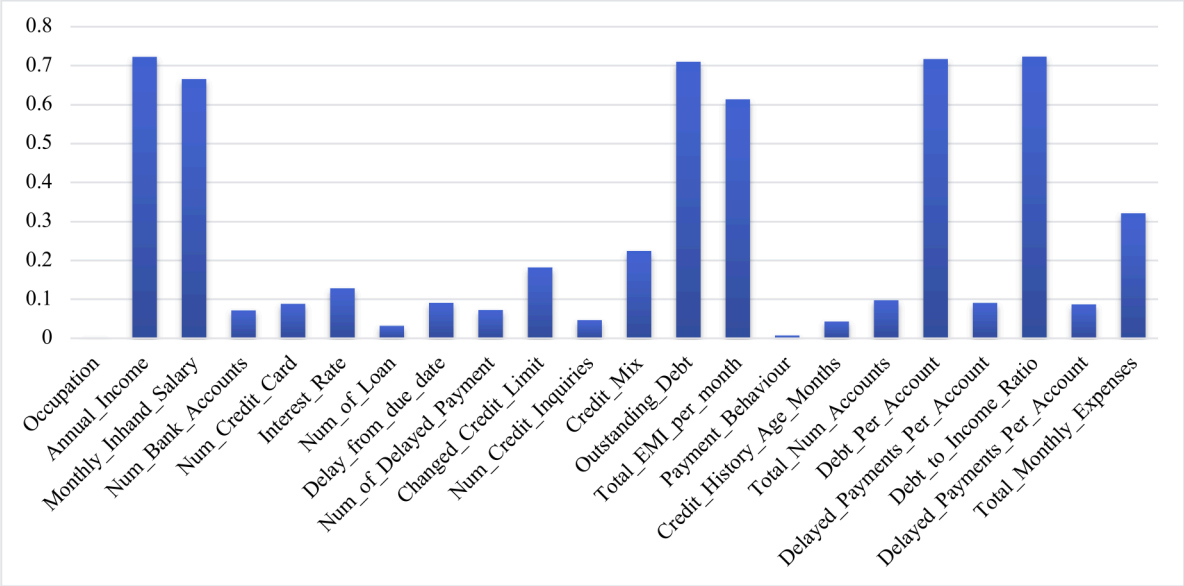
Bước	Phương pháp đề xuất xử lý	Kỹ thuật trong Python
1	Điền bổ sung dữ liệu khuyết bằng các giá trị mean (đối với dữ liệu dạng số) và mode (với dữ liệu dạng chuỗi) Xoá đi các dòng chứa dữ liệu không hợp lệ	Viết hàm xử lý trong Python
2	Gán nhãn cho các biến phân loại: Occupation, Credit_Mix, Payment_of_Min_Amount, Credit_Score	LabelEncoder()
3	Xử lý dữ liệu ngoại lệ bằng phương pháp thay thế giá trị ngoại lệ bằng các ngưỡng an toàn trong tứ phân vị	Viết hàm xử lý trong Python
4	Cân bằng dữ liệu theo ba phân lớp: tín dụng tốt, xấu và tiêu chuẩn	OverSampler()
5	Tạo ra các cột dữ liệu mới từ dữ liệu ban đầu: *1. Tính tổng số tài khoản* (['Total_Num_Accounts'] = ['Num_Bank_Accounts'] + ['Num_Credit_Card']) *2. Tính tổng dư nợ trên mỗi tài khoản* (['Debt_Per_Account'] = ['Outstanding_Debt']/['Total_Num_Accounts']) *3. Tính tỷ lệ nợ tồn đọng trên thu nhập hàng năm* (['Debt_to_Income_Ratio'] = ['Outstanding_Debt'] / ['Annual_Income']) *4. Tính tổng số khoản thanh toán chậm cho mỗi tài khoản* (['Delayed_Payments_Per_Account'] = ['Num_of_Delayed_Payment']/['Total_Num_Accounts']) *5. Tính tổng chi phí hàng tháng* (['Total_Monthly_Expenses'] = ['Total_EMI_per_month'] + ['Amount_invested_monthly'])	Viết hàm xử lý trong Python
6	Lựa chọn thuộc tính trong bộ dữ liệu đầu vào dựa trên phương pháp tính điểm tương quan MI Score (Mutual Information Score)	mutual_info_classif
7	Chuẩn hoá dữ liệu trong miền giá trị (0,1)	MinMaxScaler()
8	Chia tập huấn luyện và tập kiểm tra theo tỷ lệ 8:2	train_test_split()

(Nguồn: Đề xuất và tính toán bởi tác giả)

