So sánh hiệu quả các mô hình học máy trong đánh giá rủi ro tín dụng

Cao Văn Kiên[1,*], Vũ Thuận An[1,2]

[1]Khoa Công nghệ Thông tin, Trường Đại học Nguyễn Tất Thành, TP. Hồ Chí Minh, Việt Nam
[2]Trung tâm Dữ liệu và Công nghệ Thông tin, Trường Đại học Bách khoa TP. Hồ Chí Minh, Việt Nam
[*]cvkien@ntt.edu.vn

Tóm tắt

Trong ngành ngân hàng, quản lý rủi ro tín dụng ngày càng trở nên phức tạp và quan trọng trong bối cảnh toàn cầu hóa. Rủi ro tín dụng là một trong những thách thức chính đối diện các tổ chức tài chính, khi những người vay không thực hiện nghĩa vụ trả nợ theo cam kết. Để giảm thiểu rủi ro này, các phương pháp học máy đã trở thành một công cụ quan trọng trong việc đánh giá khả năng vay của cá nhân. Nghiên cứu này so sánh hiệu suất của bốn mô hình học máy phổ biến: "Cây quyết định", "Rừng ngẫu nhiên", "Máy véctơ hỗ trợ", và "Hồi quy logistic" trong việc đánh giá rủi ro tín dụng. Dữ liệu đã trải qua kiểm thử và phân tích cho thấy mô hình "Rừng ngẫu nhiên" vượt trội hơn so với các mô hình còn lại, với độ chính xác cao nhất là 93,22 %. Kết quả này cung cấp cái nhìn sâu sắc về khả năng ứng dụng của các mô hình học máy trong việc đánh giá rủi ro tín dụng và có thể hỗ trợ các tổ chức tài chính trong quyết định về việc cấp tín dụng cho cá nhân.

Nhận 10/03/2024
Được duyệt 20/03/2024
Công bố 29/03/2024

Từ khóa

học máy,
cây quyết định,
rừng ngẫu nhiên,
máy véctơ hỗ trợ,
hồi quy logistic

1 Đặt vấn đề

Trong xu hướng tài chính hóa toàn cầu, cá nhân và ngân hàng có mối quan hệ cộng sinh để giải quyết khó khăn tài chính. Cá nhân đạt được mục tiêu thông qua việc nhận các khoản vay dành cho các mục đích khác nhau làm tăng tính cạnh tranh trong ngành tài chính, khiến cho việc cho vay tín dụng trở thành một phần không thể thiếu. Để đáp ứng nhu cầu đó, hiện nay có nhiều tổ chức tài chính, cả ngân hàng và tổ chức tài chính không thuộc ngân hàng, cung cấp dịch vụ cho vay tín dụng. Thêm vào đó, một phần đáng kể của doanh thu của những tổ chức này đến trực tiếp từ lợi suất thu được từ các khoản vay.

Những rủi ro đáng kể liên quan đến việc cấp vay là điều khó tránh khỏi. "Rủi ro tín dụng" đề cập đến những tình huống khi người vay không thể trả lại số tiền vay theo điều kiện mà cả người cho vay và người vay đã thống nhất [1]. Mặc dù cả hai bên đều hưởng lợi nhưng giảm thiểu rủi ro trở thành một trong những mục tiêu chính của các tổ chức cho vay. Để kiểm tra người vay trong quy trình cho vay truyền thống, ngân hàng chủ yếu sử dụng "Nguyên tắc 5C" − Khả năng trả nợ, Vốn, Tính cách, Điều kiện và Tài sản thế chấp [2]. Tuy nhiên quy trình 5C này rõ ràng phụ thuộc nhiều vào cảm tính, chủ yếu là sự đánh giá chủ quan của nhân viên kiểm soát rủi ro. Ngân hàng và các tổ chức tài chính khác cấp vay sau khi xác minh và xác nhận nhưng vấn đề mấu chốt lại là không thể tuyệt đối xác định liệu người xin vay đã chọn có thể trả nợ đúng hạn hay không.

Theo truyền thống, ngân hàng thuê các chuyên viên chỉ để đánh giá hồ sơ của cá nhân và quyết định xem có an toàn để cấp vay cho họ hay không. Lúc đó, họ đánh giá độ xứng đáng của người vay bằng một điểm số số liệu, còn đư c biết đến là "Điểm tín dụng". Điểm này giúp các cơ quan quản lý ước lượng xác suất người vay trả nợ trong thời gian và điều kiện đã thỏa thuận dựa trên

https://doi.org/10.55401/6k3rsa38

lịch sử tín dụng và/hoặc lịch sử thanh toán của người xin vay cùng với nền tảng của họ [3].

