Journal of Finance - Marketing Research; Vol. 16, Issue 3; 2025
p-ISSN: 1859-3690; e-ISSN: 3030-427X
DOI: https://doi.org/10.52932/jfmr.v16i3

Journal of Finance - Marketing Research
http://jfm.ufm.edu.vn

FEATURE SELECTION AND CORPORATE DEFAULT RISK PREDICTION: AN EMPIRICAL STUDY WITH MACHINE LEARNING MODELS

Nguyen Minh Nhat[1*], Ngo Hoang Khanh Duy[1]
[1]Ho Chi Minh University of Banking, Vietnam

ARTICLE INFO

DOI:
10.52932/jfmr.v16i3.761

Received:
February 16, 2025
Accepted:
March 26, 2025
Published:
June 25, 2025

Keywords:
Default risk;
Feature Selection;
LightGBM;
Machine learning;
Recursive feature elimination

JEL codes:
C36, C45, G28

ABSTRACT

Corporate default risk prediction (RRVN) is a crucial factor in credit operations, enabling institutions to identify risks early and optimize their credit portfolios. This study focuses on analyzing the impact of four feature selection methods, including Recursive Feature Elimination (RFE), Forward Feature Selection (FFS), Backward Feature Elimination (BFE), and Permutation Feature Importance (PFI), on three machine learning models: Logistic Regression, Decision Tree, and LightGBM. Experiments on a dataset of 15,862 enterprises show that LightGBM combined with RFE achieves the highest performance, with superior accuracy and F1-score compared to other methods. The results highlight the critical role of feature selection in enhancing predictive model performance. Additionally, the study identifies key features influencing RRVN, including Borrowing Dependency, Cash Turnover Rate, Degree of Financial Leverage, and Total Assets to GNP. These findings provide a scientific basis for financial institutions to improve models, optimize credit decision-making, and enhance risk management.

*Corresponding author:
Email: nhatnm@hub.edu.vn

Tạp chí Nghiên cứu Tài chính - Marketing
http://jfm.ufm.edu.vn

LỰA CHỌN ĐẶC TRƯNG VÀ DỰ BÁO RỦI RO VỠ NỢ DOANH NGHIỆP: THỰC NGHIỆM VỚI MÔ HÌNH HỌC MÁY

Nguyễn Minh Nhật[1*], Ngô Hoàng Khánh Duy[1]
[1]Trường Đại học Ngân hàng Thành phố Hồ Chí Minh

THÔNG TIN

DOI:
10.52932/jfmr.v16i3.761

Ngày nhận:
16/02/2025
Ngày nhận lại:
26/03/2025
Ngày đăng:
25/06/2025

Từ khóa:
Loại bỏ đặc trưng đệ quy (RFE);
Lựa chọn đặc trưng;
Mô hình học máy;
Mô hình LightGBM;
Rủi ro vỡ nợ

Mã JEL:
C36, C45, G28

TÓM TẮT

Dự báo rủi ro vỡ nợ (RRVN) là yếu tố quan trọng trong hoạt động tín dụng, giúp các tổ chức xác định sớm nguy cơ và tối ưu hóa danh mục tín dụng. Nghiên cứu tập trung phân tích ảnh hưởng của 04 phương pháp lựa chọn đặc trưng (LCĐT) bao gồm Loại bỏ đặc trưng đệ quy (RFE), Loại bỏ đặc trưng chuyển tiếp (FFS), Loại bỏ đặc trưng ngược (BFE) và Loại bỏ đặc trưng dựa trên hoán vị (PFI) trên 03 mô hình học máy là Hồi quy Logistic, Cây quyết định và LightGBM. Thực nghiệm trên bộ dữ liệu gồm 15,862 doanh nghiệp cho thấy LightGBM kết hợp với RFE đạt hiệu suất cao nhất, với độ chính xác và điểm F1 vượt trội so với các phương pháp khác. Kết quả nhấn mạnh vai trò quan trọng của LCĐT trong việc nâng cao hiệu suất mô hình dự báo. Ngoài ra, nghiên cứu cũng xác định một số đặc trưng quan trọng nhất ảnh hưởng đến RRVN bao gồm Mức độ phụ thuộc vào nợ vay, Tỷ lệ vòng quay tiền mặt, Mức độ đòn bẩy tài chính và Tổng tài sản trên GNP. Những phát hiện này cung cấp cơ sở khoa học cho các tổ chức tài chính trong việc cải thiện mô hình, tối ưu hóa quyết định cấp tín dụng và kiểm soát rủi ro tốt hơn.

1. Giới thiệu

Rủi ro vỡ nợ (RRVN) được xem là một trong những thách thức lớn đối với các tổ chức tín dụng, bởi sự tác động lớn của nó đến khả năng cho vay, phân bổ nguồn vốn và kiểm soát rủi ro tín dụng. Việc nhận diện sớm các dấu hiệu rủi ro của khách hàng không những giúp các tổ chức tín dụng hạn chế được những tổn thất mà còn giúp họ xây dựng danh mục cho vay phù hợp với mục tiêu của tổ chức.

Sự phát triển mạnh mẽ của học máy (Machine learning) trong thời gian qua đã chứng tỏ cách tiếp cận hiệu quả trong dự báo RRVN, vượt qua nhiều hạn chế của các mô hình truyền thống (Shi và cộng sự, 2022). Các thuật toán học máy như Cây quyết định (Decision Tree), Rừng

*Tác giả liên hệ:
Email: nhatnm@hub.edu.vn

ngẫu nhiên (Random Forest) hay các thuật toán tăng cường như LightGBM đã mang lại kết quả ấn tượng nhờ khả năng xử lý dữ liệu phức tạp và phi tuyến (Nguyen & Ngo, 2025). Tuy nhiên, việc xây dựng một mô hình học máy phục vụ cho quá trình dự báo RRVN không chỉ phụ thuộc vào tính ưu việt của thuật toán được sử dụng mà còn phụ thuộc lớn vào tính chất của dữ liệu hay tập hợp đặc trưng (features) được lựa chọn cho mô hình dự báo. Nếu một mô hình với quá nhiều đặc trưng không cần thiết, nó có thể làm tăng mức độ nhiễu, làm phát sinh các chi phí tính toán, từ đó ảnh hưởng đến hiệu suất và mức độ dự báo chính xác của mô hình. Ngược lại, nếu những đặc trưng quan trọng bị bỏ sót, mô hình dự báo có thể không phản ánh đúng bản chất của vấn đề dự báo. Do đó, việc LCĐT trở thành một yếu tố then chốt trong việc tối ưu hóa hiệu suất của các mô hình học máy (Rudnicki và cộng sự, 2015).