Để xác định các yếu tố có độ tương quan cao với biến mục tiêu, tác giả sử dụng phương pháp tính điểm tương quan Mutual Information Score. Đây là một kỹ thuật phổ biến được sử dụng trong các mô hình máy học. Phương pháp Mutual Information Score (MIS) là một phương pháp đo lường độ tương quan giữa hai biến ngẫu nhiên. Đối với bài toán lựa chọn đặc trưng trong học máy, MIS thường được sử dụng để đo lường mức độ quan trọng của mỗi đặc trưng đối với biến mục tiêu. Trong bối cảnh lựa chọn đặc trưng, MIS có thể được sử dụng để xác định đặc trưng nào chứa nhiều thông tin hơn về biến mục tiêu. Cụ thể, MIS tính toán độ tương quan giữa mỗi đặc trưng và biến mục tiêu bằng cách đo lường mức độ tin cậy của biến mục tiêu dựa trên thông tin từ đặc trưng đó. Đối với mỗi đặc trưng, MIS cao hơn đồng nghĩa với việc đặc trưng đó cung cấp nhiều thông tin hơn về biến mục tiêu.

Kết quả thu được khi đo lường mức độ tương quan giữa các biến đầu vào so với biến mục tiêu bằng phương pháp MI cho thấy các thuộc tính Payment_Behaviour, Num_of_Loan, Occupation không tác động nhiều đến kết quả phân loại (Biểu đồ 4). Vì vậy, các yếu tố này sẽ được loại ra khỏi bộ dữ liệu huấn luyện. Kết quả cũng cho thấy, các yếu tố như thu nhập hàng năm (Annual_Income), tổng dư nợ trên mỗi tài khoản (Debt_Per_Account), tỷ lệ nợ tồn đọng trên thu nhập hàng năm (Debt_to_Income_Ratio), số tiền nợ tồn đọng (Outstanding_Debt), lương hàng tháng sau khi trừ các khoản (Monthly_Inhand_Salary), tổng số tiền trả góp hàng tháng mà khách hàng thanh toán (Total_EMI_per_month) tương quan mạnh với biến mục tiêu.

Bộ dữ liệu dự kiến đưa vào huấn luyện sau khi tiền xử lý bao gồm 18 biến: biến mục tiêu (Credit_Score) và 17 biến đầu vào với các kết quả giá trị tổng quan được trình bày tại bảng thống kê mô tả (Bảng 3). Các biến lựa chọn đưa vào huấn luyện dựa trên kết quả đo lường MI Scores ở trình



(Nguồn: Tổng hợp và tính toán bởi tác giả trên Python)

Biểu đồ 4: *Kết quả đo lường điểm số tương quan theo phương pháp MI Scores*

Bảng 3: *Thống kê mô tả các biến trong bộ dữ liệu huấn luyện*

Thuộc tính	Số lượng mẫu	Giá trị trung bình	Độ lệch chuẩn	Giá trị nhỏ nhất	25%	50%	75%	Giá trị lớn nhất
Annual_Income (X1)	72.192	168.616,4	1.424.564,0	7.006,5	21.454,8	35.235,5	63.658,3	24.198.062,0
Monthly_Inhand_Salary (X2)	72.192	3.595,0	2.503,7	249,3	1.711,8	2.828,6	5.000,9	15.101,9
Num_Bank_Accounts (X3)	72.192	4,1	2,3	0,0	3,0	4,0	6,0	10,0
Num_Credit_Card (X4)	72.192	4,6	1,8	0,0	3,0	5,0	6,0	10,0
Interest_Rate (X5)	72.192	9,5	5,9	1,0	5,0	9,0	12,0	34,0
Delay_from_due_date (X6)	72.192	13,9	9,4	0,0	7,0	12,0	20,0	60,0
Num_of_Delayed_Payment (X7)	72.192	10,3	5,8	-3,0	6,0	10,0	15,0	29,8
Changed_Credit_Limit (X8)	72.192	8,2	4,8	0,0	4,4	7,9	11,2	27,0
Num_Credit_Inquiries (X9)	72.192	3,8	2,8	0,0	2,0	3,0	5,0	12,0
Credit_Mix (X10)	72.192	1,4	0,5	0,0	1,0	1,0	2,0	2,0
Outstanding_Debt (X11)	72.192	774,8	444,9	0,2	388,9	776,6	1.182,1	1.499,9
Total_EMI_per_month (X12)	72.192	57,8	52,0	0,0	17,5	44,7	85,7	199,9
Credit_History_Age_Months (X13)	72.192	247,9	109,4	0,0	199,0	262,0	334,0	404,0
Credit_Score (X14)	72.192	0,8	0,7	0,0	0,0	1,0	1,0	2,0
Total_Num_Accounts (X15)	72.192	8,8	3,3	1,0	7,0	9,0	11,0	20,0
Debt_Per_Account (X16)	72.192	108,3	111,0	0,0	45,9	88,4	135,5	1.496,6
Debt_to_Income_Ratio (X17)	72.192	0,0	0,0	0,0	0,0	0,0	0,0	0,2
Delayed_Payments_Per_Account (X18)	72.192	1,3	0,9	-1,5	0,8	1,2	1,6	15,0