Với sự hỗ trợ của công nghệ, các nhà nghiên cứu, ngân hàng và các tổ chức tài chính khác đã bắt đầu sử dụng các thuật toán học máy và học sâu để đào tạo các mô hình có thể dự đoán khả năng đủ điều kiện của một người xin vay để nhận được khoản vay dựa trên lịch sử tín dụng và dữ liệu khác. Quá trình này có thể giúp dễ dàng lựa chọn ứng viên đủ điều kiện trước khi chấp thuận một khoản vay.

Trong lĩnh vực đánh giá rủi ro tín dụng, các phương pháp học máy đã được ứng dụng rộng rãi với nhiều nghiên cứu đánh giá về hiệu suất của các phương pháp này. Trong số đó, cây quyết định (Decision tree, DT), rừng ngẫu nhiên (Random Forest, RF), máy véctơ hỗ trợ (Support Vector Machine, SVM), và hồi quy logistic (Logistic Regression, LR) là những phương pháp được quan tâm nhiều nhất.

DT là một kỹ thuật phân loại nhanh và dễ hiểu, chia nhỏ tập quan sát thành các nhóm nhỏ hơn dựa trên một tập luật và biến mục tiêu cụ thể [4]. Nhiều nghiên cứu đã chỉ ra hiệu suất cao của DT trong đánh giá tín dụng. Davis [5] và Galindo và Tamayo [6] đều nhận thấy DT có độ chính xác tương đương hoặc cao hơn so với mạng nơ ron và các mô hình khác. Dù vậy, so với các phương pháp như SVM hay LR, DT thường không đạt hiệu suất tốt nhất [7].

Về phương pháp RF, phương pháp này xây dựng một tập hợp các DT được huấn luyện trên các tập dữ liệu khác nhau bằng kỹ thuật bootstrap, với kết quả dự đoán cuối cùng là kết quả trung bình của tất cả các cây [8]. Loureiro [9] và Xiao [10] đều nhấn mạnh phương pháp RF đạt hiệu suất phân loại tín dụng cao hơn so với các mô hình truyền thống. Trái ngược quan điểm đó, Brown và Mues [11] cũng như Butaru [12] lại không tìm thấy sự vượt trội của phương pháp RF so với các phương pháp khác.

SVM là một công cụ phổ biến trong đánh giá rủi ro tín dụng nhờ khả năng thực hiện ánh xạ phi tuyến và tránh bị kẹt tại cực trị cục bộ [13, 14]. Tuy nhiên, một số nghiên cứu khác lại chỉ ra RF đạt hiệu suất tốt hơn so với SVM [6, 7, 15].

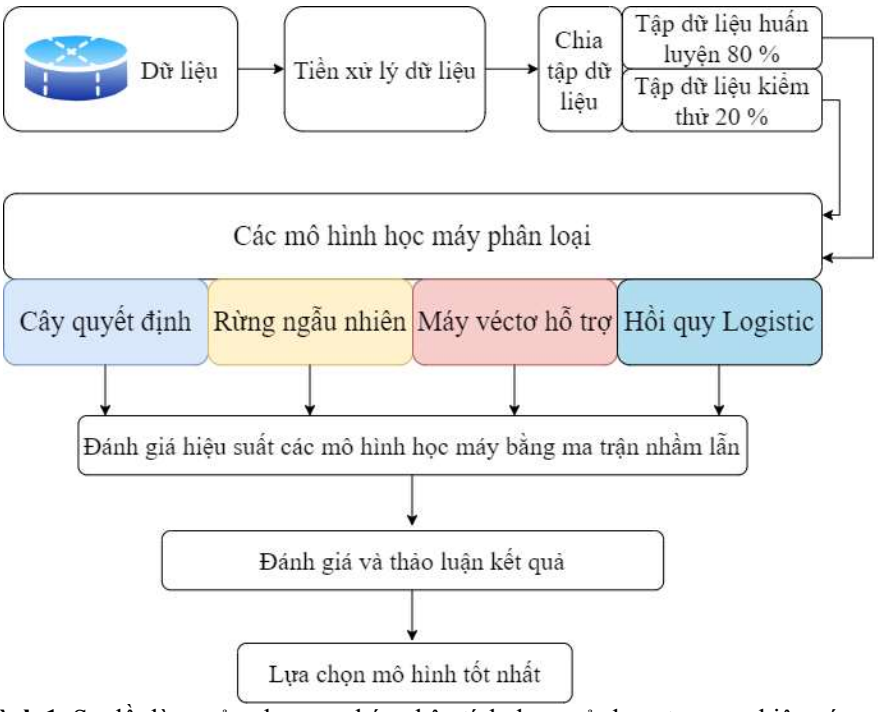
LR là một phương pháp thống kê truyền thống hiệu quả trong đánh giá tín dụng [14, 16, 17] và tính phổ biến của phương pháp này là vẫn được sử dụng rộng rãi nhờ tính đơn giản cũng như phân bố lỗi khá cân bằng [4,18, 19].

