Ứng dụng mô hình stacking kết hợp smote và tối ưu hóa Bayesian đánh giá rủi ro tín dụng

Dương Hớn Minh

Khoa Dược - Trường Đại học Nguyễn Tất Thành
dhminh@ntt.edu.vn

Tóm tắt

Dự đoán rủi ro tín dụng là nhiệm vụ quan trọng đối với các tổ chức tài chính nhằm giảm thiểu nguy cơ vỡ nợ và tối ưu hóa quyết định cho vay. Trong bối cảnh sự phát triển nhanh chóng của các kỹ thuật học máy, nhiều phương pháp phân loại đã được phát triển để cải thiện khả năng dự đoán rủi ro tín dụng. Nghiên cứu này áp dụng mô hình stacking để đánh giá rủi ro tín dụng, kết hợp dự đoán từ nhiều mô hình học máy khác nhau, bao gồm XGBoost, Random Forest, và CatBoost. Một mô hình meta, hồi quy logistic, được sử dụng để tối ưu hóa dự đoán từ các mô hình cơ sở để đưa ra dự đoán. Dữ liệu được xử lý bằng kỹ thuật SMOTE để cân bằng và các siêu tham số của các mô hình cơ sở được tối ưu hóa thông qua phương pháp tối ưu hóa Bayesian. Kết quả cho thấy mô hình stacking đạt được độ chính xác 95,50 % và chỉ số ROC-AUC đạt 98,15 %, chứng tỏ độ tin cậy cao của các dự đoán. Kết quả này cung cấp về khả năng ứng dụng của các mô hình học máy trong việc đánh giá rủi ro tín dụng, hỗ trợ các tổ chức tài chính trong việc ra quyết định cấp tín dụng cho cá nhân.

Nhận 02/09/2024
Được duyệt 03/12/2024
Công bố 28/12/2024

Từ khóa
học máy,
học máy tổ hợp, tối ưu hóa Bayesian, SMOTE, dự đoán rủi ro tín dụng

1 Giới thiệu

1.1 Đặt vấn đề

Dự đoán rủi ro tín dụng (RRTD) là một khía cạnh quan trọng trong quản lý rủi ro tài chính, đóng vai trò then chốt trong các quyết định của các tổ chức tài chính. Dự đoán RRTD liên quan đến việc đánh giá khả năng trả nợ của người vay, từ đó xác định khả năng vỡ nợ. Việc dự đoán RRTD hiệu quả giúp các tổ chức tài chính giảm thiểu tổn thất, tối ưu hóa quyết định cho vay và quản lý danh mục đầu tư một cách hiệu quả hơn [1].

Dự đoán RRTD có vai trò thiết yếu vì nhiều lý do. Thứ nhất, dự đoán RRTD giúp các tổ chức tài chính giảm thiểu tổn thất do vỡ nợ. Bằng cách đánh giá chính xác mức độ tin cậy của người vay, các tổ chức cho vay có thể đưa ra các quyết định hợp lý về việc chấp nhận hay từ chối các đơn xin vay. Thứ hai, dự đoán RRTD hiệu quả góp phần vào sự ổn định của hệ thống tài chính. Khi các ngân hàng và tổ chức cho vay có thể dự đoán chính xác khả năng vỡ nợ, các ngân hàng và tổ chức cho vay có thể quản lý dự trữ vốn tốt hơn và giảm thiểu nguy cơ phá sản. Thứ ba, đánh giá RRTD chính xác giúp cung cấp tín dụng cho những người vay xứng đáng, thúc đẩy tăng trưởng kinh tế và tăng cường tài chính [2].

Từ thập niên 1960, hệ thống điểm tín dụng đã được áp dụng để đánh giá xem một người vay có đủ điều kiện và có khả năng trả nợ đúng hạn hay không. Điểm tín dụng hỗ trợ các quyết định tín dụng bằng cách sử dụng

https://doi.org/10.55401/6tb18p40

các mô hình toán học để chuyển đổi dữ liệu thu thập từ khách hàng, hệ thống nội bộ và các cơ quan tín dụng thành một điểm số. Trong lĩnh vực tín dụng bán lẻ, phương pháp này không chỉ giảm bớt tính chủ quan trong việc đánh giá của các chủ nợ mà còn tối ưu hóa giá trị của thông tin hiện có và tiết kiệm đáng kể chi phí nhân lực [3].

Trải qua nhiều năm phát triển, ngoài hồi quy logistic, các phương pháp học có giám sát như rừng ngẫu nhiên, XGBoost và CatBoost đã phát triển nhanh chóng. Hồi quy logistic xuất hiện từ những năm 1950 và là một trong những phương pháp cơ bản trong phân tích dữ liệu. Các thuật toán tiên tiến như rừng ngẫu nhiên, XGBoost và CatBoost sau đó đã được phát triển và ứng dụng rộng rãi trong nhiều lĩnh vực. Sự hỗ trợ của công nghệ giúp các nhà nghiên cứu, ngân hàng và tổ chức tài chính sử dụng các thuật toán này để đào tạo mô hình dự đoán khả năng đủ điều kiện vay dựa trên lịch sử tín dụng và dữ liệu khác, giúp dễ dàng chọn lọc những người đủ điều kiện trước khi phê duyệt khoản vay [4, 5].

