Đặng Xuân Thọ (2024). Nghiên cứu dự đoán phá sản của doanh nghiệp sử dụng kỹ thuật học máy và SMOTEWB. *Tạp chí nghiên cứu Chính sách và Phát triển*, 02(2024), 70-80

Tạp chí Nghiên cứu Chính sách và Phát triển

© Học viện Chính sách và Phát triển 2024
© CSR,2024

Bài báo khoa học

Nghiên cứu dự đoán phá sản của doanh nghiệp sử dụng kỹ thuật học máy và SMOTEWB

Đặng Xuân Thọ (TS)
Khoa Kinh tế số, Học viện Chính sách và Phát triển
Email: thodx@apd.edu.vn

Tóm tắt:

Trong bối cảnh kinh tế hiện đại, khả năng dự đoán phá sản của doanh nghiệp ngày càng trở nên quan trọng, đóng vai trò then chốt trong việc hỗ trợ các nhà quản lý, nhà đầu tư đưa ra quyết định nhằm giảm thiểu rủi ro tài chính. Để đáp ứng nhu cầu đó, nghiên cứu này đề xuất một phương pháp mới, sử dụng kết hợp các kỹ thuật học máy và chiến lược cân bằng dữ liệu. Mục tiêu chính là tăng cường độ chính xác trong dự đoán phá sản doanh nghiệp. Nghiên cứu được tiến hành qua các bước bao gồm tiền xử lý dữ liệu và phát triển các mô hình phân lớp dữ liệu, với sự tập trung đặc biệt vào việc tích hợp mô hình học máy cùng với phương pháp cân bằng dữ liệu SMOTEWB. Kết quả thực nghiệm cho thấy mô hình đề xuất không chỉ đạt được độ chính xác cao mà còn có tiềm năng ứng dụng rộng rãi trong thực tiễn.

Ngày nhận bài: 31/7/2024
Bản sửa lại lần 1: 05/9/2024
Ngày duyệt bài: 15/9/2024

Mã số: TC070224

Từ khóa: Dự báo phá sản, Học máy, Dữ liệu mất cân bằng, SMOTEWB

Abstract:

In the context of modern economy, the ability to predict corporate bankruptcy is becoming increasingly important, playing a key role in supporting managers and investors in making decisions to minimize financial risks. To meet this need, the present study proposed a novel method, using a combination of machine learning techniques and data balancing strategies. The main goal is to improve the accuracy of corporate bankruptcy prediction. The study was conducted through steps including thorough data preprocessing and developing classification models, with a special focus on integrating machine learning models with the SMOTEWB data balancing method. The experimental results show that the proposed model not only achieves high accuracy but also has the potential for wide application in practice

Keywords: Bankruptcy Prediction, Machine Learning, Imbalanced Data, SMOTEWB

1. Giới thiệu

Dự đoán khả năng phá sản của doanh nghiệp, thường được gọi là dự đoán phá sản, là một chủ đề có tầm quan trọng đặc biệt trong lĩnh vực tài chính và kế toán. Sự ổn định tài chính của một doanh nghiệp không chỉ ảnh hưởng đến các khoản nợ, mà còn có tác động lớn đến nhà đầu tư, cổ đông, đối tác kinh doanh, khách hàng và nhà cung cấp. Do đó, các phương pháp dự đoán phá sản chính xác và nhanh chóng đã thu hút sự quan tâm của nhiều nhà nghiên cứu (Altman, 1968).

Nghiên cứu về dự đoán phá sản đã bắt đầu từ gần 50 năm trước, khi các phương pháp học máy thống kê được sử dụng để dự đoán khả năng phá sản của doanh nghiệp. Từ thập niên 1990, các mô hình học máy đã trở thành công cụ phổ biến trong dự đoán phá sản, bao gồm các thuật toán như cây quyết định, mạng nơ-ron và máy vector hỗ trợ (Lin, Hu, & Tsai, 2011; Atiya, 2001). Cũng giống như bài toán chấm điểm tín dụng, dự đoán phá sản thường được xử lý như một bài toán phân loại nhị phân, trong đó nhiệm vụ chính là dự đoán liệu một doanh nghiệp có thể phá sản hay không. Để đạt được độ chính xác cao trong dự đoán, các nhà nghiên cứu thường huấn luyện các mô hình bằng các tập dữ liệu tài chính được lấy từ báo cáo tài chính của doanh nghiệp. Beaver (1966) là người tiên phong trong việc sử dụng dữ liệu tài chính để nghiên cứu dự đoán phá sản. Quá trình huấn luyện này là nền tảng của việc áp dụng kỹ thuật học máy trong dự đoán phá sản.

Trong những năm gần đây, trí tuệ nhân tạo đã nổi lên như một công cụ mạnh mẽ với nhiều ứng dụng trong các lĩnh vực khác nhau. Các kỹ thuật học máy đã đạt được thành công lớn trong các lĩnh vực như lái xe tự động, thị giác máy tính, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, cũng như trong các bài toán phân loại trong kinh doanh và quản lý, bao gồm cả dự đoán phá sản và chấm điểm tín dụng. Trong bài viết này, chúng tôi sẽ trình bày việc áp dụng các kỹ thuật học máy nhằm cải thiện hiệu quả dự đoán phá sản.