Bài báo này nghiên cứu tập trung vào các thuật toán học máy để tìm ra mô hình phù hợp nhất hiện nay để dự đoán một khoản vay có thể xảy ra mắc nợ hay không. Các mô hình sử dụng trong bài này bao gồm: DT, RF, SVM, và LR. Mỗi mô hình sẽ được phân tích độc lập cho bộ dữ liệu, tìm ra các mẫu và rút ra kết luận từ sự phân tích này. Cuối cùng, dựa trên phân tích, nhóm nghiên cứu sẽ xác định liệu một ứng viên mới có nợ khoản vay hay không nhằm giúp ngân hàng và các tổ chức tài chính giải quyết vấn đề truyền thống.

Phần tiếp theo của bài báo được bố cục như sau: các lý thuyết nền tảng về các mô hình học máy cũng như các phương pháp nghiên cứu, bao gồm cách thức thu thập dữ liệu, quy trình phân tích và các công cụ được sử dụng trong quá trình nghiên cứu sẽ được trình bày trong Phần 2. Phần 3 trình bày cụ thể các kết quả nghiên cứu và thảo luận. Cuối cùng là một số kết luận và đề xuất được đưa ra ở Phần 4.

2 Phương pháp nghiên cứu

Hình 1 minh họa tổng quan về cấu trúc của phương pháp được đề xuất để dự đoán khả năng vay tín dụng. Nghiên cứu này tiến hành qua các giai đoạn quan trọng. Đầu tiên, dữ liệu được trích xuất từ cơ sở dữ liệu. Sau đó, giai đoạn tiền xử lý dữ liệu bao gồm loại bỏ giá trị thiếu và ngoại lệ, cũng như chuẩn hóa dữ liệu để chuẩn bị cho việc huấn luyện mô hình. Sau giai đoạn tiền xử lý, dữ liệu được phân chia thành hai phần: một để huấn luyện mô hình và một để đánh giá hiệu suất mô hình, đảm bảo tính khách quan. Bước quan trọng tiếp theo là huấn luyện các mô hình học máy khác nhau, bao gồm DT, RF, SVM, và LR. Mục tiêu chính của nghiên cứu này là kiểm tra tỉ mỉ và so sánh hiệu suất mỗi mô hình để xác định giải pháp hiệu quả nhất cho vấn đề nghiên cứu cụ thể. Cuối cùng, thực hiện phân tích so sánh độ chính xác và kết quả của mô hình để thảo luận toàn diện về hiệu quả của từng mô hình và rút ra kết luận quan trọng phù hợp với vấn đề nghiên cứu.



Hình 1 Sơ đồ dòng của phương pháp phân tích được sử dụng trong nghiên cứu này.

2.1 Tập dữ liệu

Trong phần này của nghiên cứu, tập dữ liệu được sử dụng là "Tập dữ liệu rủi ro tín dụng" (Credit Risk Dataset) [20], được công bố trên nền tảng Kaggle. Tập dữ liệu này bao gồm khoảng 300 triệu giao dịch vay được thực hiện bởi 32 581 cá nhân. Bộ dữ liệu này bao gồm tổng cộng 11 đặc trưng, mô tả hồ sơ của mỗi cá nhân, được liệt kê trong Bảng 1.

Bảng 1 Ký hiệu và định nghĩa biến theo các đặc điểm dữ liệu

Biến đầu vào	Định nghĩa biến
person_age	Tuổi của cá nhân
person_income	Thu nhập hàng năm của cá nhân.
person_home_ownership	Loại sở hữu nhà - thuê, thế chấp, thuê mua, sở hữu hoặc khác.
person_emp_length	Thời gian làm việc của cá nhân (theo năm).
loan_intent	Mục đích của khoản vay.
loan_amnt	Số tiền được hoàn trả cho người vay.
loan_int_rate	Lãi suất đối với khoản vay.
loan_status	Trạng thái thanh toán khoản vay (0 là không vi phạm, 1 là vi phạm).
loan_percent_income	Tỷ lệ phần trăm số tiền vay theo tổng thu nhập.
cb_person_default_on_file	Lịch sử các khoản nợ (nếu có) được thực hiện bởi cá nhân.
cb_person_cred_hist_length	Lịch sử tín dụng của cá nhân.

Ngoài ra, Bảng 2 mô tả chi tiết về các loại dữ liệu và các đặc điểm thống kê của tập dữ liệu.