Một trong những kỹ thuật tiên tiến và hiệu quả được sử dụng rộng rãi trong lĩnh vực học máy là phương pháp học máy tổ hợp (Ensemble Learning). Đây là một kỹ thuật mạnh mẽ nhằm kết hợp nhiều mô hình học máy để tạo ra một mô hình dự đoán có độ chính xác cao hơn so với bất kỳ mô hình đơn lẻ nào. Các phương pháp học máy tổ hợp phổ biến bao gồm Bagging (Bootstrap Aggregating), Boosting, và Stacking. Gần đây, stacking đã được sử dụng để cải thiện độ chính xác trong việc dự đoán RRTD [6, 7].

1.2 Mục tiêu nghiên cứu

Mặc dù các phương pháp truyền thống như hồi quy logistic và các thuật toán học máy tiên tiến như rừng ngẫu nhiên, XGBoost và CatBoost đã được áp dụng rộng rãi trong dự đoán RRTD, vẫn tồn tại một số hạn chế và thách thức cần giải quyết.

Thứ nhất, hầu hết các nghiên cứu hiện tại chủ yếu tập trung vào việc cải thiện độ chính xác của các mô hình đơn lẻ, nhưng ít chú ý đến khả năng tổng hợp và kết hợp các mô hình để tạo ra một mô hình dự đoán mạnh mẽ hơn. Các phương pháp học máy tổ hợp như Bagging và Boosting đã được nghiên cứu và ứng dụng trong nhiều lĩnh vực, nhưng việc áp dụng xếp chồng, một kỹ thuật kết hợp mạnh mẽ hơn, trong lĩnh vực dự đoán RRTD vẫn còn hạn chế. Điều này mở ra cơ hội nghiên cứu về việc tận dụng các mô hình cơ sở mạnh mẽ và xây dựng mô hình meta hiệu quả để cải thiện hiệu suất dự đoán.

Thứ hai, mặc dù nhiều nghiên cứu đã áp dụng các kỹ thuật tối ưu hóa mô hình, phương pháp tối ưu hóa Bayes (Bayesian Optimization – BO) chưa được khai thác triệt để trong việc tìm kiếm và lựa chọn các base learner tối ưu. BO có tiềm năng lớn trong việc tối ưu hóa quá trình huấn luyện mô hình, đặc biệt khi kết hợp với các phương pháp học máy tổ hợp như xếp chồng. Tuy nhiên, ứng dụng của BO trong việc cải thiện hiệu suất của các mô hình xếp chồng trong dự đoán RRTD vẫn chưa được khám phá đầy đủ.

Cuối cùng, các nghiên cứu hiện tại thường tập trung vào một số chỉ số đánh giá nhất định như độ chính xác, độ nhạy và chỉ số F1. Tuy nhiên, chưa có nhiều nghiên cứu đánh giá toàn diện các chỉ số quan trọng khác như ROC AUC hay khả năng tổng quát hóa của mô hình trên các tập dữ liệu thực tế. Do đó, cần có thêm nghiên cứu để đánh giá toàn diện hiệu quả của các mô hình và đề xuất các phương pháp cải tiến có khả năng ứng dụng trong thực tiễn.

Vì vậy, mục tiêu của nghiên cứu này là áp dụng kỹ thuật xếp chồng kết hợp với phương pháp tối ưu hóa Bayes để xây dựng một hệ thống dự đoán RRTD vượt trội, đồng thời đánh giá toàn diện các chỉ số hiệu suất của mô hình nhằm mang lại giá trị thực tiễn cao cho các tổ chức tài chính.

2 Cơ sở lý thuyết

2.1 Kỹ thuật xây dựng đặc trưng

Tạo đặc trưng là một bước quan trọng trong quy trình học máy, bao gồm việc tạo ra các đặc trưng mới hoặc chuyển đổi các đặc trưng hiện có để cải thiện hiệu suất của mô hình. Đầu tiên, kiểm tra và loại bỏ các giá trị bị thiếu trong tập dữ liệu để đảm bảo tính đầy đủ của thông tin.

Sau đó, tiến hành xử lý các biến phân loại bằng các kỹ thuật mã hóa phù hợp. Đối với các biến có số lượng giá trị khác nhau nhỏ hơn hoặc bằng 10, áp dụng phương pháp mã hóa nhãn để chuyển đổi các giá trị phân loại thành các số nguyên, giúp đơn giản hóa dữ liệu. Đối với

các biến phân loại có số lượng giá trị lớn hơn 10, sử dụng phương pháp mã hóa một nóng [8]. Kỹ thuật này tạo ra các cột mới đại diện cho từng giá trị riêng biệt và loại bỏ cột biến gốc, giúp tránh hiện tượng đa cộng tuyến và cải thiện độ chính xác của mô hình. Đa cộng tuyến có thể gây ra vấn đề nghiêm trọng trong quá trình huấn luyện mô hình, làm giảm khả năng dự đoán và tăng sai số của mô hình. Sau đó dùng bộ chuẩn hóa chuẩn (Stadard scaler) để chuẩn hóa các dữ liệu [9].

