ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY TRONG DỰ BÁO RỦI RO PHÁ SẢN CỦA CÁC DOANH NGHIỆP VIỆT NAM

Trương Thị Thùy Dương
Học viện Ngân hàng
Email: duongtt@hvnh.edu.vn
Lê Hải Trung
Học viện Ngân hàng
Email: trunglh@hvnh.edu.vn

Mã bài: JED - 1066
Ngày nhận bài: 26/12/2022
Ngày nhận bài sửa: 22/03/2023
Ngày duyệt đăng: 04/04/2023
DOI: 10.33301/JED.VI.1066

Tóm tắt
Dự báo rủi ro phá sản của doanh nghiệp đóng vai trò quan trọng trong việc đưa ra các cảnh báo sớm cho các doanh nghiệp. Các nghiên cứu đánh giá rủi ro phá sản sử dụng các phương pháp thống kê truyền thống và mô hình học máy. Trong nghiên cứu này sử dụng hồi quy logistic và các mô hình học máy để dự báo rủi ro phá sản của các doanh nghiệp Việt Nam. Nghiên cứu đi kiểm chứng tính hiệu quả của các mô hình học máy so với thống kê truyền thống và kiểm tra tính hiệu quả của các mô hình học máy. Kết quả cho thấy sự ưu thế của mô hình XGBoost và Random Forest so với logistic và các phương pháp khác.
Từ khóa: Phá sản, Logistic, Random Forest, Extreme Gradient Boosting, K-Nearest Neighboor, Naïve Bayses.
Mã JEL: C45, C52, C63, G33

Machine learning based bankruptcy prediction of Vietnam companies

Abstract
Bankruptcy prediction plays an important role in providing early warning for companies. Traditional statistics and machine learning methods have been used for failure prediction problems. In this study, we test the performance of machine learning methods comparing to logistic regression. The finding shows that XGBoost and Random Forest outperform than orther methods.
Keywords: *Bankruptcy prediction, Random Forest, K-Nearest Neighboor, Naïve Bayses, Extreme Gradient Boosting, Logistic.*
JEL code: C45, C52, C63, G33

1. Giới thiệu

Những biến động của kinh tế thế giới giai đoạn hậu Covid-19 đang có ảnh hưởng mạnh mẽ đến năng lực tài chính và hoạt động của các doanh nghiệp. Tình hình lạm phát tăng cao cùng với những mâu thuẫn chính trị khiến rủi ro phá sản của các doanh nghiệp trở nên rõ ràng hơn. Điều này ảnh hưởng tiêu cực đến nền kinh tế và xã hội do tác động lan truyền tới các doanh nghiệp trong chuỗi cung ứng và sự suy giảm thu nhập của người lao động. Do đó, việc đưa ra những dự báo về khả năng phá sản của doanh nghiệp có ý nghĩa quan trọng nhằm đưa ra các cảnh báo sớm về rủi ro tài chính. Ở nghiên cứu này, chúng tôi so sánh khả năng dự

báo của các phương pháp truyền thống và hiện đại đối với rủi ro phá sản của các doanh nghiệp Việt Nam nhằm tìm phương pháp dự báo phù hợp.

Altman (1968) và Beaver (1966) đã mở đầu cho phương pháp dự báo rủi ro phá sản truyền thống. Beaver (1966) sử dụng một số tỷ lệ tài chính như đòn bẩy tài chính, lợi nhuận trên tài sản và tính thanh khoản để dự báo rủi ro phá sản của doanh nghiệp. Các nghiên cứu về sau hướng tới việc cải thiện khả năng dự báo thông qua các mô hình phi tuyến (Jones & Hensher 2004). Kolari & cộng sự (2002) phát triển hệ thống cảnh báo sớm dựa trên kết hợp mô hình logit và mô hình nhận dạng đặc điểm cho các ngân hàng Mỹ. Lam & Moy (2002) đã kết hợp các mô hình phân biệt và thực hiện các mô phỏng để nâng cao độ chính xác của phân loại trong mô hình phân tích khác biệt. Cho đến nay mô hình logistic vẫn chứng tỏ được tính hiệu quả trong việc giải thích các yếu tố ảnh hưởng đến rủi ro tài chính của doanh nghiệp (Barboza & cộng sự, 2017).

Sự phát triển của công nghệ với năng lực xử lý các thuật toán phức tạp dẫn tới sự phát triển của các mô hình tính toán thông minh trong dự báo khả năng phá sản (Goldstein & cộng sự, 2019). Mô hình học máy đã được chứng minh có hiệu suất vượt trội (Florez-Lopez, 2007) do có thể xử lý hiệu quả các mối quan hệ phi tuyến cũng như các bài toán có độ phức tạp cao mà không đòi hỏi nhiều yêu cầu về dữ liệu. Các mô hình học máy bao gồm mô hình đơn và mô hình kết hợp. Mô hình kết hợp là tập hợp các mô hình để thu được mô hình tốt hơn. Mô hình kết hợp nâng cao gồm hai nhóm bao đóng (bagging) và tăng cường (boosting). Random forest là một phương pháp phân loại mạnh mẽ thuộc nhóm bao đóng có độ chính xác cao và xác định tầm quan trọng của các biến, một trong những lợi thế mà các phương pháp học máy như Neural Network, Support Vector Machine không có (Zoričák & cộng sự, 2020). Extreme Gradient Boosting (XGBoost) là một dạng của mô hình tăng cường, đã được sử dụng rộng rãi trong những năm gần đây và chứng tỏ ưu thế vượt trội (Barboza & cộng sự 2017). Bên cạnh các mô hình học máy kết hợp, mô hình K-Nearest Neighboor và Naïve Bayses được xem là những thuật toán đơn giản, dễ sử dụng và hiệu quả trong bài toán phân lớp.

