Ứng dụng học máy và học sâu trong nghiên cứu tài chính: Một nghiên cứu về dự báo khả năng hoàn trả khoản vay của khách hàng

Applying in machine learning and deep learning in finance industry: A case study on repayment prediction

Nguyễn Phát Đạt[1,2], Hồ Mai Minh Nhật[1,2], Trương Công Vinh[1,2], Lê Quang Chấn Phong[1,2], Lê Hoành Sử[1,2*]

[1]Trường Đại Học Kinh tế - Luật, Thành phố Hồ Chí Minh, Việt Nam
[2]Đại học Quốc Gia Thành Phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: sulh@uel.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.econ.vi.20.1.3828.2024

Ngày nhận: 18/10/2023

Ngày nhận lại: 16/04/2024

Duyệt đăng: 26/04/2024

Mã phân loại JEL:

G20; G23

Từ khóa:

dự báo khả năng hoàn trả khoản vay; đánh giá rủi ro; học máy; học sâu; vay ngang hàng

Keywords:

repayment prediction; risk assessment; machine learning; deep learning; peer-to-peer lending

TÓM TẮT

Trong bối cảnh cho vay ngang hàng (P2P lending) ngày càng phát triển, việc đánh giá khả năng trả nợ của khách hàng trở nên cần thiết, không chỉ giúp nhà đầu tư cá nhân hạn chế rủi ro mà còn phát hiện các cơ hội đầu tư tiềm năng. Nghiên cứu này đề xuất việc áp dụng học máy và học sâu để phân tích hành vi, thông tin nhân khẩu và lịch sử tín dụng của người vay, qua đó dự báo khả năng hoàn trả khoản vay. Các thuật toán được áp dụng trong bài nghiên cứu bao gồm: Logistic Regression (LR), K-Nearest Neighbor (KNN), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM) và học sâu: Long Short Term Memory (LSTM), Artificial Neural Network (ANN). Kết quả sau khi xử lý và tối ưu hóa cho thấy các mô hình Ensemble Learning như XGB, LGBM đem lại kết quả vượt trội so với các mô hình máy học truyền thống với độ chính xác mô hình đạt hơn 85%. Các đặc trưng như tỷ lệ lãi suất (int_rate), xếp hạng tín dụng (subgrade) và số tiền vay (loan_amnt) có ý nghĩa đặc biệt quan trọng trong việc dự đoán này. Với kết quả dự đoán, chúng tôi kỳ vọng rằng nghiên cứu sẽ cung cấp một công cụ hỗ trợ đắc lực cho nhà đầu tư cá nhân trong việc đánh giá và lựa chọn hồ sơ vay, từ đó góp phần vào việc thúc đẩy một thị trường cho vay ngang hàng minh bạch và hiệu quả hơn.

ABSTRACT

In the current era marked by the proliferation of peer-to-peer lending platforms, the imperative of ascertaining borrowers' capacity to honor their financial obligations has assumed paramount significance. This endeavor transcends mere risk mitigation for individual investors, extending to the identification of judicious investment prospects. The present inquiry advocates for the adoption of sophisticated computational methodologies, including machine learning and deep learning, to analyze borrowers' behavioral patterns, demographic profiles, and credit histories, thus facilitating the prognostication of loan repayment

likelihood. Employed techniques encompass Logistic Regression (LR), K-Nearest Neighbor (KNN), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), in conjunction with deep learning architectures such as Long Short-Term Memory (LSTM) and Artificial Neural Network (ANN). Following methodological refinement, it becomes apparent that ensemble learning approaches, exemplified by XGB and LGBM, exhibit markedly superior predictive performance, surpassing conventional models with an accuracy rate exceeding 85%. Salient predictors include interest rates, credit ratings, and loan amounts. It is anticipated that the findings of this investigation will furnish investors with a potent analytical toolset for discerning and selecting loan portfolios, thereby fostering greater transparency and efficiency within the peer-to-peer lending ecosystem.

1. Giới thiệu

Trong bối cảnh tài chính và ngân hàng hiện nay, việc cung cấp các khoản vay có khả năng thu hồi không chỉ là nền tảng cho hoạt động quản lý rủi ro tín dụng mà còn đóng góp vào sự phát triển bền vững của nền kinh tế. Đặc biệt sau sự kiện khủng hoảng tài chính toàn cầu năm 2008, tầm quan trọng của việc đánh giá khả năng trả nợ của khách hàng đã được nhấn mạnh mạnh mẽ hơn bao giờ hết (Singh, 2023). Trong bối cảnh đó, nghiên cứu này tập trung vào lĩnh vực cho vay ngang hàng nhằm mục tiêu cung cấp cho các nhà đầu tư một công cụ đánh giá khả năng trả nợ của người vay hiệu quả. Điều này không chỉ giúp các nhà đầu tư gia tăng khả năng đánh giá rủi ro mà còn hỗ trợ người vay nhận ra những yếu tố quan trọng nhất ảnh hưởng đến khả năng trả nợ của họ, đặc biệt nghiên cứu sẽ càng hữu ích nếu Chính phủ Việt Nam cho phép hoạt động cho vay ngang hàng hoạt động trong tương lai.

Nhận thấy được nhược điểm đó, nhiều nhóm nghiên cứu đã tiến hành ứng dụng các thuật toán máy học để hỗ trợ dự đoán khả năng hoàn trả khoản vay của khách hàng. Costa e Silva và cộng sự (2020) đánh giá cao khả năng dự đoán của mô hình hồi quy Logistic hay Chang và cộng sự (2018) lựa chọn XGBoost cho bài toán dự đoán của mình. Bên cạnh đó, nhiều nghiên cứu cũng ứng dụng học sâu nhằm cải thiện độ chính xác của mô hình, điển hình như công bố của Ko và cộng sự (2022), Graves (2012) cho thấy hiệu quả của thuật toán ANN, CNN và LSTM.

Chính sự phát triển vượt bậc trong công nghệ thông tin và dữ liệu lớn đã giúp việc xử lý và phân tích thông tin khách hàng trở nên thuận lợi hơn bao giờ hết. Sự kết hợp của máy học và khai thác dữ liệu, tạo điều kiện cho việc xây dựng các mô hình dự đoán hiệu quả, nhằm đánh giá khả năng trả nợ của khách hàng dựa trên dữ liệu sẵn có. Trong khuôn khổ nghiên cứu này, chúng tôi đặc biệt tập trung vào ứng dụng các thuật toán học máy và học sâu để dự đoán khả năng hoàn trả khoản vay ngang hàng, đồng thời nhấn mạnh vào việc nhận diện các đặc trưng quan trọng như tỷ lệ lãi suất (int_rate), xếp hạng tín dụng (subgrade) và số tiền vay (loan_amnt), bởi chúng có ảnh hưởng đặc biệt đến khả năng trả nợ của khách hàng. Những phân tích kỹ lưỡng này không chỉ tăng cường khả năng dự đoán chính xác mà còn góp phần vào việc tạo ra các giải pháp đánh giá tài chính hiệu quả. Nghiên cứu này là cơ hội để chúng tôi đóng góp vào lĩnh vực cho vay ngang hàng, từ đó cung cấp giá trị thực tiễn và có thể thúc đẩy sự phát triển trong ngành tín dụng và tài chính.

2. Cơ sở lý thuyết

2.1. Phương pháp Học máy

2.1.1. Hồi quy Logistic

Hồi quy Logistic là một trong những phương pháp thống kê phổ biến nhất trong lĩnh vực tài chính cho các mô hình đánh giá rủi ro tín dụng. Mô hình hồi quy Logistic được đánh giá cao nhờ sự đơn giản trong việc hiểu biết, khả năng hiệu suất mạnh mẽ và độ dễ dàng trong việc thực hiện (Phan & Nguyen, 2013; Zhao & Zou, 2021).

Hồi quy Logistic giải quyết nhược điểm của hồi quy tuyến tính bằng cách sử dụng hàm phi tuyến để thay thế hàm tuyến tính trong hồi quy. Hàm sigmoid tạo ra một phạm vi điểm từ 0 đến 1 và giới hạn đầu ra trong khoảng này, từ đó biểu thị khả năng xảy ra một sự kiện nhất định.