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

X là giá trị của đặc trưng cần chuẩn hóa.
μ là trung bình (mean) của đặc trưng.
σ là độ lệch chuẩn (standard deviation) của đặc trưng.

2.2 SMOT

SMOTE (Synthetic Minority Over-sampling Technique) là một phương pháp được sử dụng trong học máy để giải quyết vấn đề mất cân bằng lớp trong các tập dữ liệu. Nó hoạt động bằng cách tạo ra các mẫu tổng hợp cho lớp thiểu số để cân bằng phân phối các lớp. Kỹ thuật này cải thiện hiệu suất của mô hình bằng cách giảm thiểu sự thiên lệch đối với lớp đa số, điều này có thể xảy ra khi huấn luyện trên các tập dữ liệu mất cân bằng. SMOTE tạo ra các mẫu mới bằng cách nội suy giữa các ví dụ lớp thiểu số hiện có [12].

2.3 Tối ưu hóa Bayesian – Bayesian Optimization (BO)

Quá trình tối ưu tham số là một bước quan trọng không thể thiếu trong việc huấn luyện mô hình học máy với hiệu suất đánh giá cao. Việc xác định các tham số tối ưu cho mô hình không chỉ cải thiện độ chính xác mà còn giúp tăng cường khả năng tổng quát hóa của mô hình trên các tập dữ liệu khác nhau.

Tối ưu hóa Bayesian là một phương pháp hiệu quả để tối ưu hóa các hàm hộp đen (black-box functions) có chi phí đánh giá cao. BO đặc biệt hữu ích trong việc tinh chỉnh siêu tham số cho các mô hình học máy. Không giống như các kỹ thuật tối ưu hóa truyền thống, Tối ưu hóa Bayesian sử dụng một mô hình xác suất để dự đoán hiệu suất của các tổ hợp tham số khác nhau và chọn những tổ hợp hứa hẹn nhất để đánh giá. Cách tiếp cận này giảm số lượng các lần đánh giá hàm cần thiết để tìm các tham số tối ưu [13].

2.4 Mô hình học máy

Trong nghiên cứu này, ba mô hình học máy được lựa chọn để tiến hành nghiên cứu bao gồm XGBoost, rừng ngẫu nhiên, và CatBoost.

a) XGBoost (XGB)

XGBoost là một triển khai hiệu quả và có khả năng mở rộng của các máy tăng cường độ dốc. XGBoost nổi tiếng về tốc độ và hiệu suất, khả năng xử lý dữ liệu thưa, và các kỹ thuật điều chỉnh giúp ngăn ngừa hiện tượng quá khớp (overfitting). Hàm mục tiêu trong XGBoost kết hợp một hàm mất mát lồi và một thuật ngữ điều chỉnh, cải thiện cả độ chính xác dự đoán và khả năng giải thích [15].

b) Rừng ngẫu nhiên

Rừng ngẫu nhiên là một phương pháp học tập tập hợp xây dựng nhiều cây quyết định trong quá trình huấn luyện. Rừng ngẫu nhiên sử dụng phương pháp bagging, trong đó mỗi cây được huấn luyện trên một tập con ngẫu nhiên của dữ liệu, và các đặc trưng được chọn ngẫu nhiên tại mỗi điểm chia. Dự đoán cuối cùng được thực hiện bằng cách tổng hợp các dự đoán của tất cả các cây, giảm hiện tượng quá khớp của mô hình và cải thiện khả năng tổng quát [16].

c) CatBoost

CatBoost là một thuật toán tăng cường độ dốc xử lý hiệu quả các đặc trưng phân loại mà không cần tiền xử lý nhiều. CatBoost sử dụng tăng cường có thứ tự để giảm rò rỉ mục tiêu và cung cấp hiệu suất mạnh mẽ trên nhiều loại dữ liệu. CatBoost đặc biệt hữu ích trong việc xử lý các biến phân loại phổ biến trong các tập dữ liệu thực tế [17].

2.5 Mô hình xếp chồng

Mô hình xếp chồng là một kỹ thuật trong học máy, với nhiều mô hình được kết hợp để cải thiện hiệu suất dự đoán so với việc sử dụng một mô hình đơn lẻ. Các mô hình đơn lẻ được gọi là mô hình cơ sở, sử dụng một mô hình để kết hợp chúng thành mô hình meta bằng một kỹ thuật học máy khác, trong nghiên cứu này sử dụng hồi quy phi tuyến – *logistic regression*.