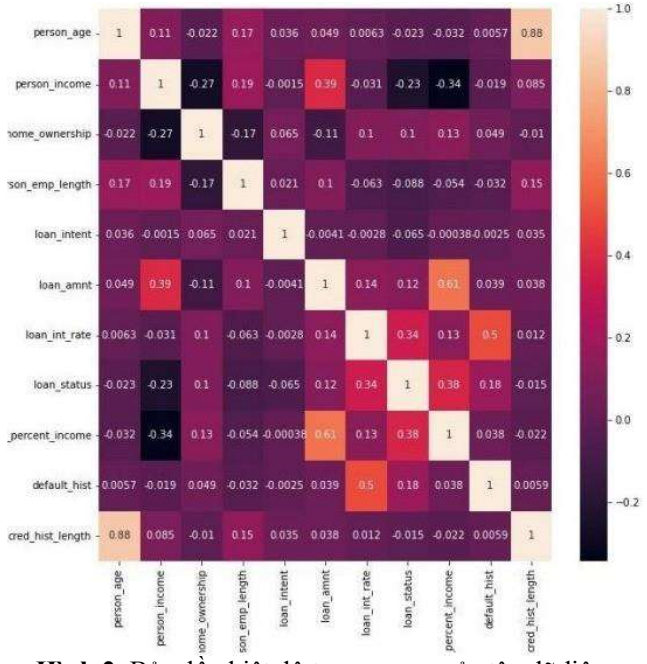
Bảng 2 Đặc điểm thống kê

No.	Attributes	Data type	Min Values	Max Values	Mean	Standard Deviation (std)
1	person_age	int64	20	144	27,73	6,31
2	person_income	int64	4 000	6 000 000	66 649,37	62 356,45
3	person_home_ownership	object	-	-	-	-

4	person_emp_length	float64	0	123	4,79	4,15
5	loan_intent	object	-	-	-	-
6	loan_amnt	int64	500	35 000	9 656,49	6 329,68
7	loan_int_rate	float64	5,42	23,22	11,04	3,23
8	loan_status	int64	0	1	0,22	0,41
9	loan_percent_income	float64	0	0,83	0,17	0,11
10	cb_person_default_on_file	object	-	-	-	-
11	cb_person_cred_hist_length	int64	2	30	5,79	4,04

Chú trọng đến các bước tiền xử lý dữ liệu không chỉ nhằm tăng cường hiệu suất của mô hình mà còn đảm bảo tính toàn vẹn và nhất quán của dữ liệu đầu vào. Điều này tạo ra một nền tảng đáng tin cậy cho quá trình huấn luyện và đánh giá mô hình.

2.2 Tiền xử lý dữ liệu



Hình 2 Bản đồ nhiệt độ tương quan của tập dữ liệu.

Trong mục này, tiến hành phân tích trên dữ liệu để chuẩn bị cho việc xây dựng một mô hình dự đoán mạnh mẽ. Quá trình phân tích dữ liệu được mô tả như sau:

Kiểm tra giá trị thiếu: một bước quan trọng là kiểm tra xem có giá trị thiếu nào trong tập dữ liệu hay không. Bỏ qua các giá trị thiếu có thể dẫn đến kết quả không chính xác. Do đó, nhóm nghiên cứu đã kiểm tra kỹ lưỡng các thuộc tính dữ liệu để xác định xem có giá trị thiếu hoặc NA nào không. Các giá trị thiếu hoặc NA sẽ được xóa hàng tương ứng.

Phân tích tương quan: trong quá trình này, việc phân tích tập dữ liệu đã được thực hiện để đánh giá mức độ tương quan giữa các thuộc tính. Các đặc trưng hoặc hệ số tương quan cao có thể có ảnh hưởng đáng kể đến hiệu suất của mô hình phân loại. Mức độ tương quan âm cao thường dẫn đến hiệu suất thấp. Hình 2 minh họa một cách trực quan về ma trận tương quan của tập dữ liệu, thể hiện mức độ tương quan giữa các cặp biến thông qua các hệ số tương quan từ −1 đến 1. Chẳng hạn như khoản vay (loan_amnt) và tỷ lệ khoản vay trên thu nhập (loan_percent_income) có mối quan hệ tích cực và có hệ số tương quan là 0,61. Ma trận tương quan thường được sử dụng trong phân tích thống kê và khoa học dữ liệu để đánh giá mức độ liên kết giữa các biến và phát hiện ra các mẫu hoặc mối quan hệ trong dữ liệu.

Chuẩn hóa đặc trưng: tập dữ liệu về khả năng cho vay tín dụng bao gồm các thuộc tính được đo trên các thang đo khác nhau. Sự khác biệt này có thể làm ảnh hưởng đến hiệu suất của mô hình. Để giải quyết vấn đề này, các thuộc tính đã được chuẩn hóa để có cùng một thang đo từ 0 đến 1 bằng công thức toán học như sau:

$$x_{scale} = \frac{x - \min(x)}{\max(x) - \min(x)},$$

trong đó, x là giá trị gốc mà ta muốn chuẩn hóa, x_{scale} là giá trị đã được chuẩn hóa của x, min(x) là giá trị nhỏ nhất trong tập dữ liệu, và max(x) là giá trị lớn nhất của tập dữ liệu.