2.1.2. K-Nearest Neighbors

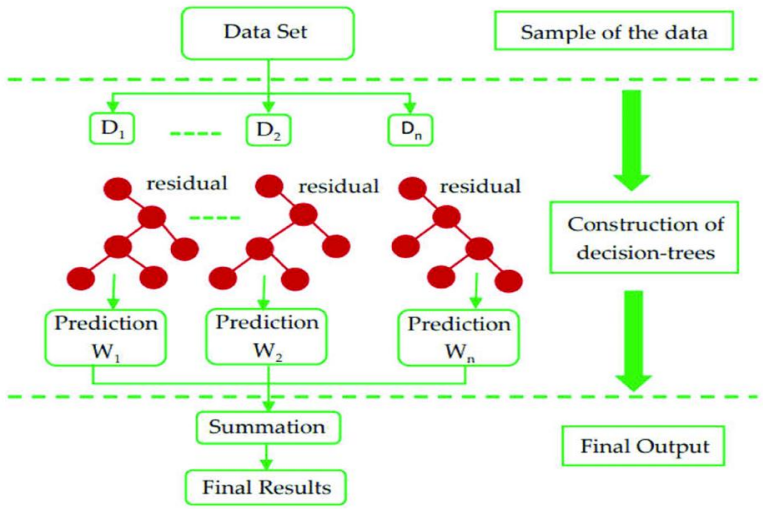
Thuật toán K-Nearest Neighbors (KNN) là một thuật toán học máy có giám sát với tính đơn giản và khả năng dễ triển khai, đã được áp dụng rộng rãi trong các bài toán phân loại và hồi quy, như đã được chỉ ra trong nghiên cứu của Laaksonen và Oja (1996). Theo Kramer (2013), KNN đánh giá các điểm dữ liệu dựa trên việc xem xét các điểm lân cận trong không gian đặc trưng. Nếu các điểm tương tự gần nhau, chúng sẽ thuộc cùng một lớp. Sau đó, KNN xác định các hàng xóm lân cận để đưa ra dự đoán và gán nhãn cho một điểm cụ thể.

Trong công trình nghiên cứu của Mucherino và cộng sự (2009), nhóm tác giả cho rằng giá trị k trong thuật toán KNN là số lượng điểm lân cận được xem xét để phân loại một điểm truy vấn. Khi giá trị k = 1, mô hình sẽ dựa vào lớp của điểm lân cận gần nhất để thực hiện phân loại. Việc xác định giá trị k tối ưu là một bước quan trọng nhằm đảm bảo độ chính xác của mô hình. Tuy nhiên, quá trình này phụ thuộc vào các đặc tính cụ thể của tập dữ liệu và yêu cầu sự thử nghiệm và điều chỉnh cẩn thận. Do đó, khi lựa chọn giá trị k thích hợp, cần xem xét cả tỷ lệ lớn/nhỏ của dữ liệu cũng như độ phức tạp của nó, nhằm đảm bảo rằng mô hình có thể đạt được độ chính xác tối ưu.

2.1.3. Extreme Gradient Boosting

Hình 1

Kiến Trúc Thuật Toán XGBoost



Nguồn: Dữ liệu từ "Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation" bởi M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński và U. Amjad, 2022, *Applied Sciences, 12*(4), Article 2126

Từ những phân tích sâu sắc trong nghiên cứu của Chen và Guestrin (2016) hay của Li và cộng sự (2021), thuật toán Extreme Gradient Boosting - XGBoost là một thuật toán tăng cường dựa trên cây quyết định, được biết đến với khả năng mở rộng và hiệu quả cao. Khác với các thuật toán tăng cường truyền thống, XGBoost có khả năng thực hiện tính toán đồng thời trên nhiều luồng, đó là kết hợp các cây mô hình học tập cơ bản yếu thành một cây mô hình học tập mạnh hơn theo kiểu tuần tự, giúp cải thiện độ chính xác của dự đoán cuối cùng. Kiến trúc của XGBoost có thể được thể hiện trong Hình 1.

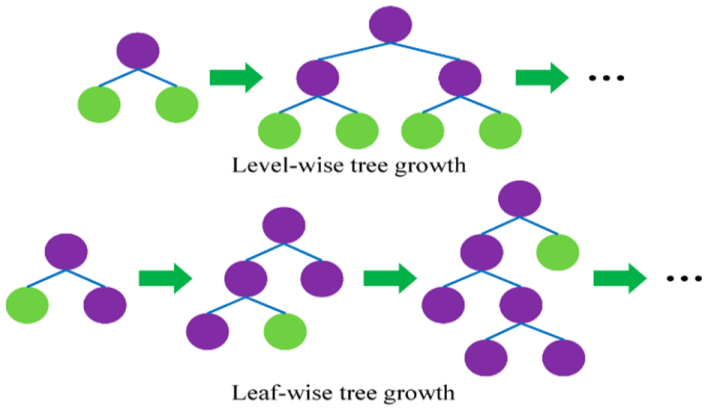
2.1.4. Light Gradient Boosting Machine

Light Gradient Boosting Machine - LightGBM là một khung công cụ (framework) tăng cường gradient dựa trên thuật toán cây quyết định được đề xuất và công bố bởi Microsoft vào năm 2017. Mục tiêu của LightGBM là cải thiện hiệu quả tính toán và giải quyết các vấn đề dự đoán với dữ liệu lớn. Trong nghiên cứu của Taha và Malebary (2020), nguyên tắc của thuật toán LightGBM được mô tả là sử dụng phương pháp giảm dần độ dốc để xác định giá trị gần đúng của phần dư bằng cách sử dụng độ dốc âm của hàm mất mát trong mô hình hiện tại, sau đó khớp với cây hồi quy. Sau nhiều vòng lặp, kết quả của tất cả các cây hồi quy được cộng dồn để đạt được kết quả cuối cùng.

Nghiên cứu của Zhang và Gong (2020) cùng Al Daoud (2019) đã chỉ ra rằng cả LightGBM và XGBoost đều hỗ trợ tính toán song song, tuy nhiên, sự khác biệt chính giữa XGBoost và LightGBM nằm ở cách xây dựng cây quyết định (Hình 2). Trong XGBoost, cây quyết định được xây dựng theo chiều ngang (theo cấp độ), trong khi cây quyết định của LightGBM được xây dựng theo chiều dọc (theo chiều lá), chính điều này đã tạo nên sự khác biệt về tốc độ huấn luyện và độ chính xác của hai thuật toán.

Hình 2

Phát Triển Theo Cấp Độ và Phát Triển Theo Chiều Lá



Nguồn: Dữ liệu từ "Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms" bởi W. Liang, S. Luo, G. Zhao và H. Wu, 2020, *Mathematics, 8*(5), Article 765

2.2. Phương pháp Học sâu

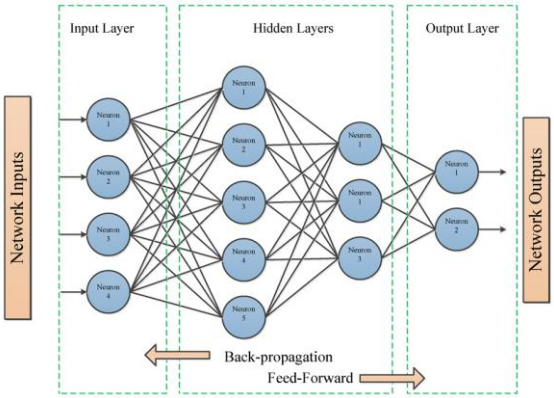
2.2.1. Mạng Nơron nhân tạo

Mạng Nơron nhân tạo (Artificial Neural Networks - ANN) là một cấu trúc được mô phỏng tế bào thần kinh sinh học trong não bộ của động vật hoặc con người. Nó được hình thành bởi các đơn vị xử lý đơn giản được gọi là tế bào thần kinh (Daoud & Mayo, 2019; Walczak, 2019). Bộ não con người chứa hàng tỷ tế bào thần kinh, chúng đóng vai trò quan trọng trong truyền tải và xử lý thông tin trong cơ thể. Những tế bào thần kinh này được kết nối với nhau thông qua một cấu trúc đặc biệt được gọi là khớp thần kinh.