Mô hình stacking kết hợp dự đoán của nhiều mô hình học máy bằng cách sử dụng một mô hình meta để học cách tối ưu từ các dự đoán đó đã được triển khai ở nhiều nghiên cứu trước đây và cho thấy kết quả rất khả quan [7, 14]. Lý do chọn cả ba mô hình này làm mô hình cơ

sở là vì chúng mang lại sự đa dạng trong phương pháp học. Mỗi mô hình có những ưu điểm khác nhau, giúp hệ thống stacking khai thác tốt hơn các khía cạnh khác nhau của dữ liệu:

- *Rừng ngẫu nhiên* tốt trong việc giảm quá khớp bằng cách học từ các cây độc lập.
- *XGBoost* và *CatBoost* có khả năng tối ưu hóa hiệu suất và tránh quá khớp thông qua boosting, một kỹ thuật học tuần tự cải thiện mô hình.
- Bằng cách kết hợp các mô hình có khả năng tổng quát hóa cao và giảm quá khớp, mô hình xếp chồng sẽ tạo ra kết quả mạnh mẽ và ổn định hơn.

Hồi quy logistic được sử dụng làm mô hình meta trong mô hình xếp chồng ở nghiên cứu này. Đây là một mô hình phi tuyến được sử dụng cho phân loại nhị phân, ước lượng xác suất rằng một điểm đầu vào thuộc về một lớp nhất định. Bằng cách lấy các dự đoán từ các mô hình cơ sở làm đặc trưng đầu vào, hồi quy logistic có thể học cách gán trọng số tối ưu cho từng dự đoán của mô hình cơ sở, từ đó cải thiện hiệu suất dự đoán tổng thể [18, 19].

2.6 Các độ đo đánh giá

Đánh giá hiệu suất là một việc quan trọng và việc lựa chọn các độ đo nào cũng quan trọng không kém.

Ma trận nhầm lẫn (confusion matrix) là một công cụ mạnh mẽ và quan trọng trong việc đánh giá hiệu suất của các mô hình học máy, đặc biệt là trong các bài toán phân loại. Ma trận nhầm lẫn cho phép thấy được số lượng dự đoán đúng và sai của mô hình cho mỗi lớp. Ở Bảng 1, nó cho biết số lượng:

- True Positives (TP): số lượng dự đoán đúng cho lớp dương.
- True Negatives (TN): số lượng dự đoán đúng cho lớp âm.
- False Positives (FP): số lượng dự đoán sai, mô hình dự đoán là dương nhưng thực tế là âm.
- False Negatives (FN): số lượng dự đoán sai, mô hình dự đoán là âm nhưng thực tế là dương.

Bảng 1 Ma trận nhầm lẫn

Chân trị	+	−
Dự đoán +	TP	FP
Dự đoán −	FN	TN

ROC AUC là một trong những chỉ số quan trọng để đánh giá hiệu suất của các mô hình phân loại nhị phân. ROC AUC đo lường khả năng phân biệt giữa các lớp của mô hình, giúp hiểu rõ hơn về hiệu suất tổng thể của mô hình trong việc xác định các trường hợp dương tính và âm tính.

$$AUC = \int_0^1 TPR\ d(FPR) \quad (2)$$

Trong đó:

TPR (True Positive Rate) hay còn gọi là Độ nhạy (Recall).

FPR (False Positive Rate) được tính bằng công thức:

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

3 Phương pháp nghiên cứu

3.1 Mô tả dữ liệu

Để dự đoán RRTD hiệu quả bằng học máy, cần có dữ liệu chất lượng cao và kỹ thuật tạo đặc trưng mạnh mẽ. Dữ liệu dùng để huấn luyện các mô hình dự đoán trong nghiên cứu này được lấy từ tập "Tập dữ liệu RRTD" (Credit Risk Dataset) trên Kaggle. Tập dữ liệu bao gồm 11 cột và 32 581 dòng dữ liệu. Trong đó, cột Loan_status là cột mục tiêu cần dự đoán, còn 10 cột còn lại là các đặc trưng để dự đoán cột mục tiêu.

Bảng 2 Mô tả các biến có trong tập dữ liệu RRTD

Biến đầu vào	**Định nghĩa biến**
person_age	Tuổi của cá nhân
person_income	Thu nhập hàng năm của người vay
person_home_ownership	Loại hình sở hữu nhà - thuê, thế chấp, thuê mua, sở hữu hoặc khác
person_emp_length	Thời gian làm việc của cá nhân (tính theo năm)
loan_intent	Mục đích của khoản vay
loan_amnt	Số tiền được vay
loan_int_rate	Lãi suất của khoản vay
loan_status	Trạng thái thanh toán khoản vay (0: không vi phạm, 1: vi phạm)
loan_percent_income	Tỷ lệ (%) số tiền vay so với tổng thu nhập
cb_person_default_on_file	Lịch sử các khoản nợ (nếu có) của người vay

cb_person_cred_hist_length	Độ dài lịch sử tín dụng của người vay

Việc xem thống kê mô tả của dữ liệu cho biết chi tiết và đặc điểm thống kê của dữ liệu. Được mô tả qua Bảng 3.