Một trong những thách thức lớn mà chúng tôi nhận thấy trong nghiên cứu này là sự mất cân bằng dữ liệu trong các bộ dữ liệu dự đoán phá sản. Cụ thể, dữ liệu về các doanh nghiệp phá sản thường chiếm tỷ lệ rất nhỏ trong toàn bộ tập dữ liệu, gọi là dữ liệu lớp thiểu số, trong khi số lượng doanh nghiệp không phá sản chiếm đa số và được gọi là dữ liệu lớp đa số. Khi áp dụng các thuật toán học máy tiêu chuẩn vào dữ liệu mất cân bằng này, mô hình thường có xu hướng dự đoán chính xác cho lớp đa số (doanh nghiệp không phá sản) nhưng kém hiệu quả với lớp thiểu số (doanh nghiệp phá sản). Điều này dẫn đến việc nhiều doanh nghiệp thực sự có nguy cơ phá sản bị dự đoán nhầm là không phá sản, gây ra những hậu quả nghiêm trọng. Do đó, nghiên cứu và giải quyết vấn đề mất cân bằng dữ liệu trong dự đoán phá sản là điều cực kỳ quan trọng. Thách thức này ngày càng phổ biến trong nhiều lĩnh vực, đặc biệt là trong dự đoán phá sản. Tuy nhiên, các thuật toán học máy truyền

thông thường không hiệu quả trong việc xử lý vấn đề này, và vẫn còn ít nghiên cứu tập trung vào giải quyết sự mất cân bằng dữ liệu trong bài toán dự đoán phá sản.

Những đóng góp chính của nghiên cứu này bao gồm: (a) Giới thiệu dự đoán phá sản như một vấn đề nghiên cứu quan trọng trong lĩnh vực tài chính, đồng thời phân tích các phương pháp giải quyết hiện tại từ các nhà nghiên cứu khác. (b) Phân tích những khó khăn và thách thức trong việc xử lý sự mất cân bằng dữ liệu trong dự đoán phá sản, cùng với các giải pháp tiềm năng để khắc phục. (c) Đề xuất việc kết hợp các kỹ thuật trí tuệ nhân tạo, học máy tiên tiến với các phương pháp giải quyết bài toán dự đoán phá sản.

Các phần tiếp theo của bài báo này được tổ chức như sau: Phần 2 giới thiệu một số nghiên cứu hiện có, đánh giá ưu và nhược điểm của các phương pháp trước đó. Phương pháp tiếp cận đề xuất của chúng tôi được mô tả chi tiết trong Phần 3, bao gồm các thành phần và quy trình triển khai thực nghiệm. Phần 4 trình bày kết quả thực nghiệm và phân tích hiệu suất của phương pháp đề xuất. Cuối cùng, Phần 5 tóm tắt các phát hiện chính, đưa ra kết luận từ kết quả nghiên cứu và đề xuất hướng phát triển trong tương lai.

2. Tổng quan nghiên cứu

Dự đoán khả năng phá sản của doanh nghiệp là một chủ đề nghiên cứu có tầm quan trọng đặc biệt trong lĩnh vực tài chính và kế toán. Việc dự đoán chính xác khả năng phá sản không chỉ có ý nghĩa lớn đối với các doanh nghiệp, mà còn ảnh hưởng sâu sắc đến các bên liên quan như nhà đầu tư, cổ đông, đối tác kinh doanh, khách hàng và nhà cung cấp. Trong bối cảnh đó, nhiều phương pháp đã được phát triển và áp dụng để nâng cao hiệu quả dự đoán, cụ thể như một số nghiên cứu điển hình sau.

Phương pháp Phân tích phân biệt đa biến (MDA) là một trong những phương pháp dự đoán phá sản sớm nhất, được Altman (1968) áp dụng để phân loại các doanh nghiệp dựa trên khả năng thanh toán và mất khả năng thanh toán, sử dụng dữ liệu báo cáo tài chính của doanh nghiệp. Altman đã sử dụng năm tỷ số tài chính quan trọng làm đầu vào, bao gồm Vốn lưu động/Tổng tài sản và Thu nhập trước lãi vay và thuế (EBIT)/Tổng tài sản. Các tỷ số này đã trở thành tiêu chuẩn và được sử dụng rộng rãi trong nhiều nghiên cứu sau này. Phương pháp MDA cho phép phân loại doanh nghiệp thành hai nhóm: có khả năng phá sản và không có khả năng phá sản. Mặc dù MDA đã tạo ra một nền tảng vững chắc cho nghiên cứu về dự đoán phá sản, nhưng phương pháp này cũng gặp phải một số hạn chế, đặc biệt khi dữ liệu không tuân theo giả định phân phối chuẩn.