Mặc dù có sự đồng thuận về tầm quan trọng của LCĐT trong các mô hình dự báo, nhưng vẫn tồn tại nhiều tranh luận liên quan đến phương pháp tối ưu nhất trong từng bối cảnh cụ thể. Ahmed và cộng sự (2024) lập luận rằng, phương pháp Loại bỏ đặc trưng đệ quy (Recursive Feature Elimination – RFE) có thể cải thiện độ chính xác của mô hình Hồi quy Logistic do khả năng loại bỏ những đặc trưng không quan trọng. Tuy nhiên, Sahu và Dash (2022) lại chỉ ra rằng, phương pháp Loại bỏ đặc trưng chuyển tiếp (Forward Feature Selection – FFS) có xu hướng hoạt động hiệu quả hơn đối với mô hình cây quyết định, do tính chất phân cấp của mô hình này cho phép khai thác tốt hơn các đặc trưng quan trọng. Liên quan đến phương pháp Loại bỏ đặc trưng ngược (Backward Feature Elimination – BFE), Cao và cộng sự (2010) phát hiện rằng, phương pháp này phù hợp hơn với các mô hình cây quyết định, do cơ chế loại bỏ dần các đặc trưng ít quan trọng giúp cải thiện kết quả dự báo.

Các tranh luận này tiếp tục đẩy lên cao khi một số nghiên cứu khác lại lập luận rằng, phương pháp Loại bỏ đặc trưng dựa trên hoán vị (PFI) có tính chất phù hợp nhất với những mô hình tăng cường (Boosting) như mô hình LightGBM và ít phù hợp với mô hình Hồi quy Logistic, do khả năng xác định mức độ quan trọng của từng đặc trưng thông qua việc hoán vị ngẫu nhiên và đánh giá ảnh hưởng của chúng lên hiệu suất mô hình (Bian và cộng sự, 2023; Lao và cộng sự, 2023). Nghiên cứu của Saarela và Jauhiainen (2021), Elghazel và Aussem (2015) đưa ra quan điểm rằng, không có phương pháp LCĐT nào vượt trội hơn mà hiệu quả của từng phương pháp phụ thuộc vào tính chất của mô hình học máy và đặc điểm của tập dữ liệu. Các tranh luận này dẫn đến một khoảng trống quan trọng đó là, liệu các phương pháp LCĐT có thật sự ảnh hưởng đến hiệu suất của các mô hình hay không, và nếu điều đó xảy ra thì phương pháp nào sẽ là tối ưu nhất trong bài toán dự báo RRVN.

Trên cơ sở phân tích, bài nghiên cứu này được thực hiện nhằm mục đích đánh giá sự tác động của các phương pháp LCĐT đến hiệu suất dự báo vỡ nợ doanh nghiệp của các mô hình học máy. Cụ thể, nghiên cứu sẽ tiến hành phân tích, so sánh tính hiệu quả của bốn phương pháp LCĐT là RFE, FFS, BFE và PFI trên ba mô hình dự báo RRVN được lựa chọn là Hồi quy Logistic, Cây quyết định và LightGBM. Từ đó, xác định phương pháp LCĐT nào giúp cải thiện mức độ chính xác và có khả năng diễn giải mô hình tốt nhất. Bên cạnh đó, nghiên cứu cũng xác định một số đặc trưng quan trọng nhất ảnh hưởng đến RRVN doanh nghiệp thông qua việc phân tích cách thức LCĐT của các phương pháp. Để đạt được mục tiêu nghiên cứu, bài viết tiến hành các thí nghiệm trên bộ dữ liệu với 15.862 doanh nghiệp được công khai trên Kaggle (2024). Kết quả nghiên cứu kỳ vọng cung cấp thêm nhiều thông tin giá trị về tầm quan trọng của việc LCĐT trong quá trình xây dựng mô hình dự báo RRVN. Trong phần nội dung tiếp theo, bài viết được cấu trúc như sau: (1) Khảo lược nghiên cứu; (2) Phương pháp nghiên cứu; (3) Kết quả nghiên cứu; và (4) Kết luận.

2. Khảo lược nghiên cứu

Loại bỏ đặc trưng đệ quy (RFE), Loại bỏ đặc trưng chuyển tiếp (FFS), Loại bỏ đặc trưng

ngược (BFE), và Loại bỏ đặc trưng dựa trên hoán vị (PFI) được xem là các phương pháp LCĐT phổ biến nhằm cải thiện mức độ dự báo chính xác của các mô hình (Xia & Yang, 2023). Trong quá khứ, RFE đã cho thấy tiềm năng vượt trội nhờ loại bỏ từng đặc trưng không quan trọng theo từng bước, từ đó tăng cường hiệu suất dự đoán của các mô hình. FFS và BFE, nhờ vào quy trình hoạt động linh hoạt, có khả năng thêm hoặc loại bỏ đặc trưng theo hướng tiến hoặc lùi, giúp chúng trở nên đa năng trong việc xác định các biến số có tầm quan trọng đặc biệt. Ngoài ra, PFI đã được đề xuất như một kỹ thuật đánh giá tầm quan trọng của từng đặc trưng trong quá trình dự đoán và là một công cụ mạnh mẽ để phân tích cấu trúc dữ liệu (Urbanowicz và cộng sự, 2018).

Bên cạnh đó, một số phương pháp khác cũng có thể được sử dụng để LCĐT trong lĩnh vực tài chính như Lasso, Ridge, ElasticNet hay k-NN. Các phương pháp điều chuẩn như Lasso và Ridge có lợi thế trong việc kiểm soát đa cộng tuyến bằng cách ràng buộc trọng số của các biến, song chúng hoạt động hiệu quả chủ yếu trong các mô hình tuyến tính và có thể không tận dụng tối đa các quan hệ phi tuyến vốn thường xuất hiện trong dữ liệu tài chính (Muthukrishnan & Rohini, 2016). Trong khi đó, các phương pháp như k-NN dựa vào khoảng cách giữa các điểm dữ liệu để chọn ra đặc trưng quan trọng, nhưng lại nhạy cảm với phân phối dữ liệu và không đảm bảo tính tổng quát khi áp dụng trên các tập dữ liệu lớn có cấu trúc phức tạp (Maleki và cộng sự, 2021).

Các phương pháp LCĐT như RFE, FFS, BFE và PFI thường có xu hướng xử lý tốt hơn trong các mối quan hệ phi tuyến, đồng thời linh hoạt và tương thích với nhiều mô hình học máy khác nhau (Ren và cộng sự, 2020; Sanz và cộng sự, 2018). Hơn nữa, các phương pháp LCĐT này cũng có thể giúp giảm nguy cơ quá khớp (overfitting) bằng cách loại bỏ các đặc trưng dư thừa, nhưng nếu không được kiểm soát chặt chẽ, chúng cũng có thể loại bỏ nhầm những đặc trưng hữu ích, dẫn đến giảm khả năng tổng quát hóa của mô hình. Đặc biệt, PFI có thể bị ảnh hưởng bởi nhiễu trong dữ liệu nếu số lượng hoán vị không đủ lớn, làm sai lệch đánh giá mức độ quan trọng của các đặc trưng. Do đó, việc kết hợp các phương pháp LCĐT với các chiến lược kiểm định chéo (cross-validation) và điều chỉnh tham số là cần thiết để cân bằng giữa hiệu suất mô hình và khả năng tổng quát hóa (Qi và cộng sự, 2019).

