



Dự báo kiệt quệ tài chính của các doanh nghiệp Việt Nam:

Ứng dụng Machine Learning

LÊ HỒNG NGỌC^a, NGUYỄN THẾ LONG^b, HỒ THỊ LAM^{*a}, HỒ THU HOÀI^c

^a Trường Đại học Tài chính - Marketing

^b Ngân hàng TMCP Xuất nhập khẩu Việt Nam

^c Đại học Kinh tế Thành phố Hồ Chí Minh

THÔNG TIN	TÓM TẮT
<p>Ngày nhận: 18/11/2023 Ngày nhận lại: 04/01/2024 Duyệt đăng: 05/01/2024</p> <p>Mã phân loại JEL: C53; G33; M21.</p> <p>Từ khóa: Kiệt quệ tài chính; Dự báo kiệt quệ tài chính; XGBoost; Machine Learning.</p> <p>Keywords: Financial Distress; Financial Distress Prediction; XGBoost; Machine Learning.</p>	<p>Dự báo kiệt quệ tài chính (KQTC) là một trong những nhiệm vụ quan trọng trong đánh giá rủi ro doanh nghiệp. Trong nghiên cứu này, nhóm tác giả ứng dụng các thuật toán Machine Learning nhằm dự báo KQTC, đồng thời xem xét các yếu tố ảnh hưởng đến nguy cơ KQTC của các doanh nghiệp tại Việt Nam. Nghiên cứu này sử dụng dữ liệu của 657 doanh nghiệp niêm yết trên hai sàn, gồm: (1) Sở Giao dịch Chứng khoán TP. Hồ Chí Minh (HOSE), và (2) Sở Giao dịch Chứng khoán Hà Nội (HNX), giai đoạn 2009–2022 với sáu thuật toán gồm: Logistic Regression, KNN, Decision Trees, Random Forests, AdaBoost, và XGBoost. Nghiên cứu đã chỉ ra rằng mô hình XGBoost là phù hợp nhất cho dự báo KQTC tại Việt Nam. Từ kết quả nghiên cứu, nhóm tác giả cũng đã đề xuất một số hàm ý quản trị và hàm ý chính sách trong việc lựa chọn mô hình dự báo KQTC và theo dõi các yếu tố tác động đến KQTC để phát triển bền vững doanh nghiệp.</p> <p>Abstract</p> <p>Forecasting financial distress is one of the important tasks in enterprise risk assessment. In this study, the authors apply machine learning algorithms to predict the financial distress and consider factors affecting financial distress ability of Vietnamese companies. This article uses data of 657 listed companies on the HOSE and HNX over the period of 2009–2022 with six algorithms, including Logistic Regression, KNN, Decision Trees, Random Forests, AdaBoost, and XGBoost. The</p>

* Tác giả liên hệ.

Email: lehongngoc2018@gmail.com (Lê Hồng Ngọc), long.nthe@gmail.com (Nguyễn Thế Long), holam@ufm.edu.vn (Hồ Thị Lam), hoaiht@ueh.edu.vn (Hồ Thu Hoài).

Trích dẫn bài viết: Lê Hồng Ngọc, Nguyễn Thế Long, Hồ Thị Lam, & Hồ Thu Hoài. (2023). Dự báo kiệt quệ tài chính của các doanh nghiệp Việt Nam: Ứng dụng Machine Learning. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, 34(12), 35–52

results show that the XGBoost algorithm is the most suitable for forecasting financial distress in Vietnam. From these results, this article proposes some management implications and policy implications in choosing a financial distress forecasting model and monitoring factors affecting financial distress for the sustainable development of the company.

1. Giới thiệu

Kiệt quệ tài chính (KQTC) được hiểu là tình huống trong đó một công ty không thể tạo ra đủ thu nhập để trang trải các nghĩa vụ tài chính của mình (Tinoco & Wilson, 2013; Vassalou & Xing, 2004) như thanh toán các khoản vay, khoản nợ hay các chi phí khi đến hạn. Kể từ những năm 1960, sự xuất hiện các tình huống KQTC hay phá sản của các doanh nghiệp đã làm gia tăng sự quan tâm đến các mô hình dự báo (Aziz & Dar, 2006). Beaver (1966) và Altman (1968) là những nghiên cứu tiên phong trong lĩnh vực này. Kể từ đó, dự báo KQTC trở thành một chủ đề thu hút sự chú ý của nhiều nhà khoa học và các nhà quản lý (Begum, 2022).

Tại Việt Nam, trải qua giai đoạn COVID-19, số doanh nghiệp không còn đủ tài chính để duy trì hoạt động tăng cao. Cụ thể, theo ghi nhận của Tổng cục Thống kê¹, trong quý I/2023, số doanh nghiệp rút lui khỏi thị trường là 60,2 nghìn doanh nghiệp. Trong khi số doanh nghiệp đăng ký thành lập mới và quay trở lại hoạt động chỉ có 57 nghìn doanh nghiệp, giảm 5,4% so với cùng kỳ năm 2022. Qua đó cho thấy, đại dịch COVID-19 có ảnh hưởng rất lớn đến các doanh nghiệp. Tình trạng các doanh nghiệp ngưng hoạt động, rơi vào tình trạng KQTC, hoặc phá sản có tác động rất lớn đến nền kinh tế. Do đó, việc phát hiện kịp thời nguy cơ KQTC tiềm ẩn của doanh nghiệp là rất quan trọng, nhằm đảm bảo cho chặng đường phát triển lâu dài của doanh nghiệp, và giúp các nhà hoạch định chính sách có những định hướng chính sách hữu hiệu để hỗ trợ doanh nghiệp phục hồi sản xuất.

Mặc dù nghiên cứu về KQTC đã được quan tâm, tuy nhiên, vẫn còn nhiều tranh luận. Đầu tiên là đa phần các nghiên cứu tại thị trường Việt Nam đều sử dụng phương pháp thống kê truyền thống dựa trên các mô hình mặc định được các tác giả đề xuất trước đây như: Altman (1968), Ohlson (1980), hay Zmijewski (1984) (xem Hoàng Thị Hồng Vân, 2020; Lê Cao Hoàng Anh & Nguyễn Thu Hằng, 2012; Lê Hoàng Vinh và cộng sự, 2022; Nguyễn Trà Ngọc Vy & Nguyễn Văn Công, 2013). Tuy nhiên, việc sử dụng mô hình truyền thống vẫn còn hạn chế khi chưa xem xét tổng quát các chỉ tiêu tài chính, mà chỉ mặc định một vài chỉ tiêu dựa trên mô hình gốc. Hơn nữa, các mô hình thống kê đòi hỏi nhiều giả định chặt chẽ không có trong đời thực, bao gồm các mối quan hệ tuyến tính, tính đồng nhất của phương sai và các giả định về tính độc lập. Vi phạm các giả định này có thể làm giảm khả năng dự đoán của các phương pháp thống kê. Tranh luận tiếp theo là các phương pháp dự đoán KQTC được sử dụng rộng rãi hơn dựa trên mô hình Z-score và O-score hầu hết được thực hiện ở các nước phát triển (Altman, 1968; Beaver, 1966; Ohlson, 1980; Shumway, 2001). Điều quan trọng cần lưu ý là các nước phát triển có cơ cấu kinh tế khác nhau, thủ tục phá sản rõ ràng và luật pháp liên quan đến phá sản được xác định rõ, trong khi các nước đang phát triển và kém phát triển lại thiếu điều này. Do

Tham khảo thêm tại đường dẫn: <https://www.gso.gov.vn/du-lieu-va-so-lieu-thong-ke/2023/03/bao-cao-tinh-hinh-kinh-te-xa-hoi-quy-i-nam-2023/>

đó, dự báo KQTC bằng cách sử dụng Z-score của Altman (1968) hoặc O-score của Ohlson (1980) có thể không hiệu quả đối với các nước đang phát triển vì những mô hình này được xây dựng dành cho các nước phát triển.