(Nguồn: Tổng hợp và tính toán bởi tác giả trên Python)

bày trước đó (Biểu đồ 4). Kết quả từ thống kê mô tả cho thấy sai số và độ biến thiên dữ liệu lớn, đặc biệt ở các thuộc tính X1 (Annual_Income), X2 (Monthly_Inhand_Salary), X11 (Outstanding_Debt), X13 (Credit_History_Age_Months) và X16 (Debt_Per_Account) có giá trị chênh lệch rất lớn so với các yếu tố còn lại. Vì vậy, dữ liệu sẽ được chuẩn hoá trước khi đưa vào huấn luyện với các thuật toán ML như đã đề xuất trước đó.

Kết quả tính toán ma trận tương quan giữa các yếu tố đầu vào như sau (bảng 4):

Kết quả tính ma trận tương quan giữa các yếu tố trong mô hình huấn luyện (Bảng 4) cho thấy mức độ tương quan của các yếu tố phần lớn nằm trong ngưỡng an toàn (<0.6). Riêng biến Num_Credit_Card và Num_Bank_Accounts có độ tương quan cao với Total_Num_Accounts. Điều này có thể xảy ra bởi giá trị Total_Num_Accounts chính là tổng của hai giá trị Num_Credit_Card và Num_Bank_Accounts

Bảng 4: *Ma trận tương quan của các biến trong mô hình*

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
X1	*1,000*	0,033	-0,013	0,008	-0,015	-0,016	-0,008	-0,009	-0,018	-0,007	-0,003	0,014	-0,007	-0,003	-0,005	-0,003	-0,096	-0,009
X2	0,033	*1,000*	-0,122	-0,072	-0,113	-0,100	-0,130	-0,066	-0,104	-0,073	-0,049	0,440	0,081	0,006	-0,126	0,033	-0,514	-0,022
X3	-0,013	-0,122	*1,000*	0,255	0,400	0,423	0,493	0,238	0,272	0,382	0,087	0,002	-0,145	-0,127	0,853	-0,396	0,184	-0,161
X4	0,008	-0,072	0,255	*1,000*	0,259	0,297	0,273	0,112	0,200	0,111	0,091	0,033	-0,093	0,026	0,723	-0,375	0,149	-0,275
X5	-0,015	-0,113	0,400	0,259	*1,000*	0,401	0,408	0,282	0,383	0,356	0,162	0,044	-0,186	-0,048	0,426	-0,124	0,256	0,060
X6	-0,016	-0,100	0,423	0,297	0,401	*1,000*	0,429	0,189	0,304	0,224	0,150	0,036	-0,154	0,018	0,463	-0,144	0,227	0,052
X7	-0,008	-0,130	0,493	0,273	0,408	0,429	*1,000*	0,241	0,284	0,402	0,104	0,014	-0,159	-0,145	0,500	-0,196	0,200	0,520
X8	-0,009	-0,066	0,238	0,112	0,282	0,189	0,241	*1,000*	0,298	0,418	0,049	0,064	-0,164	-0,155	0,231	-0,095	0,105	0,046
X9	-0,018	-0,104	0,272	0,200	0,383	0,304	0,284	0,298	*1,000*	0,198	0,159	0,079	-0,242	0,028	0,303	-0,064	0,233	0,039
X10	-0,007	-0,073	0,382	0,111	0,356	0,224	0,402	0,418	0,198	*1,000*	-0,021	0,008	-0,112	-0,449	0,333	-0,197	0,029	0,105
X11	-0,003	-0,049	0,087	0,091	0,162	0,150	0,104	0,049	0,159	-0,021	*1,000*	0,002	-0,068	0,085	0,111	0,534	0,605	0,028
X12	0,014	0,440	0,002	0,033	0,044	0,036	0,014	0,064	0,079	0,008	0,002	*1,000*	-0,057	0,015	0,019	-0,014	-0,270	0,005
X13	-0,007	0,081	-0,145	-0,093	-0,186	-0,154	-0,159	-0,164	-0,242	-0,112	-0,068	-0,057	*1,000*	0,009	-0,154	0,043	-0,123	-0,029
X14	-0,003	0,006	-0,127	0,026	-0,048	0,018	-0,145	-0,155	0,028	-0,449	0,085	0,015	0,009	*1,000*	-0,077	0,090	0,084	-0,055
X15	-0,005	-0,126	**0,853**	**0,723**	0,426	0,463	0,500	0,231	0,303	0,333	0,111	0,019	-0,154	-0,077	*1,000*	-0,486	0,212	-0,264
X16	-0,003	0,033	-0,396	-0,375	-0,124	-0,144	-0,196	-0,095	-0,064	-0,197	0,534	-0,014	0,043	0,090	-0,486	*1,000*	0,244	0,387
X17	-0,096	-0,514	0,184	0,149	0,256	0,227	0,200	0,105	0,233	0,029	0,605	-0,270	-0,123	0,084	0,212	0,244	*1,000*	0,038
X18	-0,009	-0,022	-0,161	-0,275	0,060	0,052	0,520	0,046	0,039	0,105	0,028	0,005	-0,029	-0,055	-0,264	0,387	0,038	*1,000*