2.3 Các mô hình học máy

Trong phạm vi của nghiên cứu này, bốn phương pháp học máy có giám sát phổ biến đã được đánh giá để so sánh hiệu suất của các phương pháp này trên tập dữ liệu rủi ro tín dụng. Do đó, các kỹ thuật như DT, RF, SVM, và LR đã được triển khai bằng cách so sánh hiệu suất của các phương pháp này dựa trên ma trận nhầm lẫn (Confusion Matrix), độ chính xác (Accuracy), độ chuẩn xác (Precision), độ nhạy (Recall), và điểm F1 (F1 Score). Các kỹ thuật này được đánh giá để phân tích hiệu quả của các phương pháp học máy khác nhau trên cùng một tập dữ liệu. Các thuật toán này được ưa chuộng vì dễ triển khai và có thể tạo ra kết quả tốt về hiệu suất.

2.3.1 Mô hình cây quyết định

Cây quyết định (DT) là một trong những công cụ mạnh mẽ nhất của các thuật toán học có giám sát được sử dụng cho cả các nhiệm vụ phân loại và hồi quy. DT xây dựng một cấu trúc cây giống như một biểu đồ dòng điều chỉnh, trong đó mỗi nút nội bộ biểu thị một kiểm tra trên một thuộc tính, mỗi nhánh biểu thị một kết quả của kiểm tra, và mỗi nút lá (nút cuối cùng) chứa một nhãn lớp. DT được xây dựng bằng cách chia tách đệ quy dữ liệu huấn luyện thành các tập con dựa trên các giá trị của các thuộc tính cho đến khi đáp ứng được một điều kiện dừng, chẳng hạn như độ sâu tối đa của cây hoặc số lượng mẫu tối thiểu cần thiết để chia một nút.

Trong quá trình huấn luyện, thuật toán DT chọn thuộc tính tốt nhất để chia dữ liệu dựa trên một phương pháp đánh giá như entropy hoặc độ không chắc chắn Gini, đo lường mức độ không thuần khiết hoặc ngẫu nhiên trong các tập con. Mục tiêu là tìm thuộc tính tối ưu nhất mà tăng thông tin hoặc giảm độ không thuần khiết sau khi chia. Người đọc, có thể xem các tài liệu [21-25] để có thể hiểu sâu hơn về mô hình DT. Ngoài ra, các ứng dụng của mô hình DT có thể xem ở tài liệu [26-28].

2.3.2 Mô hình rừng ngẫu nhiên

Một thuật toán RF là một thuật toán học máy giám sát cực kỳ phổ biến và được sử dụng cho các vấn đề phân loại và hồi quy trong học máy, biết rằng một khu rừng bao gồm nhiều cây, và càng nhiều cây càng mạnh mẽ hơn. Tương tự, càng nhiều cây trong một thuật toán RF, độ chính xác và khả năng giải quyết vấn đề của thuật toán đó càng cao. RF là một bộ phân loại có chứa nhiều DT trên các tập con khác nhau của tập dữ liệu đã cho và lấy trung bình để cải thiện độ chính xác dự đoán của tập dữ liệu đó. Thuật toán này dựa trên khái niệm học hợp tác, đó là quá trình kết hợp nhiều bộ phân loại để giải quyết một vấn đề phức tạp và cải thiện hiệu suất của mô hình. Người đọc, có thể xem các tài liệu [29, 30] để có thể hiểu sâu hơn về mô hình RF Ngoài ra, người đọc có thể xem các ứng dụng của mô hình RF ở tài liệu [31].

2.3.3 Mô hình máy véctơ hỗ trợ

Máy véctơ hỗ trợ (SVM) là một phương pháp trong thống kê và khoa học máy tính. Phương pháp này được sử dụng để phân loại và phân tích dữ liệu. SVM là thuật toán phân loại nhị phân, tức là phân loại dữ liệu thành hai lớp khác nhau. Thuật toán SVM xây dựng một mô hình để phân loại các ví dụ vào hai lớp đó. Mô hình SVM biểu diễn các điểm trong không gian và lựa chọn ranh giới giữa hai lớp sao cho khoảng cách từ các ví dụ luyện tập tới ranh giới là xa nhất có thể. SVM cũng có thể ánh xạ dữ liệu vào không gian mới để phân tách các điểm dữ liệu dễ dàng hơn. Trong tóm tắt, SVM là một công cụ mạnh mẽ trong học máy, giúp phân loại và phân tích dữ liệu dựa trên việc xây dựng các siêu phẳng tối ưu để phân chia các lớp dữ liệu. Người đọc, có thể xem các tài liệu [32] để có thể hiểu sâu hơn về mô hình SVM. Ngoài ra, các ứng dụng của mô hình SVM có thể xem ở tài liệu [33].