Hồi quy logistic (LR), được giới thiệu vào nghiên cứu dự đoán phá sản bởi Ohlson (1980), là một phương pháp khác được sử dụng rộng rãi. Khác với MDA, LR sử dụng hàm sigmoid để thực hiện phân loại, trong đó đầu ra là xác suất phá sản, thuộc khoảng từ 0 đến 1. LR cho phép các nhà nghiên cứu không chỉ phân loại doanh nghiệp mà còn ước tính xác suất phá sản của chúng, mang lại khả năng giải

thích xác suất cao hơn so với phương pháp MDA. Một trong những lợi thế lớn của LR là khả năng xử lý các mối quan hệ phi tuyến tính giữa các biến đầu vào và đầu ra thông qua việc sử dụng các biến giả định (dummy variables) hoặc biến tương tác. Tuy nhiên, LR cũng không tránh khỏi các hạn chế khi phải đối mặt với các tập dữ liệu mất cân bằng, nơi số lượng doanh nghiệp phá sản thường ít hơn nhiều so với số lượng doanh nghiệp không phá sản.

Trong những năm gần đây, các phương pháp dựa trên học máy và trí tuệ nhân tạo đã được áp dụng rộng rãi trong lĩnh vực dự đoán phá sản, trong đó có thể kể đến các phương pháp sử dụng tập thô (Beynon & Peel, 2001; McKee, 2003), lý luận dựa trên trường hợp (Li & Sun, 2010, 2013), và đặc biệt là SVM (Lin, Yeh, & Lee, 2011; Li & Sun, 2012). Các phương pháp này đã chứng minh được hiệu quả cao trong việc phân loại và dự đoán khả năng phá sản của doanh nghiệp. SVM, chẳng hạn, là một phương pháp học máy mạnh mẽ, đặc biệt hiệu quả trong việc phân loại dữ liệu có không gian tính toán cao và phức tạp. Tuy nhiên, SVM cũng gặp phải những thách thức khi xử lý các tập dữ liệu lớn hoặc mất cân bằng, đòi hỏi phải có các kỹ thuật điều chỉnh đặc biệt để cải thiện độ chính xác của mô hình.

Một trong những tiến bộ quan trọng trong lĩnh vực học máy là sự ra đời của Mạng thần kinh nhân tạo (NN), một mô hình được thiết kế để mô phỏng quá trình xử lý thần kinh trong não người. Zhao và các cộng sự (2015) đã phát triển một hệ thống chấm điểm tín dụng tự động bằng cách sử dụng Mạng nơ-ron đa lớp (MLPNN), đạt được độ chính xác cao (87%) trên dữ liệu tín dụng của Đức. Mạng nơ-ron có khả năng học từ dữ liệu phi tuyến và có thể xử lý các mối quan hệ phức tạp giữa các biến. Tsai và Wu (2008) đã chứng minh rằng việc kết hợp nhiều mô hình NN đơn lẻ thành một mô hình tổng hợp có thể cải thiện hiệu suất phân loại so với việc sử dụng một mô hình đơn lẻ. Điều này mở ra nhiều cơ hội cho việc áp dụng các phương pháp học sâu trong dự đoán phá sản, mặc dù còn nhiều thách thức cần phải giải quyết, đặc biệt là vấn đề mất cân bằng dữ liệu.

Từ năm 2000 đến nay, số lượng bài báo liên quan đến dự đoán phá sản trong kinh doanh đã tăng lên đáng kể, điều này phản ánh sự quan tâm ngày càng lớn đến chủ đề này trong cộng đồng nghiên cứu. Tuy nhiên, hiện nay vẫn còn nhiều hạn chế trong việc áp dụng các công cụ mạnh mẽ của học máy vào lĩnh vực này. Một vấn đề nổi cộm là sự mất cân bằng dữ liệu, khi số lượng doanh nghiệp phá sản thường chiếm tỷ lệ rất nhỏ so với tổng số doanh nghiệp. Điều này dẫn đến việc các mô hình học máy có xu hướng dự đoán chính xác hơn cho lớp đa số (doanh nghiệp không phá sản) và ít chính xác hơn cho lớp thiểu số (doanh nghiệp phá sản). Việc không giải quyết hiệu quả vấn đề này có thể dẫn đến sai sót nghiêm trọng trong việc dự đoán phá sản, gây ra hậu quả nặng nề về mặt tài chính và kinh tế.

Trong phần tiếp theo của bài báo,

chúng tôi sẽ trình bày chi tiết về phương pháp đề xuất nhằm nâng cao hiệu quả dự đoán phá sản, đặc biệt tập trung vào việc giải quyết vấn đề mất cân bằng dữ liệu. Chúng tôi sẽ khám phá cách kết hợp các kỹ thuật học máy tiên tiến với các phương pháp điều chỉnh dữ liệu để cải thiện độ chính xác của mô hình dự đoán, đồng thời đề xuất các hướng nghiên cứu tiếp theo nhằm phát triển các giải pháp toàn diện hơn trong lĩnh vực này.