(Nguồn: Tổng hợp và tính toán bởi tác giả trên Python)

(Bảng 3). Tuy nhiên, khi thực hiện tính toán MI Scores thì kết quả cho thấy từng biến này vẫn tác động nhất định với biến mục tiêu Credit_Score nên tác giả vẫn giữ các yếu tố đầu vào của mô hình huấn luyện. Nếu kết quả huấn luyện có hiện tượng quá khớp dữ liệu thì ta sẽ xem xét để loại bỏ biến ra khỏi mô hình.

3.2. Kết quả huấn luyện mô hình ứng dụng các kỹ thuật Machine Learning

Nghiên cứu thực nghiệm mô phỏng huấn luyện bằng các thuật toán Machine Learning trên bộ dữ liệu sau khi tiền xử lý thu được kết quả tương ứng như sau (Bảng 5):

Bảng 5: *Kết quả hiệu suất các thuật toán phân loại trong Machine Learning*

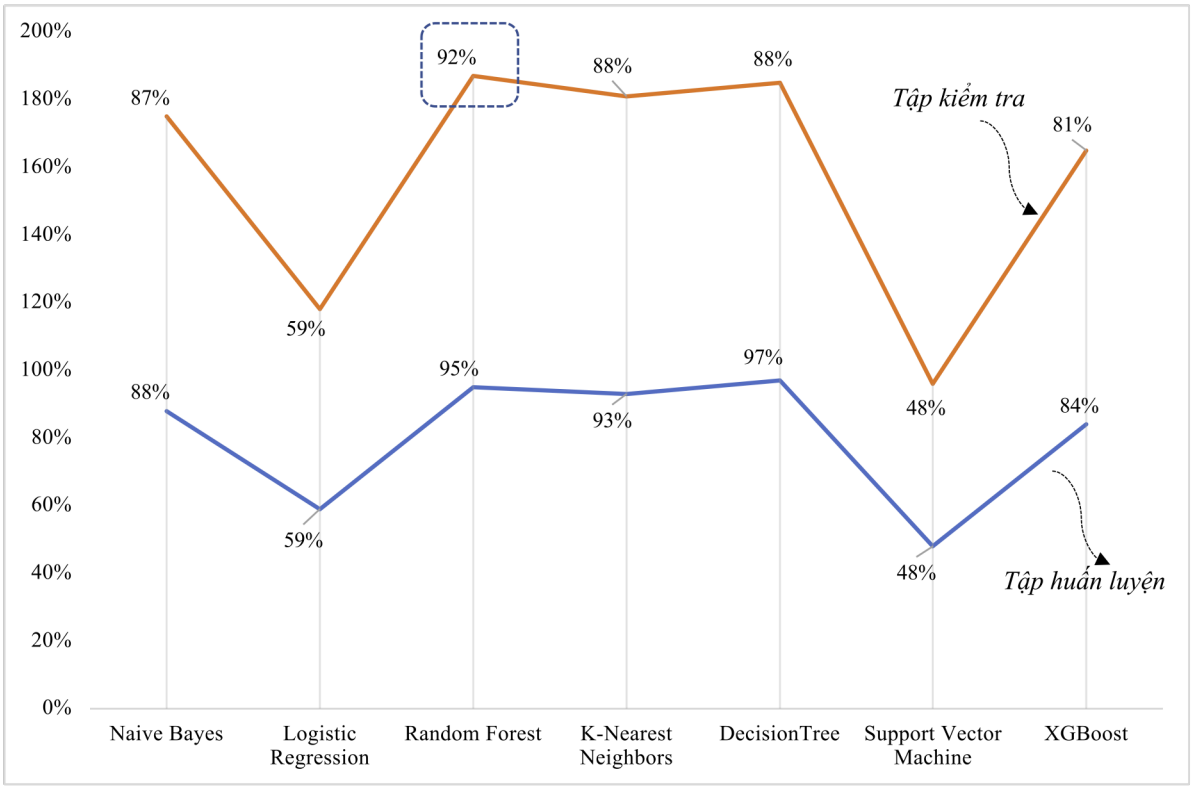
Thuật toán Machine Learning	Thời gian (giây)	Độ chính xác dự đoán trên tập huấn luyện accuracy (%)	Hiệu suất dự đoán trên tập kiểm tra (%)				
			accuracy	Phân lớp	precision	recall	f1-score
Naive Bayes	0,03	0,88	0,87	0.0	1,00	0,88	0,93
				1.0	0,78	0,96	0,86
				2.0	0,88	0,79	0,83
Logistic Regression	1,23	0,59	0,59	0.0	0,53	0,69	0,60
				1.0	0,65	0,66	0,65
				2.0	0,67	0,11	0,19
Random Forest	15,50	0,95	0,92	0.0	0,90	0,95	0,92
				1.0	0,95	0,90	0,93
				2.0	0,88	0,95	0,91
K-Nearest Neighbors	20,89	0,93	0,88	0.0	0,90	0,78	0,83
				1.0	0,89	0,97	0,93
				2.0	0,79	0,86	0,82
DecisionTree	0,70	0,97	0,88	0.0	0,90	0,79	0,84
				1.0	0,90	0,97	0,93
				2.0	0,79	0,85	0,82
Support Vector Machine	244,07	0,48	0,48	0.0	0,49	0,17	0,25
				1.0	0,48	0,90	0,63
				2.0	0,00	0,00	0,00
XGBoost	12,4	0,84	0,81	0.0	0,80	0,68	0,73
				1.0	0,81	0,91	0,85
				2.0	0,83	0,85	0,85