Giai đoạn huấn luyện của ANN điều chỉnh trọng số của các khớp thần kinh này, từ đó mô hình hóa mối quan hệ giữa đầu vào và đầu ra của hệ thống. ANN có khả năng mô hình hóa các vấn đề phi tuyến tính và phức tạp, đồng thời dễ triển khai vì có sẵn nhiều thư viện hỗ trợ cho các ngôn ngữ lập trình khác nhau. Đặc biệt, thuật toán này còn có khả năng tổng quát hóa cao, cho phép hệ thống chấp nhận dữ liệu bên ngoài tập huấn luyện. Tuy nhiên, cần kiểm chứng và xác nhận độ chính xác của lý thuyết này thông qua các tài liệu và nghiên cứu thực tế.

Hình 3

Kiến Trúc Mạng Noron Nhân Tạo



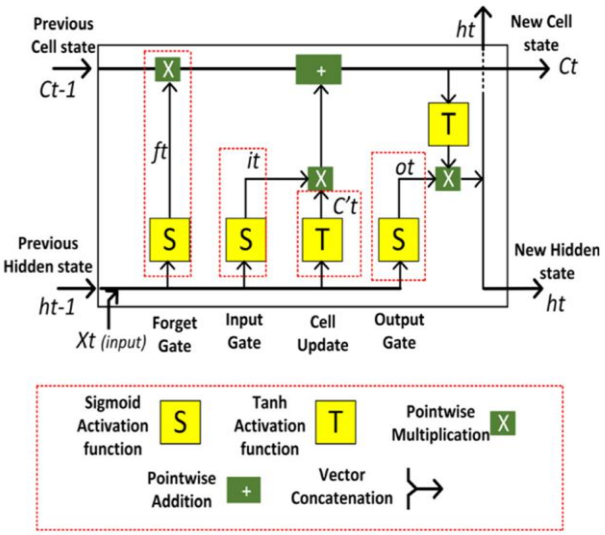
Nguồn: Dữ liệu từ "Artificial neural networks based optimization techniques: A review" bởi M. G. Abdolrasol, S. M. Hussain, T. S. Ustun, M. R. Sarker, M. A. Hannan, R. Mohamed, ... A. Milad, 2021, *Electronics*, *10*(21), Article 2689

2.2.2. Long short-term memory

Theo Graves (2012), Long Shot-Term Memory - LSTM là một mô hình học sâu được tạo ra từ mạng hồi quy RNN. Nó là một thuật toán được thiết kế để xử lý dữ liệu tuần tự như văn bản, lời nói và chuỗi thời gian. Hochreiter và Schmidhuber (1997) đã đề xuất thuật toán LSTM nhằm giải quyết vấn đề về sự phụ thuộc dài hạn của RNN, trong đó RNN không thể dự đoán được thông tin lưu trữ trong bộ nhớ dài hạn nhưng có thể cung cấp dự đoán chính xác hơn từ thông tin gần đây.

Hình 4

Kiến Trúc của Mạng LSTM



Nguồn: Dữ liệu từ "CNN-LSTM vs. LSTM-CNN to predict power flow direction: A case study of the high-voltage subnet of Northeast Germany" bởi F. Aksan, Y. Li, V. Suresh và P. Janik, 2023, *Sensors, 23*(2), Article 901

Trong nghiên cứu của Aksan và cộng sự (2023), mạng LSTM bao gồm các khối bộ nhớ được gọi là ô. Mỗi ô có hai trạng thái: trạng thái ô và trạng thái ẩn. LSTM được lựa chọn để xử lý và dự đoán các sự kiện quan trọng trong chuỗi thời gian với khoảng thời gian dài và độ trễ. Điều này cũng là một lợi thế lớn so với RNN, vì RNN bị hạn chế trong khả năng lưu trữ thông tin dài hạn và gặp vấn đề biến mất độ dốc, khiến trọng số của mạng nơ-ron ở các tầng sâu hơn không thể được cập nhật trong quá trình lan truyền ngược. Kiến trúc hoạt động của LSTM được mô tả trong Hình 4.

2.3. Phương pháp đánh giá hiệu quả mô hình

Một bước quan trọng sau khi xây dựng mô hình là đánh giá hiệu suất và chất lượng của mô hình bằng các phương pháp đánh giá. Trong nghiên cứu này, nhóm tác giả sử dụng một số tiêu chí đánh giá quan trọng và phổ biến trong việc đánh giá độ chính xác mô hình phân loại.

Trong ngữ cảnh của dữ liệu ứng dụng trong lĩnh vực cho vay trong nghiên cứu này, nhóm tác giả phân loại khách hàng có khả năng chi trả khoản vay là Positive và Negative nếu không có khả năng chi trả khoản vay. Mỗi dự đoán có thể thuộc vào một trong bốn kết quả, dựa trên cách nó khớp với giá trị thực tế:

- True Positive (TP): Khách hàng được dự đoán là có khả năng trả nợ và thực tế cũng đã trả nợ.
- True Negative (TN): Khách hàng được dự đoán là không có khả năng trả nợ và thực tế cũng không trả nợ.
- False Positive (FP - Sai loại 1): Khách hàng được dự đoán là có khả năng trả nợ nhưng thực tế họ không trả nợ.
- False Negative (FN - Sai loại 2): Khách hàng được dự đoán là không có khả năng trả nợ nhưng thực tế họ đã trả nợ.

Độ chính xác (Accuracy) là một phương pháp tiêu chuẩn được sử dụng để đánh giá các thuật toán học tập, đo lường số lượng quan sát dự báo đúng trên tổng số quan sát . Công thức được thể hiện trong nghiên cứu của (Halagundegowda & ctg., 2023):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision phản ánh mức độ đáng tin cậy của mô hình trong việc phân loại các mẫu là Positive. Khi giá trị Precision càng cao, các mẫu tin được phân loại bởi mô hình sẽ có tỉ lệ chính xác cao. Giá trị Precision càng gần với 1 thì mô hình có tỷ lệ dự đoán đúng thuộc vào lớp True Positives so với tổng số dự đoán Positive càng cao. Tuy nhiên, Precision cao không có nghĩa là mô hình hoàn hảo và có thể gặp phải các vấn đề khác như Recall thấp hoặc quá khớp (overfitting). Công thức được thể hiện trong nghiên cứu của (Halagundegowda & ctg., 2023):

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity) xác định có bao nhiêu mẫu tin thực sự ở lớp Positive được mô hình phân lớp đúng vào lớp Positive. Khi giá trị Recall càng cao thì mô hình có khả năng tìm ra nhiều hơn các mẫu tin thuộc lớp True Positive, điều đó chỉ ra rằng mô hình có độ chính xác cao. Giá trị Recall càng gần với 1 cho thấy mô hình không bỏ sót mẫu tin quá nhiều mẫu tin thực sự thuộc lớp Positive. Công thức được thể hiện trong nghiên cứu của (Halagundegowda & ctg., 2023):

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score là một đo lường kết hợp giữa Recall và Precision. Tuy nhiên, có sự đánh đổi giữa Precision và Recall, do đó F1-Score có thể được sử dụng để đo lường mức độ hiệu quả mà mô hình của chúng ta thực hiện sự đánh đổi đó. Công thức được thể hiện trong nghiên cứu của (Halagundegowda & ctg., 2023):

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Specificity xác định có bao nhiêu mẫu tin thực sự ở lớp Negative được mô hình phân lớp đúng vào lớp Negative. Khi giá trị Specificity càng cao thì mô hình có khả năng tìm ra nhiều hơn các mẫu tin thuộc lớp True Negative, điều đó chỉ ra rằng mô hình có độ chính xác cao trong việc xác định những trường hợp không có khả năng chi trả cho khoản vay. Giá trị Specificity càng gần với 1 cho thấy mô hình không phân loại sai quá nhiều mẫu tin thực sự thuộc lớp Negative. Công thức được chi ra trong công trình nghiên cứu của (Peterson & ctg., 1999):

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

Đường cong ROC hoạt động như một chỉ số trực quan, minh họa cho khả năng hoạt động của mô hình phân loại như đã được mô tả trong nghiên cứu của Van der Schouw và cộng sự (1992). Nó cung cấp cho chúng ta cái nhìn sâu sắc về cách mô hình ra quyết định dựa trên các mức độ tin cậy khác nhau. Đường cong ROC bao gồm hai thành phần chính, một là Tỷ lệ Dương tính Thực (True Positive Rate - TPR), thể hiện tần suất mô hình dự đoán chính xác các trường hợp dương tính; và hai là Tỷ lệ Dương tính Giả (False Positive Rate - FPR), thể hiện tần suất mô hình dự đoán sai các trường hợp âm tính thành dương tính. Qua việc phân tích biểu đồ này, chúng ta có thể đánh giá được mức độ hiệu quả của mô hình và xác định ngưỡng phù hợp để đạt được sự cân đối tốt nhất giữa các dự đoán đúng và sai.