3. Phương pháp nghiên cứu

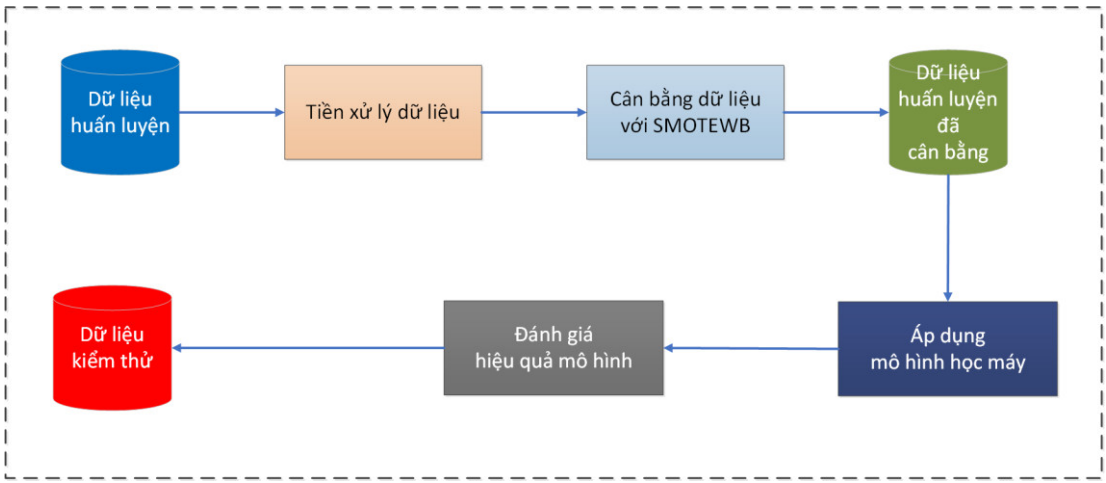
Trong nghiên cứu này, chúng tôi đề xuất một quy trình thực nghiệm toàn diện gồm bốn bước để dự đoán phá sản doanh nghiệp. Mục tiêu của quy trình này là xây dựng và đánh giá các mô hình dự đoán chính xác, đồng thời giải quyết các thách thức liên quan đến chất lượng dữ liệu và mất cân bằng dữ liệu. Các bước được đề xuất trong quy trình thực nghiệm bao gồm:

Bước 1: Thu thập và tiền xử lý dữ liệu

Bước đầu tiên là thu thập các dữ liệu tài chính doanh nghiệp, sau đó làm sạch dữ liệu bằng cách loại bỏ các giá trị thiếu, ngoại lai và dư thừa. Mục tiêu là tạo ra tập dữ liệu sạch và nhất quán để mô hình học máy hoạt động hiệu quả.

Bước 2: Giải quyết vấn đề mất cân bằng dữ liệu

Doanh nghiệp phá sản ít hơn nhiều so với không phá sản, dẫn đến mất cân bằng dữ liệu. Chúng tôi sử dụng phương pháp SMOTEWB để tạo thêm mẫu d ữ liệu cho lớp thiểu số, giúp cân bằng dữ liệu và cải thiện độ chính xác của mô hình.



Hình 1: Quy trình thực nghiệm dự đoán phá sản

Bước 3: Xây dựng mô hình học máy

Sau khi cân bằng dữ liệu, chúng tôi áp dụng các thuật toán học máy như NN, SVM, LR và các phương pháp tích hợp để xây dựng mô hình dự đoán phá sản, với mục tiêu tìm ra mô hình tối ưu nhất.

Bước 4: Kiểm thử và đánh giá hiệu quả mô hình

Cuối cùng, chúng tôi kiểm thử và đánh giá mô hình bằng cách sử dụng dữ liệu kiểm thử để đo lường độ chính xác, độ nhạy, và các chỉ số hiệu suất khác, nhằm xác định tính khả thi của mô hình trong thực tế và điều chỉnh nếu cần.

Quy trình thực nghiệm này được minh họa chi tiết trong Hình 1, thể hiện rõ các bước từ tiền xử lý dữ liệu đến việc đánh giá mô hình. Việc áp dụng quy trình này không chỉ giúp tối ưu hóa hiệu quả dự đoán mà còn góp phần giải quyết những thách thức phức tạp trong việc xử lý dữ liệu mất cân bằng, một vấn đề thường gặp trong dự đoán phá sản.

Thuật toán SMOTEWB (Synthetic Minority Over-sampling Technique with Boosting), được đề xuất bởi Sağlam và Cengiz (2022), là một kỹ thuật mới nhằm cân bằng dữ liệu, đặc biệt hữu ích trong các tình huống mà dữ liệu bị mất cân bằng nghiêm trọng. SMOTEWB mở rộng phương pháp SMOTE truyền thống bằng cách thêm các quy trình phát hiện nhiễu và tăng cường hiệu suất thông qua boosting. Quá trình này bao gồm ba bước cụ thể như sau:

Bước 1. Phát hiện nhiễu: Trước khi áp dụng SMOTE, dữ liệu được kiểm tra để xác định và loại bỏ các điểm dữ liệu gây nhiễu. Điều này giúp đảm bảo rằng các mẫu tổng hợp được tạo ra không bị ảnh hưởng bởi các giá trị bất thường hoặc không đại diện.