(Nguồn: Tổng hợp và tính toán bởi tác giả trên Python)

Sau khi huấn luyện mô hình bằng các thuật toán ML, thu được một số kết quả (các chỉ số đánh giá hiệu suất mô hình được trình bày chi tiết tại bảng 5 và được trực quan hoá tại biểu đồ 6) như sau:

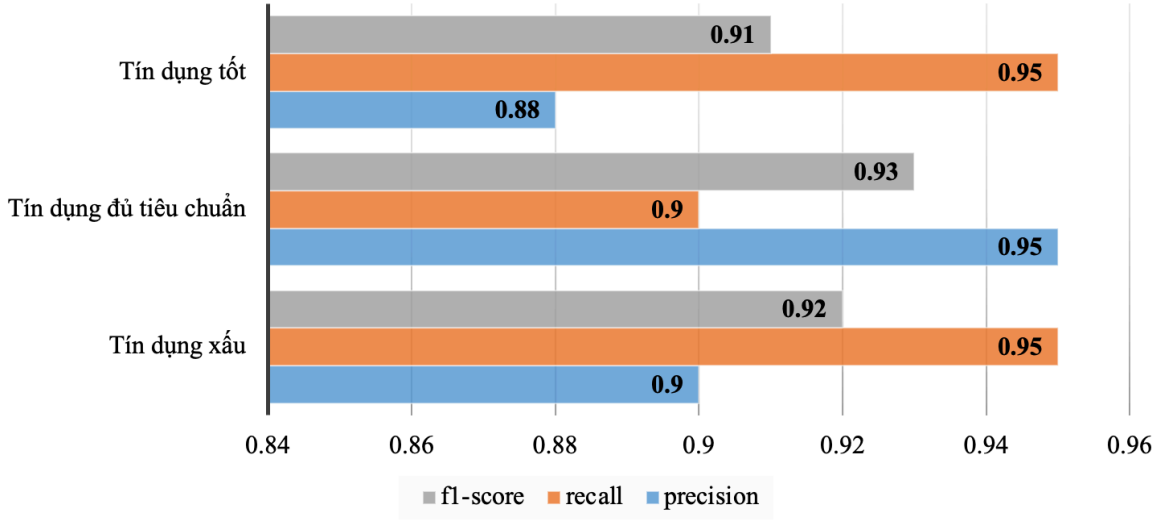
Thứ nhất, thuật toán Random Forest đạt hiệu suất cao nhất với độ chính xác trên tập kiểm tra lên đến 92%, giá trị các chỉ số precision, recall và f1-score đa số đều trên 90%, phân lớp "tín dụng đủ tiêu chuẩn". Mặt khác, kết quả chi tiết các chỉ số precision, recall và f1-score khi huấn luyện bằng Random Forest có độ biến thiên tương đối ổn định trên cả ba phân lớp "tín dụng tốt", "tín dụng xấu", "tín dụng đủ tiêu chuẩn" (Biểu đồ 6).

Thứ hai, các thuật toán Naive Bayes, Decision Tree, KNN hay XGBoost cũng có hiệu suất lần



(Nguồn: Tổng hợp và tính toán bởi tác giả)

Biểu đồ 5: *Kết quả đánh giá hiệu suất các thuật toán bằng chỉ số accuracy*



(Nguồn: Tổng hợp và tính toán bởi tác giả)

Biểu đồ 6: *Kết quả đánh giá chi tiết hiệu suất thuật toán Random Forest*

lượt là 97%, 88%, 88% và 81%. Trong đó, các chỉ số precision, recall và f1-score phân bố biến thiên tuy không nhiều nhưng có giá trị dưới 80%.