Trong những năm gần đây, thuật toán Machine Learning đã được sử dụng rộng rãi trong lĩnh vực dự báo KQTC (Zeng và cộng sự, 2020) và thu hút được nhiều sự chú ý của giới học thuật (Begum, 2022; Heo & Yang, 2014; Kim & Upneja, 2014; Shetty và cộng sự, 2022; Sun và cộng sự, 2011). Tuy vậy, Machine Learning vẫn chưa được sử dụng phổ biến tại thị trường Việt Nam trong các nghiên cứu dự báo KQTC. Đó là một phần động lực thúc đẩy nhóm tác giả thực hiện nghiên cứu này nhằm mục tiêu dự báo nguy cơ KQTC của các doanh nghiệp Việt Nam.

Nghiên cứu này đóng góp cho các tài liệu liên quan hiện có trên một số điểm như sau:

- *Đầu tiên*, nghiên cứu sử dụng các thuật toán Machine Learning cho dự báo KQTC đối với các doanh nghiệp Việt Nam. Đây cũng là một trong những nghiên cứu hiếm hoi sử dụng Machine Learning cho dự báo KQTC tại Việt Nam. Bên cạnh đó, nhóm tác giả sử dụng đồng thời nhiều thuật toán khác nhau và đánh giá hiệu suất dự báo của các thuật toán để từ đó có đề xuất hữu ích trong việc áp dụng thuật toán dự báo tối ưu cho các doanh nghiệp Việt Nam.

- *Thứ hai*, nghiên cứu này xem xét khoảng thời gian dài hơn so với các nghiên cứu trước và bao gồm cả giai đoạn COVID-19 (từ năm 2009 đến năm 2022) với số lượng mẫu lớn hơn (gồm 657 công ty phi tài chính niêm yết trên hai sàn HOSE và HNX). Điều này giúp nhóm tác giả có được số lượng các doanh nghiệp gặp tình trạng kiệt quệ lớn hơn và dự báo tốt hơn nguy cơ KQTC của doanh nghiệp.

- *Thứ ba*, những phát hiện trong nghiên cứu này sẽ là một cơ sở học thuật đáng tin cậy cho các nhà quản lý trong việc điều hành doanh nghiệp, không chỉ giúp các doanh nghiệp chủ động có biện pháp phòng ngừa rủi ro thông qua áp dụng phương pháp đề xuất để có thể dự báo được nguy cơ KQTC mà còn giúp các công ty xếp hạng tín nhiệm có được phương pháp để dự báo nguy cơ vỡ nợ của các tổ chức phát hành trái phiếu trên thị trường, cũng như các tổ chức tín dụng trong việc đánh giá rủi ro và ra quyết định cấp tín dụng cho các doanh nghiệp. Nghiên cứu cũng cung cấp công cụ cảnh báo sớm cho các nhà hoạch định chính sách về rủi ro của công ty đại chúng nhằm xây dựng các biện pháp bảo vệ nhà đầu tư cá nhân trước nguy cơ vỡ nợ trái phiếu.

Phần còn lại của nghiên cứu cấu trúc như sau, tổng quan nghiên cứu được trình bày ở phần 2. Tiếp theo, phần 3 trình bày về dữ liệu và phương pháp nghiên cứu. Kết quả nghiên cứu được trình bày ở phần 4, và phần 5 là kết luận.

2. Cơ sở lý thuyết

2.1. Quan điểm kiệt quệ tài chính

KQTC là một khái niệm rộng bao gồm một số tình huống trong đó các công ty gặp khó khăn về tài chính. Các quốc gia khác nhau có các thủ tục, quy định kế toán khác nhau và định nghĩa về KQTC do các học giả khác nhau đưa ra không phải lúc nào cũng giống nhau (Geng và cộng sự, 2015). Các thuật ngữ phổ biến nhất được sử dụng để mô tả những tình huống này là “phá sản”, “thất bại”, và “mất khả năng thanh toán”. Altman và Hotchkiss (1993) đã đưa ra một mô tả và định nghĩa đầy đủ về KQTC và chỉ ra rằng phá sản gần nhất với định nghĩa pháp lý về KQTC. Zmijewski (1984) định nghĩa KQTC là hành động nộp đơn xin phá sản. Tuy nhiên, nhiều công ty gặp khó khăn về tài chính

không bao giờ nộp đơn xin phá sản do mua lại hoặc tư nhân hóa, trong khi các công ty khỏe mạnh thường nộp đơn xin phá sản để tránh thuế và các vụ kiện tốn kém (Theodossiou và cộng sự, 1996). Wruck (1990) cho rằng “KQTC là tình trạng dòng tiền hoạt động của công ty không đủ để đáp ứng nghĩa vụ hiện tại (như tín dụng thương mại hoặc chi phí lãi vay) và công ty buộc phải thực hiện hành động khắc phục”. Trong nghiên cứu này, nhóm tác giả xác định KQTC là tình huống mà dòng tiền hiện tại không đủ để thanh toán cho các nghĩa vụ trước mắt hay không thể đáp ứng hoặc gặp khó khăn trong việc thanh toán các nghĩa vụ tài chính. Đây cũng là quan điểm được nhiều nghiên cứu sử dụng, chẳng hạn như Geng và cộng sự (2015); Powell và cộng sự (2023); Vu và cộng sự (2023); và Wruck (1990).

2.2. Tổng quan các nghiên cứu trước

Dự báo KQTC là vấn đề đã được nghiên cứu trong nhiều thập kỷ (Sun và cộng sự, 2011). Từ những năm 1930, Fitzpatrick (1932) đã tiên phong trong việc so sánh các tỷ số tài chính của các doanh nghiệp thành công và các công ty thất bại. Từ đó trở đi, dự báo KQTC tiếp tục được nghiên cứu rộng rãi do vai trò quan trọng của nó trong việc hỗ trợ ra quyết định. Beaver (1966) đã sử dụng phân tích đơn biến để dự báo sự phá sản của doanh nghiệp. Altman (1968) đã phát triển mô hình Z-score nổi tiếng bằng cách sử dụng phân tích phân biệt đa biến (Multiple Discriminant Analysis – MDA) kết hợp năm tỷ số tài chính. Ohlson (1980) đã áp dụng mô hình hồi quy logistic cho dự báo KQTC, mô hình này có thể chỉ ra khả năng xảy ra KQTC. Các mô hình dự báo KQTC có thể được phân loại thành mô hình dựa trên kế toán (Altman, 1968; Ohlson, 1980; Zmijewski, 1984), mô hình dựa trên thị trường (Bharath & Shumway, 2008; Merton, 1974) hoặc mô hình hỗn hợp (Chava & Jarrow, 2004; Trabelsi và cộng sự, 2015).

Dự báo KQTC bước vào giai đoạn bùng nổ vào cuối những năm 1980 nhờ sự tiến bộ nhanh chóng của trí tuệ nhân tạo và kỹ thuật khai thác dữ liệu. Frydman và cộng sự (1985) thực hiện dự báo KQTC thông qua cây quyết định (Decision Trees – DT). Theo kết quả của họ, kỹ thuật này cho phép dễ dàng xác định các đặc điểm quan trọng nhất trong dự báo KQTC. Odom và Sharda (1990) bắt đầu xây dựng mô hình mạng nơron (Neural Network – NN) cho dự báo KQTC. Kể từ đó, NN đã trở thành một trong những phương pháp Machine Learning được sử dụng rộng rãi nhất cho dự báo KQTC (Odom & Sharda, 1990; Serrano-Cinca, 1996).