Bước 2. Tạo mẫu tổng hợp với SMOTE: Sau khi loại bỏ nhiễu, SMOTE được áp dụng để tạo ra các mẫu tổng hợp cho lớp thiểu số. Quá trình này sử dụng các kỹ thuật như k-nearest neighbors (k-NN) để xác định các hàng xóm gần nhất của từng điểm dữ liệu trong lớp thiểu số, từ đó tạo ra các điểm dữ liệu mới giữa các hàng xóm này.

Bước 3. Tăng cường hiệu suất với Boosting: Cuối cùng, SMOTEWB tích hợp một quy trình boosting, giúp tăng cường độ chính xác của mô hình dự đoán. Boosting là một kỹ thuật học máy nâng cao, hoạt động bằng cách kết hợp nhiều mô hình yếu thành một mô hình mạnh, tối ưu hóa dự đoán của mô hình tổng hợp này dựa trên các trọng số điều chỉnh.

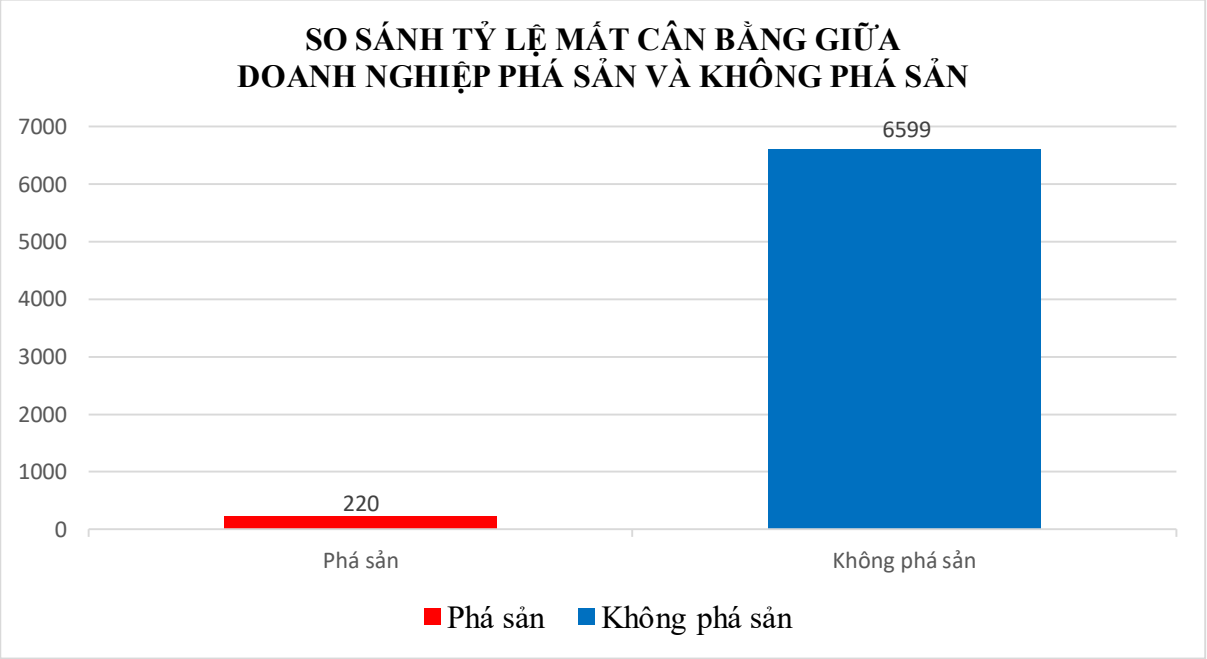
Kết hợp ba bước này, SMOTEWB không chỉ cải thiện khả năng xử lý các tập dữ liệu mất cân bằng mà còn nâng cao độ chính xác tổng thể của mô hình bằng cách giảm thiểu tác động của nhiễu và tối ưu hóa quá trình học máy. Phương pháp này được chứng minh là vượt trội so với các kỹ thuật cân bằng dữ liệu truyền thống, đặc biệt trong các tình huống dữ liệu có tính phức tạp cao.

Dữ liệu

Trong nghiên cứu này, chúng tôi đã tiến hành thu thập và phân tích dữ liệu từ Tạp chí Kinh tế Đài Loan, bao gồm khoảng thời gian từ năm 1999 đến 2009, như được ghi nhận trong nguồn dữ liệu của Taiwanese Bankruptcy Prediction (2020). Quá trình xác định tình trạng phá sản của các công ty được thực hiện dựa trên các quy định và tiêu chuẩn kinh doanh được đề ra bởi Sở Giao dịch Chứng khoán Đài Loan, nhằm đảm bảo tính chính xác và đồng nhất của kết quả nghiên cứu.