$$TPR = \frac{TP}{TP + FN} = Recall \quad (6)$$

$$FPR = \frac{FP}{FP+TN} = 1 - Specificity \quad (7)$$

AUC, hay diện tích dưới đường cong ROC, trong nghiên cứu của Van der Schouw và cộng sự (1992), nhóm tác giá nhận định AUC là một chỉ số tổng hợp đánh giá hiệu suất của mô hình phân loại qua các ngưỡng phân loại khác nhau. Khi giá trị AUC tiến gần tới 1, chúng ta có thể kết luận rằng mô hình phân loại hoạt động rất hiệu quả. Trong trường hợp AUC = 0.5, mô hình không thể phân biệt được giữa lớp Positive và Negative. Đặc biệt, nếu AUC < 0.5, điều này có nghĩa là mô hình đang dự đoán ngược, tức là dự đoán lớp Positive là Negative và ngược lại.

2.4. Khái niệm về cho vay

Khái niệm về cho vay được đưa ra trong Luật số 47/2010/QH12 của Quốc hội: Luật các tổ chức tín dụng, cho vay là một dạng tín dụng, trong đó bên cho vay sẽ cung cấp một lượng tiền nhất định cho một cá nhân hoặc tổ chức nào đó để sử dụng cho mục đích cụ thể trong một khoảng thời gian đã được thỏa thuận, với điều kiện phải trả lại cả số tiền gốc và lãi suất (Chính phủ Việt Nam, 2010). Các ngân hàng thương mại thông qua hoạt động cho vay đóng góp quan trọng vào việc tạo ra lợi nhuận kinh tế cho chính họ, đồng thời đáp ứng nhu cầu vay vốn của khách hàng và đẩy mạnh sự tăng trưởng của nền kinh tế.

Ngoài ra, cùng với sự phát triển của công nghệ thông tin, hoạt động cho vay ngang hàng cũng đang ngày càng mở rộng nhanh chóng, đây là một mô hình dịch vụ tài chính sử dụng công nghệ Fintech để tạo ra một kết nối trực tiếp giữa người vay và nhà đầu tư, mà không cần sự can thiệp của các tổ chức tài chính trung gian như ngân hàng (Bachmann & ctg., 2011).

2.5. Nghiên cứu liên quan

Theo Costa e Silva và cộng sự (2020), các nhà nghiên cứu và chuyên gia tài chính có thể đưa ra dự đoán về xác suất rủi ro tín dụng dựa trên các biến đầu vào bằng cách áp dụng mô hình hồi quy Logistic. Mô hình này đã được chứng minh là hữu ích trong việc phân loại các khoản vay thành các nhóm rủi ro khác nhau và đưa ra quyết định về việc cấp vay, quản lý rủi ro và xây dựng các chiến lược tín dụng hiệu quả.

Bên cạnh đó, thuật toán Extreme Gradient Boosting (XGBoost) đã nhanh chóng lan truyền và được sử dụng rộng rãi kể từ khi ra đời và đã đạt được kết quả dự đoán khả quan trong nhiều nhiệm vụ dự đoán và phân tích dữ liệu không cân bằng. Đặc biệt trong lĩnh vực tài chính, việc áp dụng thuật toán XGBoost để dự đoán rủi ro tín dụng đang trở nên ngày càng phổ biến hơn. Nhóm tác giả Chang và cộng sự (2018) trong công trình nghiên cứu của mình đã ứng dụng thuật toán XGBoost để xây dựng mô hình đánh giá rủi ro tín dụng cho các tổ chức tài chính. Từ kết quả cho thấy, thuật toán XGBoost hoạt động tốt và tối ưu hơn các thuật toán khác được sử dụng như Máy vectơ hỗ trợ (Support Vector Machines) và Hồi quy Logistic (Logistic Regression).

Để tối ưu hóa khả năng dự đoán của mô hình, nhiều nhóm tác giả đã thực hiện các nghiên cứu tối ưu hóa các tham số của mô hình. Theo nghiên cứu của Xia và cộng sự (2017), nhóm tác giả đã tiến hành cải thiện các tham số trong thuật toán XGBoost bằng cách vận dụng phương pháp tối ưu hóa siêu tham số Bayesian để xây dựng mô hình đánh giá rủi ro tín dụng. Từ kết quả, có thể kết luận rằng sau khi áp dụng phương pháp tối ưu hóa siêu tham số Bayes đưa ra kết quả dự đoán mô hình tốt hơn so với các mô hình chuẩn khác. Nhìn chung, kết quả dự đoán của mô hình đã được cải thiện đáng kể khi áp dụng những phương pháp tối ưu hóa tham số của mô hình.

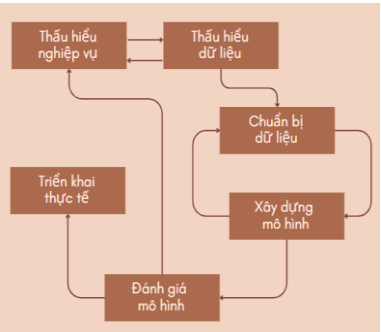
Ko và cộng sự (2022) đã kiểm nghiệm sự hiệu quả của ba mô hình thống kê: Hồi quy logistic, Phân loại Bayes và Linear Discriminant Analysis (LDA); cùng với năm mô hình AI: Cây quyết định, Rừng ngẫu nhiên, LightGBM, ANN và CNN thông qua việc xây dựng các mô hình dự đoán từ dữ liệu Lending Club. Mục tiêu là tạo điều kiện cho các nhà đầu tư có thể đánh giá sớm mức tín dụng của người vay, đồng thời tăng khả năng minh bạch thông tin và giảm rủi ro vỡ nợ cũng như tình trạng không cân xứng thông tin. Để đo lường hiệu suất dự đoán, các thước đo chính trong nghiên cứu này bao gồm độ chính xác, giá trị AUC, giá trị KS, thước đo F, Kappa và một số độ đo khác. Kết quả cho thấy LightGBM vượt trội hơn so với các mô hình còn lại. Nhóm tác giả cũng tiến hành phân tích để chứng minh rằng việc áp dụng LightGBM có thể tạo ra lợi ích đáng kể cho doanh thu của Lending Club so với việc sử dụng các mô hình còn lại.

3. Phương pháp nghiên cứu và thực nghiệm mô hình

3.1. Quy trình thực nghiệm

Hình 5

Mô Hình CRISP-DM



Nguồn: Dữ liệu từ "CRISP-DM: Towards a standard process model for data mining" bởi R. Wirth và J. Hip, 2000, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 1*, pp. 29-39, The Practical Application Company

Trên cơ sở mô hình CRISP-DM Wirth và Hipp (2000), nghiên cứu này được xây dựng với quy trình gồm 04 giai đoạn chính phù hợp với bài toán đã đặt ra như sau:

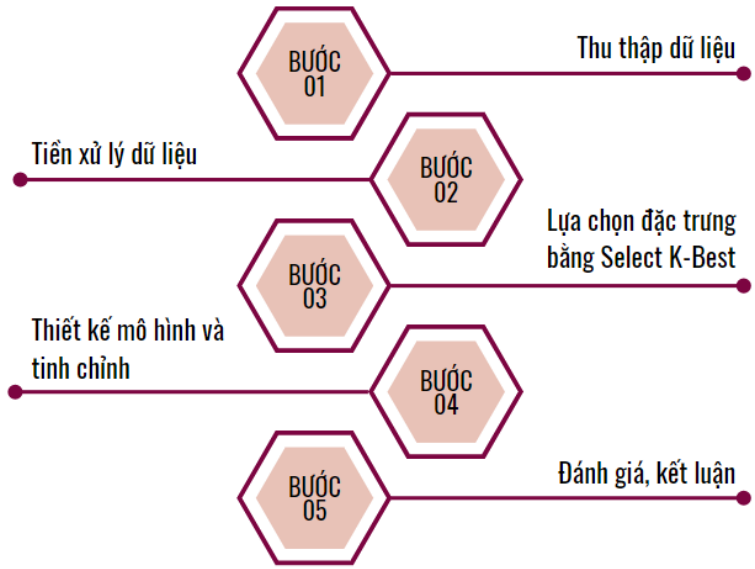
- Bước 1: Thu thập dữ liệu: Thu thập dữ liệu sẵn có;
- Bước 2: Tiền xử lý dữ liệu;
- Bước 3: Lựa chọn đặc trưng bằng thuật toán Select K-Best;
- Bước 4: Thiết kế mô hình và tinh chỉnh;

  Mô hình Máy học: Logistic Regression, KNN, XGB, LGBM

  Mô hình Học sâu: LSTM, ANN
- Bước 5: Đánh giá, kết luận.