Thứ ba, thuật toán Logistic Regression và SVM có hiệu suất không cao, lần lượt là 59% và 48%. Trong đó, thuật toán SVM hầu như không phân loại được các hồ sơ tín dụng thuộc phân lớp "tín dụng đủ tiêu chuẩn", các chỉ số precision, recall và f1-score phân bố biến thiên lớn. Thuật toán Logistic Regression cho kết quả các chỉ số recall và f1-score rất thấp (dưới 20%).

Từ các kết quả trên, có thể kết luận thuật toán Random Forest đạt hiệu suất tốt nhất so với các thuật toán còn lại trên bộ dữ liệu nghiên cứu. Bên cạnh đó, khi huấn luyện dữ liệu bằng Random Forest thu được kết quả về các chỉ số precision, recall và f1-score cao, khá tương đồng và mức độ chênh lệch giá trị không đáng kể cho thấy phương pháp này hoạt động ổn định, ít rủi ro và có độ chính xác cao khi phân loại HSTD. Ngoài ra, trong quá trình thử nghiệm trước khi huấn luyện, tác giả nhận thấy rằng việc kết hợp các kỹ thuật tiền xử lý dữ liệu đã cải thiện đáng kể hiệu suất mô hình. Việc tạo ra các biến mới như đã đề xuất (Bảng 2) giúp cải thiện độ chính xác của mô hình huấn luyện như: giảm mức độ tự tương quan giữa các biến đầu vào với nhau; giảm các nguy cơ quá khớp dữ liệu do nhiều thông tin không cần thiết; các biến thay thế vừa làm giảm số lượng biến gốc ban đầu vừa mang các thông tin mới dễ diễn giải hơn với biến mục tiêu (Biểu đồ 4).

4. Kết luận và một số hàm ý đề xuất

Với bộ dữ liệu ban đầu gồm 150.000 dòng, nghiên cứu tiến hành phân tích và thực hiện các thao tác tiền xử lý dữ liệu phù hợp, thu được bộ dữ liệu kết quả 72.192 dòng. Nghiên cứu thực nghiệm mô phỏng trên bảy thuật toán phân loại trong Machine Learning: Naive Bayes, Logistic Regression, Random Forest, Decision Tree, KNN, Support Vector Machine, XGBoost. Trên cơ sở các kết quả thu được (Bảng 5) có thể cho thấy thuật toán Random Forest mang lại kết quả vượt trội so với sáu thuật toán còn lại trong việc phân lớp và dự đoán kết quả thẩm định HSTD dựa trên tập hợp các yếu tố thông tin tín dụng đầu vào của người đi vay. Đối với đặc trưng của bộ dữ liệu huấn luyện trong nghiên cứu, hai thuật toán Logistic Regression và SVM được đánh giá là không phù hợp, cho độ chính xác thấp, khả năng dự báo trên các phân lớp độ rủi ro cao. Các thuật toán còn lại cũng cho kết quả độ chính xác trên 80%. Căn cứ vào kết quả huấn luyện bảy thuật toán, tác giả đề xuất một số hàm ý nhằm cải thiện hiệu suất cũng như nâng cao hiệu quả hoạt động cho mô hình thẩm định HSTD như sau:

Thứ nhất, việc lựa chọn các yếu tố đầu vào và các kỹ thuật tiền xử lý dữ liệu là vô cùng quan trọng để quyết định hiệu suất mô hình huấn luyện vì vậy cần hết sức thận trọng. Không có thuật toán nào là tối ưu cho mọi bộ dữ liệu nghiên cứu. Hay nói cách khác, mỗi thuật toán sẽ cho kết quả hiệu suất khác nhau trên các bộ dữ liệu riêng biệt. Vì vậy, việc lựa chọn dữ liệu phù hợp vào từng yêu cầu cụ thể, việc phân tích cấu trúc của dữ liệu, các chức năng của ứng dụng, mức độ tách biệt các lớp là vô cùng quan trọng để quyết định thuật toán nào là phù hợp nhất. Đối với bài toán mà biến mục tiêu có ba phân lớp, cần theo dõi thận trọng các chỉ số kết quả precision, recall và f1-score, mức độ biến thiên giữa các phân lớp là căn cứ để đánh giá độ chính xác và độ ổn định của thuật toán khi huấn luyện. Đồng thời, cần có sự am hiểu về nguyên tắc huấn luyện của các thuật toán máy học để điều chỉnh các tham số cũng như theo dõi các biến động dữ liệu trong quá trình huấn luyện để khắc phục kịp thời và hạn chế các rủi ro quá khớp dữ liệu.