Trong một thập kỷ gần đây, nghiên cứu dự báo KQTC bằng Machine Learning lại càng trở nên phổ biến. Sử dụng dữ liệu từ năm 1985 đến năm 2013, Barboza và cộng sự (2017) đã so sánh độ chính xác của năm mô hình Machine Learning để dự đoán KQTC. Kết quả đã cho thấy các mô hình truyền thống có khả năng dự đoán thấp hơn (từ 52% đến 77%) so với các mô hình Machine Learning (từ 71% đến 87%). Altman và cộng sự (2020) đánh giá hiệu quả của các phương pháp ước tính khác nhau trong một mẫu lớn các công ty Phần Lan trong khoảng thời gian 10 năm. Nghiên cứu đã áp dụng năm phương pháp khác nhau, bao gồm: DT, GBM (Gradient Boosting Machine), LR (Logistic Regression), NN, và SVM (Support Vector Machine), để dự báo KQTC bằng cách sử dụng cả biến tài chính và phi tài chính. Kết quả cho thấy hồi quy logistic và NN vượt trội hơn so với các phương pháp khác. Begum (2022) sử dụng dữ liệu được lấy từ Tạp chí Kinh tế Đài Loan trong khoảng thời gian 10 năm (1999–2009). Nghiên cứu dùng bốn thuật toán phân loại gồm RF (Random Forest), XGBoost, LR, và ANN (Artificial Neural Network). Kết quả cho thấy thuật toán RF vượt trội hơn các thuật toán khác với điểm chính xác là 96,53%. Mặc dù các nghiên cứu trên đã cho thấy hiệu suất

vượt trội của Machine Learning, nhưng vẫn còn hạn chế khi chưa đề cập các yếu tố nào quan trọng trong dự báo KQTC.

Các kỹ thuật Machine Learning ngày càng được phát triển với những kỹ thuật mới mang lại hiệu suất dự báo cao. Trong đó, nổi bật nhất phải kể đến là thuật toán XGBoost và đã được sử dụng trong các nghiên cứu gần đây, chẳng hạn như trong nghiên cứu của Climent và cộng sự, 2019; Huang & Yen, 2019 và Pham & Ho, 2021. Climent và cộng sự (2019) sử dụng dữ liệu của các ngân hàng khu vực đồng Euro giai đoạn 2006–2016. Kết quả cho thấy XGBoost có hiệu suất dự báo vượt trội. Độ chính xác của các mô hình trong nghiên cứu của Climent dao động từ 70% đến 90%. Huang và Yen (2019) đã so sánh XGBoost với các thuật toán khác như SVM, HACT, GA-Fuzzy, DBN, DBN-SVM, và cho thấy XGBoost là tốt nhất. Độ chính xác dự đoán của XGBoost trong nghiên cứu này dao động từ 82,8% đến 90,6%. Pham và Ho (2021) cho rằng việc sử dụng các thuật toán boosting trong dự báo phá sản là một chủ đề quan trọng và cần được nghiên cứu rộng rãi. Nghiên cứu sử dụng ba mô hình là Gradient Boosting Machine (GBM), AdaBoost và XGBoost. Kết quả cho thấy XGBoost cung cấp hiệu suất dự đoán tuyệt đối, trong khi GBM và AdaBoost vẫn bị sai lệch.

Tại Việt Nam, dự báo KQTC cũng là vấn đề đã được nhiều học giả nghiên cứu. Nghiên cứu của Lê Cao Hoàng Anh và Nguyễn Thu Hằng (2012) dự báo thất bại cho 293 doanh nghiệp niêm yết trên sàn HOSE, kết quả cho thấy mô hình Z-score có độ chính xác đạt 91%. Nguyễn Trà Ngọc Vy và Nguyễn Văn Công (2013) đã sử dụng mô hình Z-score để dự báo rủi ro phá sản của các công ty được phẩm trên thị trường chứng khoán Việt Nam. Lê Hoàng Vinh và cộng sự (2022) đã đánh giá sự khác biệt kết quả đo lường giữa các mô hình Z-score, S-score, O-score, X-score, Z-Taffler, H-score, và G-score. Kết quả cho thấy S-score là phù hợp nhất với tỷ lệ chính xác đạt 86,24%. Tuy nhiên, các nghiên cứu trên vẫn sử dụng mô hình dự báo truyền thống là chính. Gần đây, nghiên cứu của Trương Thị Thùy Dương và Lê Hải Trung (2023) đã sử dụng các thuật toán Machine Learning để dự báo phá sản cho 300 doanh nghiệp Việt Nam giai đoạn 2017–2019. Kết quả cho thấy mô hình XGBoost có độ chính xác cao nhất với 87% và lỗi loại II là 10,81%. Đây cũng là nghiên cứu hiếm hoi khi sử dụng Machine Learning dự báo KQTC tại Việt Nam. Tuy nhiên, nghiên cứu trên vẫn còn hạn chế ở mặt số lượng doanh nghiệp và thời gian nghiên cứu. Điều này có thể ảnh hưởng đến kết quả nghiên cứu. Hơn nữa, nghiên cứu này cũng chưa chỉ ra được các yếu tố ảnh hưởng nghiêm trọng đến tình trạng KQTC của các doanh nghiệp để có cơ sở đề xuất giải pháp tránh KQTC cho doanh nghiệp.

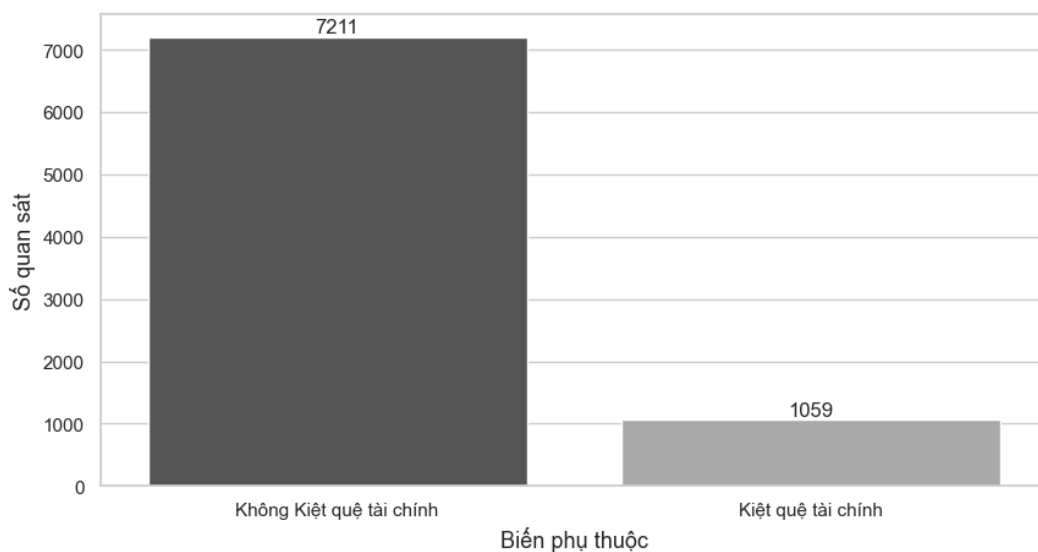
3. Dữ liệu và phương pháp nghiên cứu

3.1. Dữ liệu nghiên cứu

Dữ liệu cho nghiên cứu được thu thập từ cơ sở dữ liệu Refinitiv Eikon² gồm các công ty niêm yết trên Sở Giao dịch Chứng khoán TP. Hồ Chí Minh (HOSE) và Sở Giao dịch Chứng khoán Hà Nội (HNX) giai đoạn 2009–2022. Tiêu chí đầu tiên là, công ty lựa chọn phải niêm yết ít nhất ba năm tính đến 2022. Tiêu chí tiếp theo là các công ty thuộc lĩnh vực tài chính/bảo hiểm/ngân hàng bị loại khỏi mẫu nghiên cứu vì cơ cấu bảng cân đối kế toán của các công ty này khác với cấu trúc của các ngành

2 Hiện nay, cơ sở dữ liệu Refinitiv Eikon được xem là lớn nhất và đầy đủ nhất, bao gồm một lượng lớn dữ liệu về kinh tế, quản lý Nhà nước, dữ liệu doanh nghiệp, quản trị công ty, phân tích tài chính và thị trường. Tham khảo thêm tại <https://eikon.refinitiv.com/>.

khác (Powell và cộng sự, 2023). Cuối cùng, các công ty bị thiếu nhiều dữ liệu cũng bị loại khỏi mẫu. Mẫu dữ liệu cuối cùng bao gồm 657 công ty với 8.270 quan sát.



Hình 1. Phân phối biến phụ thuộc

Các công ty KQTC được xác định dựa trên tiêu chí tỷ lệ thanh toán lãi vay (Interest Coverage Ratio – ICR), đây cũng là tiêu chí được nhiều nghiên cứu sử dụng, như Asquith và cộng sự, (1994); Dinh và cộng sự, (2021); Ninh và cộng sự, (2018); Powell và cộng sự, (2023). ICR là tỷ lệ giữa thu nhập trước lãi vay và thuế (EBIT), và chi phí lãi vay (IE). Nếu ICR của một công ty thấp hơn 1 thì công ty này được phân loại là KQTC và được gán giá trị là “1”; ngược lại, các công ty được phân loại là không KQTC và gán giá trị là “0”.

Dựa trên các nghiên cứu trước (du Jardin, 2015; du Jardin và cộng sự, 2019; Heo & Yang, 2014; Inam và cộng sự, 2018; Liang và cộng sự, 2016; Lin và cộng sự, 2014; Powell và cộng sự, 2023), nhóm tác giả đã tổng hợp, lựa chọn các chỉ tiêu tài chính phù hợp với nghiên cứu và được phân thành bốn nhóm gồm tỷ suất sinh lời, tỷ lệ thanh khoản, đòn bẩy và tỷ lệ tăng trưởng. Việc phân tích theo nhóm cho chúng ta biết các khía cạnh khác nhau của tài chính và hoạt động của doanh nghiệp (Lin và cộng sự, 2014). Thông tin các chỉ tiêu tài chính sử dụng trong mô hình dự báo được trình bày ở Bảng 1.

Bảng 1.

Tóm tắt các biến nghiên cứu

Nhóm tỷ lệ tài chính	TT	Ký hiệu	Định nghĩa
Tỷ suất sinh lời	N1	EBITTA	Thu nhập trước lãi vay và thuế/Tổng tài sản
(Profitability Ratios)	N2	NITA	Thu nhập ròng/Tổng tài sản
	N3	TSTA	Tổng doanh thu/Tổng tài sản
	N4	RETA	Lợi nhuận giữ lại/Tổng tài sản
	N5	NITS	Thu nhập ròng/Tổng doanh thu

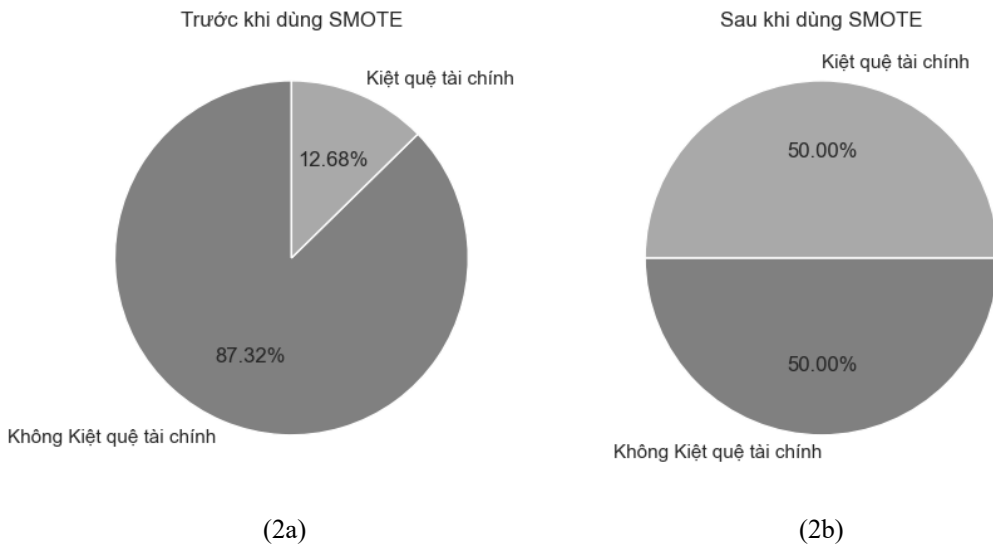
Nhóm tỷ lệ tài chính	TT	Ký hiệu	Định nghĩa
	N6	EBITDATA	Thu nhập trước lãi vay, thuế, khấu hao và phí trả dần/Tổng tài sản
	N7	EBITTS	Thu nhập trước lãi vay và thuế/Tổng doanh thu
	N8	NISE	Thu nhập ròng/Vốn cổ đông
	N9	EBITDATL	Thu nhập trước lãi vay, thuế, khấu hao và phí trả dần/Tổng nợ phải trả
Tỷ lệ thanh khoản (Liquidity Ratios)	N10	CACL	Tài sản ngắn hạn/Nợ ngắn hạn
	N11	WCTA	Vốn lưu động/Tổng tài sản
	N12	CLTA	Nợ ngắn hạn/Tổng tài sản
	N13	QACL	Tài sản nhanh/Nợ ngắn hạn
	N14	CATA	Tài sản hiện tại/Tổng tài sản
	N15	CTA	Tiền mặt và tương đương tiền/Tổng tài sản
	N16	NOCREDINT	Tỷ lệ của khoảng thời gian không có tín dụng
	N17	CATL	Tài sản ngắn hạn/Tổng nợ phải trả
	N18	QATA	Tài sản thanh khoản nhanh/Tổng tài sản
	N19	WCTS	Vốn lưu động/Tổng doanh thu
	N20	CLTS	Nợ ngắn hạn/Tổng doanh thu
	N21	CLTL	Nợ ngắn hạn/Tổng nợ
	N22	CCA	Tiền mặt và tương đương tiền/Tài sản ngắn hạn
N23	ITA	Hàng tồn kho/Tổng tài sản	
Đòn bẩy (Leverage)	N24	TLTA	Tổng nợ/Tổng tài sản
	N25	CCL	Tiền mặt và tương đương tiền/Nợ ngắn hạn
	N26	TLSE	Tổng nợ phải trả/Vốn cổ đông
Tỷ lệ tăng trưởng (Growth Ratios)	N27	GNI	Tăng trưởng thu nhập ròng
	N28	GTA	Tăng trưởng tổng tài sản
	N29	GOI	Tăng trưởng thu nhập thông thường
	N30	GRTA	Tăng trưởng của Tỷ suất sinh lời trên tổng tài sản
	N31	GSE	Tăng trưởng vốn cổ đông
	N32	GTS	Tăng trưởng tổng doanh thu

Tỷ lệ tập huấn luyện và thử nghiệm trong nghiên cứu này là 70:30. Nghiên cứu này sử dụng các thuật toán Machine Learning để dự báo KQTC sau một năm (dữ liệu năm t-1 được sử dụng để dự báo tình trạng KQTC của công ty ở năm t).

3.2. Phương pháp nghiên cứu

3.2.1. Sử dụng kỹ thuật SMOTE xử lý dữ liệu mất cân bằng

SMOTE (Synthetic Minority Oversampling Technique) khắc phục hiện tượng quá khớp của mô hình trong tập dữ liệu mất cân bằng bằng cách tạo ra dữ liệu nhân tạo (Shrivastava và cộng sự, 2020).



Hình 2. Dữ liệu trước và sau khi sử dụng SMOTE

Sau khi sử dụng SMOTE, các thuật toán Machine Learning được áp dụng để dự báo KQTC.

3.2.2. Mô hình hồi quy logistic (LR)

Từ những năm 1980, một số nghiên cứu đã phát triển các mô hình xác suất có điều kiện sử dụng kỹ thuật hồi quy Logistic (Dimitras và cộng sự, 1996; Keasey & Watson, 1991; Liao, 1994; Ohlson, 1980; Zavgren, 1983). Hồi quy logistic là phương pháp đầu tiên được sử dụng thay cho MDA để thiết kế mô hình phá sản (Ohlson, 1980). Phương trình hồi quy Logistic như sau:

$$\ln\left(\frac{p_n}{1-p_n}\right) = \alpha + \sum_{j=1}^J \beta_j x_{nj} \quad (1)$$

Trong đó,

p_n : Xác suất KQTC của công ty n

$x_n = (x_{i,1}, \dots, x_{i,J})$: Vec-tơ J chiều chứa các giá trị mà các biến giải thích J giả định cho công ty n

α : Tham số đại diện cho phần chặn

β_j : Hệ số hồi quy thứ j .

Khi các tham số α và β_j được ước tính bằng cách sử dụng dữ liệu có sẵn, xác suất KQTC có thể được ước tính bằng cách đảo ngược mô hình hồi quy logistic:

$$p_n = \left(1 + \exp(\alpha + \sum_{j=1}^J \beta_j x_{nj})\right)^{-1} \quad (2)$$

Thông thường, ngưỡng 0,5 được sử dụng để xác định xem công ty nên được phân loại là KQTC (giá trị $p_n \geq 0,5$) hay không KQTC (giá trị $p_n < 0,5$) (Zavgren, 1983).

3.2.3. Mô hình K-Nearest Neighbors

K-Nearest Neighbors (KNN) là một trình phân loại không theo tham số, xác định nhãn lớp cho từng phiên bản chưa được nhìn thấy theo k (số lượng lân cận) gần nhất trong tập dữ liệu huấn luyện. Ưu điểm của KNN là nó mạnh mẽ ngay cả khi tập dữ liệu huấn luyện bị nhiễu (El Gayar và cộng sự, 2006). Hạn chế chính của KNN là chi phí tính toán cao vì nó cần tính toán khoảng cách của mọi trường hợp chưa được nhìn thấy cho tất cả dữ liệu huấn luyện (Son và cộng sự, 2014).

3.2.4. Mô hình cây quyết định (Decision Trees)

Decision Trees (DT) bao gồm một loạt các quyết định logic, được biểu diễn dưới dạng cấu trúc cây (Breiman và cộng sự, 1984). Các quyết định này tương tự như các quy tắc nếu-thì, vì ở đây, đầu vào là một điều kiện được mô tả bởi một tập hợp các thuộc tính (Hussin Adam Khatir & Bee, 2022). DT không yêu cầu bất kỳ giả định thống kê nào liên quan đến dữ liệu trong mẫu huấn luyện. DT rất hữu ích trong việc khám phá dữ liệu nhằm tìm ra mối quan hệ giữa một số lượng lớn các biến đầu vào và biến mục tiêu.

3.2.5. Mô hình rừng ngẫu nhiên (Random Forests)

Random Forest (RF) là sự tổng quát hóa của DT, vì RF là một tập hợp các DT (Breiman, 2001). RF là một mô hình Machine Learning phổ biến, nó đã được sử dụng rộng rãi trong việc dự báo KQTC. RF được phát triển lần đầu bởi Breiman (2001). Kỹ thuật này kết hợp một khung tổng hợp và cây quyết định để cải thiện độ chính xác hiệu suất mô hình. Theo Katuwal và cộng sự (2020), hàm đầu ra thu được bằng cách lấy trung bình như sau:

$$Z = \operatorname{argmax}_T \frac{1}{T} \sum_{t=1}^T p_t(y|x) \quad (3)$$

Trong đó, $p_t(y|x)$ là phân bố xác suất của mỗi cây (t) và x là tập hợp các mẫu thử nghiệm. RF là một kỹ thuật hiệu quả để xây dựng khung cây quyết định nhằm xử lý dữ liệu phức tạp và nhiễu. Hơn nữa, RF được đề xuất vì khả năng triển khai và diễn giải đơn giản, giúp cải thiện khả năng dự báo.

3.2.6. Mô hình AdaBoost

AdaBoost (Adaptive Boosting) là một thuật toán Machine Learning lặp lại để xây dựng một tập hợp phân loại (Freund & Schapire, 1997). Nó chỉ sử dụng một thuật toán phân loại để xây dựng các bộ phân loại cơ sở yếu đa dạng, được huấn luyện trên các tập dữ liệu được lấy mẫu có chọn lọc từ tập dữ liệu huấn luyện ban đầu. Mỗi mẫu được gán một trọng số thể hiện khả năng được chọn làm mẫu huấn luyện và tất cả các mẫu đều có trọng số bằng nhau trong vòng lặp đầu tiên. Trong các lần lặp tiếp theo, các mẫu được phân loại chính xác bởi bộ phân loại cơ sở của lần lặp cuối cùng sẽ có trọng số thấp hơn và các mẫu bị phân loại sai sẽ có trọng số cao hơn.

3.2.7. Mô hình XGBoost

Trong những năm gần đây, XGBoost (eXtreme Gradient Boosting) đã thu hút sự chú ý của các nhà nghiên cứu về tài chính, đặc biệt trong các nghiên cứu dự báo KQTC (Climent và cộng sự, 2019; Huang & Yen, 2019; Pham & Ho, 2021; Shetty và cộng sự, 2022). Kỹ thuật XGBoost được phát triển bởi Chen và Guestrin (2016). XGBoost là một biến thể của phương pháp tăng cường độ dốc với hiệu

suất vượt trội sử dụng hình thức hóa mô hình chính quy nhằm kiểm soát overfitting để đạt được hiệu suất tốt hơn (Chen & Guestrin, 2016). Khung kết hợp sử dụng sơ đồ lấy mẫu ngẫu nhiên để cải thiện khả năng dự đoán và giảm phương sai của mô hình cuối cùng:

$$Z = F(x_i) = \sum_{t=1}^T f_t(x_i) \quad (4)$$

Trong đó, x_i biểu thị các biến giải thích và $f_t(x_i)$ là hàm đầu ra của mỗi cây.

3.3. Đánh giá hiệu suất mô hình

Biến phụ thuộc được sử dụng trong nghiên cứu này chứa các công ty không KQTC và KQTC được gán nhãn là 0 và 1. Do đó, nhóm tác giả xây dựng ma trận nhầm lẫn và xác định một số tính năng thể hiện hiệu suất phân loại của mô hình.

Bảng 2.

Ma trận nhầm lẫn

		Giá trị thực tế	
		(1)	(0)
Giá trị dự báo	(1)	True Positive (TP)	False Positive (FP)
	(0)	False Negative (FN)	True Negative (TN)

Ghi chú: True Positive (dương tính thực): dự báo công ty KQTC và thực tế KQTC; False Positive (dương tính giả): Dự báo công ty KQTC và thực tế không KQTC; False Negative (âm tính giả): Dự báo công ty không KQTC và thực tế KQTC; True Negative (âm tính thực): Dự báo công ty không KQTC và thực tế không KQTC.

Độ chính xác (Accuracy) là thước đo phổ biến nhất được sử dụng để đánh giá hiệu suất phân loại (Tharwat, 2021). Theo đó, hầu hết các nghiên cứu dự báo KQTC đều sử dụng độ chính xác làm thước đo để đánh giá hiệu suất mô hình (Soui và cộng sự, 2020). Accuracy được định nghĩa là tỷ lệ các trường hợp được phân loại chính xác (cả doanh nghiệp KQTC và không KQTC).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Tuy nhiên, Accuracy là một thước đo không phù hợp cho các vấn đề phân loại không cân bằng. Có thể đạt được Accuracy cực cao bằng cách đưa ra dự đoán chính xác về lớp đa số (không KQTC), ngay cả khi mô hình hoạt động cực kỳ tệ trong việc dự đoán lớp thiểu số (KQTC). Do đó, cần phải xem xét thêm các chỉ tiêu khác gồm Precision, Recall, và F1-score cùng với các số liệu khác.

Precision cho biết tỷ lệ mà dự báo công ty KQTC và thực tế là kiệt quệ.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

Độ bao phủ (Recall) hay độ nhạy (Sensitivity) tương ứng với tỷ lệ phần trăm doanh nghiệp KQTC được dự báo đúng.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

F1-score là giá trị trung hoà của Precision và Recall. F1-score càng cao thì mô hình càng tốt.

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Tiếp theo, nhóm tác giả sử dụng lỗi loại I (Type I error) và lỗi loại II (Type II error) cho các mô hình dự báo. Lỗi loại I có nghĩa là số công ty hoạt động bình thường được phân loại là KQTC. Lỗi loại II có nghĩa là số lượng công ty KQTC được phân loại là bình thường. Trong nghiên cứu này, lỗi loại II nghiêm trọng hơn bởi vì chi phí tổn thất khi dự báo nhầm một công ty KQTC là khoẻ mạnh cao hơn nhiều so với chi phí dự đoán một công ty KQTC nhưng thực tế khoẻ mạnh (Shrivastava và cộng sự, 2020).

$$\text{Lỗi loại I} = \frac{FP}{FP + TP} \quad (9)$$

$$\text{Lỗi loại II} = \frac{FN}{TN + FN} \quad (10)$$

Bởi vì bộ dữ liệu sử dụng mất cân bằng đáng kể, cho nên ACC có thể không phù hợp khi đo lường hiệu suất mô hình. Có thể đạt được ACC cực cao bằng cách đưa ra dự báo chính xác về nhóm đa số (không kiệt quệ tài chính), ngay cả khi mô hình hoạt động cực kỳ kém trong việc dự đoán nhóm thiểu số (kiệt quệ tài chính). Do đó, nhóm tác giả tính diện tích dưới đường cong (Area Under the Curve – AUC) của từng mô hình. So sánh AUC và Accuracy, người ta thấy rằng các nghiên cứu gần đây có xu hướng chọn AUC để xác định độ phù hợp của mô hình (Jabeur và cộng sự, 2021, 2023; Kuiziniene và cộng sự, 2022; Pham & Ho, 2021). Mô hình có AUC lớn hơn 0,9 được xem là có hiệu suất vượt trội (Cheng và cộng sự, 2014; Tserng và cộng sự, 2011).

4. Kết quả nghiên cứu và thảo luận

Bảng 3 trình bày kết quả hiệu suất của các mô hình Machine Learning sử dụng. Accuracy của tất cả các mô hình trong nghiên cứu này đều đạt kết quả khá cao, thấp nhất là LR (70,82%) và cao nhất là XGBoost (93,67%). Ngoài ra, mô hình RF cũng có Accuracy cao (93,15%), chỉ đứng sau XGBoost. Từ kết quả Bảng 3, nghiên cứu đã xác định được 3 mô hình có độ chính xác cao nhất là RF, AdaBoost, và XGBoost.

Bảng 3.

Kết quả dự báo

Hiệu suất mô hình	Logistic Regression	K-Nearest Neighbors	Decision Trees	Random Forests	AdaBoost	XGBoost
TP	168	203	257	267	283	271
FN	157	122	68	58	42	54
FP	567	462	166	112	178	103
TN	1.589	1.694	1.990	2.044	1.978	2.053
Accuracy (%)	70,82	76,46	90,57	93,15	91,13	93,67
Precision (%)	22,86	30,53	60,76	70,45	61,39	72,46
Recall (%)	51,69	62,46	79,08	82,15	87,08	83,38

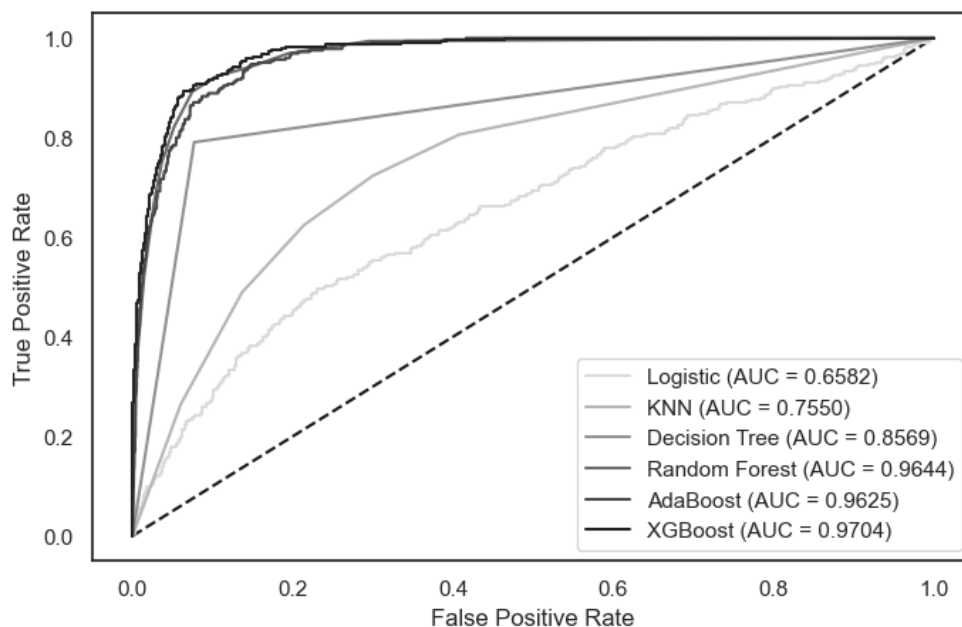
Hiệu suất mô hình	Logistic Regression	K-Nearest Neighbors	Decision Trees	Random Forests	AdaBoost	XGBoost
F1-score (%)	31,70	41,01	68,72	75,85	72,01	77,54
Lỗi loại I (%)	77,14	69,47	39,24	29,55	38,61	27,54
Lỗi loại II (%)	8,99	6,72	3,30	2,76	2,08	2,56

Tiếp theo, các chỉ số đánh giá hiệu suất khác như Precision, Recall và F1-score cũng được đề cập trong Bảng 3. Mô hình XGBoost có chỉ số Precision và F1-score cao nhất khi có tỷ lệ lần lượt là 72,46% và 77,54%. Mô hình AdaBoost có Recall cao nhất (87,08%), mô hình cao thứ hai là XGBoost (83,38%).

Kết quả Bảng 3 còn so sánh về tỷ lệ lỗi loại I và lỗi loại II. Dựa vào Bảng 3, tỷ lệ lỗi loại I của XGBoost thấp nhất (27,54%), tiếp theo là RF (29,55%). Ngoài ra, kết quả cũng cho thấy tỷ lệ lỗi loại II của mô hình AdaBoost là thấp nhất (2,08%), XGBoost đứng thứ hai (2,567%). Các mô hình còn lại gồm LR, KNN, DT, và RF có tỷ lệ lỗi loại II lần lượt là 8,99%; 6,72%; 3,3%; và 2,76%.

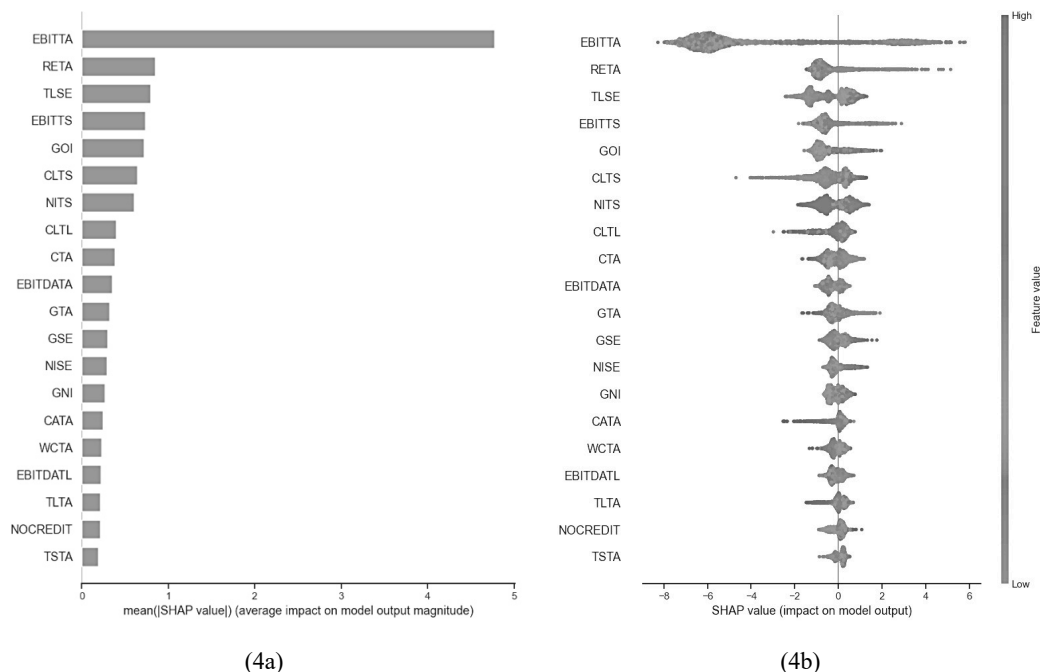
Trong thực tế, cái giá phải trả của lỗi loại II có thể lớn hơn nhiều so với lỗi loại I. Trong kết quả trên, AdaBoost cho thấy tỷ lệ lỗi loại II thấp nhất. Tuy nhiên, vẫn chưa kết luận AdaBoost là mô hình phù hợp nhất bởi vì các đại lượng đo lường hiệu suất mô hình trước đó đều cho thấy XGBoost vượt trội hơn. Ngoài ra, mô hình XGBoost cũng có tỷ lệ lỗi loại II rất thấp. Do đó, nhóm tác giả đã quyết định xem xét thêm về chỉ số AUC.

Trong Hình 3, hầu hết các mô hình đều có AUC cao. Mô hình XGBoost có AUC cao nhất (AUC = 97,04%), tiếp theo là RF (AUC = 96,44%) và đứng thứ 3 là AdaBoost (AUC = 96,25%). Ngược lại, mô hình LR, KNN và DT lại có kết quả không tốt khi AUC lần lượt là 65,82%; 75,50%; và 85,69%.



Hình 3. So sánh đường cong (Receiver Operating Curve – ROC) giữa các mô hình

Từ kết quả trên, có thể thấy mô hình XGBoost là mô hình phù hợp nhất cho dự báo KQTC của các doanh nghiệp tại Việt Nam. Kết quả về XGBoost là mô hình phù hợp để nghiên cứu này dự báo KQTC cũng đã được tìm thấy ở các nghiên cứu trước (Huang & Yen, 2019; Perboli & Arabnezhad, 2021; Pham & Ho, 2021; Shetty và cộng sự, 2022). Qua đó, nghiên cứu này cũng cho thấy tính ưu việt của các thuật toán Machine Learning khi ứng dụng vào dự báo KQTC.



Hình 4. Tầm quan trọng của các biến đặc tính trong XGBoost với SHAP

Cuối cùng, phương pháp (Shapley Additive Explanations – SHAP) được sử dụng để diễn giải kết quả đầu ra của mô hình XGBoost. Kết quả cho thấy các biến ảnh hưởng lớn đến KQTC gồm: EBITTA, RETA, TLSE, EBITTS, GOI, CLTS, NITS, CLTL, CTA, EBITDATA, và GTA. Kết quả này cũng tương đồng với các phát hiện của các nghiên cứu trước đây (xem Jabeur và cộng sự, 2021, 2023; Tian và cộng sự, 2015; Tian & Yu, 2017; Tran và cộng sự, 2022; Vu và cộng sự, 2023; Yang và cộng sự, 2021) và cho thấy phù hợp với thực tế ở các doanh nghiệp Việt Nam với các biến chính ảnh hưởng đến KQTC là các biến liên quan đến khả năng tăng trưởng lợi nhuận và mức độ sử dụng đòn bẩy.

5. Kết luận

Dự đoán KQTC phần lớn đã được nghiên cứu trong vài thập kỷ qua từ cả quan điểm lý thuyết lẫn quan điểm thực nghiệm. Với số lượng doanh nghiệp đóng cửa và không quay trở lại thị trường ngày càng tăng sau COVID-19, việc dự báo KQTC là rất cần thiết tại Việt Nam hiện nay. Mục tiêu của nghiên cứu này là dự báo KQTC của các doanh nghiệp phi tài chính tại Việt Nam. Sử dụng dữ liệu của 657 công ty giai đoạn 2009–2022, kết hợp sáu thuật toán Machine Learning (gồm LR, KNN, DT, RF, AdaBoost, và XGBoost) để so sánh hiệu quả của các mô hình dự báo nhằm tìm ra phương pháp cảnh báo sớm tốt nhất. Kết quả đã chỉ ra rằng thuật toán XGBoost là tốt nhất cho dự báo KQTC. Các

chỉ số đánh giá hiệu suất của XGBoost đều có kết quả tốt như Accuracy rất cao (93,67%), lỗi loại II rất thấp (2,56%) và AUC cao nhất trong các mô hình sử dụng (97,04%).

Việc theo dõi liên tục tình hình tài chính của công ty là vô cùng quan trọng để xác định các dấu hiệu KQTC. Do đó, ban lãnh đạo của bất kỳ công ty nào cũng phải liên tục đối mặt với nhiệm vụ tìm ra phương pháp dự báo KQTC để đưa ra đánh giá chính xác và kịp thời nhất về tình trạng tài chính của doanh nghiệp. Nhìn chung, nghiên cứu về dự báo KQTC sử dụng Machine Learning có tiềm năng đóng góp cho lĩnh vực quản lý rủi ro tài chính, đóng vai trò là hệ thống cảnh báo sớm.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế nhất định. Trong tương lai, việc đưa thêm nhiều dữ liệu kế toán và các biến số vào các nghiên cứu sâu hơn có thể tạo ra các mô hình dự đoán tốt hơn. Các biến liên quan đến quản lý công ty và kinh tế vĩ mô cần được xem xét để cải thiện độ chính xác của mô hình trong dự báo KQTC. Ngoài ra, nghiên cứu này chưa đề cập đến các yếu tố can thiệp như hành vi con người, tâm lý đám đông hay đầu cơ ảnh hưởng đến việc tăng giảm rủi ro KQTC của các doanh nghiệp niêm yết tại Việt Nam.

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.
- Altman, E. I., & Hotchkiss, E. (1993). *Corporate Financial Distress and Bankruptcy* (Vol. 1998). New York: John Wiley & Sons.
- Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2020). A race for long horizon bankruptcy prediction. *Applied Economics*, 52(37), 4092–4111.
- Asquith, P., Gertner, R., & Scharfstein, D. (1994). Anatomy of financial distress: An examination of junk-bond issuers. *The Quarterly Journal of Economics*, 109(3), 625–658.
- Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: Where we stand? *Corporate Governance*, 6(1), 18–33.
- Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111.
- Begum, S. (2022). A detailed study for bankruptcy prediction by machine learning technique. *Intelligent Sustainable Systems: Selected Papers of Worlds4 2021*, 2, 201–213.
- Jabeur, S. B., Stef, N., & Carmona, P. (2023). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics*, 61(2), 715–741.
- Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton Distance to Default model. *The Review of Financial Studies*, 21(3), 1339–1369.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

- Chava, S., & Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects. *Review of Finance*, 8(4), 537–569.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco.
- Cheng, M. Y., Hoang, N. D., Limanto, L., & Wu, Y. W. (2014). A novel hybrid intelligent approach for contractor default status prediction. *Knowledge-Based Systems*, 71, 314–321.
- Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. *Journal of Business Research*, 101, 885–896.
- Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90(3), 487–513.
- Dinh, D. V., Powell, R. J., & Vo, D. H. (2021). Forecasting corporate financial distress in the Southeast Asian countries: A market-based approach. *Journal of Asian Economics*, 74, 101293.
- du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, 242(1), 286–303.
- du Jardin, P., Veganzones, D., & Séverin, E. (2019). Forecasting corporate bankruptcy using accrual-based models. *Computational Economics*, 54(1), 7–43.
- El Gayar, N., Schwenker, F., & Palm, G. (2006). A study of the robustness of KNN classifiers trained using soft labels. *Lecture Notes in Computer Science*, 4087, 67–80.
- Fitzpartrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed companies. *Certified Public Accountant*, 10, 598–605.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
- Frydman, H., Altman, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, 40(1), 269–291.
- Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, 241(1), 236–247.
- Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, 24, 494–499.
- 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 & Đào tạo Ngân hàng*, 217, 43–51.
- Huang, Y. P., & Yen, M. F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing*, 83, 105663.
- Hussin Adam Khatir, A. A., & Bee, M. (2022). Machine learning models and data-balancing techniques for credit scoring: What is the best combination?. *Risks*, 10(9), 1–22.
- Inam, F., Inam, A., Mian, M. A., Sheikh, A. A., & Awan, H. M. (2018). Forecasting Bankruptcy for organizational sustainability in Pakistan: Using artificial neural networks, logit regression, and discriminant analysis. *Journal of Economic and Administrative Sciences*, 35(3), 183–201.

- Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166, 120658.
- Katuwal, R., Suganthan, P. N., & Zhang, L. (2020). Heterogeneous oblique random forest. *Pattern Recognition*, 99, 107078.
- Keasey, K., & Watson, R. (1991). Financial distress prediction models: A review of their usefulness. *British Journal of Management*, 2(2), 89–102.
- Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, 36, 354–362.
- Kuiziniene, D., Krilavičius, T., Damaševičius, R., & Maskeliūnas, R. (2022). Systematic review of financial distress identification using artificial intelligence methods. *Applied Artificial Intelligence*, 36(1), 2138124.
- Lê Cao Hoàng Anh, & Nguyễn Thu Hằng. (2012). Kiểm định mô hình chỉ số Z của Altman trong dự báo thất bại doanh nghiệp tại Việt Nam. *Tạp chí Công nghệ Ngân hàng*, 74, 3–9.
- Lê Hoàng Vinh, Phạm Lê Quang, & Bùi Kim Dung. (2022). Mô hình nào phù hợp để đo lường kiệt quệ tài chính cho công ty phi tài chính niêm yết tại Việt Nam?. *Tạp chí Quản lý và Kinh tế Quốc tế*, 144, 39–52.
- Liao, T. F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. (Quantitative Applications in the Social Sciences, Vol. 101). SAGE Publishing.
- Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2), 561–572.
- Lin, F., Liang, D., Yeh, C. C., & Huang, J. C. (2014). Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, 41(5), 2472–2483.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449–470.
- Nguyễn Trà Ngọc Vy, & Nguyễn Văn Công. (2013). Vận dụng mô hình Z-Score của GS. E.I. Altman để dự báo rủi ro phá sản của các công ty dược phẩm trên thị trường chứng khoán Việt Nam. *Tạp chí Kinh tế & Phát triển*, 10, 63–70.
- Ninh, B. P. V., Thanh, D. T., & Duc, V. H. (2018). Financial distress and bankruptcy prediction: An appropriate model for listed firms in Vietnam. *Economic Systems*, 42(4), 616–624.
- Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *1990 IJCNN International Joint Conference on Neural Networks*, 2, 163–168.
- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.
- Perboli, G., & Arabnezhad, E. (2021). A machine learning-based DSS for mid and long-term company crisis prediction. *Expert Systems with Applications*, 174, 114758.
- Pham, X. T. T., & Ho, T. H. (2021). Using boosting algorithms to predict bank failure: An untold story. *International Review of Economics & Finance*, 76, 40–54.

- Powell, R. J., Dinh, D. V., Vu, N. T., & Vo, D. H. (2023). Accounting-based variables as an early warning indicator of financial distress in crisis and non-crisis periods. *International Journal of Finance & Economics*, 1–20.
- Serrano-Cinca, C. (1996). Self organizing neural networks for financial diagnosis. *Decision Support Systems*, 17(3), 227–238.
- Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy Prediction using machine learning techniques. *Journal of Risk and Financial Management*, 15(1), 35.
- Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics & Finance*, 8(1), 1–17.
- Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101–124.
- Son, H., Kim, C., Hwang, N., Kim, C., & Kang, Y. (2014). Classification of major construction materials in construction environments using ensemble classifiers. *Advanced Engineering Informatics*, 28(1), 1–10.
- Soui, M., Smiti, S., Mkaouer, M. W., & Ejbali, R. (2020). Bankruptcy prediction using stacked Auto-Encoders. *Applied Artificial Intelligence*, 34(1), 80–100.
- Sun, J., Jia, M. Y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems with Applications*, 38(8), 9305–9312.
- Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192.
- Theodossiou, P., Kahya, E., Saidi, R., & Philippatos, G. (1996). Financial distress and corporate acquisitions: Further empirical evidence. *Journal of Business Finance and Accounting*, 23(5–6), 699–719.
- Tian, S., & Yu, Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. *International Review of Economics & Finance*, 51, 510–526.
- Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52, 89–100.
- Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30, 394–419.
- Trabelsi, S., He, R., He, L., & Kusy, M. (2015). A comparison of Bayesian, Hazard, and Mixed Logit model of bankruptcy prediction. *Computational Management Science*, 12(1), 81–97.
- Tran, K. L., Le, H. A., Nguyen, T. H., & Nguyen, D. T. (2022). Explainable machine learning for financial distress prediction: Evidence from Vietnam. *Data*, 7(11), 160.
- Trương Thị Thùy Dương, & Lê Hải Trung. (2023). Ứng dụng phương pháp học máy trong dự báo rủi ro phá sản của các doanh nghiệp Việt Nam. *Tạp chí Kinh tế & Phát triển*, 310, 44–53.
- Tserng, H. P., Lin, G. F., Tsai, L. K., & Chen, P. C. (2011). An enforced support vector machine model for construction contractor default prediction. *Automation in Construction*, 20(8), 1242–1249.

- Vassalou, M., & Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, 59(2), 831–868.
- Vu, N. T., Nguyen, N. H., Tran, T., Le, B. T., & Vo, D. H. (2023). A LASSO-based model for financial distress of the Vietnamese listed firms: Does the COVID-19 pandemic matter?. *Cogent Economics & Finance*, 11(1).
- Wruck, K. H. (1990). Financial distress, reorganization, and organizational efficiency. *Journal of Financial Economics*, 27(2), 419–444.
- Yang, H., Li, E., Cai, Y. F., Li, J., & Yuan, G. X. (2021). The extraction of early warning features for predicting financial distress based on XGBoost model and shap framework. *International Journal of Financial Engineering*, 8(03), 2141004.
- Zavgren, C. (1983). The prediction of corporate failure: The state of the art. *Journal of Accounting Literature*, 2(1), 1–38.
- Zeng, S., Li, Y., Yang, W., & Li, Y. (2020). A financial distress prediction model based on sparse algorithm and support vector machine. *Mathematical Problems in Engineering*, 2020, 1–11.
- Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59–82.