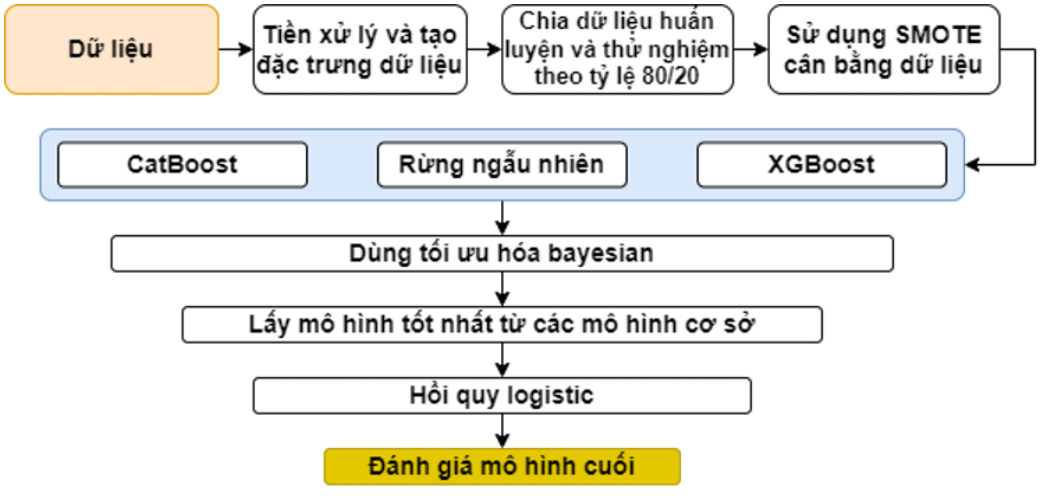
Bảng 3 Thống kê mô tả

Thống kê	person_age	person_ income	person_ emp_length	loan_amnt	loan_int_ rate	loan_ status	loan_ percent_ income	cb_person_ cred_hist_ length
Số lượng	32581,0	32581,0	31686,0	32581,0	29465,0	32581,0	32581,0	32581,0
Trung bình	27,7346	66745,26	4,793856	9593,371	11,01169	0,211364	0,170283	5,894211
Độ lệch chuẩn	6,340878	62358,45	4,14263	6322,085	3,24205	0,408396	0,106702	4,055001
Nhỏ nhất	20,0	4000,0	0,0	500,0	5,42	0,0	0,0	2,0
25%	23,0	40000,0	2,0	4000,0	7,9	0,0	0,09	3,0
50%	26,0	65000,0	4,0	8000,0	10,9	0,0	0,15	4,0
75%	30,0	90500,0	7,0	12000,0	13,47	0,0	0,23	7,0
Lớn nhất	144,0	600000,0	123,0	35000,0	23,22	1,0	0,83	30,0

Nghiên cứu này tập trung trên một bộ dữ liệu được công bố công khai trên Kaggle nhằm đánh giá và so sánh năng lực của các mô hình học máy trên cùng một nền tảng dữ liệu, giúp đảm bảo tính khách quan và công bằng khi nghiên cứu. Ngoài ra, các nhà nghiên cứu khác có thể dễ dàng truy cập, tái hiện và xác minh kết quả của nghiên cứu này, như rất nhiều các công bố khác đã sử dụng bộ dữ liệu này để tiến hành thử nghiệm các mô hình học máy khác; từ đó giúp chứng minh phương pháp đề xuất của nghiên cứu đạt được kết quả khả quan và có giá trị.

3.2 Phương pháp đề xuất

Quy trình dự đoán RRTD bắt đầu với việc thu thập và tiền xử lý dữ liệu, bao gồm làm sạch dữ liệu, xử lý giá trị thiếu và mã hóa các biến phân loại. Tiếp theo, kỹ thuật SMOTE được sử dụng để cân bằng dữ liệu, giải quyết vấn đề mất cân bằng lớp. Sau đó, ba mô hình cơ sở là CatBoost, rừng ngẫu nhiên và XGBoost được huấn luyện và tối ưu hóa bằng tối ưu hóa Bayesian. Kết quả từ các mô hình cơ sở được kết hợp lại bằng hồi quy logistic để tạo ra mô hình tổng hợp cuối cùng. Cuối cùng, hiệu suất của mô hình tổng hợp được đánh giá.



Hình 1 Sơ đồ phương pháp được sử dụng trong nghiên cứu

Các bước cụ thể triển khai trong nghiên cứu này:
- Bước 1: tiền xử lý và tạo dữ liệu đặc trưng.
- Bước 2: sử dụng SMOTE trên tập dữ liệu huấn luyện nhằm cân bằng số lượng nhãn.
- Bước 3: áp dụng Bayesian Optimization trên 3 mô hình học máy khác nhau gồm CatBoost, Rừng ngẫu nhiên và XGBoost.
- Bước 4: xây dựng mô hình xếp chồng từ 3 mô hình cơ sở ở bước 3.
- Bước 5: đánh giá mô hình.

Các bước triển khai này được giới thiệu ngay sau trong phần 3.3. Bước đánh giá mô hình được thể hiện ở phần 4 khi tiến hành đo đạc và so sánh kết quả với các nghiên cứu khác.