Thứ hai, việc sử dụng kết hợp các kỹ thuật tiền xử lý có thể giúp tăng cường hiệu suất mô hình và cũng như tính chính xác trong dự đoán tốt hơn. Tuy nhiên, trong quá trình vận dụng cần phải xem xét đến mức độ ảnh hưởng của dữ liệu sau khi

biến đổi phải đảm bảo phù hợp với các chính sách và quy định của các tổ chức tài chính ngân hàng trong thực tế. Đối với việc thẩm định HSTD, cần hết sức thận trọng trong các mẫu dữ liệu ngoại lệ. Nguyên nhân là có những khách hàng chưa có lịch sử tín dụng trước đó nên có thể thiếu sót thông tin. Các nghiên cứu gần đây đang tiếp cận với các nguồn thông tin khác từ lịch sử giao dịch đối với các khách hàng thiếu thông tin lịch sử tín dụng trước đó. Vì vậy, cần cân nhắc xây dựng hệ thống thẩm định tín dụng có các xử lý riêng, tích hợp các thông tin bổ sung thay vì loại bỏ các mẫu ngoại lệ ra khỏi bộ dữ liệu huấn luyện. Điều này có thể bỏ sót hay làm mất đi một lượng lớn các khách hàng tiềm năng do kết quả thẩm định HSTD sai.

Thứ ba, hành động thêm biến mới hay loại bỏ bớt các biến không cần thiết cần được cân nhắc kỹ lưỡng. Việc thêm biến có thể làm dư thừa thông tin không cần thiết cũng như việc loại bỏ bớt biến có thể làm mất mát thông tin tiềm năng làm ảnh hưởng đến hiệu suất cũng như kết quả huấn luyện mô hình. Chính vì vậy, không hoàn toàn phụ thuộc vào thuật toán huấn luyện mà cần phải có sự hiểu biết sâu về chuyên môn nghiệp vụ thẩm định để lựa chọn các yếu tố phù hợp.

Thứ tư, một trong những hạn chế của nghiên cứu là chưa sử dụng bộ dữ liệu thực tế của các ngân hàng vì hiện tại tác giả chưa tìm được bộ dữ liệu nào được công bố công khai tại Việt Nam. Chính vì vậy, kết quả nghiên cứu thực nghiệm được tiến hành trên bộ dữ liệu phục vụ cho các nghiên cứu mô phỏng máy học. Mục tiêu nghiên cứu của tác giả là cung cấp thêm bằng chứng thực nghiệm về hiệu quả của các thuật toán máy học khác nhau trên cùng một bộ dữ liệu nghiên cứu. Đồng thời, một số kỹ thuật tiền xử lý đề xuất được vận dụng để cải thiện hiệu suất mô hình cũng được thể hiện trong kết quả nghiên cứu. Không có một thuật toán hay phương pháp nào là tối ưu cho tất cả các bộ dữ liệu hay tất cả các trường hợp. Việc lựa chọn các kỹ thuật, các thuật toán ML khác nhau sẽ ảnh hưởng trực tiếp đến kết quả hiệu suất mô hình. Vì vậy, để đảm bảo tính phù hợp, chính xác và tăng cường hiệu suất trong quá trình xây dựng các mô hình thẩm HSTD, cần nghiên cứu thận trọng trong việc lựa chọn yếu tố quan trọng đánh giá được khả năng tín dụng của một khách hàng. Kết quả thực nghiệm cho thấy các biến mới được tạo từ dữ liệu thu thập ban đầu cũng quyết định rất lớn đến hiệu quả thẩm định. Vì vậy, quá trình phát triển mô hình thẩm định tín dụng không thể thiếu thông tin tư vấn từ các chuyên gia trong việc xác định các thông tin nào là hiệu quả cho việc xây dựng hệ thống các chỉ tiêu (tiêu chuẩn) căn cứ thẩm định hồ sơ tín dụng. Từ đó, xây dựng bộ tiêu chí phù hợp với chính sách, đặc trưng vận hành của ngân hàng nhằm thu thập đúng dữ liệu cần thiết.

Cuối cùng, vấn đề các tổ chức tín dụng rất quan tâm là làm sao xét duyệt các hồ sơ mới mà khách hàng chưa có lịch sử tín dụng tại ngân hàng hiệu quả nhất. Để tránh mất đi các khách hàng tiềm năng, tổ chức tín dụng cần xây dựng chính sách phù hợp để linh hoạt xử lý thẩm định HSTD khi lượng thông tin về khách hàng chưa nhiều, thậm chí là chưa có lịch sử giao dịch trước đó. Các thông tin giao dịch khác (lịch sử trả nợ và thanh toán trên các sàn thương mại điện tử, thu nhập và thói quen thanh toán trong quá khứ, các yếu tố nhân khẩu học,...) cần được cân nhắc để đưa ra quyết định xét duyệt HSTD cho khách hàng mới. Hiệu quả thẩm định HSTS chắc chắn không thể thiếu sự minh bạch trong khâu xét duyệt, đảm bảo phù hợp với chính sách, quy định của tổ chức tín dụng. Bên cạnh đó, cần có sự lãnh đạo sáng suốt của các nhà quản lý, đảm bảo hoạt động thẩm định HSTD vận hành sự thống nhất của tất cả các bộ phận có liên quan.◆

***Tài liệu tham khảo*:**

Anderson, R. A. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* London: Oxford University Press.