Bộ dữ liệu được sử dụng trong nghiên cứu bao gồm tổng cộng 6819 mẫu dữ liệu, với 96 thuộc tính khác nhau, đại diện cho các khía cạnh tài chính và hoạt động của doanh nghiệp. Trong số các thuộc tính

này, thuộc tính "Phá sản?" đóng vai trò là nhãn lớp, chứa thông tin cụ thể về việc doanh nghiệp có rơi vào tình trạng phá sản hay không, tạo cơ sở để thực hiện các phân tích sâu hơn về nguy cơ phá sản.



Hình 2: So sánh tỷ lệ mất cân bằng giữa số lượng doanh nghiệp phá sản và doanh nghiệp không phá sản

Qua quá trình phân tích dữ liệu, chúng tôi phát hiện rằng tỷ lệ mất cân bằng giữa các doanh nghiệp phá sản và không phá sản là rất đáng kể. Cụ thể, số lượng doanh nghiệp phá sản so với doanh nghiệp không phá sản theo thuộc tính "Phá sản?" được ghi nhận là 1:29.99. Điều này cho thấy sự chênh lệch lớn trong dữ liệu, tạo ra thách thức đối với các phương pháp phân tích và dự đoán, đồng thời yêu cầu các kỹ thuật xử lý dữ liệu phù hợp để đảm bảo tính hiệu quả và độ chính xác của kết quả nghiên cứu.

Ngoài thuộc tính "Phá sản?" dùng để xác định tình trạng phá sản của doanh nghiệp, bộ dữ liệu còn bao gồm 95 thuộc tính khác cung cấp thông tin chi tiết về từng doanh nghiệp. Các thuộc tính này thể hiện nhiều khía cạnh tài chính và hoạt động kinh doanh khác nhau, cụ thể bao gồm: chi phí nợ chịu lãi, tỷ lệ tái đầu tư tiền mặt, tỷ lệ chi phí lãi vay trên tổng doanh thu, tỷ lệ tổng nợ trên vốn chủ sở hữu, tỷ lệ nợ trên vốn chủ sở hữu chịu lãi, thu nhập trước thuế trên vốn, tỷ lệ vốn lưu động trên tổng tài sản, tỷ lệ tài sản nhanh trên tổng tài sản, tỷ lệ tiền mặt trên tổng tài sản, cùng với nhiều chỉ số tài chính khác.

Chi tiết về từng thuộc tính và cách chúng được tính toán, cùng với các ví dụ minh họa cụ thể, được trình bày chi tiết trong tài liệu tham khảo của nghiên cứu (Taiwanese Bankruptcy Prediction, 2020). Việc nắm vững và hiểu rõ các thuộc tính này là bước đầu tiên để thực hiện các phân tích chuyên sâu và áp dụng các phương pháp học máy nhằm dự đoán chính xác

tình trạng phá sản của các doanh nghiệp.