3.3 Các bước tiến hành

3.3.1 Tạo dữ liệu đặc trưng

Tính toán tỷ lệ khoản vay trên thu nhập và tỷ lệ lãi suất khoản vay trên thu nhập từ khoản vay.

Tỷ lệ khoản vay trên thu nhập: đo lường số tiền vay so với thu nhập hàng năm của người vay. Chỉ số này giúp các tổ chức tài chính đánh giá khả năng trả nợ của người vay, từ đó quyết định có nên phê duyệt khoản vay hay không. Tỷ lệ thấp cho thấy người vay có khả năng tài chính tốt hơn để trả nợ [10].

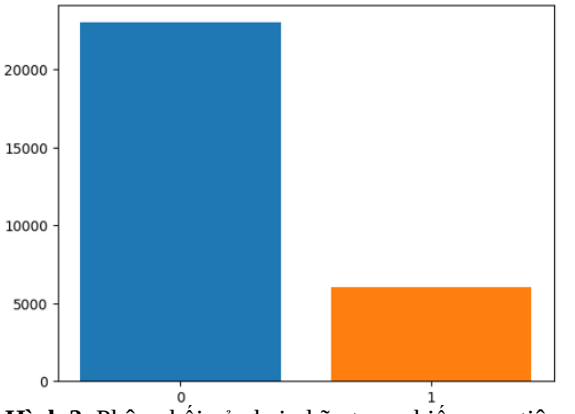
$$Loan\_to\_income\_ratio = \frac{Loan\_amnt}{Person\_income}$$

Tỷ lệ lãi suất trên thu nhập: kết hợp lãi suất của khoản vay với tỷ lệ phần trăm khoản vay so với thu nhập. Chỉ số này cung cấp thông tin về chi phí lãi suất mà người vay phải trả tương ứng với thu nhập của họ, giúp phân tích khả năng tài chính và quản lý RRTD hiệu quả hơn. Việc theo dõi cả hai tỷ lệ này giúp các tổ chức tài chính duy trì một danh mục cho vay lành mạnh và giảm thiểu rủi ro [11].

$$Interest\_income\_ratio = Loan\_int\_rate \times Loan\_person\_income$$

3.3.2 Sử dụng SMOTE để cân bằng số lượng nhãn

Trong Hình 2, hiện tượng mất cân bằng dữ liệu xuất hiện khi số lượng mẫu của nhãn 0 vượt trội so với nhãn 1. Số lượng nhãn 0 gấp xấp xỉ 4 lần so với nhãn 1, điều này làm cho quá trình học của các mô hình học máy có thiên hướng chỉ học tập trung nhãn 0. Vì vậy sau khi SMOTE được áp dụng, số lượng nhãn 1 được làm giàu bằng với số lượng nhãn 0 nhằm cân bằng nhãn.



Hình 2 Phân phối của hai nhãn trong biến mục tiêu Loan_status.

3.3.3 Bayesian Optimization

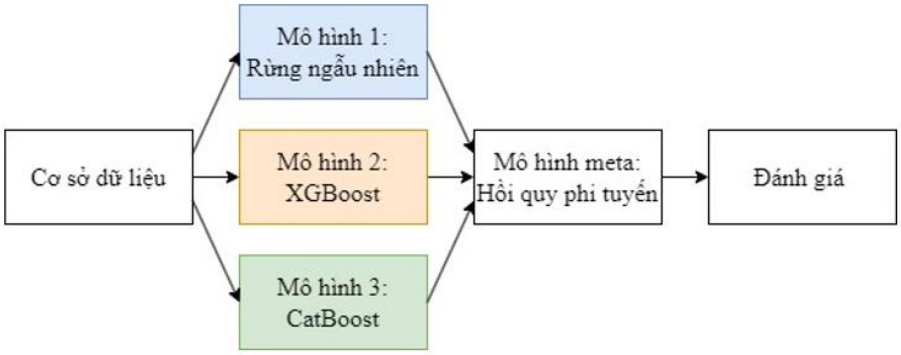
Việc lựa chọn không gian siêu tham số để tiến hành tìm kiếm rất quan trọng, vì vậy nghiên cứu này tập trung vào các tham số chính được sử dụng ở hầu hết các nghiên cứu trước đây là số cây (estimators hoặc iteratons) và chiều sâu tối đa của cây (max_depth hoặc depth), tất nhiên không bỏ qua tỉ lệ học (learning_rate) một chỉ số quan trọng trong siêu tham số huấn luyện mô hình. Bảng 4 mô tả không gian siêu tham số tìm kiếm đã áp dụng trên các mô hình lựa chọn là cơ sở.

Bảng 4 Không gian tìm kiếm được chọn để tìm siêu tham số trong các mô hình XGBoost, rừng ngẫu nhiên, và CatBoost đều nhằm mục đích tối ưu hóa hiệu suất của mô hình.