Hình 6

Quy Trình Nghiên Cứu



Nguồn: Nhóm tác giả

Nhóm sử dụng ngôn ngữ lập trình Python - ngôn ngữ lập trình được yêu thích nhất năm 2023 bởi tính trực quan, dễ hiểu và tận dụng các thư viện mạnh mẽ, hỗ trợ xây dựng các mô hình học máy.

3.2. Thu thập dữ liệu

Theo tìm hiểu và nghiên cứu của tác giả, hoạt động cho vay có nhiều hình thức như: ngân hàng cho vay, công ty tài chính cho vay, hoặc cho vay ngang hàng. Trong nghiên cứu này, tác giả sử dụng tập dữ liệu từ Lending Club - một trang website hỗ trợ người vay kết nối với nhà đầu tư qua Internet để tiến hành dự báo khả năng hoàn trả khoản vay của khách hàng. Tập dữ liệu với 28 thuộc tính gồm 113,997 mẫu được thu thập từ Kaggle trong năm 2016. Dưới đây là thông tin mô tả tập dữ liệu.

Bảng 1

Thông Tin Mô Tả Tập Dữ Liệu

Số thứ tự	Tên thuộc tính	Kiểu dữ liệu	Mô tả thuộc tính	Giá trị đại diện
1	loan_amount	float64	Số tiền đã vay của khách hàng	1,000, 8,000, 24,375
2	term	object	Thời gian cho khoản vay theo tháng.	36 months, 60 months
3	int_rate	float64	Lãi suất cho vay	11.44, 17.27, 11.99
4	installment	float64	Số tiền trung bình hàng tháng phải trả nếu khoản vay được kích hoạt	329.48, 265.68, 609.33
5	grade	object	Nhóm khách hàng được xếp loại theo Lending Club	B, A, C
6	sub_grade	object	Nhóm khách hàng được xếp loại chi tiết	B4, A2, C5
7	zip_code	object	Ba số đầu tiên mã zip của người vay	
8	emp_title	object	Nghề nghiệp của khách hàng	Marketing, Credit analyst, Pilot
9	emp_length	object	Số năm làm việc của khách hàng	10+ years, < 01 year, 03 years
10	home_ownership	object	Tình trạng sở hữu nhà ở của người vay khi đăng ký.	RENT, OWN, MORTGAGE
11	annual_inc	float64	Thu nhập bình quân hằng năm của người đi vay	117,000, 46,000, 24,000
12	verification status	object	Xác định tình trạng thu nhập của khách hàng: gồm 02 giá trị verified: xác thực, not verified: chưa xác thực	Not Verified, Verified, Source Verified
13	issue_d	object	Ngày kích hoạt khoản vay	Jan-15, Oct-14, Apr-13
14	purpose	object	Mục đích của khoản vay của khách hàng như mua nhà, mua xe, sinh hoạt phí, ...	Vacation, credit_card, small_business
15	title	object	Tiêu đề do người đi vay cung cấp cho khoản vay nhằm cung cấp mục đích vay của khách hàng	Vacation, Credit card Refinance, Business

Số thứ tự	Tên thuộc tính	Kiểu dữ liệu	Mô tả thuộc tính	Giá trị đại diện
16	dti	float64	Hệ số nợ trên thu nhập (Debt to Income Ratio) là tỉ lệ phần trăm của tổng thu nhập hàng tháng để trả các khoản thanh toán nợ hàng tháng	26.24, 22.05, 12.79
17	earlies_cr_line	object	Ngày đầu tiên mà khách hàng mở hạn mức tín dụng	Jun-90, Aug-07, Mar-82
18	open_acc	float64	Số tài khoản mà khách hàng đã mở	16, 8, 13
19	pub_rec	float64	Thông tin về lịch sử tín dụng của bên vay (báo cáo tín dụng) mà bên cho vay có thể sử dụng hợp pháp, để bác bỏ yêu cầu vay hoặc đơn xin vay tín dụng. Bao gồm những việc như phá sản, thanh toán trễ hạn và các khoản vay đã xóa bỏ trước đây	0, 1
20	revol_bal	float64	Số dư tài khoản tín dụng liên quan của khách hàng	36,369, 20,131, 5,472
21	revol_until	float64	Thời gian hiệu lực của khoản tín dụng	41.8, 100.6, 4.9
22	total_acc	float64	Tổng số hạn mức tín dụng hiện tại của khách hàng được khi nhận	25, 40, 17
23	initial_list_status	object	Trạng thái ghi nhận ban đầu, giá trị phù hợp là W, F	W, F
24	application_type	object	Loại đăng ký, bao gồm Individual: cá nhân, Join: kết hợp và Direct_pay: trả trực tiếp	INDIVIDUAL, JOINT, DIRECT_PAY
25	mort_acc	float64	Tổng số tài khoản thế chấp của khách hàng	0, 3, 4
26	pub_rec_bankruptcies	float64	Khoản ghi nhận phá sản công khai	0, 1
27	address	object	Địa chỉ của người vay	0174 Michelle Gateway, USCGC Tran, 3390 Luis Rued
28	loan_status	object	Trạng thái của khoản vay. Gồm 2 giá trị: Charged Off - không có khả năng hoàn trả hoàn toàn và Fully Paid - hoàn trả toàn bộ khoản vay	Fully Paid, Charged Off,

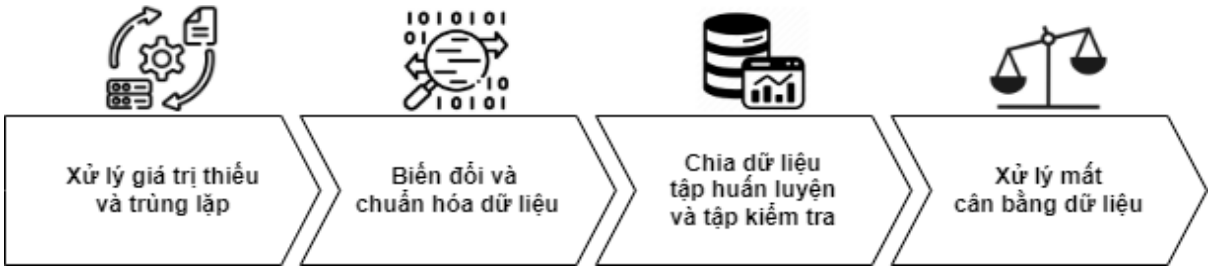
Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

3.3. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một phần quan trọng trong quá trình làm việc với dữ liệu và xây dựng các mô hình học máy. Nếu không tiền xử lý dữ liệu tốt, có thể gây ra nhiều vấn đề và ảnh hưởng đến hiệu suất và độ tin cậy của mô hình. Hình 7 dưới đây là các bước tiền xử lý dữ liệu trước khi đưa vào lựa chọn thuộc tính đầu vào mô hình.

Hình 7

Các Bước Tiền Xử Lý Dữ Liệu



Nguồn: Nhóm tác giả

Xử lý giá trị thiếu và trùng lặp: Trong bài nghiên cứu, phương thức dropna(), drop_duplicates() được sử dụng để loại bỏ các giá trị không cần thiết.

Biến đổi và chuẩn hóa dữ liệu: Nhóm tác giả lấy Logarit một vài biến có giá trị rất lớn và biến động mạnh nhằm giảm sự biến động và sự không đồng nhất phương sai. Đồng thời phương thức MinMaxScaler() có thể giúp giải quyết vấn đề các mô hình yêu cầu đặc trưng phải có cùng phạm vi và Label_Encoder() giúp biến đổi dữ liệu phân loại và dữ liệu văn bản thành dạng phù hợp cho huấn luyện mô hình.