Hồi quy Logistic là một trong nhiều mô hình thống kê truyền thống thường được sử dụng trong dự báo RRVN nhờ tính đơn giản và khả năng diễn giải cao. Tuy nhiên, sức mạnh dự đoán của Hồi quy Logistic (LR) lại phụ thuộc nhiều vào chất lượng dữ liệu và quá trình LCĐT của mô hình (Chen và cộng sự, 2023). Vishraj và cộng sự (2023) đã phát hiện ra rằng, RFE và FFS là các phương pháp phù hợp để tinh chỉnh mô hình LR, giúp cải thiện độ chính xác và giảm lỗi dự báo. Nhờ vậy, các hạn chế tự nhiên của LR có thể được khắc phục thông qua việc LCĐT, qua đó nâng cao đáng kể khả năng dự đoán của mô hình.

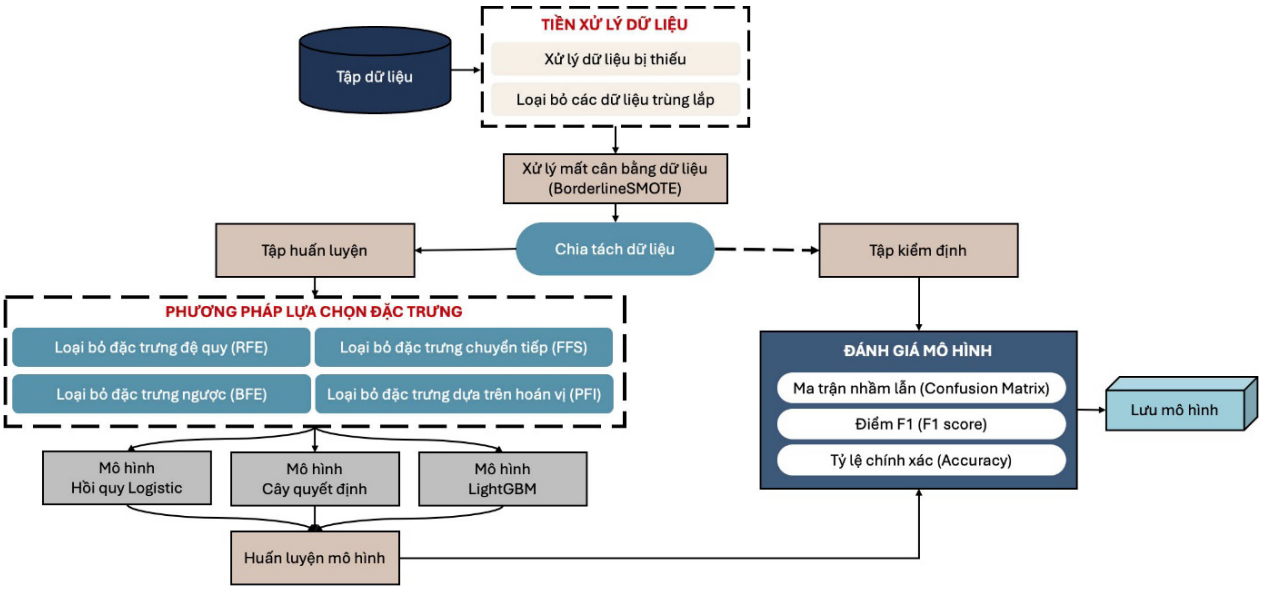
Guo và Zhou (2022) nhận thấy rằng, các thuật toán như LightGBM hay XGBoost, vốn rất hiệu quả trong việc tự động hóa quá trình lựa chọn và xử lý đặc trưng, mang lại kết quả vượt trội so với các mô hình cây truyền thống. Tuy nhiên, hiệu quả của mô hình này phụ thuộc chủ yếu vào việc lựa chọn các thuộc tính được sử dụng, trong đó RFE và PFI thường được áp dụng để tối ưu hóa đầu vào dữ liệu. Guo và Zhou (2022) cũng khẳng định rằng, mô hình Cây quyết định, khi kết hợp với các phương pháp LCĐT, vẫn đủ ổn định để cạnh tranh với các mô hình máy học tiên tiến trong nhiệm vụ dự đoán RRVN.

Tại Việt Nam, các nghiên cứu liên quan đến việc kết hợp các phương pháp LCĐT và mô hình học máy trong dự báo RRVN còn khá hạn chế và cần được nghiên cứu rộng rãi hơn. Kết quả khảo lược cho thấy rằng, vẫn còn tồn tại một khoảng trống nghiên cứu lớn về việc đánh giá một cách toàn diện mức độ hiệu quả của các phương pháp LCĐT trong việc xây dựng các mô hình dự báo RRVN. Việc lựa chọn RFE, FFS, BFE và PFI trong nghiên cứu này nhằm tập trung vào các phương pháp đã được chứng

minh là có tính ứng dụng cao, đồng thời so sánh khả năng của chúng trong việc tối ưu hóa hiệu suất mô hình dự báo rủi ro vỡ nợ doanh nghiệp.

3. Phương pháp nghiên cứu

3.1. Thiết kế nghiên cứu



Hình 1. Quá trình thực hiện nghiên cứu

Quy trình nghiên cứu trong Hình 1 bắt đầu với bước thu thập và tiền xử lý dữ liệu. Dữ liệu được làm sạch bằng cách loại bỏ các giá trị bị thiếu và các giá trị trùng lặp, đồng thời cân bằng phân phối giữa các trường hợp vỡ nợ và không vỡ nợ bằng kỹ thuật BorderlineSMOTE. Quá trình này nhằm đảm bảo rằng dữ liệu đầu vào có chất lượng tốt, giảm thiểu các sai lệch trong kết quả dự báo. Sau khi hoàn tất tiền xử lý, dữ liệu được phân ra hai tập riêng biệt: tập huấn luyện và tập kiểm định, đảm bảo rằng mô hình được đánh giá trên dữ liệu chưa từng thấy trước đó.

Từ tập huấn luyện, các phương pháp LCĐT được áp dụng để xác định những yếu tố quan trọng nhất cho dự đoán. Các phương pháp bao gồm Loại bỏ đặc trưng đệ quy (RFE), Loại bỏ đặc trưng chuyển tiếp (FFS), Loại bỏ đặc trưng ngược (BFE), và Loại bỏ đặc trưng dựa trên hoán vị (PFI). Mục tiêu của bước này là giảm bớt số lượng đặc trưng dư thừa, tập trung vào các đặc trưng có ảnh hưởng lớn đến hiệu suất dự đoán. Kết quả là một tập hợp các đặc trưng tối ưu được sử dụng cho việc huấn luyện các mô hình khác nhau.

Ba mô hình chính được sử dụng trong nghiên cứu này là Hồi quy Logistic, Cây quyết định, và LightGBM. Các mô hình này được huấn luyện trên tập dữ liệu với các đặc trưng đã được lựa chọn ở bước trước. Mỗi mô hình được tối ưu hóa để đưa ra dự đoán chính xác nhất dựa trên dữ liệu đầu vào. Quá trình huấn luyện đảm bảo rằng các mô hình có thể học được mối quan hệ giữa các đặc trưng và biến mục tiêu một cách hiệu quả.

Cuối cùng, các mô hình được đánh giá dựa trên tập kiểm định thông qua ba tiêu chí chính: ma trận nhầm lẫn (Confusion Matrix), điểm F1 (F1 score), và tỷ lệ chính xác (Accuracy). Kết quả đánh giá này cho phép so sánh hiệu suất giữa các mô hình và lựa chọn mô hình phù hợp nhất cho mục tiêu nghiên cứu. Mô hình đạt hiệu suất tốt nhất sẽ được lưu lại để thực hiện các bước phân tích đánh giá tiếp theo.

3.2. Phương pháp lựa chọn đặc trưng

LCĐT là một cơ chế để chọn lọc các yếu tố quan trọng nhất ảnh hưởng đến mục tiêu dự báo, từ đó tối ưu hóa tính nhất quán của mô hình, giảm tải tính toán và tăng cường khả năng

diễn giải. Nghiên cứu này sử dụng bốn phương pháp LCĐT bao gồm: Loại bỏ đặc trưng đệ quy (RFE), Loại bỏ đặc trưng chuyển tiếp (FFS), Loại bỏ đặc trưng ngược (BFE), và Loại bỏ đặc trưng dựa trên hoán vị (PFI). Các đặc trưng quan trọng được lựa chọn dựa trên các phương pháp này sẽ được sử dụng để huấn luyện trên ba mô hình dự báo bao gồm: Hồi quy Logistic, Cây quyết định và LightGBM (*xem Phụ lục 2 online*).

3.3. Các mô hình dự báo

3.3.1. Mô hình hồi quy Logistic

Mô hình hồi quy Logistic đóng vai trò quan trọng trong việc dự báo nguy cơ vỡ nợ của các doanh nghiệp. Biến phụ thuộc Y được mã hóa với giá trị 1 khi doanh nghiệp rơi vào trạng thái vỡ nợ và 0 khi không xảy ra tình trạng này. Các biến độc lập $X_1,\ldots, X_n$ biểu thị các đặc điểm tài chính quan trọng của doanh nghiệp (Han và cộng sự, 2012). Xác suất xảy ra vỡ nợ được tính toán dựa trên công thức của mô hình Logistic như sau:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

Trong quá trình huấn luyện mô hình, các tham số β được tối ưu hóa bằng cách cực đại hóa hàm hợp lý (Maximum Likelihood Estimation – MLE). Hàm hợp lý trong bối cảnh này được xác định bằng:

$$L(\beta) = \prod^{N}_{i=1} P(Y_i|X_i)^{Y_i}(1-P(Y_i|X_i))^{1-Y_i} \quad (2)$$

Mỗi Y_i tương ứng với trạng thái vỡ nợ hay là không vỡ nợ của doanh nghiệp i, còn $P(Y_i|X_i)$ là xác suất được tính toán từ mô hình. Mục tiêu là hàm L(β) đạt giá trị cực đại, đảm bảo mô hình dự báo phù hợp với thực tế và giảm thiểu sai sót trong quá trình dự báo rủi ro của các doanh nghiệp.

Mô hình hồi quy Logistic cũng cho phép phân tích tác động của từng đặc trưng tài chính trong mô hình đến xác suất vỡ nợ của các doanh nghiệp thông qua công thức sau:

$$logit(P) = log\ \frac{P}{1-P} = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n \quad (3)$$

Mô hình hồi quy Logistic nổi bật với tính đơn giản, dễ hiểu và khả năng triển khai hiệu quả, đặc biệt trong các trường hợp mối quan hệ giữa các biến đặc trưng và biến mục tiêu có tính tuyến tính. Hồi quy Logistic thường đạt hiệu suất cao đối với các tập dữ liệu có quy mô vừa và nhỏ, đồng thời cung cấp khả năng giải thích rõ ràng về tác động của từng biến đặc trưng đến kết quả dự báo, từ đó giúp xác định các yếu tố có ảnh hưởng lớn đến nguy cơ vỡ nợ. Tuy nhiên, mô hình gặp hạn chế khi xử lý các tập dữ liệu phi tuyến tính hoặc khi mối quan hệ giữa các biến mang tính phức tạp. Bên cạnh đó, hiệu quả của mô hình có thể bị suy giảm trong trường hợp dữ liệu chứa nhiều nhiễu hoặc số lượng lớn các đặc trưng không liên quan.

3.3.2. Mô hình Decision Tree

Decision Tree là một mô hình học máy dựa trên cấu trúc cây, thường được sử dụng để dự đoán xác suất vỡ nợ của doanh nghiệp. Tại mỗi nút của cây, dữ liệu được phân chia thông qua các tiêu chí như Entropy hoặc Gini Index. Trong nghiên cứu này, Gini Index là tiêu chí được sử dụng do khả năng xử lý tốt các tập dữ liệu bị mất cân bằng, đồng thời có sự cân bằng tốt giữa độ chính xác trong phân chia và thời gian tính toán (Hajek & Michalak, 2013). Gini Index đo lường mức độ hỗn loạn tại mỗi nút và được tính bằng công thức:

$$G(t) = 1 - \sum^{k}_{i=1} p_i^2 \quad (4)$$

Trong đó, p_i là xác suất của các quan sát thuộc lớp i (vỡ nợ hoặc không vỡ nợ), và k là số lớp (trong nghiên cứu này k=2). Giá trị Gini càng thấp thì dữ liệu tại nút càng đồng nhất.

Để sự phân chia tốt nhất, mô hình Decision Tree cần phải tối ưu hóa RI (Reduction in Impurity) tại mỗi nút t. Công thức ước tính RI như sau:

$$RI(t) = G(t) - \sum_{j=1}^{m} \frac{|t_j|}{|t|} G(t_j) \quad (5)$$

Trong đó, t_j là các tập con sau khi chia, G(t) là giá trị Gini Index trước khi phân chia và $|t|$, $|t_j|$ lần lượt là số quan sát tại nút trước và sau khi chia. Mục tiêu là lựa chọn các biến đặc trưng và ngưỡng phân chia sao cho giá trị RI là lớn nhất nhằm đảm bảo rằng các tập con sau phân chia trở nên đồng nhất hơn.

Khi đạt đến các nút lá, mô hình đưa ra dự đoán xác suất dựa trên tỷ lệ quan sát thuộc từng lớp. Nếu tại nút lá có s_1 doanh nghiệp vỡ nợ và s_0 doanh nghiệp không vỡ nợ thì xác suất vỡ nợ được ước tính là:

$$P(Y = 1) = \frac{s_1}{s_1 + s_0} \quad (6)$$

Mô hình Decision Tree được đánh giá cao nhờ tính trực quan và dễ diễn giải, với cấu trúc cây rõ ràng cho phép chúng ta hiểu được các quy tắc phân chia được sử dụng. Một lợi thế quan trọng của mô hình là không yêu cầu dữ liệu phải tuân theo các giả định về tính tuyến tính hay cần được chuẩn hóa, khiến nó trở nên phù hợp với nhiều loại dữ liệu trong thực tiễn. Tuy nhiên, một nhược điểm đáng chú ý là Decision Tree dễ gặp hiện tượng quá khớp (overfitting), đặc biệt khi không kiểm soát độ sâu của cây hoặc kích thước tối thiểu của các nút lá. Để giải quyết hạn chế này, các phương pháp cắt tỉa cây (pruning) hoặc các mô hình tổng hợp như Gradient Boosting thường được áp dụng nhằm nâng cao hiệu quả dự báo và tính ổn định của mô hình.