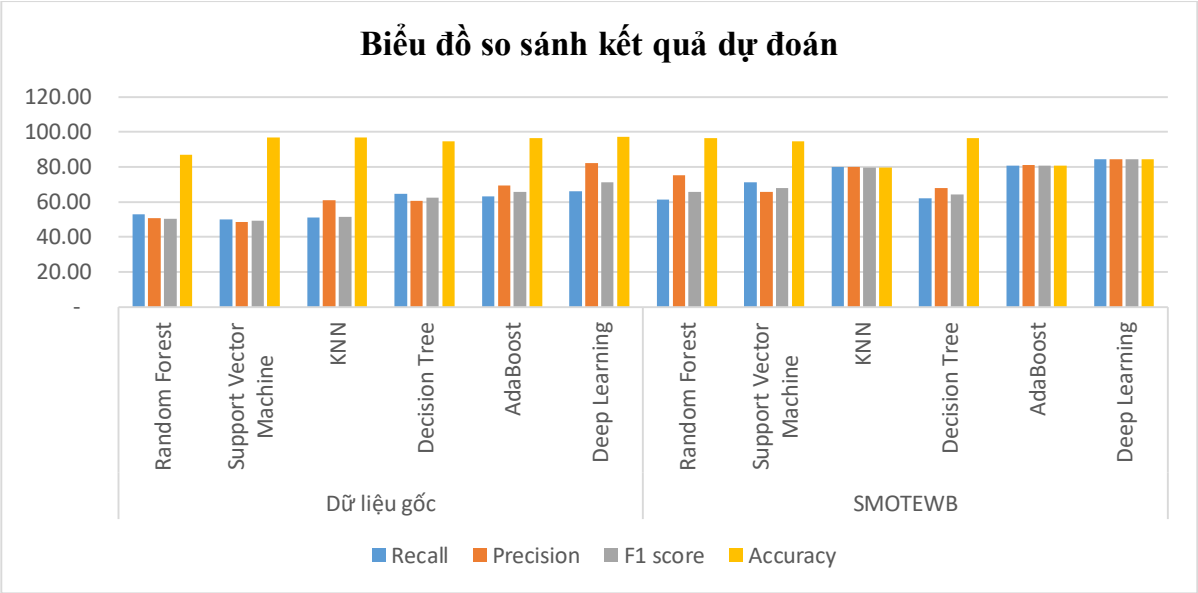
4. Kết quả nghiên cứu và thảo luận

Trong nghiên cứu này, chúng tôi tập trung vào việc so sánh hiệu suất giữa các mô hình học máy khác nhau, bao gồm: Random Forest (RF; Rigatti, 2017), Support Vector Machine (SVM; Pisner & Schnyer, 2020), K Nearest Neighbors (KNN; Zhang, 2016), Decision Tree (Kotsiantis, 2013), AdaBoost (Ying, Qi-Guang, Jia-Chen, & Lin, 2013) và Deep Learning (LeCun, Bengio, & Hinton, 2015; Shinde & Shah, 2018). Các mô hình này được chọn vì chúng đại diện cho những phương pháp phổ biến và hiệu quả trong việc dự đoán và phân loại trong lĩnh vực tài chính.

Dữ liệu thu thập được đã trải qua quá trình tiền xử lý kỹ lưỡng để loại bỏ các yếu tố gây nhiễu và chuẩn hóa các thuộc tính nhằm đảm bảo tính nhất quán. Sau đó, dữ liệu được chia nhỏ và sử dụng phương pháp kiểm định chéo 10-fold để tạo ra các tập dữ liệu bao gồm tập huấn luyện (train) và tập kiểm tra (test). Phương pháp kiểm định chéo này giúp đảm bảo rằng mỗi phần của dữ liệu đều được sử dụng để đánh giá mô hình, giảm thiểu nguy cơ quá khớp và cung cấp một cái nhìn chính xác về hiệu suất của các mô hình.

Bảng 1. So sánh kết quả dự đoán giữa các phương pháp

	Phương pháp	Recall	Precision	F1 score	Accuracy
Dữ liệu gốc	Random Forest	52.80	50.88	50.18	87.02
	Support Vector Machine	50.00	48.42	49.20	96.85
	KNN	51.05	60.96	51.29	96.70
	Decision Tree	64.58	60.75	62.32	94.57
	AdaBoost	63.31	69.53	65.76	96.48
	Deep Learning	66.01	82.25	71.20	**97.36**
SMOTEWB	Random Forest	61.48	75.37	65.54	96.55
	Support Vector Machine	71.27	65.80	68.06	94.77
	KNN	80.09	79.90	79.68	79.69
	Decision Tree	62.11	67.77	64.34	96.33
	AdaBoost	80.90	80.95	80.90	80.91
	Deep Learning	**84.42**	**84.55**	**84.44**	84.47



Hình 3. Biểu đồ so sánh kết quả dự đoán giữa các phương pháp

Bằng cách áp dụng phương pháp kiểm định chéo và so sánh kết quả dự đoán trên các tập dữ liệu kiểm tra, chúng tôi đã có thể đánh giá chi tiết hiệu suất của từng mô hình. Các chỉ số như độ chính xác dự đoán, độ nhạy, độ đặc hiệu và các chỉ số khác đã được tính toán để hiểu rõ hơn về khả năng phân loại và dự đoán của mỗi mô hình trong bối cảnh bài toán phá sản doanh nghiệp. Phân tích này không chỉ giúp xác định mô hình nào hoạt động tốt nhất mà còn cung cấp thông tin quan trọng về cách thức các mô hình xử lý dữ liệu phức tạp và mất cân bằng, từ đó hỗ trợ việc đưa ra các quyết định hợp lý trong thực tiễn.

Bảng kết quả trên cung cấp một cái nhìn tổng quan về hiệu suất của các mô hình khác nhau khi được áp dụng trên hai bộ dữ liệu: dữ liệu gốc và dữ liệu đã qua xử lý SMOTEWB. Khi phân tích hiệu suất trên dữ liệu gốc, có thể thấy rằng mô hình Random Forest đạt độ chính xác (Accuracy) là 87.02%, nhưng các chỉ số khác như Recall, Precision, và F1 Score đều ở mức tương đối thấp, đặc biệt là F1 Score chỉ đạt 50.18%. Điều này cho thấy mô hình có thể chưa tối ưu trong việc nhận diện các mẫu khó phân loại.

Mô hình Support Vector Machine (SVM), mặc dù có độ chính xác rất cao (96.85%), lại có các chỉ số Recall và Precision đều dưới 50%. Điều này cho thấy mặc dù SVM có thể dự đoán chính xác một số lượng lớn các mẫu, nhưng nó có thể gặp khó khăn trong việc cân bằng giữa việc phát hiện đúng các mẫu thuộc các lớp khác nhau. Ngược lại, KNN trên dữ liệu gốc lại có Precision cao nhất (60.96%), nhưng với Recall và F1 Score không quá cao, điều này cho thấy mô hình có xu hướng thiên về việc dự đoán đúng các mẫu dương tính, nhưng không ổn định trong việc phân loại đúng toàn bộ dữ liệu.

Decision Tree và AdaBoost cho thấy hiệu suất tương đối tốt trên dữ liệu gốc. Decision Tree đạt được Recall cao nhất (64.58%), trong khi AdaBoost có F1 Score cao nhất (65.76%), cho thấy sự cân

bằng tốt giữa Precision và Recall. Tuy nhiên, mô hình Deep Learning vượt trội hơn so với các phương pháp khác trên dữ liệu gốc, với các chỉ số hiệu suất cao nhất, đặc biệt là F1 Score (71.20%) và Accuracy (97.36%). Điều này cho thấy mô hình này có khả năng mạnh mẽ trong việc xử lý dữ liệu và đưa ra dự đoán chính xác.

Khi dữ liệu được xử lý qua SMOTEWB, các mô hình đều có sự cải thiện đáng kể về hiệu suất. Random Forest có sự gia tăng đáng kể về Recall và Precision, dẫn đến F1 Score tăng lên 65.54%. SVM cũng cải thiện đáng kể, với F1 Score đạt 68.06%. KNN có sự cải thiện mạnh mẽ nhất về tất cả các chỉ số, với F1 Score lên đến 79.68%, cho thấy mô hình này hoạt động rất hiệu quả trên dữ liệu cân bằng.

AdaBoost và Deep Learning tiếp tục thể hiện sự vượt trội khi sử dụng dữ liệu qua xử lý SMOTEWB. AdaBoost đạt F1 Score cao nhất (80.90%) trong các mô hình, trong khi Deep Learning không chỉ có F1 Score cao nhất (84.44%) mà còn đạt độ chính xác cao nhất (84.47%). Những kết quả này cho thấy rằng việc sử dụng SMOTEWB giúp cải thiện đáng kể hiệu suất của các mô hình, đặc biệt là trong việc xử lý dữ liệu mất cân bằng, từ đó cung cấp các dự đoán chính xác hơn trong các tình huống thực tế.

5. Kết luận

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp tiên tiến kết hợp các kỹ thuật học máy và trí tuệ nhân tạo với chiến lược cân bằng dữ liệu nhằm cải thiện đáng kể độ chính xác trong dự đoán phá sản doanh nghiệp. Quá trình nghiên cứu bao gồm thu thập, tiền xử lý dữ liệu và phát triển mô hình dựa trên nhiều thuật toán phân loại kết hợp với phương pháp SMOTEWB. Kết quả thực nghiệm cho thấy mô hình không chỉ đạt độ chính xác cao mà còn phù hợp để áp dụng trong thực tế. Ngoài ra, việc mở rộng và tinh chỉnh các biến đầu vào, cũng như tối ưu hóa tham số của mô hình có thể nâng cao hiệu quả hơn nữa. Việc so sánh mô hình này với các phương pháp tiên tiến khác sẽ được nghiên cứu thêm để đánh giá toàn diện hiệu quả của phương pháp đề xuất.

TÀI LIỆU THAM KHẢO

1. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609.

2. Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks, 12*(4), 929-935.

3. Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71-111.

4. Beynon, M. J., & Peel, M. J. (2001). Variable precision rough set theory and data discretisation: An application to corporate failure prediction. *Omega, 29*(6), 561-576.

5. Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261-283.

6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

7. Li, H., & Sun, J. (2010). Forecasting business failure in China using case-based

reasoning with hybrid case representation. *Journal of Forecasting, 29*(5), 486-501.

8. Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples–Evidence from the Chinese hotel industry. *Tourism Management, 33*(3), 622-634.

9. Li, H., & Sun, J. (2013). Predicting business failure using an RSF-based case-based reasoning ensemble forecasting method. *Journal of Forecasting, 32*(2), 180-192.

10. Lin, F., Yeh, C. C., & Lee, M. Y. (2011). The use of hybrid manifold learning and support vector machines in the prediction of business failure. *Knowledge-Based Systems, 24*(1), 95-101.

11. Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2011). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 421-436.

12. Manju, B. R., & Nair, A. R. (2019, December). Classification of cardiac arrhythmia of 12 lead ECG using combination of SMOTEWB, XGBoost and machine learning algorithms. In *2019 9th International Symposium on Embedded Computing and System Design (ISED)* (pp. 1-7). IEEE.

13. McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signalling rates. *Journal of Forecasting, 22*(8), 569-586.

14. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-131.

15. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101-121). Academic Press.

16. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine, 47*(1), 31-39.

17. Sağlam, F., & Cengiz, M. A. (2022). A novel SMOTE-based resampling technique through noise detection and the boosting procedure. *Expert Systems with Applications, 200*, 117023.

18. Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

19. Taiwanese Bankruptcy Prediction. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.

20. Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639-2649.

21. Wang, L., & Wu, C. (2017). Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. *Knowledge-Based Systems, 121*, 99-110.

22. Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica, 39*(6), 745-758.

23. Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine, 4*(11).