Assef, F., Teresinha, M., & Steiner, A. (2020). Machine Learning Techniques in Bank Credit Analysis. *International Journal of Economics and Management Engineering, V.14(7)*, 517-520.

Benton E. Gup, James W. Kolari, & Donald R. Fraser. (2005). *Commercial Banking: The Management of Risk, 3rd Edition.* London: Wiley.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications.* 3302 - 3309.

Benjamin, T. S. (2017). Can Financial Technology Innovate Benefit Distribution in Payments for Ecosystem Services and REDD+?, *Ecological Economics*, 150 - 157.

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy, V. 37(3)*, 585-617.

CIC.org. (2022, 2 14). *Trung tâm thông tin dữ liệu tín dụng quốc gia Việt Nam*. Được truy lục từ Trung tâm thông tin dữ liệu tín dụng quốc gia Việt Nam: https://faq.cic.org.vn/category/question_faq/?id=259

Crook, J., Edelman, D., & Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, V. 183*, 1447-1465.

Coffman, J. (1986). The Proper Role of Tree Analysis in the Forecasting the Risk Behaviour of Borrowers. *MDS Reports, Management Decision Systems, Atlanta*, 47 - 59.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, V. 210(2)*, 368-378.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2020). How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm. *BIS Working Papers*, 1-29.

Ghosh, M. (2017). Disruptive Innovation and Academy Library management. *Disruptive innovation and academic library management.* India.

Gómez, A., Rosado , A., & Espinosa, O. (2024). Data preprocessing to improve fairness in machine learning models: An application to the reintegration process of demobilized members of armed groups in Colombia. *Applied Soft Computing*.

Goh, R., & Lee, L. (2019). Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. *Advances in Operations Research*, 1-31.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications, V. 42*, 741–750.

Henley, W., & Hand, D. (1997). Construction of a knearest-neighbour credit-scoring system. *IMA Journal of Management Mathematics, V. 8 (4)*, 305-321.

Hué, S., Hurlin, C., & Tokpavi, S. (2018). Machine Learning for Credit Scoring: ImprovingLogistic Regression with Non Linear Decision TreeE□ects. *European Journal of Operational Research 297(1)*, 1-29.

Huang, J., Li, Y., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology, V.67*, 108-127.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, V. 40 (13)*, 5125-5131.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-

the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 124-136.

Machado, M., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, 1168-1189.

Mukid, M., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k nearest neighbor. *Journal of Physics Conference Series 1025(1)*, 1 - 7.

Obaid, H. S., Dheyab, A. S., & Sabry, S. S. (2019). The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning. *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*. Jaipur, India: IEEE.

Porter, M. E. (1998). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press; Illustrated edition (June 1, 1998).

Teles, G., Rodrigues, J., Saleem, K., Kozlov, S., & Rabêl, R. (2020). Machine learning and decision support system on credit scoring. *Neural Computing and Applications*, 9809-9826.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and Its Applications*. London: Society for Industrial and Applied Mathematics.

Vieira, J., Barboza, F., Sobreiro, V., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing 83(1)*.

Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science, V.199*, 1128-1135.

Zhou, Y., Shen, L., & Laura, B. (2023). A two-stage credit scoring model based on random forest: Evidence from Chinese small firms. *International Review of Financial Analysis*, 1027-1055.

Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing*, 1040-1056.

Wonglimpiyarat, J. (2017). FinTech banking industry: a systemic approach. *The journal of future studies, strategic thinking and policy*, 590 - 603.

Summary

Application of Machine Learning algorithms to evaluate credit records is considered to bring many strengths in processing financial data. Algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, XGBoost are applied to simulate the ability to classify credit records at banks into three categories: good, bad and standard. The results obtained show that Random Forest provides the best performance with an accuracy of over 92%; Naive Bayes, K-Nearest Neighbors, Decision Tree achieve prediction performance over 80%; Logistic Regression and Support Vector Machine yield low performance (59% and 48%). In order to increase the suitability of training input data, the research also uses a combination of data preprocessing techniques such as: creating new variables that match the evaluation criteria from the original data set, assigning labels, outlier handling, best feature selection analysis, data normalization, data balancing, etc. The results show that data preprocessing techniques improve training performance. The obtained results are expected to add reliable experimental evidence to other studies related to the topic of credit profile appraisal using machine learning algorithms.