Sự phát triển của các nhóm mô hình với hướng tiếp cận khác nhau dẫn đến câu hỏi về sự so sánh giữa các mô hình về mức độ hiệu quả trong việc dự báo rủi ro phá sản của doanh nghiệp (Duénez-Guzmán, & Vose, 2013). Điều này là quan trọng bởi việc lựa chọn mô hình dự báo rủi ro phá sản phụ thuộc vào đặc điểm của các doanh nghiệp trong từng quốc gia và đặc biệt là mức độ sẵn có của chuỗi dữ liệu để dự báo. Nghiên cứu của chúng tôi đóng góp vào lý luận và thực tiễn về dự báo rủi ro phá sản của doanh nghiệp bằng việc so sánh hiệu năng dự báo của các mô hình học máy và mô hình truyền thống đối với dữ liệu của các doanh nghiệp Việt Nam. Cụ thể, trong nghiên cứu này so sánh phương pháp hồi quy Logistic, Random forest, Decision tree, K-Nearest Neighboor, Naïve Bayses, XGBoost, kết quả của bài nghiên cứu ủng hộ quan điểm của các nghiên cứu trước về tính ưu thế hơn của học máy so với phương pháp truyền thống và mô hình XGBoost có hiệu quả cao nhất.

2. Tổng quan nghiên cứu dự báo rủi ro phá sản

2.1. Các nghiên cứu dự báo rủi ro phá sản sử dụng mô hình truyền thống

Các phương pháp nghiên cứu truyền thống khởi đầu từ Alman (1968), Beaver (1966) sử dụng chỉ tiêu tài chính để dự báo rủi ro phá sản của doanh nghiệp. Lin (2009) kiểm tra khả năng dự đoán khó khăn tài chính của các mô hình phân tích khác biệt, logit, probit đối với các công ty Đài Loan sau cuộc khủng hoảng tài chính năm 2009 cho thấy kết quả khả quan của các phương pháp truyền thống. Serrano-Cinca & Gutiérrez-Nieto (2013) sử dụng phân tích khác biệt với bình phương nhỏ nhất từng phần để dự báo cuộc khủng hoảng tài chính của các ngân hàng Mỹ năm 2008 và cho thấy hiệu suất dự báo tương đương với hiệu suất khi sử dụng mô hình học máy. Liang & cộng sự (2015) đã sử dụng các mô hình phân tích khác biệt và hồi quy logistic để lựa chọn các biến dự báo kiệt quệ tài chính, sử dụng đầu vào cho các mô hình học máy. Ưu điểm chính của các phương pháp truyền thống là tính giải thích đối với các biến dự báo và rủi ro phá sản của doanh nghiệp nhưng lại đòi hỏi chặt chẽ về dữ liệu.

2.2. Các nghiên cứu dự báo rủi ro phá sản sử dụng mô hình thông minh

Các mô hình thông minh được phát triển tương đối sớm, trong đó, mô hình mạng thần kinh được phát triển đầu tiên từ những năm 1990 (Serrano-Cinca, 1996). Sự tiến bộ của công nghệ cho phép xử lý các thuật toán phức tạp trong thời gian ngắn cho phép phát triển các mô hình học máy với khả năng tự cải thiện hiệu suất, cho phép xử lý nhiều bài toán có độ phức tạp cao với hiệu suất cao mà không đòi hỏi nhiều về yêu cầu của dữ liệu. Random forest được đề xuất bởi Breiman (2001), trong đó tập hợp cây quyết định được tạo ra trong quá trình bootstrap và đưa ra kết quả dựa trên biểu quyết đa số. Trong lĩnh vực tài chính, Random

forest đã được ứng dụng để phát hiện thành công gian lận tín dụng (Whitrow & cộng sự, 2009) và dự báo rời bỏ của khách hàng đối với các ngân hàng (Xie & cộng sự, 2009).

Nghiên cứu của Zhao & cộng sự (2009) đã chứng tỏ mô hình học máy cho hiệu quả cao hơn so với truyền thống. Tương tự, Barboza & cộng sự (2017) đã chỉ ra Random forest, bagging và boosting hiệu quả vượt trội hơn so với SVM, logit và phân tích khác biệt.

2.3. Các nghiên cứu dự báo rủi ro phá sản ở Việt Nam

Tại Việt Nam, dự báo khả năng phá sản của doanh nghiệp cũng thu hút được nhiều quan tâm. Bùi Phúc Trung (2012) sử dụng phương pháp truyền thống Z-score để đánh giá nguy cơ phá sản của các công ty niêm yết. Nguyễn Thị Cành & Phạm Chí Khoa (2014) xét các khách hàng doanh nghiệp của Vietcombank để dự báo xác suất phá sản bằng phương pháp KVM-Merton. Huỳnh Thị Cẩm Hà & cộng sự (2017) đã áp dụng mô hình cây phân lớp trong học máy để dự báo kiệt quệ tài chính của các công ty Việt Nam, kết quả thu được độ chính xác trên 90%. Nghiên cứu sử dụng Z-score của Alman cho 60 doanh nghiệp của Việt Nam được thể hiện trong nghiên cứu của Hoàng Thị Hồng Vân (2020) cho kết quả dự báo chính xác đến 76.67% sử dụng các chỉ tiêu gồm tài sản trung bình, ROA và ROE.

Tuy vậy, các nghiên cứu về rủi ro phá sản của các doanh nghiệp Việt Nam chủ yếu đang sử dụng các mô hình truyền thống, đối với các mô hình học máy chưa được sử dụng nhiều. Vì vậy, ở nghiên cứu này chúng tôi tiến hành so sánh hiệu suất dự báo khả năng phá sản của doanh nghiệp Việt Nam đối với cả các phương pháp truyền thống và mô hình học máy hiện đại.

3. Dữ liệu và phương pháp nghiên cứu

3.1. Dữ liệu

Bảng 1: Các biến độc lập trong mô hình

Tên biến	Miêu tả biến
APSALE	Khoản phải trả / doanh thu
CASHMTA	Tiền mặt và đầu tư ngắn hạn / (vốn chủ sở hữu thị trường + tổng nợ phải trả)
CHLCT	Tiền mặt / nợ ngắn hạn
(EBIT + DP)/AT	(Thu nhập trước lãi và thuế + khấu hao và khấu hao) / tổng tài sản
EBITSALE	Thu nhập trước lãi và thuế / doanh thu
INVCHINVT	Tăng trưởng hàng tồn kho / hàng tồn kho
INVTSALE	Hàng tồn kho / bán hàng
LCTAT	Nợ ngắn hạn / tổng tài sản
LCTLT	Nợ ngắn hạn / tổng nợ phải trả
LCTSALE	Nợ ngắn hạn / doanh thu
LTAT	Tổng nợ phải trả / tổng tài sản
LOG(AT)	log (tổng tài sản)
LOG(SALE)	log (bán)
MB	Tỷ lệ thị trường trên sổ sách
NISALE	Thu nhập ròng / doanh thu
OIADPSALE	Thu nhập hoạt động / bán hàng
PRICE	log (giá)
QALCT	Tài sản nhanh / nợ ngắn hạn
REAT	Thu nhập giữ lại / tổng tài sản
RELCT	Thu nhập giữ lại / nợ hiện tại
RSIZE	log (vốn hóa thị trường)
SALEAT	Doanh thu / tổng tài sản
SIGMA	Biến động cổ phiếu
WCAPAT	Vốn lưu động / tổng tài sản

Nguồn: Tian & Yu (2017)

Để dự báo rủi ro phá sản đối với các doanh nghiệp Việt Nam, chúng tôi sử dụng các chỉ tiêu tài chính của 300 doanh nghiệp Việt Nam trong thời gian 2017-2019, thu thập từ cơ sở dữ liệu của FiinGroup. Barboza & cộng sự (2017) đã cho thấy hiệu quả phân lớp của doanh nghiệp có rủi ro và không rủi ro khi sử dụng các

chỉ tiêu tài chính đặc trưng cho nhóm đòn bẩy tài chính, tính thanh khoản, nhóm lợi nhuận, quy mô công ty và tăng trưởng. Khẳng định này cũng được thể hiện trong nghiên cứu thực nghiệm của Zoričák & cộng sự (2020) trong dự báo rủi ro phá sản của các công ty vừa và nhỏ. Tổng hợp lại, Tian & Yu (2017) đã sử dụng 26 chỉ tiêu tài chính để dự báo khả năng phá sản của doanh nghiệp, kết quả chỉ ra nhóm các chỉ tiêu về khả năng thanh khoản và đòn bẩy tài chính có ảnh hưởng lớn nhất đến dự báo rủi ro tài chính. Nghiên cứu này lựa chọn các biến tài chính để dự báo rủi ro phá sản của các doanh nghiệp Việt Nam được nghiên cứu và tham khảo từ các nghiên cứu tổng quan của các nghiên cứu Tian & Yu (2017) và được trình bày trong Bảng 1.

Dữ liệu gồm 846 quan sát với 24 đặc tính đại diện cho biến giải thích. Các doanh nghiệp được phân lớp vào hai nhóm gồm 130 doanh nghiệp phá sản và 170 doanh nghiệp không phá sản tương ứng với giá trị mã hóa là 0 và 1. Trong thực tế, số doanh nghiệp phá sản ít hơn số doanh nghiệp không phá sản, do đó dữ liệu thường không cân bằng. Tuy nhiên, tỷ lệ chênh lệch trong tập dữ liệu nghiên cứu không nhiều, do đó dữ liệu được sử dụng để thực hiện mô hình mà không sử dụng thêm các kỹ thuật làm cân bằng dữ liệu. Dữ liệu sau khi được làm sạch được chia ngẫu nhiên thành hai tập huấn luyện và tập kiểm tra theo tỷ lệ 75% và 25%. Tập dữ liệu huấn luyện được sử dụng để xây dựng mô hình, tập kiểm tra để đánh giá hiệu quả của mô hình. Trong bài báo này, các mô hình được sử dụng để dự báo khả năng phá sản của doanh nghiệp sau một năm, các biến giải thích là các chỉ tiêu của doanh nghiệp trong năm trước, biến phụ thuộc là tình trạng của doanh nghiệp trong năm kế tiếp.

Bảng 2 cung cấp thống kê mô tả của các biến giải thích sử dụng trong mô hình. Hầu hết các biến giải thích trong mô hình đều có giá trị độ lệch chuẩn tương đối cao so với giá trị bình quân. Điều này cho thấy mức độ đa dạng trong bộ dữ liệu doanh nghiệp sử dụng để tiến hành dự báo rủi ro phá sản. Điều này thể hiện ở cả các chỉ tiêu thể hiện về khả năng sinh lời của doanh nghiệp như thu nhập trước thuế và lãi vay, thu nhập ròng trên tổng doanh thu hay các chỉ tiêu phản ánh hiệu quả quản lý tăng trưởng hàng tồn kho, thu nhập hoạt động trên doanh thu bán hàng. Mức độ đa dạng này cho phép đánh giá một cách chính xác và phù hợp hơn về mô hình dự báo rủi ro phá sản và mang tính đại diện đối với các doanh nghiệp của Việt Nam.

Bảng 2: Thống kế mô tả các biến

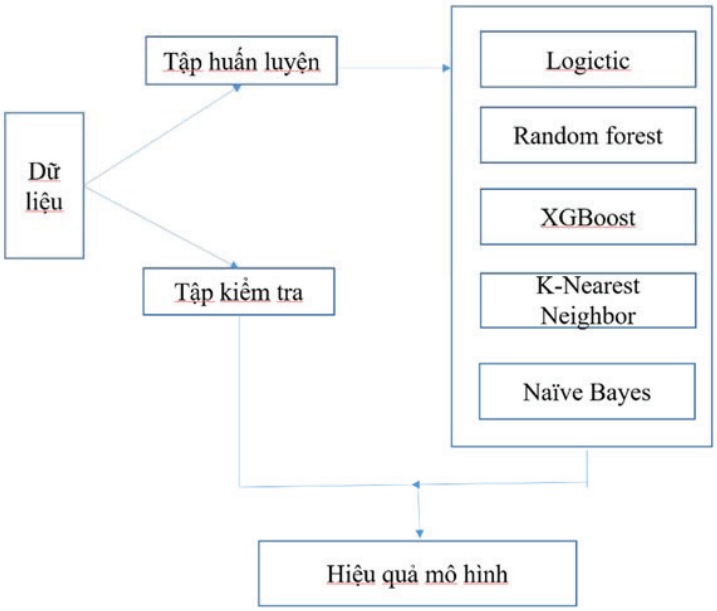
	Số quan sát	mean	std	min	25%	50%	75%	max
APSALE	846	0.3850	1.9925	0.0000	0.0683	0.1347	0.2642	46.5241
CASHMTA	846	0.0739	0.0910	0.0003	0.0181	0.0444	0.0916	0.7682
CHLCT	846	0.1433	0.6380	0.0001	0.0229	0.0540	0.1160	15.8496
(EBIT + DP)/AT	846	0.0704	0.0811	-0.3454	0.0215	0.0553	0.1043	0.4600
EBITSALE	846	0.0777	1.0308	-21.3821	0.0224	0.0644	0.1197	18.2854
INVCHINVT	846	0.7093	6.4287	-1.0000	-0.1602	0.0371	0.3178	108.4948
INVTSALE	846	1.6314	10.2026	0.0004	0.1104	0.2489	0.5550	168.3800
LCTAT	846	0.4530	0.2283	0.0091	0.2741	0.4421	0.6393	0.9597
LCTLT	846	0.8083	0.2281	0.0308	0.7021	0.9021	0.9821	1.0000
LCTSALE	846	1.9297	8.0041	0.0068	0.3808	0.6891	1.2584	182.1145
LTAT	846	0.5665	0.2275	0.0113	0.4041	0.6073	0.7382	1.1729
LOG(AT)	846	11.9319	0.6934	9.9191	11.5022	11.9215	12.3200	14.6061
LOG(SALE)	846	11.6461	0.7322	7.9271	11.2396	11.6730	12.0899	14.1145
MB	846	1.6295	1.5023	0.0190	0.7854	1.2037	1.9519	15.4179
NISALE	846	0.0488	1.0229	-23.6481	0.0118	0.0439	0.1177	7.5867
OIADPSALE	846	0.0634	1.0578	-23.6466	0.0131	0.0537	0.1330	9.6697
PRICE	846	4.0082	0.3947	2.6021	3.7848	4.0128	4.2524	5.3277
QALCT	846	2.1931	4.9801	0.1591	1.0960	1.3469	2.1056	105.7035
REAT	846	0.0306	0.1631	-1.3392	0.0133	0.0413	0.0832	0.4554
RELCT	846	0.2451	1.5852	-10.3294	0.0227	0.0942	0.2831	38.5537
RSIZE	846	11.3811	0.8008	9.3173	10.8565	11.3205	11.8672	14.5881
SALEAT	846	0.8450	0.9489	0.0004	0.2948	0.5879	1.0645	8.3236
SIGMA	846	0.1462	0.6405	-0.8862	-0.1662	0.0258	0.2775	7.4783
WCAPAT	846	0.1864	0.2391	-0.6830	0.0518	0.1618	0.3472	0.9853

Nguồn: Tính toán của nhóm tác giả

3.2. Phương pháp nghiên cứu

Mục đích nghiên cứu này là dự báo các doanh nghiệp có rủi ro hoặc không có rủi ro. Mô hình logistic là mô hình phân loại truyền thống phổ biến và hiệu quả nhất. Các mô hình học máy được sử dụng trong nghiên cứu này là Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor (KNN) và Naïve Bayes (NB). Các bước thực hiện dự báo được trình bày ngắn gọn trong Hình 1.

Hình 1: Các bước thực hiện mô hình



3.2.1. Hồi quy logistic

Hồi quy logistic là một trong những phương pháp thống kê phổ biến nhất dùng để phân lớp các biến nhị phân. Mô hình hồi quy logistic thể hiện dưới dạng sau:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Trong đó p là xác suất một doanh nghiệp có rủi ro phá sản, hoặc có rủi ro tín dụng, $x_1, x_2, ..., x_n$ là các biến độc lập. Phương pháp ước lượng hợp lý cực đại được sử dụng để tìm các hệ số.

Mỗi doanh nghiệp được tính xác suất rủi ro phá sản p và so sánh với một ngưỡng cho trước, dựa vào ngưỡng phân loại thì doanh nghiệp sẽ được xếp vào nhóm có rủi ro hay không rủi ro tương ứng với giá trị nhị phân 0 và 1.

3.2.2. Mô hình Random Forest

Random Forest là một trong kỹ thuật bao đóng hoạt động dựa trên cây quyết định phát triển từ thuật toán Bagging. Bagging là một kỹ thuật lấy mẫu từ tập dữ liệu lấy ra ngẫu nhiên các tập con thay thế. Kết quả cuối cùng là sự tổng hợp từ các mô hình dự báo trên các mẫu thay thế. Kỹ thuật này giúp xây dựng tính ổn định của mô hình, giảm phương sai, cải thiện độ chính xác và tránh quá mức.

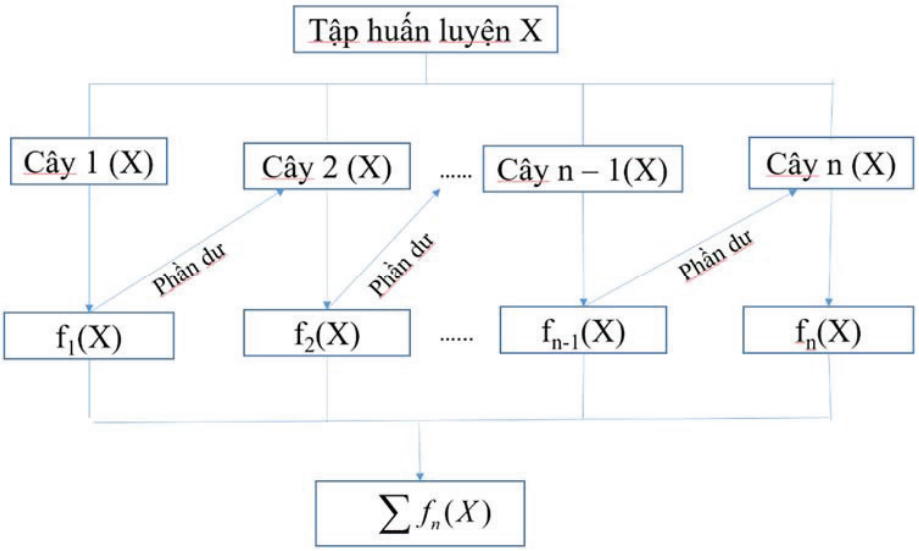
Ưu điểm của phương pháp Random Forest là khả năng xử lý được với các giá trị ngoại lai và các nhiễu (Yeh & cộng sự, 2014). Ngoài phân lớp hay dự báo, rừng ngẫu nhiên có thể xác định được tầm quan trọng của các biến trong mô hình. Điều này giúp đưa ra các yếu tố quyết định trong việc phân lớp hay dự báo (Maione & cộng sự, 2016). Các bước cơ bản của rừng ngẫu nhiên là:

- Tạo ngẫu nhiên các tập con khác nhau có các tính năng khác nhau.
- Các phần tử của tập hợp được dãn nhãn (thất bại hoặc không thất bại) và được chia vào các cây quyết định.
- Đối với mỗi bản ghi, lớp được bình chọn nhiều nhất được phân lớp.

3.2.3. Mô hình XGBoost

Thuật toán Extreme Gradient Boosting là một trong những thuật toán mới và hiệu quả cao trong học máy. Thuật toán này là sự mở rộng của thuật toán Gradient Tree Boosting được đề xuất bởi Friedman (2001). Nguyên lý của mô hình này là đào tạo các mô hình mới tốt hơn từ việc kết hợp các mô hình yếu trước đó để bù đắp các thiếu sót trong các mô hình trước.

Hình 2: Thuật toán XGBoost



Nguồn: Trương Việt Hùng & Hà Mạnh Hùng (2020)

Hình 2 thể hiện các bước của thuật toán XGBoost. Từ tập huấn luyên ban đầu X với n quan sát và đầu ra là y. Tại bước đầu tiên, mô hình huấn luyện tạo ngẫu nhiên một cây học tập với giá trị đầu ra $f_1(X)$ và sai số là e_1. Để có mô hình tốt hơn, cây học tập tiếp theo được huấn luyện để ước lượng sai số e_1 và đồng thời ước lượng giá trị $f_2(X)$ cùng với sai số e_2. Quá trình tuần tự cho đến cây học tập thứ n, giá trị ước lượng là $\sum_{i=1}^{n} f_i(X)$ và sai số nhỏ hơn sau mỗi quá trình huấn luyện.

3.2.4. Mô hình K-Nearest Neighbor

K-Nearest Neighbor (KNN) là một trong những thuật toán phân lớp đơn giản nhất dựa trên hàm khoảng cách. Thuật toán này không khai thác thông tin từ tập dữ liệu học, mọi tính toán được thực hiện khi cần dự đoán nhãn của tập dữ liệu mới dựa vào nhãn của các hàng xóm có khoảng cách gần nhất. Do đó KNN gọi là kỹ thuật dựa trên bộ nhớ. Các bước cơ bản của thuật toán như sau:

- Giả sử có một tập học và một tập dữ liệu cần phân lớp.
- Chọn K là số lân cận cần tính toán.
- Tìm khoảng cách từ tập dữ liệu mới đến các điểm trong tập học, tìm K điểm gần tập dữ liệu nhất, nhãn của tập dữ liệu mới là nhãn của các tập cùng nhãn gần nó nhất.

KNN là dễ thực hiện, đào tạo nhanh và không bị nhạy đối với các nhiễu, có thể phân lớp với dữ liệu có nhiều nhãn nhưng nhạy cảm với cấu trúc bộ dữ liệu, tốn bộ nhớ và nhiều thời gian hoạt động.

3.2.5. Mô hình Naïve Bayes

Mô hình Naïve Bayes là một trong những thuật toán phân loại Bayes cổ điển, có cấu trúc thuật toán đơn giản và hiệu quả tính toán cao.

Các bước cơ bản của thuật toán là (Soria & cộng sự, 2011): giả sử các quan sát chứa các thuộc tính là $\{A_1,...,A_n\}$ độc lập với nhau, các lớp phân loại là $C=\{C_1,...,C_m\}$. Với mỗi quan sát X, X được phân lớp vào lớp C_i tương ứng với xác suất hậu nghiệm lớn nhất theo công thức Bayes:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}, i = 1,...,m.$$

3.3. Phương pháp đánh giá khả năng dự báo

Nghiên cứu này sử dụng ma trận nhầm lẫn đo độ chính xác của mô hình. Đây là phương pháp đánh giá hiệu suất phân loại các quan sát vào hai lớp phá sản (nhận giá trị 1) hay không phá sản (nhận giá trị 0). Ma trận được trình bày tại Bảng 3.

Bảng 3: Ma trận nhầm lẫn

		Giá trị thật	
		1	0
Kết quả dự báo	1	TP	FP
	0	FN	TN

TP (true positive) là số dự đoán tích cực, nghĩa là số lượng công ty phá sản được dự báo đúng là phá sản. TN (true negative) là số lượng doanh nghiệp không phá sản được dự báo không phá sản, FP (false positive) là số lượng các doanh nghiệp không phá sản nhưng dự báo phá sản, FN (false negative) là số lượng doanh nghiệp phá sản nhưng được dự báo không phá sản.

Độ chính các của mô hình là tỷ lệ dự báo đúng, được tính theo công thức sau:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Trong đánh giá khả năng phá sản của doanh nghiệp, việc dự báo được doanh nghiệp phá sản quan trọng hơn dự báo doanh ngiệp không phá sản do hệ lụy lớn từ việc dự báo sai doanh nghiệp phá sản thành không phá sản. Vì vậy để nâng cao hiệu quả dự báo của mô hình, thường sử dụng thêm tiêu chí sai số loại I (Type I error) và sai số loại II (Type II error), độ hội tụ (Precision) và độ bao phủ (Recall), được xác định như dưới đây:

$$Type\ I\ \text{error} = \frac{FP}{FP + TP}, Type\ II\ \text{error} = \frac{FN}{TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}$$

Sai số loại I chỉ mô hình dự báo sai của doanh nghiệp không phá sản, sai số loại 2 chỉ doanh nghiệp phá sản được dự báo sai. Một mô hình tốt khi có độ chính xác cao và sai số thấp. Precision cho biết tỷ lệ dự đoán doanh nghiệp phá sản thực sự là phá sản, Recall cho biết tỷ lệ dự báo đúng phá sản trên tổng doanh nghiệp phá sản.

4. Kết quả và thảo luận kết quả

Dữ liệu được huấn luyện với 5 phương pháp dự báo bao gồm gồm logistic (LR), Random Forest (RF), XGBoost, K-Nearest Neighbor, Naïve Bayes (NB). Các chỉ tiêu đánh giá được sử dụng đối với các mô hình trên tập kiểm tra để kiểm tra hiệu quả dự báo, chúng được trình bày trong Bảng 4. Các mô hình huấn luyện và kiểm tra hiệu quả dự báo trên tập kiểm tra đều mất khoảng thời gian ngắn, do đó nghiên cứu không xem xét so sánh về mặt thời gian của mô hình.

Bảng 4: Kết quả dự báo

Mô hình	TP	FP	FN	TN	Accuracy (%)	Type I error (%)	Type II error (%)
LR	80	21	20	91	80,66	20,79	18,02
RF	79	22	13	98	83,49	21,78	11,71
XGboost	87	14	12	99	87,73	13,86	10,81
KNN	76	25	23	88	77,35	24,75	20,72
NB	68	33	13	98	78,3	32,67	11,71

Nguồn: Tính toán của nhóm nghiên cứu

Các chỉ tiêu Recall (Precision) của năm phương pháp LR, RF, XGBoost, KNN, NB theo thứ tự là 80%; 85,87%; 87,88%; 76,77%; 83,95% (79,2%; 78,2%; 86,14%; 75,24%; 67,72%).

Từ kết quả trên tập kiểm tra, các mô hình đều cho độ chính xác trên 77%, cao nhất là mô hình XGBoost với dự đoán đúng trên 87% doanh nghiệp. Tiếp theo là mô hình RF với trên 83% doanh nghiệp. Điều đáng

ngạc nhiên là mô hình LR dự báo chính xác về tổng số doanh nghiệp phá sản và không phá sản cao hơn mô hình KNN và NB trong tập dữ liệu này với tỷ lệ dự đoán đúng 80,66% so với 77,35% và 78,3%. Tuy nhiên sai số loại II của mô hình NB thấp hơn so với LR, tương ứng 11,71% và 18,02%.

Trong dự báo sức khỏe của doanh nghiệp, sai số loại II quan trọng hơn sai số loại I vì những tổn thất khi dự báo một doanh nghiệp phá sản thành không phá sản lớn hơn nhiều so với dự đoán doanh nghiệp không phá sản thành phá sản. Với tiêu chí này, mô hình KNN cho hiệu quả thấp nhất vì độ chính xác thấp nhất và sai số loại II cao nhất. Kết quả này có sự khác biệt so với dự báo của Le & Vivian (2018) đối với các doanh nghiệp Mỹ khi cho rằng KNN có hiệu suất cao hơn LR, tuy nhiên lại tương thích với Zhao & cộng sự (2009) đối với dữ liệu trên FDIC (fdic.gov). Mặc dù LR có độ chính xác cao hơn NB, nhưng sai số loại II cũng cao hơn, xét về chi phí cơ hội thì NB có hiệu quả cao hơn.

Kết quả cũng cho thấy mô hình dự báo hiệu quả nhất là XGBoost với độ chính xác, độ hội tụ và bao phủ cao nhất và sai số thấp nhất tương ứng hai loại sai số là 13,86% và 10,81%, tiếp đến là RF tương ứng độ chính xác, độ hội tụ và bao phủ cao thứ 2 và hai loại sai số là 21,78% và 11,71%. Kết quả này phù hợp với nghiên cứu của Barboza & cộng sự (2017). Các kết quả này chỉ ra rằng mô hình học máy có ưu thế hơn trong các bài toán dự báo so với phương pháp truyền thống như logit. Đồng thời dạng thuật toán "boosting" như XGBoost thường cho hiệu quả cao hơn các phương pháp thông minh khác đối với các vấn đề dự báo.

5. Kết luận

Dự đoán phá sản là vấn đề quan trọng liên quan đến khả năng thanh toán của doanh nghiệp. Các tổ chức tài chính, các cơ quan quản lý cũng như các doanh nghiệp cần dự báo các rủi ro có thể xảy ra nhằm có các chính sách ngăn chặn và giảm thiểu tổn thất. Mặc dù các mô hình thống kê đã được ứng dụng rộng rãi trong thực tế (Altman, 1968), tuy nhiên hạn chế của nó vẫn là hiệu suất do đòi hỏi chặt chẽ về điều kiện của dữ liệu. Sự phát triển của khoa học tính toán đã thúc đẩy các mô hình học máy phát triển và chúng đã được chứng minh là hiệu quả hơn với độ chính xác cao hơn và sai số thấp hơn so với các phương pháp truyền thống (Zhao & cộng sự, 2009; Yeh & cộng sự, 2014; Le & Vivian, 2018). Mô hình được học từ dữ liệu do đó không đòi hỏi nhiều về cấu trúc dữ liệu và có thể áp dụng linh hoạt.

Nghiên cứu này so sánh 6 phương pháp dự báo sử dụng dữ liệu là 300 doanh nghiệp Việt Nam trong thời kỳ 2017 – 2019, dự báo phá sản của doanh nghiệp sau thời gian một năm. Kết quả dự báo ủng hộ quan điểm của các nghiên cứu trước về tính ưu thế hơn của học máy so với phương pháp truyền thống và mô hình XGBoost có hiệu quả cao nhất. Chúng tôi cho rằng, kết quả này có ý nghĩa đối với các doanh nghiệp, các cơ quan quản lý, các cổ đông, các chủ nợ của doanh nghiệp cũng như các nhà đầu tư tiềm năng bởi khả năng cảnh báo sớm và phân biệt rủi ro phá sản của các doanh nghiệp sẽ giúp các bên liên quan đưa ra các quyết định tài chính phù hợp.

Nghiên cứu này của chúng tôi vẫn còn một số điểm hạn chế có thể được cải thiện ở các nghiên cứu trong tương lai. Thứ nhất, bộ số liệu của chúng tôi chưa bao gồm các quan sát từ năm 2020 trở lại đây với những sự thay đổi lớn trong môi trường kinh tế vĩ mô với bối cảnh của đại dịch Covid-19. Thứ hai, trong phạm vi của mục tiêu nghiên cứu, chúng tôi mới chỉ dừng lại ở việc chỉ ra phương pháp dự báo rủi ro phá sản phù hợp nhất đối với các doanh nghiệp của Việt Nam nhưng chưa chỉ ra được chỉ tiêu tài chính nào có tác động lớn nhất tới rủi ro phá sản của các doanh nghiệp trong mẫu thống kê. Các điểm hạn chế này sẽ là các câu hỏi mà chúng sẽ hướng tới giải quyết ở các nghiên cứu trong tương lai.

Tài liệu tham khảo

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

Barboza, F., Kimura, H. & Alman, E. (2017), 'Machine learning models and bankruptcy prediction', *Expert Systems with Applications*, 83, 405-417.

Beaver, W.H. (1966), 'Finance ratios as predictors of failure', *Finance of Accounting Research*, 4, 71-111.

Breiman (2001), 'Random Forest', *Machine Learning*, 45, 5-32.

Bùi Phúc Trung (2012), 'Đánh giá nguy cơ phá sản của doanh nghiệp niêm yết trên sàn chứng khoán Việt Nam bằng hàm phân biệt, *Tạp chí khoa học đại học mở Thành phố Hồ Chí Minh – Kinh tế và Quản trị kinh doanh*, 7(1), 41-47.

Duénez-Guzmán, E. A., & Vose, M. D. (2013), 'No free lunch and benchmarks', *Evolutionary Computation*, 21(2), 293-312.

Friedman, J.H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics*, 29(5), 1189-1232.

Florez-Lopez, R. (2007), 'Modelling of insurers' rating determinants. An application of machine learning techniques and statistical models', *European Journal of Operational Research*, 183(3), 1488-1512.

Goldstein, I., Jiang, W., & Karolyi, G.A. (2019), 'To fintech and beyond', *The Review of Financial Studies*, 32(5), 1647–1661.

Huỳnh Thị Cẩm Hà, Nguyễn Thị Uyên Uyên &Lê Đào Tuyết Mai (2017), 'Sử dụng các mô hình cây phân lớp dự báo kiệt quệ tài chính cho doanh nghiệp Việt Nam', Tạp chí khoa học đại học mở Thành phố Hồ Chí Minh – Kinh tế và Quản trị kinh doanh, 12(3), 62-76.

Hoàng Thị Hồng Vân (2020), 'Vận dụng mô hình Z-score trong dự báo khả năng phá sản doanh nghiệp tại Việt Nam', *Tạp chí khoa học và đào tạo Học viện Ngân hàng*, 217, 43-51.

Jones, S. & Hensher, D.A. (2004), 'Predicting firm finance distress: a mixed logit model', *Accounting Review*, 79(4), 1011-1038.

Kolari, J., Glennon, D., Shin, H. & Caputo, M. (2002), 'Predicting large US commercial bank failures', *Journal of Economics and Business*, 54(4), 361 – 387.

Lam, K.F. & Moy, J.W. (2002), 'Combining discriminant methods in solving classification problems in two-group discriminant analysis', *European Journal of Operational Research*, 180(1), 1- 28.

Le, H.H. & Viviani, J.L. (2018), 'Predicting bank failure: An improvement implementing a machine-learning approach to classical financai ratios', *Research in International Bussiness and Finance*, 44, 16-25.

Liang, D., Tsai, C.F., & Wu, H.T. (2015), 'The effect of feature selection on financial distress prediction', *Knowledge-Based Systems*, 73, 289-297.

Lin, T.H. (2009), 'A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models', *Neurocomputing*, 72(16-18), 3507-3516.

Maione, C., Batista, B.L., Campiglia, A.D., Barbosa, F., Jr, & Barbosa, R.M. (2016), 'Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry', *Computers and Electronics in Agriculture*, 121, 101–107.

Nguyễn Thị Cành & Phạm Chí Khoa (2014), 'Áp dụng mô hình KMV–Merton dự báo rủi ro tín dụng khách hàng doanh nghiệp và khả năng thiệt hại của ngân hàng', *Tạp chí Phát triển Kinh tế*, 289, 39-57.

Serrano-Cinca, C. (1996), 'Set organizing neural networks for financial diagnosis', *Decision Support Systems*, 17(3), 227-238.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2013), 'Partial least square discriminant analysis for bankruptcy prediction', *Decision Support Systems*, 54(3), 1245-1255.

Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., & Ellis, I.O. (2011), 'A 'non-parametric'version of the naive Bayes classifier', *Knowledge-Based Systems*, 24(6), 775-784.

Tian, S. & Yu, Y. (2017), 'Financial ratios and bankruptcy predictions: international evidence', *International Review of Economics and Finance*, 51, 510-526.

Trương Việt Hùng & Hà Mạnh Hùng (2020), 'Ước lượng khả năng chịu tải của giàn thép sử dụng phân tích trực tiếp và thuật toán XGBoost', *Tạp chí Xây dựng*, 2, 91-94.

Yeh, C.C., Chi, D.J., & Lin, Y.R. (2014), 'Going-concern prediction using hybrid random forests and rough set approach', *Information Sciences*, 254, 98-110.

Whitrow, C., Hand, D.J., Juszczak, P., Weston, D. & Adam, N.M. (2009), 'Transaction aggregation as a strategy for credit card fraud detection', *Data Mining and Knowledge Discovery*, 18, 30-55.

Xie, E., Li, X., Ngai, E. & Ying, W. (2009), 'Customer churn prediction using improved balanced random forest', *Expert Systems with Applications*, 36, 5445-5449.

Zhao, H., Sinha, A.P. & Ge, W. (2009), 'Effects of feature construction on classification performance: an empirical study in bank failure prediction', *Expert Systems with Applications*, 36(2), 2633 – 2644.

Zoričák, M., Gnip, P., Drotár, P., & Gazda, V. (2020), 'Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets', *Economic Modelling*, 84, 165-176.