2.3.4 Mô hình hồi quy logistic

LR là một thuật toán phân loại khác, thường được sử dụng để phân loại quan sát vào một tập hợp các lớp riêng biệt. Thuật toán này được suy ra từ lý thuyết xác suất và là một loại thuật toán dự đoán. Giả thuyết của LR có xu hướng giới hạn hàm chi phí. Hàm này chuyển đổi bất kỳ giá trị thực nào thành một phạm vi từ 0 đến 1 được biết đến với tên gọi là hàm sigmoid. Hàm sigmoid được sử dụng để ánh xạ dự đoán thành xác suất. Phương trình của LR được biểu diễn như sau:

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n.$$

Trong đó, y là biến phụ thuộc thường là xác suất để một sự kiện xảy ra, $x_1, x_2, \ldots, x_n$ là các biến độc lập, và $b_1, b_2, \ldots, b_n$ là các hệ số của mô hình. Người đọc, có thể xem các tài liệu [34, 35] để có thể hiểu sâu hơn về mô hình LR. Ngoài ra, các ứng dụng của mô hình LR có thể xem ở tài liệu [36].

3 Kết quả và thảo luận

Trong phần này, đề cập đến việc so sánh và thảo luận về hiệu suất của bốn thuật toán học máy được giám sát như các bộ phân loại, bao gồm DT, RF, SVM, và LR. Tập huấn luyện và kiểm tra đư c chọn ngẫu nhiên với tỷ lệ 80 % dữ liệu huấn luyện và 20 % dữ liệu kiểm tra dựa trên dữ liệu gốc để nghiên cứu về độ chính xác và hiệu suất của bộ phân loại.

3.1 Môi trường thực nghiệm

Trong nghiên cứu này, các thí nghiệm đã đư c thực hiện trên máy tính MacBook Air chạy hệ điều hành Windows 10 Professional, với CPU Intel Core i5 5250U 1,60 GHz, card đồ họa tích hợp Intel HD Graphics 6000, và bộ nhớ RAM DDR3 4 GB. Mã nguồn được viết bằng ngôn ngữ lập trình Python phiên bản 3.10.5.

3.2 Đánh giá hiệu suất các mô hình học máy

Đánh giá hiệu suất là một phần quan trọng của một kỹ thuật phân loại. Các độ đo hiệu suất giúp xác định mô

hình phân loại tốt nhất. Hiệu suất của một kỹ thuật phân loại có thể đư c đo dựa trên ma trận nhầm lẫn, độ chính xác, độ chính xác, độ nhạy, và điểm F1.

Ma trận nhầm lẫn là một cấu trúc dữ liệu mô tả và tổng hợp kết quả dự đoán trong các vấn đề phân loại. Trong ma trận nhầm lẫn, hàng sự kiện đư c gán như "Dự đoán +" và hàng không có sự kiện đư c gán như "Dự đoán -". Sau đó, cột sự kiện của các dự đoán đư c gán như "True" và không có sự kiện nào đư c gán như "False", và biểu diễn của ma trận nhầm lẫn đư c thể hiện trong Bảng 3. Ở đây, True Positive (TP) có nghĩa là kết quả dự đoán là đúng và kết quả thực tế cũng là đúng. False Positive (FP) có nghĩa là kết quả dự đoán là đúng nhưng kết quả thực tế lại là sai. Khi kết quả dự đoán là sai nhưng kết quả thực tế lại là đúng, tình huống này đư c gọi là False Negative (FN). Nếu kết quả dự đoán là sai và kết quả thực tế cũng là sai, điều này đư c gọi là True Negative (TN).

Bảng 3 Ma trận nhầm lẫn đối với phân lớp nhị phân

	Chân trị +	Chân trị -
Dự đoán +	TP	FP
Dự đoán -	FM	TN

Độ chính xác: độ chính xác đo lường tỉ lệ các dự đoán chính xác trên tổng số dự đoán. Công thức tính độ chính xác là:

$$\text{Độ chính xác} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Độ chuẩn xác: độ chuẩn xác đo lường tỉ lệ các "Dự đoán +" là đúng trong số các "Dự đoán +". Công thức tính độ chuẩn xác là:

$$\text{Độ chuẩn xác} = \frac{TP}{TP + FP}.$$

Độ nhạy: độ nhạy đo lường tỉ lệ các dự đoán positive là đúng trong số các mẫu thực sự là positive. Công thức tính độ nhạy là:

$$\text{Độ nhạy} = \frac{TP}{TP + FN}.$$

Điểm F1: điểm F1 là trung bình điều hòa của độ chuẩn xác và độ nhạy. Điểm F1 cung cấp một phép đo tổng thể về hiệu suất của mô hình. Công thức tính điểm F1 là:

$$\text{Điểm } F1 = 2 \times \frac{\text{Độ chuẩn xác} \times \text{Độ nhạy}}{\text{Độ chuẩn xác} + \text{Độ nhạy}}$$