Chia tập huấn luyện và tập kiểm tra: Tùy thuộc vào tính chất của bộ dữ liệu mà có thể phân chia ứng với từng chức năng huấn luyện, kiểm tra. Trong bài nghiên cứu, nhóm chia dataset thành tập huấn luyện và kiểm tra với tỉ lệ 8:2.

Xử lý mất cân bằng: Đối với bài toán phân loại, khi biến mục tiêu có các giá trị quá chênh lệch, hiện tượng overfitting rất dễ xảy ra. Synthetic Minority Oversampling Technique (SMOTE) là một phương pháp bổ sung kích thước mẫu được nhóm nghiên cứu tiến hành nhằm giải quyết hiện tượng mất cân bằng dữ liệu**.**

3.4. Lựa chọn đặc trưng

Độ hiệu quả của mô hình phụ thuộc rất lớn vào các thuộc tính được chọn từ tập dữ liệu. Nhóm nghiên cứu tiến hành kiểm tra tập dữ liệu và xóa những đặc trưng chỉ có một giá trị. Các đặc trưng chỉ có một giá trị cho tất cả các quan sát không cung cấp bất kỳ sự biến đổi nào trong dữ liệu và do đó không cung cấp thông tin hữu ích cho việc phân tích. Chúng làm phức tạp dữ liệu mà không đóng góp vào sự hiểu biết. Ngoài ra những đặc trưng không phù hợp cũng được loại bỏ nhằm giúp các mô hình tập trung vào thông tin quan trọng, đạt được kết quả tốt hơn. Sau các quá trình xử lý trên, tập dữ liệu còn lại 18 đặc trưng làm đầu vào cho mô hình dự đoán. Tiếp theo, nhóm nghiên cứu tiến hành chọn lọc đặc trưng. Việc này có thể giúp cải thiện hiệu suất của mô hình giảm thiểu số lượng đặc trưng trong dữ liệu mà vẫn giữ lại những thông tin quan trọng nhất, từ đó cải thiện độ chính xác của mô hình và giảm thiểu sự quá khớp và giảm thời gian huấn luyện mô hình. Phương pháp Select K-Best được đánh giá là một trong những phương pháp cải thiện mô hình và đạt hiệu suất tốt (Desyani & ctg., 2020). Thay vì chọn số lượng đặc trưng cụ thể, nhóm tác giả cải tiến phương pháp bằng cách thực hiện vòng lặp nhằm tự động tìm ra số

lượng đặc trưng tốt nhất ứng với từng mô hình. Kết quả ở Bảng 2 chọn ra được các biến tác động nhất (giá trị "TRUE") ứng với từng mô hình, từ đó nhóm tác giả tiến hành xếp hạng các biến quan trọng dùng cho quá trình thực nghiệm. Có thể thấy rằng, một số đặc trưng luôn được lựa chọn cho các mô hình như: *int_rate, subgrade, term, ...* các biến còn lại, tùy thuộc vào mỗi mô hình mà có các kết quả khác nhau.

Bảng bên dưới mô tả các thư viện mà nhóm nghiên cứu sử dụng để xây dựng các mô hình học máy:

Bảng 2

Mô Hình và Thư Viện Sử Dụng

STT	Tên mô hình	Tên thư viện
1	XGBoost	XgboostClassifier
2	LightGBM	LGBMClassifier
3	KNN	Sklearn
4	Logistic Regression	Sklearn
5	ANN	Keras
6	LSTM	Keras

Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

Bên cạnh những thư viện XgboostClassifier, LGBMClassifier, nhóm tác giả sử dụng những thư viện mã nguồn mở phổ biến như Sklearn và Keras. Keras là một thư viện học sâu mã nguồn mở trên TensorFlow, cung cấp giao diện đơn giản cho việc xây dựng và huấn luyện mô hình mạng nơ-ron. Trong khi đó, scikit-learn là một thư viện máy học Python chuyên sâu vào tiền xử lý dữ liệu, lựa chọn mô hình, và đánh giá hiệu suất.

Bảng 3

Lựa Chọn Đặc Trưng Đầu Vào được Chọn theo Từng Mô Hình

Tên thuộc tính	*XGB*	*LGBM*	*LR*	*KNN*	*Xếp hạng*
int_rate	TRUE	TRUE	TRUE	TRUE	1
subgrade	TRUE	TRUE	TRUE	TRUE	1
term	TRUE	TRUE	TRUE	TRUE	1
loan_amnt	TRUE	TRUE	TRUE	TRUE	1
dti	TRUE	TRUE	TRUE	TRUE	1
installment	TRUE	TRUE	TRUE	TRUE	1
mort_acc	TRUE	TRUE	TRUE	TRUE	1
revol_util	TRUE	TRUE	TRUE	TRUE	1
log_annual_inc	TRUE	TRUE	TRUE	TRUE	1
purpose	TRUE	TRUE	TRUE	FALSE	2
emp_length	TRUE	TRUE	TRUE	FALSE	2

Tên thuộc tính	***XGB***	***LGBM***	***LR***	***KNN***	***Xếp hạng***
pub_rec_bankruptcies	TRUE	FALSE	TRUE	TRUE	2
total_acc	FALSE	FALSE	TRUE	TRUE	3
home_ownership	TRUE	FALSE	TRUE	FALSE	3
verification_status	TRUE	FALSE	TRUE	FALSE	3
pub_rec	FALSE	FALSE	TRUE	TRUE	3
log_revol_bal	TRUE	FALSE	FALSE	FALSE	4
open_acc	FALSE	FALSE	FALSE	FALSE	5
Số thuộc tính chọn	**15**	**11**	**16**	**12**	

Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

4. Kết quả thực nghiệm

4.1. Kết quả thực nghiệm mô hình máy học

Các mô hình máy học được triển khai bao gồm Logistic Regression, KNN, XGBoost và LightGBM. Tổng quan kết quả phân loại của các thuật toán trước và sau khi lựa chọn đặc trưng có sự khác biệt. Để tối ưu hiệu suất, nhóm tác giả đã sử dụng phương pháp Optuna - một kỹ thuật tự động tinh chỉnh tham số. Cụ thể kết quả được so sánh dưới đây:

Bảng 4

Kết Quả Thực Nghiệm các Mô Hình Máy Học

Mô hình		***XGB***	***LightGBM***	***LG***	***KNN***
Trước khi lựa chọn đặc trưng	*Accuracy*	85.67%	85.80%	68.75%	73.15%
	Precision	86.71%	87.24%	68.75%	73.22%
	Recall	85.63%	85.80%	68.75%	73.16%
	F1-Score	85.56%	85.66%	68.74%	73.13%
	ROC	85.63%	85.76%	68.75%	73.16%
Sau khi lựa chọn đặc trưng	*Accuracy*	85.74%	85.96%	68.75%	73.72%
	Precision	86.83%	87.48%	68.75%	73.84%
	Recall	85.71%	85.92%	68.75%	73.73%
	F1-Score	85.63%	85.80%	68.75%	73.69%
	ROC	85.71%	85.92%	68.75%	73.16%
Sau khi tinh chỉnh	*Accuracy*	86.28%	86.37%	68.83%	78.60%
	Precision	87.23%	87.67%	68.83%	78.69%
	Recall	86.25%	86.37%	68.83%	78.61%
	F1-Score	86.18%	86.25%	68.83%	78.59%
	ROC	86.06%	86.33%	68.83%	78.61%

Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

Các mô hình học máy đã được cải thiện đáng kể sau khi áp dụng kỹ thuật lựa chọn đặc trưng. XGBoost và LightGBM đã tăng độ chính xác từ 85.67% lên 85.74% và 85.80% lên 85.96% sau khi sử dụng phương pháp Select K-Best để loại bỏ các đặc trưng không quan trọng; tương tự, mô hình KNN cũng đã cải thiện, cụ thể độ chính xác tăng từ 73.15% lên 73.84%. Mô hình Logistic Regression tuy kết quả không thể hiện rõ ràng cải thiện, song thông qua các mô hình, điều này cho thấy rằng việc loại bỏ các đặc trưng không quan trọng đã giúp mô hình tập trung vào thông tin quan trọng hơn và cải thiện khả năng dự đoán chính xác của chúng.