Mô hình	Tên tham số	Khoảng Tìm Kiếm
XGBoost	n_estimators	50-300
	max_depth	3-10
	learning_rate	0,01-0,30
Rừng ngẫu nhiên	n_estimators	50-300
	max_depth	3-20
CatBoost	iterations	50-300
	depth	3-10
	learning_rate	0,01-0,30

3.3.4 Mô hình xếp chồng

Sau khi áp dụng kỹ thuật BO để tìm ra mô hình tốt nhất làm mô hình cơ sở, tiến hành xếp chồng bằng phương pháp hồi quy phi tuyến để tạo mô hình meta hoàn tất kỹ thuật tạo mô hình xếp chồng, theo mô tả ở Hình 3.



Hình 3 Sơ đồ cấu trúc của mô hình xếp chồng

Bằng cách kết hợp các mô hình có khả năng tổng quát hóa cao và giảm quá khớp, mô hình xếp chồng sẽ tạo ra kết quả mạnh mẽ và ổn định hơn. Khi làm mô hình meta, hồi quy phi tuyến sẽ giúp tổng hợp kết quả từ các mô mô hình cơ sở một cách tuyến tính, mang lại sự chính xác và đơn giản trong việc phân loại, đặc biệt phù hợp khi kết quả từ các mô hình cơ sở đã có độ chính xác cao. Kết quả hiệu suất học của các mô hình cơ sở và mô hình xếp chồng được đánh giá cụ thể ở Bảng 5.

4 Kết quả

Môi trường thực nghiệm được thực hiện trên nền tảng Kaggle, một trong những nền tảng phổ biến nhất cho cộng đồng học máy và khoa học dữ liệu. Trong nghiên cứu này, sử dụng Kaggle Notebooks (trước đây gọi là Kaggle Kernels) để phát triển và chạy các mô hình học máy.

Mô hình xếp chồng đạt được kết quả vượt trội so với các mô hình cơ sở bình thường kết quả này còn thể hiện không chỉ thông qua các chỉ số như độ chính xác còn thể hiện qua chỉ số ROC AUC.

Bảng 5 Kết quả sau khi huấn luyện mô hình

Mô hình	Độ chính xác	Độ chuẩn xác	Độ nhạy	F1	ROC AUC
Rừng ngẫu nhiên	0,9338	0,9633	0,7288	0,8316	0,9459
XGBOOST	0,9347	0,9624	0,7377	0,8352	0,9500
CatBOOST	0,9355	0,9816	0,7268	0,8351	0,9474
Stacking model	**0,9503**	**0,9796**	**0,9198**	**0,9487**	**0,9815**

Từ Bảng 5, các mô hình cơ sở đạt độ chính xác nhỏ nhất là rừng ngẫu nhiên là 0,9338 và lớn nhất là 0,9355, với độ chuẩn xác trên 0,96 và độ nhạy ở khoảng 0,72 đến 0,74. Tuy nhiên, mô hình stacking vượt trội với độ chính xác 0,9503 và độ nhạy 0,9198, cao hơn so với bất kỳ mô hình cơ sở nào. Điều này cho thấy mô hình stacking cải thiện rõ rệt khả năng phát hiện các trường hợp dương tính.

Sự cải thiện ở cả độ chuẩn xác và độ nhạy dẫn đến chỉ số F1 của mô hình stacking cũng cao hơn, phản ánh sự cân bằng tốt hơn trong dự đoán. Đặc biệt, chỉ số ROC AUC của mô hình stacking đạt 0,9815, cho thấy khả năng phân biệt rất tốt giữa các lớp và ít nhầm lẫn. Trong bối cảnh đánh giá RRTD, chỉ số ROC AUC cao giúp các tổ chức tài chính đưa ra quyết định tín dụng chính xác, giảm thiểu rủi ro và tối ưu hóa quy trình quản lý.

5 Kết luận

Nghiên cứu này đã tổng hợp và phân tích hiệu suất của các mô hình học máy trong đánh giá RRTD, đồng thời đưa ra những kết luận quan trọng.

• Mô hình stacking đạt hiệu suất cao nhất với độ chính xác 95,503 % và chỉ số F1 là 0,9487. Độ nhạy của mô hình stacking cũng tăng lên đáng kể, đạt 91,98 %, cao hơn các mô hình cơ sở, cho thấy sự cải thiện toàn diện về các chỉ số đánh giá.

• Chỉ số ROC AUC của mô hình stacking đạt 98,15 %, chứng minh khả năng phân loại tốt và ít bị nhầm lẫn

giữa các nhãn. Tuy nhiên, việc áp dụng mô hình stacking cần cân nhắc về độ phức tạp và thời gian tính toán. Mô hình rừng ngẫu nhiên và CatBoost cũng đạt hiệu suất ấn tượng với độ chính xác trên 93,55 %, nhưng có thể gặp hiện tượng quá khớp. Mô hình XGBoost ổn định với các chỉ số về độ chuẩn xác và độ nhạy cao, nhưng cần tinh chỉnh thêm.

• Không chỉ áp dụng trong thực tiễn, nghiên cứu này còn chỉ ra rằng phương pháp tối ưu hóa Bayes (Bayesian Optimization – BO) đóng vai trò quan trọng trong việc tìm ra các base learner tốt nhất, từ đó xây dựng được meta learner có khả năng tổng hợp và tối ưu hóa hiệu quả hơn. BO giúp tự động hóa quá trình điều chỉnh siêu tham số của các mô hình nền tảng (base learner), cải thiện đáng kể hiệu suất của từng mô hình thành phần, từ đó tối ưu hóa toàn bộ mô hình stacking. Điều này chứng tỏ rằng, việc kết hợp BO với các mô hình học máy không chỉ giúp nâng cao khả năng dự đoán mà còn tối ưu hóa thời gian và tài nguyên tính toán, góp phần xây dựng những mô hình phân loại RRTD chất lượng cao.

• Về mặt thực tiễn, mô hình stacking có thể được áp dụng hiệu quả tại các tổ chức tín dụng nhằm cải thiện độ chính xác trong việc phân loại và đánh giá RRTD. Các chỉ số cao về độ nhạy và ROC AUC của mô hình cho phép các tổ chức đưa ra quyết định tín dụng chính xác hơn, giảm thiểu rủi ro nợ xấu và tối ưu hóa danh mục cho vay. Tuy nhiên, việc triển khai các mô hình phức hợp như stacking đòi hỏi cần có cơ sở hạ tầng tính toán mạnh mẽ và quy trình giám sát chặt chẽ để đảm bảo hiệu quả lâu dài. Các mô hình khác như rừng ngẫu nhiên và CatBoost, với hiệu suất cao nhưng có nguy cơ quá khớp có thể phù hợp với những tập dữ liệu nhỏ hơn hoặc các trường hợp có cấu trúc dữ liệu ít phức tạp hơn. Tóm lại, nghiên cứu này cho thấy tiềm năng của các mô hình học máy trong việc nâng cao hiệu quả đánh giá RRTD, giúp các tổ chức tài chính đưa ra các quyết định tín dụng chính xác và kịp thời hơn, từ đó cải thiện chất lượng dịch vụ và giảm thiểu tổn thất tài chính.

Tài liệu tham khảo

1. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, *34*(17), 14327-14339.

2. Chen, S., Wang, Q., & Liu, S. (2019, June). Credit risk prediction in peer-to-peer lending with ensemble learning framework. In *2019 Chinese Control and Decision Conference (ccdc)* (pp. 4373-4377). IEEE.

3. Li, Y. (2019). Credit risk prediction based on machine learning methods. In *2019 14th International Conference on Computer Science & Education (ICCSE)* (pp. 1011-1013). IEEE.

4. Ben Jabeur, S., Stef, N., & Carmona, P. (2023). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics*, *61*(2), 715-741.

5. Wang, X., Qiao, Y., Cui, Y., Ren, H., Zhao, Y., Linghu, L., ... & Qiu, L. (2023). An explainable artificial intelligence framework for risk prediction of COPD in smokers. *BMC Public Health*, *23*(1), 2164.

6. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129-99149.

7. Dey, R., & Mathur, R. (2023). Ensemble learning method using stacking with base learner, a comparison. In *International Conference on Data Analytics and Insights* (pp. 159-169). Singapore: Springer Nature Singapore.

8. Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381-114391.

9. Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42.

10. Avgouleas, E. (2015). Bank leverage ratios and financial stability: A micro-and macroprudential perspective. *Levy Economics Institute of Bard College Working Paper*, (849).
11. Bierut, B. K., Chmielewski, T., Glogowski, A., Stopczyński, A., & Zajączkowski, S. (2015). Implementing loan-to-value and debt-to-income ratios: learning from country experiences. The case of Poland. *The Case of Poland (October 8, 2015). National Bank of Poland Working Paper*, (212).
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.
13. Victoria, A. H., & Maragatham, G. (2021). Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems*, *12*(1), 217-223.
14. Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems With Applications*, *34*(2), 1434-1444.
15. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
16. Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.
17. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*.
18. Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, *37*(9), 6233-6239.
19. Sun, H., & Guo, M. (2015). Credit risk assessment model of small and medium-sized enterprise based on logistic regression. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1714-1717). IEEE.

Application of Stacking Model Combined with SMOTE and Bayesian Optimization for Credit Risk Assessment

Duong Hon Minh
Pharmacy Department, Nguyen Tat Thanh University
dhminh@ntt.edu.vn

Abstract Credit risk prediction is a critical task for financial institutions to minimize the risk of default and optimize lending decisions. In the context of rapid advancements in machine learning techniques, many classification methods have been developed to improve credit risk prediction capabilities. This study applies a stacking model to assess credit risk, combining predictions from various machine learning models, including XGBoost, Random Forest, and CatBoost. A meta-model, logistic regression, is used to optimize predictions from base models to generate the final prediction. Data is processed using the SMOTE technique for balancing, and the hyperparameters of the base models are optimized through Bayesian optimization. The results show that the stacking model achieves an accuracy of 95.503 % and an ROC-AUC score of 98.15 %, demonstrating the high reliability of the predictions. These results highlight the applicability of machine learning models in credit risk assessment, supporting financial institutions in making individual credit decisions.

Keywords Machine learning, ensemble learning, Bayesian optimization, SMOTE, credit risk prediction.