3.3 Phân tích kết quả

Kết quả của cuộc so sánh hiệu quả mô hình học máy trong đánh giá rủi ro tín dụng đã đư c thực hiện và tổng hợp trong Bảng 4 dưới đây:

Bảng 4 So sánh kết quả đạt đư c từ nghiệm của 4 mô hình học máy

	Độ chính xác	Độ chuẩn xác	Độ nhạy	Điểm F1
Cây quyết định	0,8852	0,7306	**0,7640**	0,7470
Rừng ngẫu nhiên	**0,9322**	**0,9631**	0,7218	**0,8252**
Máy véctơ hỗ trợ	0,9085	0,9182	0,6450	0,7577
Hồi quy Logistic	0,8614	0,7626	0,5446	0,6354

Nhìn chung, mô hình RF đã đạt đư c hiệu suất cao nhất với độ chính xác đạt 93,22 % và điểm F1 là 0,8252. Mặc dù mô hình này cũng có độ nhạy tương đối cao, nhưng mô hình DT thể hiện hiệu suất tốt nhất với độ nhạy là 76,40 %. Mô hình SVM cũng cho thấy kết quả ấn tượng, nhưng vẫn thấp hơn so với RF. Trong khi đó, mô hình LR có hiệu suất thấp nhất trong số các mô hình, đặc biệt là đối với độ nhạy và điểm F1. Tuy nhiên, điều này không làm mất đi sự quan trọng của mô hình LR trong một số tình huống cụ thể trong thực tế.

4 Kết luận và đề xuất

Trong nghiên cứu này, kết quả phân tích hiệu suất của các mô hình học máy trong đánh giá rủi ro tín dụng đư c tổng hợp lại và rút ra một số kết luận quan trọng cùng với đề xuất như sau:

Hiệu suất của các mô hình: kết quả thử nghiệm cho thấy mô hình RF đạt đư c hiệu suất cao nhất với độ chính xác đạt 93,22 % và điểm F1 là 0,8252. Mô hình này cũng có độ nhạy tương đối cao, đạt 72,18 %.

Ưu điểm và hạn chế của mô hình: mô hình RF đã chứng minh sức mạnh của mình thông qua hiệu suất ấn tượng, tuy nhiên, để áp dụng trong thực tế, cần phải cân nhắc kỹ về độ phức tạp và thời gian tính toán. Trong khi đó, mô hình DT đã thể hiện hiệu suất tốt nhất với tỷ lệ độ nhạy lên đến 76,40 %. Tuy nhiên, một hạn chế đáng lưu ý của mô hình DT là khả năng dễ bị quá khớp (overfitting). Hiện tượng này xảy ra khi mô hình "học" dữ liệu huấn luyện quá mức, dẫn đến việc hiểu nhầm và nhiễu các dữ liệu mới, ảnh hưởng đến khả năng tổng quát hóa của mô hình.

Đề xuất cho tương lai: để nâng cao hiệu suất của các mô hình, nhóm đề xuất tiếp tục nghiên cứu và thử nghiệm các kỹ thuật mới như tối ưu hóa siêu tham số, kỹ thuật xử lý dữ liệu mất cân bằng, và việc sử dụng các mô hình kết hợp. Đồng thời, cần thực hiện thêm các nghiên cứu và thử nghiệm trên các tập dữ liệu lớn và đa dạng để đánh giá sự tổng quát và tính linh hoạt của các mô hình