3.3.3. Mô hình LightGBM (Light Gradient Boosting Machine)

LightGBM là một thuật toán tăng cường gradient (Gradient Boosting) được thiết kế để giải quyết các bài toán phân loại và hồi quy với hiệu suất cao. Điểm nổi bật của LightGBM là khả năng xử lý bộ dữ liệu có nhiều đặc trưng mà không làm giảm hiệu quả. Thuật toán hoạt động dựa trên việc xây dựng các cây quyết định tuần tự, trong đó mỗi cây sau được đào tạo để giảm thiểu lỗi của cây trước (Wang & Zhao, 2022). Hàm Loss (Loss Function) được sử dụng trong LightGBM thường có dạng:

$$L = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \lambda \sum_{k=1}^{K} \Omega(T_k) \quad (7)$$

Trong đó, $\ell(y_i, \hat{y}_i)$ là hàm Loss dựa trên sai số dự đoán, $\Omega(T_k)$ là độ phức tạp của cây T_k và λ là tham số điều chỉnh nhằm kiểm soát hiện tượng quá khớp.

Ưu điểm lớn của LightGBM là khả năng xây dựng cây dựa trên thuật toán Leaf-wise Growth thay vì Level-wise Growth như trong các phương pháp truyền thống. Cụ thể, trong Leaf-wise Growth, tại mỗi bước, thuật toán sẽ chọn mở rộng lá có mức độ giảm lỗi lớn nhất, giúp tăng hiệu quả học tập. Công thức tính mức độ giảm lỗi (Δ) cho một nhánh phân chia là:

$$\Delta = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \quad (8)$$

Trong đó, G_L, G_R là tổng gradient của lá trái và phải, H_L, H_R là tổng hessian (giá trị đạo hàm bậc hai) của lá trái và phải, và λ là tham số điều chỉnh. Phương pháp này giúp LightGBM tạo ra các cây tối ưu nhanh hơn và hiệu quả hơn trên dữ liệu lớn.

LightGBM còn được đánh giá cao nhờ khả năng hỗ trợ các tính năng tối ưu hóa như Histogram-based Binning để giảm chi phí tính toán. Thay vì sử dụng giá trị gốc của dữ liệu, thuật toán chia các giá trị thành các nhóm (bins) và sử dụng giá trị đại diện để tính toán. Điều này giúp giảm kích thước dữ liệu mà không làm mất thông tin quan trọng. Tuy nhiên, mô hình cũng yêu cầu thời gian tính toán và tinh chỉnh hyperparameter để đạt hiệu suất tối ưu.

3.4. Tiêu chí đánh giá khả năng dự báo rủi ro vỡ nợ của các mô hình

Ma trận nhầm lẫn (Confusion matrix), tỷ lệ chính xác (Accuracy) và điểm F1 (F1 Score) là 03 tiêu chí được sử dụng để đánh giá kết quả dự báo RRVN của các mô hình. Các tiêu chí này được mô tả cụ thể như sau:

(i) Ma trận nhầm lẫn (Confusion matrix)

Ma trận nhầm lẫn gồm có 04 thành phần chính True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN) được diễn giải cụ thể trong Bảng 1, cung cấp cái nhìn chi tiết về khả năng dự báo RRVN của mô hình với từng lớp trong tập dữ liệu. Hơn nữa, việc phân tích ma trận nhầm lẫn giúp các nhà nghiên cứu hiểu rõ hơn về sự hiệu quả cũng như mức độ sai số của mô hình, từ đó có những điều chỉnh phù hợp nhằm nâng cao mức độ chính xác và khả năng dự báo của mô hình (Tharwat, 2021). Giả sử, giá trị 0 biểu thị cho những doanh nghiệp không vỡ nợ, giá trị 1 biểu thị cho những doanh nghiệp vỡ nợ thì ma trân nhầm lẫn sẽ được biểu diễn dưới dạng như sau:

Bảng 1. Ma trận nhầm lẫn (Confusion matrix)

	Mô hình dự đoán		
Dữ liệu thực tế	Các lớp	Không vỡ nợ (0)	Vỡ nợ (1)
	Không vỡ nợ (0)	**True Negative (TN) –** Số lượng doanh nghiệp không vỡ nợ được mô hình dự báo đúng	**False Positive (FP) –** Số lượng doanh nghiệp không vỡ nợ được mô hình dự báo là vỡ nợ
	Vỡ nợ (1)	**False Negative (FN) –** Số lượng doanh nghiệp vỡ nợ nhưng được mô hình dự báo là không vỡ nợ	**True Positive (TP) –** Số lượng doanh nghiệp vỡ nợ và được mô hình dự báo đúng

Ghi chú: *(ii) Tỷ lệ chính xác (Accuracy)*

Tiêu chí này phản ánh khả năng dự báo của mô hình trong việc phân biệt chính xác giữa các doanh nghiệp vỡ nợ và không vỡ nợ. Đây được xem là thước đo hiệu quả tổng thể của mô hình (Lessmann và cộng sự, 2015). Tiêu chí này được xác định bởi công thức sau:

$$Accuracy = (TN + TP) / (TN + FN + TP + FP) \quad (9)$$

Trong đó: TN, FN, TP và FP là các thông tin được xác định từ ma trận nhầm lẫn.

(ii) Điểm F1 (F1 Score)

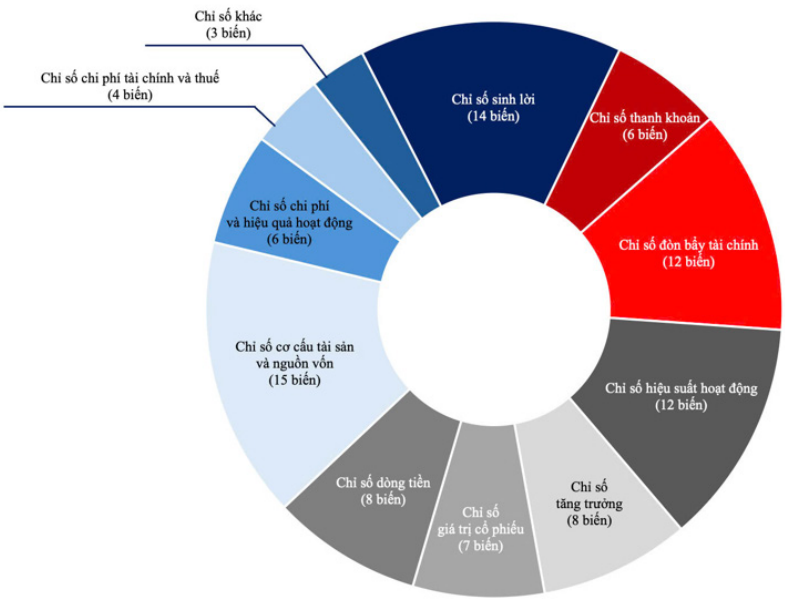
Tiêu chí này quan trọng trong việc xác định rằng, mô hình không chỉ dự đoán chính xác doanh nghiệp vỡ nợ mà còn hạn chế các trường hợp báo động sai. Từ đó có thể cung cấp một thước đo tối ưu giúp các nhà nghiên cứu điều chỉnh mô hình để đạt được trạng thái cân bằng giữa việc phát hiện rủi ro và giảm thiểu sai số dự báo (Hegde và cộng sự, 2023). Với các giá trị TN, FN, TP và FP được xác định từ ma trận nhầm lẫn, điểm F1 được tính toán như sau:

$$F_1 = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \quad (10)$$

4. Kết quả nghiên cứu

4.1. Dữ liệu nghiên cứu

Bộ dữ liệu nghiên cứu được thu thập từ 15,862 doanh nghiệp và công khai trên Kaggle (2024). Dữ liệu đầu vào bao gồm 95 biến giải thích là các chỉ số tài chính của doanh nghiệp và 01 biến mục tiêu được trình bày chi tiết trong phần phụ lục 01. Trong đó, các biến giải thích có thể chia làm 11 nhóm chỉ số tài chính quan trọng bao gồm Nhóm chỉ số sinh lời, Nhóm chỉ số thanh khoản, Nhóm chỉ số đòn bẩy tài chính, Nhóm chỉ số hiệu suất hoạt động, Nhóm chỉ số tăng trưởng, Nhóm chỉ số giá trị cổ phiếu, Nhóm chỉ số dòng tiền, Nhóm chỉ số cơ cấu tài sản và nguồn vốn, Nhóm chỉ số chi phí và hiệu quả hoạt động, Nhóm chỉ số chi phí tài chính và thuế, và Nhóm chỉ số khác. Số lượng các biến giải thích trong từng nhóm được mô tả cụ thể trong Hình 2.



Hình 2. Số lượng các biến giải thích trong từng nhóm chỉ số tài chính

Bên cạnh đó, biến mục tiêu là biến phân loại với hai giá trị là 0 và 1 phản ánh tình trạng vỡ nợ của doanh nghiệp. Trong đó, giá trị 0 đại diện cho tình trạng không vỡ nợ của doanh nghiệp chiếm phần lớn với số lượng là 15,284 tương ứng với tỷ lệ là 96,36% trong bộ dữ liệu. Số lượng dữ liệu có giá trị 1, đại diện cho doanh nghiệp vỡ nợ có số lượng tương đối thấp chỉ khoảng 578 doanh nghiệp, tương ứng với tỷ lệ 3,64% trong tệp dữ liệu nghiên cứu. Dựa trên thống kê này, chúng ta có thể thấy bộ dữ liệu nghiên cứu có sự mất cân bằng lớn giữa hai lớp nghiên cứu, điều này đòi hỏi các nhà nghiên cứu cần phải xử lý vấn đề này trước khi huấn luyện mô hình. Đây cũng là vấn đề rất đặc trưng trong lĩnh vực quản trị rủi ro tín dụng, khi số lượng các doanh nghiệp vỡ nợ chiếm tỷ lệ rất thấp khi so sánh với số lượng các doanh nghiệp không vỡ nợ (*xem Phụ lục 3 online*).

4.2. Phân tích kết quả nghiên cứu

Kết quả về khả năng dự báo RRVN doanh nghiệp của bốn phương pháp LCĐT khi kết hợp với ba mô hình dự báo được trình bày cụ thể trong Bảng 2. Cụ thể như sau:

Kết quả cho thấy, LightGBM đạt hiệu suất cao nhất trong ba mô hình khi không áp dụng phương pháp LCĐT, với tỷ lệ chính xác 80,7% và điểm F1 là 80,0%. Trong khi đó, Decision Tree có hiệu suất thấp hơn với độ chính xác 76,0% và Hồi quy Logistic có độ chính xác thấp nhất, chỉ 67,2%. Điều này phản ánh rằng, LightGBM có khả năng khái quát tốt hơn trên bộ dữ liệu so với Decision Tree và Hồi quy Logistic. Tuy nhiên, khi kết hợp với các phương pháp LCĐT, hiệu suất của tất cả các mô hình đều cải thiện đáng kể, đặc biệt là với phương pháp LCĐT Loại bỏ đặc trưng đệ quy (RFE).

Khi kết hợp với RFE, tất cả các mô hình đều đạt kết quả cao nhất so với các phương pháp khác. LightGBM + RFE đạt tỷ lệ chính xác 98,7% và điểm F1 là 98,0%, cải thiện đáng kể so với LightGBM ban đầu. Tương tự, Decision Tree + RFE đạt 95,0% độ chính xác, trong khi Logistic + RFE có độ chính xác 96,4%, tăng mạnh từ mức thấp ban đầu của Hồi quy Logistic. Điều này cho thấy, phương pháp RFE giúp mô hình chọn ra các đặc trưng quan trọng nhất, giảm bớt nhiễu và cải thiện khả năng phân loại chính xác.

So sánh ba phương pháp Loại bỏ đặc trưng chuyển tiếp (FFS), Loại bỏ đặc trưng ngược (BFE), và Loại bỏ đặc trưng dựa trên hoán vị

(PFI), ta thấy rằng FFS mang lại hiệu suất cao hơn một chút so với BFE và PFI. Chẳng hạn như, LightGBM + FFS có độ chính xác 98,2%, cao hơn BFE (97,9%) và PFI (97,4%). Xu hướng này cũng đúng với Decision Tree và Hồi quy Logistic, trong đó FFS mang lại kết quả tốt hơn một chút so với hai phương pháp còn lại. Điều này cho thấy, phương pháp LCĐT chuyển tiếp (FFS) có khả năng tối ưu hóa tốt hơn trong việc chọn biến quan trọng so với BFE và PFI.

Khi phân tích ma trận nhầm lẫn, ta thấy rằng việc áp dụng LCĐT giúp giảm đáng kể số lượng False Positive (FP) và False Negative (FN). Cụ thể, đối với Decision Tree ban đầu, số FP là 3805, nhưng khi sử dụng RFE, con số này giảm xuống còn 613, tức là mô hình phân loại sai ít hơn đáng kể. Tương tự, với Hồi quy Logistic, số FN giảm mạnh từ 5884 xuống còn 587 khi áp dụng RFE, giúp gia tăng tỷ lệ chính xác tổng thể. Điều này nhấn mạnh vai trò quan trọng của LCĐT trong việc giảm sai sót phân loại.

Sau khi áp dụng LCĐT, LightGBM vẫn là mô hình hoạt động tốt nhất, nhưng đáng chú ý là Hồi quy Logistic cải thiện đáng kể, đạt hiệu suất gần tương đương với Decision Tree. Cụ thể, Logistic + RFE đạt 96,4% độ chính xác, so với Decision Tree + RFE là 95,0%. Điều này chứng tỏ rằng Hồi quy Logistic có thể hoạt động rất hiệu quả nếu được LCĐT phù hợp. Nhìn chung, RFE là phương pháp tốt nhất cho tất cả các mô hình, giúp cải thiện độ chính xác và giảm sai sót phân loại đáng kể.

Bảng 2. Kết quả dự báo RRVN của các mô hình khi kết hợp với các phương pháp LCĐT

Mô hình dự báo kết hợp với các phương pháp LCĐT	**Tiêu chí đánh giá**					
	Ma trận nhầm lẫn				***Tỷ lệ chính xác***	***Điểm F1***
	TN	***FP***	***FN***	***TP***		
LightGBM	12850	3012	3110	12752	0,80703	0,80023
LightGBM + RFE	15701	161	241	15621	0,98735	0,98008
LightGBM + FFS	15618	244	324	15538	0,98209	0,97475
LightGBM + BFE	15575	287	368	15494	0,97934	0,97196
LightGBM + PFI	15504	358	438	15424	0,97492	0,96767
Decision Tree	12057	3805	3803	12059	0,76019	0,76033
Decision Tree + RFE	15249	613	955	14907	0,95056	0,9216
Decision Tree + FFS	15076	786	1132	14730	0,93953	0,91067
Decision Tree + BFE	15175	687	1031	14831	0,94587	0,91691
Decision Tree + PFI	15125	737	1082	14780	0,94266	0,91372
Logistic	9757	6105	5884	9978	0,67210	0,61012
Logistic + RFE	15317	545	587	15275	0,96434	0,96055
Logistic + FFS	14993	869	828	15034	0,94652	0,95015
Logistic + BFE	14842	1020	978	14884	0,93700	0,94067
Logistic + PFI	14528	1334	1206	14656	0,91992	0,93089

4.3. Phân tích các đặc trưng quan trọng được lựa chọn trên bốn phương pháp RFE, FFS, BFE, và PFI

Để hiểu rõ hơn sự khác biệt về hiệu quả của bốn phương pháp RFE, FFS, BFE, và PFI khi kết hợp với các mô hình học máy, chúng ta hãy cùng phân tích cách thức mà các phương pháp LCĐT quan trọng. Kết quả mô tả về 20 đặc trưng quan trọng nhất mà bốn phương pháp này lựa chọn trong 95 đặc trưng ban đầu để kết

hợp với các mô hình trong quá trình dự báo khả năng vỡ nợ của doanh nghiệp (*xem Phụ lục 4 online*). Cụ thể như sau:

Trong cả bốn phương pháp, đặc trưng Mức độ phụ thuộc vào nợ vay (Borrowing Dependency) xuất hiện ở vị trí hàng đầu, thể hiện tầm quan trọng vượt trội của đặc trưng này trong việc dự báo RRVN. Điều này cho thấy rằng, mức độ phụ thuộc vào nợ vay của doanh nghiệp, được đo lường bằng tổng nợ vay trên tổng tài sản, là một yếu tố cần được lựa chọn trong mô hình dự báo. Bên cạnh đó, cả bốn phương pháp cũng đồng thuận về tầm quan trọng của 04 đặc trưng khác bao gồm tỷ lệ vòng quay tiền mặt (Cash Turnover Rate), mức độ đòn bẩy tài chính (Degree Of Financial Leverage), tỷ lệ tiền mặt trên tổng tài sản (Cash To Total Assets), và tổng tài sản trên giá trị GNP (Total Assets To GNP Price).

Phân tích sâu hơn về sự khác biệt trong LCĐT quan trọng của bốn phương pháp, chúng ta có thể phát hiện rằng phương pháp RFE có xu hướng lựa chọn các đặc trưng thuộc nhóm chỉ số sinh lời và nhóm chỉ số đòn bẩy tài chính hơn là lựa chọn các đặc trưng thuộc nhóm chỉ số thanh khoản, hiệu suất hoạt động hay nhóm chỉ số dòng tiền. Trong khi đó, phương pháp FFS lại tập trung vào việc lựa chọn các đặc trưng thuộc nhóm chỉ số hiệu suất hoạt động và ít tập trung vào các nhóm chỉ số khác. Phương pháp BFE lại có xu hướng ưu tiên các đặc trưng thuộc nhóm chỉ số thanh khoản, đòn bẩy tài chính và hiệu suất hoạt động hơn các đặc trưng thuộc nhóm chỉ số tài chính khác. Cuối cùng, phương pháp PFI tập trung chủ yếu vào các đặc trưng thuộc nhóm chỉ số đòn bẩy tài chính và hiệu suất hoạt động.

Nhìn chung, sự khác biệt giữa các phương pháp không chỉ nằm ở đặc trưng được chọn mà còn ở cách chúng ưu tiên các nhóm đặc trưng. RFE và BFE phù hợp với các mô hình cần cái nhìn toàn diện, trong khi FFS và PFI mạnh mẽ trong việc tối ưu hóa các chỉ số cụ thể. Tùy thuộc vào mục tiêu nghiên cứu và bản chất dữ liệu, việc chọn phương pháp phù hợp có thể mang lại kết quả dự báo tốt nhất.

5. Kết luận

Bài viết đã phân tích và so sánh hiệu quả của bốn phương pháp LCĐT bao gồm RFE, FFS, BFE và PFI trong việc cải thiện kết quả dự báo RRVN của các mô hình học máy. Kết quả thực nghiệm cho thấy rằng, LCĐT đóng vai trò quan trọng trong việc tối ưu hóa hiệu suất, gia tăng mức độ chính xác và giảm thiểu sai số trong quá trình dự báo của các mô hình.

Trong ba mô hình được nghiên cứu, LightGBM tỏ ra vượt trội hơn cả về khả năng dự báo, đặc biệt khi kết hợp với phương pháp RFE. RFE đã cho thấy hiệu quả cao nhất trong việc loại bỏ các đặc trưng không cần thiết, giúp mô hình đạt độ chính xác cao nhất với điểm F1 cải thiện đáng kể. Điều này chứng minh rằng, phương pháp LCĐT phù hợp có thể giúp mô hình tập trung vào các đặc trưng quan trọng nhất, giảm nhiễu và nâng cao hiệu suất dự đoán. Bên cạnh đó, các phương pháp FFS, BFE và PFI cũng mang lại những cải thiện đáng kể nhưng chưa đạt mức tối ưu như RFE.

Ngoài ra, nghiên cứu cũng chỉ ra rằng, Mức độ phụ thuộc vào nợ vay, Tỷ lệ vòng quay tiền mặt hay Mức độ đòn bẩy tài chính là những đặc trưng quan trọng trong dự báo RRVN doanh nghiệp. Trong đó, đặc trưng Mức độ phụ thuộc vào nợ vay có sự ảnh hưởng lớn nhất, điều này có thể là những gợi ý quan trọng cho các tổ chức tín dụng trong quá trình xây dựng và đánh giá rủi ro tín dụng.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế nhất định. *Thứ nhất*, mặc dù bộ dữ liệu được công khai trên Kaggle (2024), tuy nhiên việc thiếu thông tin về nguồn gốc địa lý cụ thể của dữ liệu, chẳng hạn như quốc gia, khu vực, hay bối cảnh kinh tế - tài chính nơi dữ liệu được thu thập có thể ảnh hưởng đáng kể đến tính khái quát hóa của kết quả nghiên cứu. Bên cạnh đó, việc không biết bộ dữ liệu được thu thập từ một hay nhiều quốc gia cũng có thể ảnh hưởng đến tính đồng nhất của tập dữ liệu, từ đó không những tác động đến khả năng diễn giải kết quả mà còn hạn chế tính ứng dụng trong thực tiễn của nghiên cứu. *Thứ hai*, nghiên cứu chỉ tập

trung đánh giá tính hiệu quả của các phương pháp LCĐT khi kết hợp với các mô hình học máy mà chưa xem xét đến tính vững của mô hình. Điều này đặt ra yêu cầu về việc kiểm định tính ổn định của các phương pháp LCĐT trên các bộ dữ liệu đa dạng hơn, cũng như đánh giá hiệu suất của mô hình trong các điều kiện thay đổi của thị trường tài chính.

Tóm lại, bài viết khẳng định LCĐT không chỉ giúp nâng cao hiệu quả dự báo của mô hình học máy mà còn giúp xác định các yếu tố tài chính quan trọng đối với RRVN doanh nghiệp. Trong tương lai, các nghiên cứu có thể mở rộng bằng cách thử nghiệm các phương pháp LCĐT trên các bộ dữ liệu mới hoặc áp dụng trên các mô hình tiên tiến hơn như mô hình học sâu để kiểm định tính ổn định của các kết quả đạt được, đồng thời đánh giá kỹ hơn tác động của LCĐT đến nguy cơ quá khớp và khả năng tổng quát hóa trong nhiều bối cảnh dữ liệu khác nhau.

Tài liệu tham khảo

Ahmed, R., Fahad, N., Miah, M. S. U., Hossen, M. J., Morol, M. K., Mahmud, M., & Rahman, M. M. (2024). A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction. *Healthcare Analytics*, *6*. https://doi.org/10.1016/j.health.2024.100362

Bian, L., Qin, X., Zhang, C., Guo, P., & Wu, H. (2023). Application, interpretability and prediction of machine learning method combined with LSTM and LightGBM-a case study for runoff simulation in an arid area. *Journal of Hydrology, 625*. https://doi.org/10.1016/j.jhydrol.2023.130091

Chen, D., Ye, J., & Ye, W. (2023). Interpretable selective learning in credit risk. *Research in International Business and Finance, 65*. https://doi.org/10.1016/j.ribaf.2023.101940

Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Chen, X., & Li, H.-D. (2010). Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemometrics and Intelligent Laboratory Systems, 103*(2), 129–136. https://doi.org/10.1016/j.chemolab.2010.06.008

Elghazel, H., & Aussem, A. (2015). Unsupervised feature selection with ensemble learning. *Machine Learning, 98*, 157–180. https://doi.org/10.1007/s10994-013-5337-8

Fallahpour, S., Lakvan, E. N., & Zadeh, M. H. (2017). Use of combined approach of support vector machine and feature selection for financial distress prediction of listed companies in Tehran stock exchange market. *Financial Research Journal, 19*(1), 139–156. https://doi.org/10.22059/jfr.2015.52758

Guo, W., & Zhou, Z. Z. (2022). A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction. *Journal of Forecasting, 41*(6), 1248–1313. https://doi.org/10.1002/for.2856

Han, C., Kang, H., Kim, G., & Yi, J. (2012). Logit regression-based bankruptcy prediction of Korean firms. *Asia-Pacific Journal of Risk and Insurance, 7*(1). https://doi.org/10.1515/2153-3792.1159

Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems, 51*, 72–84. https://doi.org/10.1016/j.knosys.2013.07.008

Hegde, S. K., Hegde, R., R, K. P., S, S. S., Marthanda, A. V. G. A., & Logu, K. (2023). Performance analysis of machine learning algorithm for the credit risk analysis in the banking sector. *Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication* (ICCMC) (pp. 57-63). IEEE. https://doi.org/10.1109/ICCMC56507.2023.10083580

Kaggle. (2024). Corporate Bankruptcy Dataset. https://www.kaggle.com/datasets/kandie2908/corporate-default

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys* (CSUR), *50*(6), 1–45. https://doi.org/10.1145/3136625

Lao, Z., He, D., Wei, Z., Shang, H., Jin, Z., Miao, J., & Ren, C. (2023). Intelligent fault diagnosis for rail transit switch machine based on adaptive feature selection and improved LightGBM. *Engineering Failure Analysis*, *148*. https://doi.org/10.1016/j.engfailanal.2023.107219

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*, *247*(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

Maleki, N., Zeinali, Y., & Niaki, S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications, 164*. https://doi.org/10.1016/j.eswa.2020.113981

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. *Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA)* (pp. 18-20). IEEE. https://doi.org/10.1109/ICACA.2016.7887916

Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. *Cogent Economics & Finance, 13*(1). https://doi.org/10.1080/23322039.2025.2465971

Qi, C., Diao, J., & Qiu, L. (2019). On estimating model in feature selection with cross-validation. *IEEE Access, 7*, 33454-33463. https://doi.org/10.1109/ACCESS.2019.2892062

Ren, K., Fang, W., Qu, J., Zhang, X., & Shi, X. (2020). Comparison of eight filter-based feature selection methods for monthly streamflow forecasting – Three case studies on CAMELS data sets. *Journal of Hydrology, 586*. https://doi.org/10.1016/j.jhydrol.2020.124897

Rudnicki, W. R., Wrzesień, M., & Paja, W. (2015). All relevant feature selection methods and applications. In U. Stańczyk, L. Jain (Eds.), *Feature selection for data and pattern recognition* (pp. 11-28). Studies in Computational Intelligence, vol 584. Springer. https://doi.org/10.1007/978-3-662-45620-0_2

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*. https://doi.org/10.1007/s42452-021-04148-9

Sahu, M., & Dash, R. (2022). A classification model for multispectral forest datatype with the help of a decision tree and wrapper-based forward feature selection technique. In J. P. Sahoo, A. K. Tripathy, M. Mohanty, K. C. Li, & A. K. Nayak (Eds.), *Advances in distributed computing and machine learning* (pp. 444–456). *Lecture Notes in Networks and Systems, vol 302*. Springer. https://doi.org/10.1007/978-981-16-4807-6_42

Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics, 19*. https://doi.org/10.1186/s12859-018-2451-4

Shi, S., Tse, R., Luo, W., & D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systematic review. *Neural Computing and Applications, 34*, 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, *85*, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Vishraj, R., Gupta, S., & Singh, S. (2023). Evaluation of feature selection methods utilizing random forest and logistic regression for lung tissue categorization using HRCT images. *Expert Systems*, *40*(8). https://doi.org/10.1111/exsy.13320

Wang, D.-n., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, 259-268. https://doi.org/10.1016/j.ins.2022.04.058

Xia, S., & Yang, Y. (2022). An iterative model-free feature screening procedure: Forward recursive selection. Knowledge-Based Systems, *246*. https://doi.org/10.1016/j.knosys.2022.108745

Xia, S., & Yang, Y. (2023). A model-free feature selection technique of feature screening and random forest-based recursive feature elimination. *International Journal of Intelligent Systems, 2023*. https://doi.org/10.1155/2023/2400194