So sánh kết quả của các thuật toán sau khi tinh chỉnh, mô hình LightGBM cho kết quả tốt nhất với các chỉ số Accuracy, Precision, Recall, F1-Score và AUC ROC lần lượt là 86.37%, 87.67%, 86.37%, 86.25% và 86.33%. Tiếp theo là XGBoost và thấp nhất là KNN với các chỉ số Accuracy, Precision, Recall, F1-Score và AUC ROC xấp xỉ 69%.

4.2. Kết quả thực nghiệm mô hình học sâu

Các mô hình học sâu được thực hiện bao gồm ANN và LSTM. Tương tự các mô hình máy học, nhóm tác giả sử dụng Optuna nhằm tìm kiếm kiến trúc mạng nơ-ron tốt nhất, từ đó cải thiện hiệu suất mô hình.

Bảng 5

Kết Quả Thực Nghiệm các Mô Hình Học Sâu

Thuật toán		***ANN***	***LSTM***
Trước khi tinh chỉnh	*Accuracy*	77.03%	**77.75%**
	Precision	77.32%	**77.85%**
	Recall	77.03%	**77.74%**
	F1-Score	76.96%	**77.72%**
	ROC	77.01%	**77.74%**
Sau khi tinh chỉnh	*Accuracy*	77.93%	**78.44%**
	Precision	78.25%	**75.54%**
	Recall	77.93%	**84.34%**
	F1-Score	77.86%	**79.70%**
	ROC	77.91%	**78.41%**

Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

Kết quả từ Bảng 3 thể hiện mô hình LSTM cho các chỉ số Accuracy, Precision, Recall, F1-Score và AUC ROC cao hơn khi so sánh với ANN. Điều này cho thấy mô hình LSTM dự đoán tốt hơn trên tập dữ liệu được chọn.

4.3. Tổng kết thực nghiệm

Kết quả nghiên cứu đã mở ra một phương pháp hữu ích để xác định nhóm đặc trưng quan trọng nhất và nhóm đặc trưng không quan trọng đã bị loại bỏ. Bằng cách phân tích thứ hạng của các đặc trưng trong quá trình lựa chọn, chúng ta có thể hiểu rõ hơn về sự ảnh hưởng của từng nhóm đặc trưng đối với hiệu suất của mô hình, đồng thời có thể đưa ra các khuyến nghị cụ thể để cải thiện hiệu suất và hiệu quả của mô hình, từ việc tối ưu hóa nhóm đặc trưng quan trọng nhất đến việc loại bỏ nhóm đặc trưng không quan trọng.

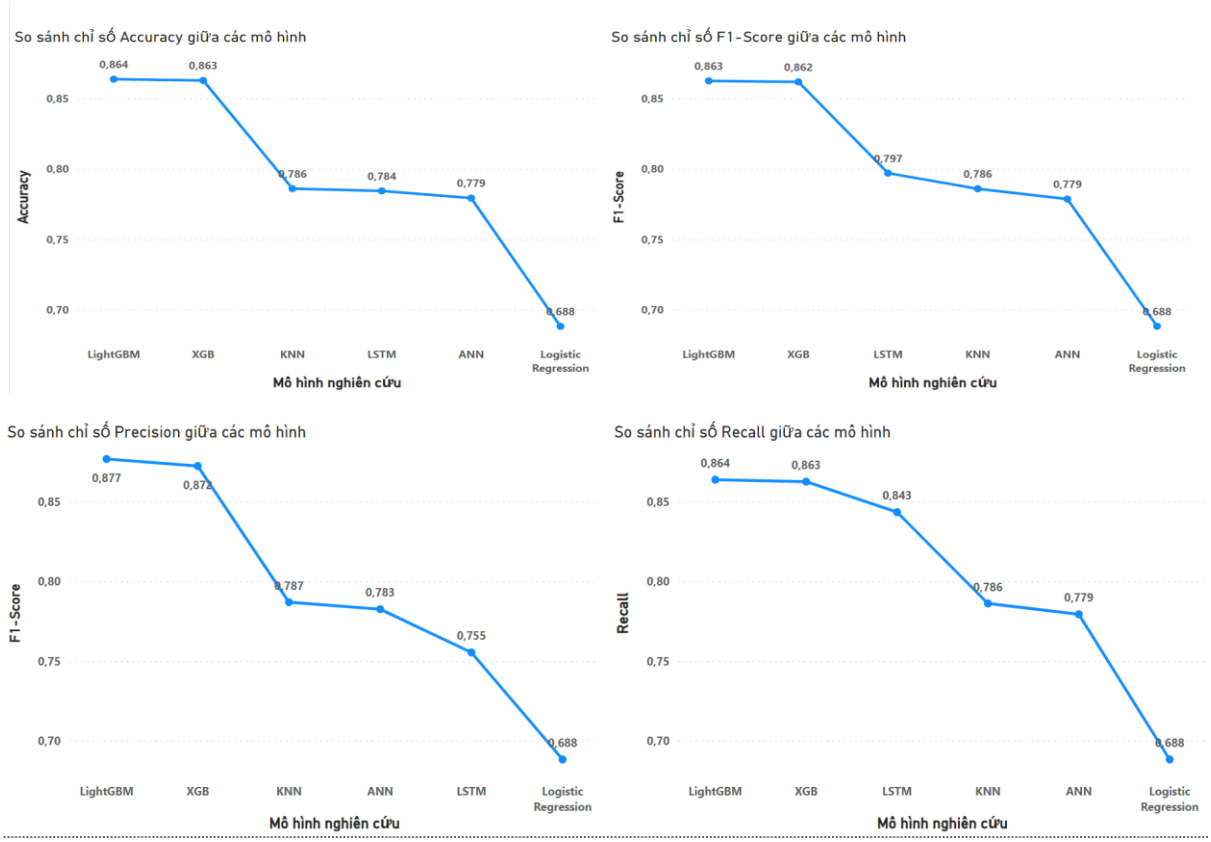
Nhóm đặc trưng quan trọng nhất có thể được xác định bằng cách xem xét những đặc trưng xuất hiện nhiều lần với giá trị "TRUE" cho các mô hình khác nhau. Những đặc trưng này thường mang lại ảnh hưởng lớn đến khả năng dự đoán chính xác của mô hình và có thể là điểm tập trung của các nỗ lực tối ưu hóa tiếp theo. Dựa vào Bảng 3, một số thuộc tính như int_rate, subgrade, term, loan_amnt, dti, installment, mort_acc, revol_util, log_annual_inc được đánh giá là tác động mạnh mẽ khi trong chúng đều có tác động đến mô hình.

Nhóm đặc trưng không quan trọng và nên bị loại bỏ có thể được xác định bằng cách xem xét những đặc trưng có giá trị "FALSE" cho nhiều mô hình. Dựa trên Bảng 3, một số thuộc tính chỉ có tác động đến một, thậm chí không có tác động đến mô hình như log_revol_bal hoặc open_acc cần được loại bỏ. Việc loại bỏ các đặc trưng này có thể giúp đơn giản hóa mô hình và giảm thời gian huấn luyện mà không ảnh hưởng đến hiệu suất của mô hình.

Ngoài ra, các chỉ số hiệu suất như Accuracy, Precision, Recall và F1-Score có vai trò quan trọng để đánh giá và so sánh hiệu suất của các mô hình dự đoán. Dựa trên các kết quả thực nghiệm, với các chỉ số Accuracy, Precision, Recall, F1-Score và AUC ROC của LightGBM lần lượt là 86.37%, 87.672%, 86.37%, 86.25% và 86.06%, đây là mô hình dự báo khả năng hoàn trả khoản vay của khách hàng tốt nhất trên tập dữ liệu của nhóm nghiên cứu. Mô hình Logistic Regression được đánh giá là kém hiệu quả nhất trên tập dữ liệu khi các chỉ số Accuracy, Precision, Recall, F1-Score và AUC ROC đều sấp xỉ 69%. Hình 8 và Hình 9 thể hiện sự so sánh các chỉ số đánh giá giữa các mô hình nghiên cứu. Tổ chức tài chính, gân hàng có thể xem xét các thuật toán nhóm nghiên cứu đề xuất nhằm dự đoán và lập kế hoạch xem xét các khách hàng được phép vay, tránh những mất mát như không có khả năng thanh toán hoặc vỡ nợ của khách hàng, từ đó ảnh hưởng đến lợi nhuận của công ty.

Hình 8

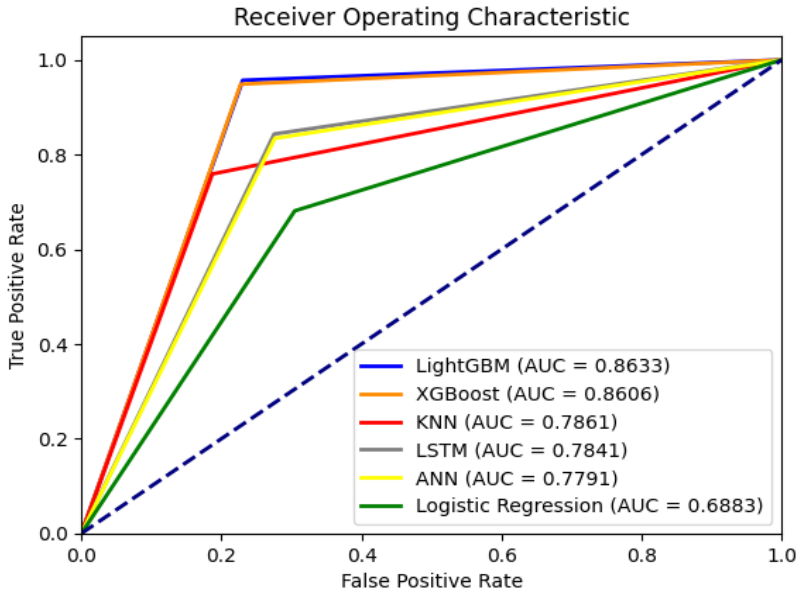
So Sánh Chỉ Số Accuracy, F1-Score, Precision và Recall giữa các Mô Hình



Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

Hình 9

So Sánh Chỉ Số AUC ROC giữa các Mô Hình



Nguồn: Kết quả phân tích dữ liệu của nhóm nghiên cứu

5. Kết quả và hướng phát triển

5.1. Kết luận

Việc áp dụng các kỹ thuật khoa học dữ liệu trong lĩnh vực Tài chính, Ngân hàng đang ngày càng được chú trọng quan tâm. Các doanh nghiệp đặc thù như ngân hàng, công ty tài chính và các tổ chức tín dụng ngày càng chú tâm tới vấn đề quản lý rủi ro từ đó đem lại lợi ích thương mại kinh tế. Các nghiên cứu được thực hiện trong lĩnh vực dự đoán khả năng thanh toán khoản nợ nói riêng và các dự đoán liên quan tín dụng nói chung là cơ hội và thách thức tới các tổ chức doanh nghiệp nhằm đưa ra quyết định chính xác nhất với khách hàng, tránh những rủi ro không đáng có nếu như đưa ra quyết định sai lầm.

Trong nghiên cứu xây dựng mô hình dự đoán khả năng hoàn trả khoản vay của khách hàng, chúng tôi để xuất các phương pháp khác nhau để cải thiện khả năng dự đoán hoàn trả vay của khách hàng cũng như tìm hiểu các yếu tố tác động nhất đến yếu tố thanh toán đó nhằm cải thiện và tránh rủi ro vỡ nợ. Phân tích dữ liệu cũng như cải tiến phương pháp Select K-Best nhằm tự động tìm ra các thuộc tính thích hợp để lựa chọn cho mô hình. Các chỉ số đánh giá như Accuracy, Precision, Recall, F1-score hay AUC ROC cũng được áp dụng nhằm so sánh lựa chọn mô hình phù hợp nhất với tập dữ liệu dự đoán khoản vay khách hàng.

Tuy nhiên, các công ty tổ chức tài chính, ngân hàng có những khắt khe về quy định bảo mật, vì vậy, việc tiếp cận được đặc biệt với các tập dữ liệu trong thời gian gần đây tương đối khó khăn. Ngoài ra, các thuật toán thực nghiệm trong nghiên cứu này còn hạn chế và có thể vẫn còn các thuật toán khác phù hợp hơn với tập dữ liệu.

5.2. Hướng phát triển

Từ bài toán này, các nghiên cứu tiếp theo có thể phối kết hợp các phương pháp máy học truyền thống, học sâu cùng các mô hình kết hợp (ensemble learning) nhằm cải thiện độ chính xác, từ đó các ngân hàng, tổ chức tín dụng có thể xây dựng một nền tảng ứng dụng để triển khai sử dụng thực tế.

Tài liệu tham khảo

Abdolrasol, M. G., Hussain, S. M., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef S., & Milad, A. (2021). Artificial neural networks based optimization techniques: A review. *Electronics, 10*(21), Article 2689.

Aksan, F., Li, Y., Suresh, V., & Janik, P. (2023). CNN-LSTM vs. LSTM-CNN to predict power flow direction: A case study of the high-voltage subnet of Northeast Germany. *Sensors*, *23*(2), Article 901.

Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, *13*(1), 6-10.

Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., & Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation. *Applied Sciences*, *12*(4), Article 2126.

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., & Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, *16*(2), Article 1.

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics, 47*(13/15), 2879-2894.

Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*(6), 914-920.

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). ACM.

Chính phủ Việt Nam. (2010). *Law No. 47/2010/QH12 by the National Assembly: Law on credit institutions.* https://vanban.chinhphu.vn/default.aspx?pageid=27160&docid=96074

Daoud, M., & Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artificial Intelligence in Medicine*, *97*, 204-214.

Desyani, T., Saifudin, A., & Yulianti, Y. (2020). Feature selection based on naive bayes for caesarean section prediction. *IOP Conference Series: Materials Science and Engineering*, *879*(1), Article 012091.

George, N. (2021). *All lending club loan data.* https://www.kaggle.com/datasets/wordsforthewise/lending-club/data

Graves, A. (2012). Long short-term memory. In A. Graves (Ed.), *Supervised sequence labelling with recurrent neural networks* (pp. 37-45). Springer.

Halagundegowda, G. R., Abhishek, S., Mohan, K. T., & Naveena, K. (2023). Evaluation of classification ability of Support Vector Machine (SVM) in binary classification problems. *Training*, *8*(5), 7-13.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy*, *24*(6), Article 801.

Kramer, O. (2013). Dimensionality reduction with unsupervised nearest neighbors. In J. Kacprzy & K. C. Jain (Eds.), *Intelligent systems reference library* (Vol. 51, pp. 13-23). Springer.

Laaksonen, J., & Oja, E. (1996). Classification with learning k-nearest neighbors. *Proceedings of International Conference on Neural Networks (ICNN'96), 3*, 1480-1483. IEEE.

Li, Z., Li, S., Li, Z., Hu, Y., & Gao, H. (2021). Application of XGBoost in P2P default prediction. *Journal of Physics: Conference Series, 1871*(1), Article 012115.

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics, 8*(5), Article 765.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). K-nearest neighbor classification. In A. P. Mucherino (Ed.), *Data mining in agriculture* (pp. 83-106). Springer.

Peterson, K., Söderström, C., Kiani-Anaraki, M., & Levy, G. (1999). Evaluation of the ability of thermal and electrical tests to register pulp vitality. *Dental Traumatology, 15*(3), 127-131.

Phan, A. D., & Nguyen, N. H. (2013). Combining option approach with logistic regression analysis to measure default risk of listed companies on Vietnamese stock market. *Journal of Economic Development, 217*(2013), 92-109.

Singh, M. (2023). *Financial crisis in review*. https://www.investopedia.com/articles/economics/09/financial-crisis-review.asp

Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access, 8*(1), 25579-25587.

Van der Schouw, Y. T., Verbeek, A. L., & Ruijs, J. H. (1992). ROC curves for the initial assessment of new diagnostic tests. *Family Practice*, 9(4), 506-511.

Walczak, S. (2019). Artificial neural networks. In D. Mehdi Khosrow-Pour (Ed.), *Advanced methodologies and technologies in artificial intelligence, computer simulation, and human-computer interaction* (pp. 40-53). IGI Global.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol 1, pp. 29-39). The Practical Application Company.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications, 78*, 225-241.

Zhang, D., & Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. *IEEE Access, 8*, 220990-221003.

Zhao, S., & Zou, J. (2021). Predicting loan defaults using logistic regression. *Journal of Student Research, 10*(1), 1-14.