Tài liệu tham khảo

1. Aslam, U., Aziz, H. I. T., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience, 16*, 3483–8.
2. Li, Y. (2019). Credit risk prediction based on machine learning methods. In *The 14th Int. Conf. On Computer Science & Education (ICCSE)* (pp. 1011–3).
3. Ahmed, M. S. I., & Rajaleximi, P. R. (2019). An empirical study on credit scoring and credit scorecard for financial institutions. *Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET), 8*, 275–9.
4. Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications, 38*(10), 13274-13283.
5. Davis, R. H., Edelman, D., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics, 4*(1), 43-51.
6. Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics, 15*(1-2), 107-143.
7. Abellan, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications, 73*, 1-10.
8. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.
9. Loureiro, A. L., Torgo, L., & Soares, C. (2018). Outstanding issues in consumer credit risk prediction. *Progress in Artificial Intelligence, 7*(3), 199-209.
10. Xiao, H., Xiao, Z., & Wang, Y. (2020). Ensemble extreme learning machine with supervised rotation for credit scoring. *Knowledge-Based Systems, 189*, 105072.
11. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446-3453.
12. Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218-239.
13. Yu, L., Wuyi, Y., Shouyang, W., & Lai, K. K. (2010). Credit risk evaluation with a least squares fuzzy support vector machines classifier. *Discrete Dynamics in Nature and Society, 2010*, 1-14.
14. Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Hybrid neural net and memory-based techniques for advanced credit risk analysis. *Journal of Management Information Systems, 20*(1), 117-138.
15. Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61-68.
16. West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11-12), 1131-1152.
17. Henley, W. E. (1995). Statistical aspects of credit scoring. *The Statistician, 44*(1), 5-26.
18. Finlay, S. M. (2012). Credit risk modelling: An application perspective. In G. B. Di Pillo (Ed.), *Machine learning: Concepts, Methodologies, Tools and Applications* (pp. 193-224). IGI Global.
19. 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.
20. Kaggle. (n.d.). Analyzing Credit Default. Retrieved from https://www.kaggle.com/code/juniorbueno/analyzing-credit-default/notebook#DataExploration
21. Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: From efficient prediction to responsible AI. *Frontiers in Artificial Intelligence, 6*, 1124553.
22. Kotsiantis, S.B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261–283.

23. Paluszek, M., & Thomas, S. (2024). Data Classification with Decision Trees. In *MATLAB Machine Learning Recipes*. Apress.
24. Zhou, H. (2023). Decision Trees. In *Learn Data Mining Through Excel*. Apress. https://doi.org/10.1007/978-1-4842-9771-1_10
25. Zollanvari, A. (2023). Decision Trees. In *Machine Learning with Python*. Springer. https://doi.org/10.1007/978-3-031-33342-2_7
26. Siddiqui, E. F., Ahmed, T., & Nayak, S. K. (2024). A decision tree approach for enhancing real-time response in exigent healthcare unit using edge computing. *Measurement: Sensors, 32*.
27. Karalis, G. (2020). Decision Trees and Applications. In *GeNeDis 2018*. Advances in Experimental Medicine and Biology, 1194. Springer. https://doi.org/10.1007/978-3-030-32622-7_21
28. Stankovski, V., & Trnkoczy, J. (2006). Application of Decision Trees to Smart Homes. In *Designing Smart Homes*. Lecture Notes in Computer Science, 4008. Springer. https://doi.org/10.1007/11788485_8
29. Schlenger, J. (2024). Random Forest. In *Computer Science in Sport*. Springer. https://doi.org/10.1007/978-3-662-68313-2_24
30. Doan, TP., Choi, B.J., Hong, K., Park, J., & Jung, S. (2023). Random Forest in Federated Learning Setting. In *Advances in Computer Science and Ubiquitous Computing*. CUTECSA 2022. Springer. https://doi.org/10.1007/978-981-99-1252-0_1
31. Fan, G. (2023). Random Forest Algorithm for Forest Fire Prediction. In *Proceedings of 2nd International Conference on Artificial Intelligence, Robotics, and Communication*. ICAIRC 2022. Springer. https://doi.org/10.1007/978-981-99-4554-2_15
32. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273-297.
33. Abramovych, A., Zaitsev, I., Piddubnyi, V., & Bereznychenko, V. (2024). Application of the support vector machines method for metal analyze by an eddy current system. *Journal of Engineering, e12346*.
34. Moscarelli, M. (2023). Logistic Regression. In *Biostatistics With 'R': A Guide for Medical Doctors*. Springer. https://doi.org/10.1007/978-3-031-33073-5_10
35. Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2023). Logistic Regression. In *Multivariate Analysis*. Springer Gabler. https://doi.org/10.1007/978-3-658-40411-6_5
36. Szafraniec-Siluta, E., Zawadzka, D., & Strzelecka, A. (2022). Application of the logistic regression model to assess the likelihood of making tangible investments by agricultural enterprises. *Procedia Computer Science, 207*, 3894-3903.

Comparing the Effectiveness of Machine Learning Models in Credit Risk Assessment

Cao Van Kien[1,*], Vu Thuan An[1,2] – *cvkien@ntt.edu.vn
[1]Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam
[2]Center for Data and Information Technology, Ho Chi Minh City University of Technology, Viet Nam

Abstract In the banking sector, credit risk management is becoming increasingly complex and crucial in the context of globalization. Credit risk is one of the primary challenges for financial institutions when borrowers fail to fulfill debt repayment obligations as promised. To mitigate this risk, machine learning methods have become important tools in assessing individual borrowing capabilities. In this study, we compare the performance of four popular machine learning models: Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression in credit risk assessment. The data underwent testing and analysis, showing that the Random Forest model outperformed the others, with the highest accuracy of 93.22 %. These results provide profound insights into the applicability of machine learning models in credit risk assessment and may assist financial institutions in making decisions regarding individual credit issuance.

Keywords Machine Learning, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression.