TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

BÁO CÁO TỔNG HỢP

KẾT QUẢ NGHIÊN CỨU KHOA HỌC

TÊN ĐỀ TÀI: XÂY DỰNG MÔ HÌNH KẾT HỢP CÁC THUẬT TOÁN HỌC MÁY ĐỂ DỰ ĐOÁN BÁO CÁO TÀI CHÍNH GIAN LẬN - ÁP DỤNG CHO THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

Tác giả:

Phan Thị Cẩm Tú

Đà Nẵng, tháng 11 năm 2021

TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

BÁO CÁO TỔNG HỢP

KẾT QUẢ NGHIÊN CỨU KHOA HỌC

TÊN ĐỀ TÀI: XÂY DỰNG MÔ HÌNH KẾT HỢP CÁC THUẬT TOÁN HỌC MÁY ĐỂ DỰ ĐOÁN BÁO CÁO TÀI CHÍNH GIAN LẬN - ÁP DỤNG CHO THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

Tác giả:

Phan Thị Cẩm Tú

Giáo viên hướng dẫn: TS. Đoàn Thị Ngọc Trai

Khoa: Kế toán

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thuộc Chương trình sinh viên nghiên cứu khoa học thành phố Đà Nẵng năm 2021

Lĩnh vực đề tài: Khoa học kỹ thuật và công nghệ

Đà Nẵng, tháng 11 năm 2021

MỤC LỤC

Danh mục hình

Danh mục bảng

Danh mục từ viết tắt, thuật ngữ

ACFE	Association of Certified Fraud Examiners	Hiệp hội các nhà điều tra gian lận - Một tổ chức được thành lập để chống lừa gạt và gian lận trong thực tiễn kinh doanh. Hiệp hội các nhà điều tra gian lận được chứng nhận là cơ quan quản lý của những người kiểm tra gian lận trên toàn cầu. Hiệp hội cung cấp cho các thành viên những công cụ, giáo dục và đào tạo nhằm hỗ trợ cho công việc của họ.
SEC	Securities and Exchange Commission	Ủy ban chứng khoán và sàn giao dịch mỹ
KTV	Kiểm toán viên	
BCTC	Báo cáo tài chính	Các thông tin kinh tế được kế toán viên trình bày dưới dạng bảng biểu, cung cấp các thông tin về tình hình tài chính, tình hình kinh doanh và các luồng tiền của doanh nghiệp đáp ứng các cầu cho những người sử dụng chúng trong việc đưa ra các quyết định về kinh tế
ML	Machine learning	Học máy - Một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.
LR	Logistic Regression	Thuật toán Hồi quy logistic
NN	Neural network	Mạng nơ-ron
DT	Decision Tree	Thuật toán Cây quyết định
SVM	Support Vector Machine	Thuật toán Máy vecto hỗ trợ
RF	Random Forest	Thuật toán Rừng ngẫu nhiên
KNN	K-Nearest Neighbor	Thuật toán K láng giền lân cận
XGBoost	XGBoost	Thuật toán eXtreme Gradient Boosting
LightGBM	LightGBM	Thuật toán Light Gradient Boosted Machine
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
Probabilistic NN	Probabilistic neural network	Một mạng nơ ron xác suất – Một mạng nơ ron phản hồi, được sử dụng rộng rãi trong các vấn đề nhận dạng và phân loại. Trong đó, hàm phân phối xác suất cha của mỗi lớp được xấp xỉ bằng cửa sổ Parzen và hàm không tham số.
	Unweighted probit	Xác suất không dựa vào tỷ trọng
	Bayesian Belief Network	Mạng Bayes - Cách biểu diễn đồ thị

		của sự phụ thuộc thống kê trên một tập hợp các biến ngẫu nhiên, trong đó các nút đại diện cho các biến, còn các cạnh đại diện cho các phụ thuộc có điều kiện.
	Genetic algorithms	Giải thuật di truyền - Một kỹ thuật của khoa học máy tính nhằm tìm kiếm giải pháp thích hợp cho các bài toán tối ưu tổ hợp. Giải thuật di truyền là một phân ngành của giải thuật tiến hóa vận dụng các nguyên lý của tiến hóa như di truyền, đột biến, chọn lọc tự nhiên, và trao đổi chéo.
Bagging	Mô hình kết hợp đóng gói	Một kỹ thuật học tập kết hợp xây dựng một lượng lớn các mô hình (thường là cùng loại) trên những tập con khác nhau từ tập huấn luyện một cách song song nhằm đưa ra dự đoán tốt hơn.
Boosting	Mô hình kết hợp tăng cường	Một kỹ thuật học tập kết hợp xây dựng một lượng lớn các mô hình (thường là cùng loại). Tuy nhiên quá trình huấn luyện trong phương pháp này diễn ra tuần tự theo chuỗi (sequence). Trong chuỗi này mỗi mô hình sau sẽ học cách sửa những lỗi của mô hình trước (hay nói cách khác là dữ liệu mà mô hình trước dự đoán sai)
Stacking	Mô hình kết hợp xếp chồng	Một kỹ thuật học tập kết hợp xây dựng một số mô hình (thường là khác loại) và một mô hình giám sát, mô hình này sẽ học cách kết hợp kết quả dự báo của một số mô hình một cách tốt nhất.
Ensemble Learning	Mô hình kết hợp	là phương pháp tổng hợp kết quả dự đoán của nhiều thuật toán khác nhau thành mô hình cuối cùng
Under sampling `	Kỹ thuật lấy mẫu dưới	Một kỹ thuật xử lí các tập dữ liệu mất cân bằng bằng cách giảm số lượng các quan sát của nhóm đa số (nhóm chiếm ưu thế trong tập dữ liệu) để nó trở nên cân bằng với số quan sát của nhóm thiểu số.
Over sampling	Kỹ thuật lấy mẫu quá mức	Một kỹ thuật xử lí các tập dữ liệu mất cân bằng bằng cách gia tăng kích thước mẫu thuộc nhóm thiểu số bằng các kĩ thuật khác nhau.

LỜI MỞ ĐẦU

1. Tính cấp thiết của đề tài

Báo cáo tài chính là công cụ thể hiện tình hình tài chính và kết quả hoạt động kinh doanh của doanh nghiệp. Báo cáo tài chính cung cấp các thông tin cần thiết về tình hình tài chính của doanh nghiệp cho các bên liên quan như chủ doanh nghiệp, nhà đầu tư, chủ nợ hay các cơ quan chức năng. Do đó, vấn đề liên quan đến sự minh bạch và trung thực của báo cáo tài chính của các công ty niêm yết trên thị trường chứng khoán luôn được quan tâm hàng đầu. Trong hơn 2 thập kỷ qua, gian lận trên báo cáo tài chính đã trở thành vấn đề nghiêm trọng không chỉ đối với doanh nghiệp, các nhà đầu tư, các chủ nợ và các bên liên quan khác, mà còn ảnh hưởng đến toàn bộ nền kinh tế do làm giảm niềm tin của các nhà đầu tư trong và ngoài nước, kéo theo việc giảm nguồn vốn đầu tư vào thị trường chứng khoán.Theo báo cáo của ACFE [2] vào năm 2016, các doanh nghiệp ở Hoa Kỳ đã mất 5% doanh thu hàng năm do gian lận. Thêm vào đó, tổng thiệt hại của 2.410 trường hợp gian lận là 6,3 tỷ USD và 23% trong số các trường hợp đã tốn thất 1 triệu USD [63].

Cũng theo ACFE, gian lận gây thiệt hại cho nền kinh tế Mỹ khoảng 6% trên tổng thu nhập quốc nội của năm 2004 (tương đương gần 12.000 tỷ USD) [1], tức là mỗi năm nền kinh tế này bị thiệt hại trên 720 tỷ USD do gian lận. Điển hình nhất là vụ gian lận báo cáo tài chính lớn nhất lịch sử của công ty Enron vào năm 2000, gây thiệt hại hơn 70 tỉ USD và kéo theo đó là sự sụp đổ của Arthur Andersen, một trong 5 công ty kiểm toán và kế toán lớn nhất toàn cầu ở thời điểm đó [7].

Tại Việt Nam, có thể kể đến những vụ gian lận báo cáo tài chính trong những năm qua như Bibica, Bông Bạch Tuyết, Gỗ Trường Thành. Chẳng hạn như đối với công ty Bông Bạch Tuyết (BBT), theo kết quả thanh tra số 1850/UBCK-QLPH ngày 12/9/2008 [65] của Ủy ban Chứng khoán Nhà nước, cho thấy công ty này đã công bố thông tin kết quả kinh doanh trong nhiều năm có sự chênh lệch trước và sau khi thực hiện kiểm toán, vi phạm quy định về việc công bố thông tin trên thị trường chứng khoán. Cụ thể, năm 2006, BCTC năm và các quý do BBT lập đều có lãi, nhưng thực tế thì công ty đã lỗ hàng tỉ đồng, trong khi đó, ý kiến của KTV độc lập đối với BCTC năm chỉ là ngoại trừ nhiều khoản mục trọng yếu. Do thua lỗ kéo dài mà cổ phiếu BBT bị chính thức hủy niêm yết trên sàn HoSE kể từ ngày 07/08/2009.

Đáng chú ý, những vụ gian lận báo cáo tài chính nói trên sau nhiều năm mới bị phát hiện, gây ra tâm lý nghi ngờ cho nhà đầu tư, ảnh hưởng tới hoạt động của thị trường. Các gian lận chậm bị phát hiện đã có tác động tiêu cực đến việc ra quyết định của các nhà đầu tư. Điều này có thể dẫn đến việc dòng tiền lưu thông không hiệu quả, gây ra nhiều hệ lụy cho cả nền kinh tế nếu không có những biện pháp khắc phục tình trạng này.

Ngoài ra, hiện nay cả MSCI và FTSE Russell - hai nhà cung cấp chỉ số uy tín nhất - đều đang theo dõi từng bước phát triển của thị trường Việt Nam để thực hiện đánh giá nâng hạng thị trường. Tháng 9/2018, Việt Nam đã được đưa vào danh sách theo dõi nâng hạng lên mức Thị trường mới nổi thứ cấp (Secondary Emerging Market). Điều đó đặt ra sự cần thiết phải cải thiện hệ thống giao dịch của các sàn giao dịch

chứng khoán và tính minh bạch của thị trường thông qua việc phát hiện các BCTC gian lận và có biện pháp xử lý kịp thời.

Hiện nay, trách nhiệm của kiểm toán viên liên quan đến gian lận trong báo cáo tài chính và yêu cầu kiểm toán viên đánh giá rủi ro gian lận đã được quy định tại Chuẩn mực kiểm toán số 240 (VSA 240) do Bộ Tài chính đã ban hành. Gian lận trong báo cáo tài chính dựa trên các yếu tố như động cơ / áp lực, cơ hội và thái độ hoặc khả năng hợp lý hóa (Bộ Tài chính, 2012). Tuy nhiên, vẫn còn nhiều khó khăn trong thực tế áp dụng các tiêu chuẩn này. Việc đánh giá rủi ro gian lận trong báo cáo tài chính phụ thuộc phần lớn vào xét đoán của kiểm toán viên. Trên thực tế các kiểm toán viên cần có một công cụ có hiệu quả cao và có thể dễ dàng hỗ trợ việc đưa ra phán đoán khả năng có gian lận trên báo cáo tài chính. Trong tình hình dịch COVID-19 vẫn còn diễn biến phức tạp, kiểm toán viên buộc phải làm việc tại nhà là chủ yếu, hạn chế đi đến công ty khách hàng, thì việc có một công cụ mạnh hỗ trợ cho việc xét đoán và phát hiện gian lận báo cáo tài chính càng trở nên cấp thiết.

Ngày nay, với sự phát triển của kỹ thuật và công nghệ đã giúp ngành công nghiệp tài chính đạt được sự tăng trưởng bùng nổ trong nhiều thập kỷ, và một lượng lớn dữ liệu đã được thu thập. Việc sử dụng các mô hình truyền thống khi phân tích gặp rất nhiều khó khăn khi xử lý dữ liệu lớn, do vậy nhiều dữ liệu thường bị bỏ qua gây lãng phí tài nguyên. Các thuật toán học máy (machine learning) có khả năng giải quyết điểm yếu của mô hình truyền thống vì nó có thể trích xuất thông tin hữu ích từ các tập dữ liệu lớn. Do đặc điểm này, machine learning có thể được áp dụng để tạo điều kiện thuận lợi cho kiểm toán viên trong việc trích xuất và khám phá các thông tin ẩn được ghi nhận trong khối lượng lớn dữ liệu một cách tối ưu và hiệu quả [23]. Bằng cách tiếp cận bài toán phân loại báo cáo tài chính gian lận dựa trên các thuật toán học máy, dữ liệu sẽ được khai thác và phân tích sâu hơn đồng thời mang lại nhiều thông tin hơn [33].

Vì những lý do trên, tác giả quyết định chọn đề tài "Xây dựng mô hình kết hợp các thuật toán học máy để dự đoán báo Báo cáo tài chính gian lận - Áp dụng cho thị trường chứng khoán Việt Nam" để nghiên cứu. Mục đích của đề tài là cung cấp một công cụ hiệu quả hơn để dự đoán khả năng gian lận báo cáo tài chính, định hướng kiểm tra các khoản mục có rủi ro gian lận trên báo cáo tài chính đã được xác định có khả năng gian lận cao nhằm tăng tính hữu hiệu và hiệu quả trong công việc của các kiểm toán viên độc lập. Ngoài ra, các nhà đầu tư, các cơ quan quản lý cũng có thể sử dụng công cụ này để đánh giá lại về báo cáo tài chính sau kiểm toán để đưa ra các quyết định phù hợp.

2. Tổng quan tình hình nghiên cứu về dự đoán và phát hiện gian lận báo cáo tài chính

Trên thế giới hiện nay đã có nhiều nghiên cứu về hành vi gian lận, cũng như xây dựng mô hình phát hiện gian lận trên báo cáo tài chính dựa trên các thuật toán học máy cho kết quả tốt và hiệu quả hơn so với các mô hình truyền thống [29, P.115)]. Do đó, tác giả tập trung tổng kết các nghiên cứu về phát hiện gian lận trên báo cáo tài chính bằng việc sử dụng các thuật toán học máy. Các nghiên cứu đi trước bao gồm cả các kĩ thuật phân tích định tính và định lượng để giải quyết bài toán phát hiện gian lận trên báo cáo tài chính [36], [62], như được trình bày ở bảng 1. Nghiên

cứu này tập trung vào phân tích định lượng để phát hiện và dự báo báo tài chính gian lận bằng sử dụng các kĩ thuật học máy.

Bảng 1. *Tổng quan các nghiên cứu trước*

Tác giả	**Tập dữ liệu sử dụng trong nghiên cứu** *(Số công ty gian lận/ Số công ty không gian lận)*	**Thuật toán tốt nhất**	**Đánh giá** *(Độ chính xác/ Độ nhạy)*
Fanning, Cogger (1998) [26]	102 / 102 (từ SEC)	ANN	62,5/66
Feroz et al (2000) [28]	38 / 38 (từ Hy Lạp)	ANN	69,72/72,2
Beneish (1999) [12]	50 / 1708	Unweighted profit	89,5/54,2
Kirkos et al (1997) [40]	38 / 38 (từ Hy Lạp)	Bayesian Belief Network	90,3/91,7
Hoogs et al (2007) [34]	51 / 339 (từ SEC)	Genetic algorithms	95/63
Humpherys et al. (2011) [35]	101 / 101 (từ Mỹ)	SVM	65,8/71,3
Ravisankar (2011) [55]	101 / 101 (từ Trung Quốc)	Probabilistic NN	90,77/87,53
Kotsiantis et al. (2006) [41]	41 / 123 (từ Hy Lạp)	Stacking	95,1/90,02
Cecchini et al (2010) [16]	205 / 6427 (từ Mỹ)	SVM	90,4/80

Từ bảng 1, chúng ta có thể thấy rằng có rất nhiều nỗ lực để giải quyết vấn đề phát hiện gian lận tài chính. Tuy nhiên, các nghiên cứu đi trước chỉ cho kết quả tốt với tập dữ liệu nhỏ dẫn đến khó khăn khi ứng dụng mô hình vào giải quyết các vấn đề phát hiện gian lận tài chính, đặc biệt là các vấn đề trong thế giới thực nơi có hàng trăm, hàng nghìn báo cáo tài chính niêm yết cần được đưa ra dự đoán. Do đó, tác giả đã đề xuất sử dụng các kỹ thuật khác để cải thiện hiệu suất giải các vấn đề phát hiện gian lận tài chính đã cho, chẳng hạn như thuật toán Boosting và ý tưởng về phương pháp học tập kết hợp (Ensemble model).

3. Khoảng trống nghiên cứu và điểm mới của bài nghiên cứu

Một số đặc điểm khác biệt làm cho lĩnh vực gian lận báo cáo tài chính trở nên đặc biệt là sự mất cân bằng trong việc phân loại và chi phí khi phân loại nhầm một báo

cáo tài chính gian lận thành không gian lận [53]. Một cách tiếp cận ở cấp độ dữ liệu để giải quyết vấn đề này là sử dụng các phương pháp dựa trên lấy mẫu để cân bằng số lượng ví dụ giữa các lớp đa số và thiểu số.

Ngoài ra, có thể thấy được rằng các nghiên cứu đi trước đều sử dụng tập dữ liệu là các công ty đã được xác định là gian lận bởi các tổ chức đáng tin cậy như SEC, tuy nhiên lượng dữ liệu được dùng để đào tạo mô hình vẫn còn khá nhỏ. Hiện tại ở Việt Nam không có sẵn tập dữ liệu thống kê các vụ gian lận báo cáo tài chính đã được phát hiện. Do đó cần sử dụng các chỉ báo đã được nghiên cứu và xác nhận về mức độ phù hợp trên thế giới để dán nhãn cho một công ty là có gian lận hay không.

Nghiên cứu này cũng cố gắng khắc phục hạn chế của nghiên cứu đi trước bằng cách đề xuất một phương pháp tổng hợp sử dụng SMOTEENN, một phương pháp kết hợp để xử lý sự mất cân bằng giữa các lớp trong dữ liệu. Đồng thời đề xuất một phương pháp gắn nhãn đáng tin cậy sẽ được trình bày ở chương 3 để xây dựng tập dữ liệu gian lận dành cho thị trường Việt Nam. Trong nghiên cứu này, tác giả đã đánh giá kết quả của việc áp dụng các thuật toán bagging như DT, SVM, KNN và cả việc sử dụng các thuật toán boosting như XGBoost và LightGBM. Tác giả nhận thấy rằng XGBoost, LightGBM và RF đạt được hiệu suất vượt trội hơn hẳn các thuật toán còn lại trong việc phát hiện báo cáo tài chính gian lận. Từ đó, tác giả tiến hành xây dựng một mô hình ensemble dựa trên ba thuật toán là XGBoost [18], LightGBM [38] và RF [42]. Mô hình ensemble đã được huấn luyện và được áp dụng để đưa ra dự đoán cho tập dữ liệu gồm báo cáo tài chính của các công ty chưa kiểm toán trên các sàn HNX, HoSE và UpCOM trong năm 2020.

Các điểm mới trong đề tài nghiên cứu của tác giả bao gồm:

(1) Tích hợp, phân tích, nghiên cứu từ nhiều mô hình, lý thuyết của các tác giả đi trước để có thể xây dựng mô hình tổng quan phù hợp với những đặc điểm riêng của thị trường Việt Nam.
(2) Nghiên cứu đề xuất phương pháp gắn nhãn các báo cáo tài chính một cách đáng tin cậy để làm đầu vào cho mô hình học máy.
(3) Nghiên cứu đã sử dụng nguồn dữ liệu với quy mô lớn nhằm mang tính dự báo tốt hơn để đánh giá gian lận trên báo cáo tài chính.
(4) Nghiên cứu đã đánh giá và so sánh các thuật toán học máy và lựa chọn các thuật toán tối ưu, phù hợp với dữ liệu của thị trường Việt Nam. Từ đó, kết hợp các thuật toán đơn lẻ lại với nhau thành mô hình ensemble cuối cùng nhằm tăng khả năng nhận diện báo cáo tài chính gian lận tốt hơn.
(5) Nghiên cứu đã đưa ra các chỉ số tài chính quan trọng được trích chọn từ mô hình học máy nhằm gợi ý các khoản mục quan trọng mà các KTV cần chú trọng khi thực hiện kiểm toán.

4. Mục tiêu nghiên cứu

Mục tiêu nghiên cứu được xác định là xây dựng một mô hình học máy có khả năng dự đoán một công ty có gian lận hay không dựa trên báo cáo tài chính trước kiểm toán với hiệu quả cao, và loại trừ các khả năng bị overfitting hay underfitting.

Cụ thể, nghiên cứu hướng đến các vấn đề cụ thể như sau:

- Hệ thống hóa các cơ sở lý thuyết liên quan đến gian lận trên báo cáo tài chính, các cách tiếp cận máy học từ các nghiên cứu trong và ngoài nước.
- Xây dựng mô hình dựa trên các thuật toán học máy, tập trung vào việc phát hiện gian lận với hiệu quả cao.
- Cải thiện mô hình đến khi đạt khả năng dự báo cao.
- Giải thích cách ứng dụng kết quả nghiên cứu.

5. Đối tượng và phạm vi nghiên cứu

Đối tượng: mô hình dự đoán gian lận trên báo cáo tài chính dựa trên thuật toán học máy

Phạm vi nghiên cứu bao gồm không gian và thời gian nghiên cứu:

- Không gian: Nhóm nghiên cứu xây dựng mô hình phát hiện gian lận trên báo cáo tài chính chưa được kiểm toán của các công ty có niêm yết trên sàn chứng khoán HOSE, HNX và UpCOM. Nhóm nghiên cứu sử dụng dữ liệu thuộc hệ thống Fiinpro của công ty FiinGroup JSC (trước đây là StoxPlus JSC), được Trường Đại học Kinh tế - Đại học Đà Nẵng cung cấp cho Giảng viên và Sinh viên nhằm đáp ứng yêu cầu giảng dạy và nghiên cứu khoa học.
- Thời gian: Nghiên cứu thu thập báo cáo tài chính chưa được kiểm toán của các doanh nghiệp trong giai đoạn 2010-2019.

6. Ý nghĩa khoa học và thực tiễn của đề tài

Kết quả nghiên cứu của đề tài này sẽ cung cấp một công cụ dự đoán báo cáo tài chính gian lận với chi phí thấp để hỗ trợ kiểm toán viên (KTV) trong công việc của mình. Với báo cáo tài chính chưa kiểm toán đã được mô hình dự báo là có gian lận, kiểm toán viên sẽ được gợi ý để tập trung kiểm tra các khoản mục có khả năng gian lận cao, từ đó tăng tính hữu hiệu và hiệu quả của công việc kiểm toán. Những người sử dụng báo cáo tài chính sau kiểm toán như các nhà đầu tư, chủ nợ, các cơ quan quản lý cũng có thể sử dụng mô hình như công cụ để đánh giá báo cáo tài chính còn có khả năng gian lận hay không, từ đó có thể đưa ra quyết định đúng đắn và kịp thời.

7. Kết cấu của đề tài nghiên cứu

Bên cạnh lời mở đầu, kết luận, mục lục, danh mục từ viết tắt, danh mục bảng, hình vẽ, danh mục tài liệu tham khảo, và các phụ lục, đề tài bao gồm 4 chương:

Chương 1: Cơ sở lý luận về gian lận báo cáo tài chính và các nghiên cứu nhằm nhận diện báo cáo tài chính gian lận

Chương 2: Quy trình nghiên cứu

Chương 3: Kết quả nghiên cứu và thảo luận các kết quả

Chương 4: Kết luận

CHƯƠNG 1: CƠ SỞ LÝ LUẬN VỀ GIAN LẬN BÁO CÁO TÀI CHÍNH VÀ CÁC NGHIÊN CỨU NHẰM NHẬN DIỆN BÁO CÁO TÀI CHÍNH GIAN LẬN

1.1. Báo cáo tài chính gian lận

Tính xác thực và minh bạch của báo cáo tài chính công ty niêm yết cho phép người sử dụng báo cáo tài chính có được thông tin đáng tin cậy và đưa ra các quyết định hợp lý. Gian lận báo cáo tài chính khiến cho các nhà đầu tư, các chủ nợ đưa ra các quyết định sai lầm và chịu thiệt hại nặng nề về tài chính. Các vụ gian lận báo cáo tài chính của các công ty niêm yết bị phát hiện sau khi báo cáo tài chính đã được kiểm toán và công bố đã gây mất lòng tin vào các công ty kiểm toán liên quan và tính minh bạch của thị trường chứng khoán. Vì vậy, nghiên cứu nhằm dự đoán, phát hiện báo cáo tài chính gian lận luôn là một chủ đề được quan tâm nhiều nhất từ trước đến nay. Chương này trình bày định nghĩa về gian lận báo cáo tài chính, các nghiên cứu nhằm nhận diện gian lận báo cáo tài chính và các thuật toán học máy đã được sử dụng nhằm dự đoán hoặc phát hiện báo cáo tài chính gian lận.

1.2. Khái niệm

Gian lận báo cáo tài chính có thể được mô tả trong nhiều bài báo và sách đã xuất bản, hầu hết là "một nỗ lực cố ý của các công ty nhằm đánh lừa hoặc gây hiểu lầm cho người sử dụng báo cáo tài chính đã công bố, đặc biệt là các nhà đầu tư và chủ nợ, bằng cách lập và công bố báo cáo tài chính có sai sót trọng yếu" [25]. Gian lận báo cáo tài chính được thực hiện bởi một nhóm thủ phạm khéo léo, có thể là quản trị viên cấp cao và KTV, với mục đích gây hiểu lầm và làm sai lệch thông tin tài chính với các kế hoạch đã được hoạch định trước nhằm huy động vốn đầu tư hoặc kiểm soát công ty một cách gian dối. VSA 240, đoạn 11 định nghĩa rằng “gian lận là hành vi cố ý do một hay nhiều người trong Ban quản trị, Ban Giám đốc, các nhân viên hoặc bên thứ ba thực hiện bằng các hành vi gian dối để thu lợi bất chính hoặc bất hợp pháp”.

Theo chuẩn mực kiểm toán AS2401 do PCAOB ban hành [6], một báo cáo tài chính gian lận có thể thuộc các trường hợp sau:

(1) Làm sai lệch, thay đổi hoặc thao túng trọng yếu hoạt động tài chính bằng các tài liệu hỗ trợ đã được sửa đổi;

(2) Cố ý trình bày sai, bỏ sót hoặc trình bày sai bản chất của các giao dịch, tài khoản tài chính và dữ liệu liên quan để điều chỉnh thông tin tài chính;

(3) Cố ý áp dụng sai, hiểu sai các chuẩn mực kế toán và chính sách ngành liên quan đến báo cáo kết quả hoạt động kinh doanh;

(4) Cố ý bỏ qua và tiết lộ hoặc trình bày sai lệch về tài chính đối với các chuẩn mực và chính sách kế toán liên quan đến thông tin tài chính liên quan;

(5) Quản lý bất hợp pháp thu nhập bằng cách áp dụng các phương pháp kế toán.

Gian lận báo cáo tài chính gây thiệt hại nặng nề cho công ty, các nhà đầu tư, các chủ nợ và cả sự phát triển của nền kinh kế quốc gia. Các báo cáo tài chính gian lận ngay

lập tức làm giảm lợi ích của các cổ đông đầu tư do họ có xu hướng phải gánh chịu các khoản thất thoát tài chính nếu báo cáo gian lận đó dẫn đến việc công ty gặp khó khăn về tài chính hoặc thậm chí phá sản. Về mặt thống kê, trong khi các khoản lỗ tài chính trực tiếp được báo cáo hợp pháp, các chi phí gián tiếp khác lại đang bị đánh giá thấp hơn, cụ thể là chi phí pháp lý, lợi ích của khách hàng và người lao động, niềm tin của nhà cung cấp và các phản ứng của thị trường với thông tin của công ty. Các nghiên cứu trước đây đã chỉ ra rằng tổng chi phí tiềm ẩn của các báo cáo tài chính gian lận được ước tính một cách định lượng, chủ yếu là vì bên cạnh các gian lận tài chính đã được công bố, không phải tất cả các gian lận đều được phát hiện sớm và được báo cáo hợp pháp.

1.3. Các nghiên cứu nhằm nhận diện gian lận trên báo cáo tài chính

Gian lận báo cáo tài chính là một vấn đề được các nhà hoạch định chính sách và nhà đầu tư rất quan tâm, vì nó ảnh hưởng trực tiếp đến môi trường kinh doanh và đầu tư. Nhiều nghiên cứu chỉ ra rằng, gian lận báo cáo tài chính khá phổ biến, và được thực hiện ngày càng tinh vi. Việc tìm cách phát hiện báo cáo tài chính gian lận luôn là thách thức lớn đối với chuyên gia trong lĩnh vực tài chính, kế toán, trước hết là các kiểm toán viên độc lập.

Theo [24], khoảng 20% công ty niêm yết tại Mỹ thường xuyên thực hiện các thủ thuật nhằm nâng cao số liệu lợi nhuận, trong khi theo [14] thậm chí còn cho rằng tỷ lệ các công ty thao túng dữ liệu báo cáo tài chính trên thị trường Mỹ cao gần 38%.

Theo [57], số liệu báo cáo tài chính bị sai lệch xuất phát từ hai nguyên nhân chính. Thứ nhất, báo cáo tài chính có nhiều nội dung phải sử dung ước tính kế toán nên rất dễ xảy ra sai lệch số liệu. Thứ hai, sai lệch về số liệu kế toán đến từ việc Ban Giám đốc chủ động lựa chọn phương pháp kế toán dồn tích để tăng lợi nhuận hoặc che giấu tình hình tài chính tồi tệ của công ty. Các công ty niêm yết có thể trì hoãn hoặc đẩy nhanh việc ghi nhận chi phí, lãi vay và doanh thu hoặc thay đổi phương pháp ghi nhận hàng tồn kho, định giá tài sản và ghi nhận tín dụng khác với các chuẩn mực kế toán. Ngoài ra, các công ty còn có thể chuyển lợi nhuận sang các kỳ kế toán trong tương lai để thu nhập ổn định hơn qua các năm, từ đó nâng cao vị thế và hình ảnh của công ty trên thị trường chứng khoán.

Các báo cáo tài chính gian lận làm cho thị trường chứng khoán kém minh bạch, và các quyết định của nhà đầu tư dựa trên phân tích tài chính trở nên kém hiệu quả. Ở một mức độ xa hơn, những công ty có hành vi gian lận gây ra thiệt hại nghiêm trọng cho nhà đầu tư và cổ đông khi không được phát hiện kịp thời. Theo [31], chỉ có 20,2% sai phạm kế toán được phát hiện bởi kiểm toán viên nội bộ hoặc kiểm toán viên độc lập, cho thấy tính hữu hiệu của các kỹ thuật kiểm toán chưa cao.

Các nghiên cứu nổi bật nhằm nhận diện báo cáo tài chính gian lận trong vài thập kỷ qua bao gồm công trình nghiên cứu của các tác giả Edward Altman [3], [4], [5] và Beneish [12]. Các tác giả đã xây dựng các chỉ số hữu ích trong việc nhận diện báo cáo tài chính gian lận như được trình bày dưới đây.

1.3.1. Altman's Z-score

Z-Score được Edward Altman đưa ra vào năm 1968 [3] để dự đoán khả năng phá sản và mất khả năng thanh toán của một công ty trong 2 năm tới. Một loạt các nghiên cứu tiếp theo đã được thực hiện trong suốt 30 năm cho đến năm 1985, Z-Score đã được chấp nhận rộng rãi bởi các KTV, kế toán quản lý và hệ thống xếp hạng tín dụng. Vào năm 1999, 80-90% các công ty bị phá sản được dự báo bằng cách sử dụng chỉ số Z-Score. Ngoài việc dự đoán khả năng phá sản của một công ty, sau này, nhiều nghiên cứu đã chỉ ra rằng Z-Score được sử dụng hiệu quả trong việc xác định khả năng gian lận trên báo cáo tài chính với độ chính xác cao.

Theo [3], các nhà quản lý có thể sử dụng mô hình Z-Score để dự đoán các vấn đề quản lý, đặc biệt là đưa ra các quyết định liên quan đến quản lý tài chính kịp thời nhằm khắc phục các vấn đề phát sinh và ngăn chặn tình trạng phá sản của doanh nghiệp. Thứ nhất, các nghiên cứu trước đây chỉ dự đoán đúng 72% rủi ro phá sản của một doanh nghiệp và tập trung vào các doanh nghiệp sản xuất. Một loạt các nghiên cứu tiếp theo sau đó đã diễn ra và xác nhận rằng Z-Score cũng có thể dự đoán được đối với các công ty phi sản xuất. Altman và Hotchkiss [5] đã nghiên cứu những thay đổi trong chỉ số Z-Score để dự đoán chính xác bất kể ngành nghề và loại hình kinh doanh nào và phát hiện hiệu quả khả năng gian lận trên báo cáo tài chính. Balcaen và Oogle đã tuyên bố rằng mặc dù đã ra đời cách đây nhiều năm, mô hình chỉ số Z vẫn là công cụ dự đoán được cả giới học thuật và cộng đồng thực hành trên thế giới công nhận và sử dụng rộng rãi nhất [8].

Công thức tính chỉ số Z-Score cho thị trường mới nổi như Việt Nam như sau (Altman và Hotchkiss, 2006) [5]:

$$Z - Score = 3.25 + 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4$$

Trong đó,

X_1 : Vốn lưu động / tổng tài sản
X_2 : Lợi nhuận giữ lại / tổng tài sản
X_3: Thu nhập trước lãi vay và thuế (EBIT) / tổng tài sản
X_4: Giá trị thị trường của vốn chủ sở hữu / tổng tài sản
Ý nghĩa của độ lớn điểm Z:
Z > 5.85: Công ty ít có khả năng đi đến phá sản và được coi là có sức khỏe tài chính tốt.
4.35 < Z ≤ 5.85: Công ty được coi là có nguy cơ phá sản vừa phải ở cả lĩnh vực phi sản xuất và các công ty ở các thị trường mới nổi.
Z ≤ 4.35: Công ty có nguy cơ lớn dẫn đến phá sản trong vài năm tới hoặc cho thấy tình trạng tài chính kém.

1.3.2. Beneish's M-score

Mô hình Beneish (1999) [12] là một mô hình kinh tế lượng tài chính, bao gồm 8 tỷ số tài chính được tính toán từ các báo cáo tài chính. Mô hình này được coi là đưa ra dự báo về rủi ro gian lận báo cáo tài chính với độ chính xác xấp xỉ 76%.
Mô hình cung cấp một công thức toán học để tính điểm M như sau:

$$M - Score = -4.84 + 0.920 \times DSRI + 0.528 \times GMI + 0.404 \times AQI + 0.892 \times SGI + 0.115 \times DEPI - 0.172 \times SGAI + 4.679 \times TATA - 0.327 \times LEVI$$

In which:

- *Chỉ số phải thu khách hàng so với doanh thu* (DSRI)

$$\text{DSRI} = \frac{\text{Khoản phải thu}_t \Big/ \text{Doanh thu thuần}_t}{\text{Khoản phải thu}_{t-1} \Big/ \text{Doanh thu thuần}_{t-1}}$$

- *Chỉ số tỷ lệ lãi gộp (GMI)*

$$\text{GMI} = \frac{(\text{Doanh thu thuần}_{t-1} - \text{Giá vốn hàng bán}_{t-1}) \Big/ \text{Doanh thu thuần}_{t-1}}{(\text{Doanh thu thuần}_t - \text{Giá vốn hàng bán}_t) \Big/ \text{Doanh thu thuần}_t}$$

- *Chỉ số chất lượng tài sản (AQI)*

$$\text{AQI} = \frac{1 - (\text{TSNH}_t + \text{PP\&E}_t + \text{Chứng khoán nắm giữ}_t) \Big/ \text{Tổng tài sản}_t}{1 - (\text{TSNH}_{t-1} + \text{PP\&E}_{t-1} + \text{Chứng khoán nắm giữ}_{t-1}) \Big/ \text{Tổng tài sản}_{t-1}}$$

- *Chỉ số tăng trưởng doanh thu bán hàng (SGI)*

$$\text{SGI} = \frac{\text{Doanh thu thuần}_t}{\text{Doanh thu thuần}_{t-1}}$$

- *Chỉ số tỷ lệ khấu hao (DEPI)*

$$\text{DEPI} = \frac{\text{Khấu hao}_{t-1} \Big/ (\text{PP\&E}_{t-1} + \text{Khấu hao}_{t-1})}{\text{Khấu hao}_t \Big/ (\text{PP\&E}_t + \text{Khấu hao}_t)}$$

- *Chỉ số chi phí bán hàng và quản lý doanh nghiệp (SGAI)*

$$\text{SGAI} = \frac{\text{Chi phí SG\&A}_t \Big/ \text{Doanh thu thuần}_t}{\text{Chi phí SG\&A}_{t-1} \Big/ \text{Doanh thu thuần}_{t-1}}$$

- *Chỉ số đòn bẩy tài chính (LVGI)*

$$\text{LVGI} = \frac{(\text{Nợ ngắn hạn}_t + \text{Nợ dài hạn}) \Big/ \text{Tổng tài sản}_t}{(\text{Nợ ngắn hạn}_{t-1} + \text{Nợ dài hạn}_{t-1}) \Big/ \text{Tổng tài sản}_{t-1}}$$

- *Chỉ số biến dồn tích so với tổng tài sản (TATA)*

$$\text{TATA} = \frac{\text{Thu nhập từ hoạt động kinh doanh liên tục}_t - \text{Lưu chuyển tiền thuần từ hoạt động kinh doanh}_t}{\text{Tổng tài sản}_t}$$

Các biến được xây dựng từ dữ liệu trong báo cáo tài chính của công ty và sau khi được tính toán, điểm số M được sử dụng để đo lường mức độ thao túng báo cáo tài chính. Trong khi giá trị điểm M dưới -2,22 cho thấy công ty sẽ không có nguy cơ bị thao túng, giá trị lớn hơn -2,22 cho thấy công ty có khả năng lập báo cáo tài chính sai lệch. Điều này có nghĩa là có khả năng cao công ty có hành vi thao túng báo cáo tài chính.

Các chỉ số Z-score và M-score đã được sử dụng khá rộng rãi trong các nghiên cứu về gian lận báo cáo tài chính và chứng minh được khả năng nhận diện báo cáo tài chính gian lận với kết quả khá tốt. Charalambos T. Spathis [40] đã kiểm tra tính hữu ích của các chỉ số tài chính và tỷ lệ rủi ro phá sản (Z-Score) trong việc phát hiện gian lận trên báo cáo tài chính. Mô hình đã xác định các chỉ số tài chính sau hữu ích trong việc phát hiện các báo cáo tài chính gian lận: tỷ lệ hàng tồn kho trên doanh thu, lợi nhuận ròng trên tổng tài sản, tỷ lệ vốn lưu động trên tổng tài sản, tổng nợ trên tổng tài sản và tỷ lệ rủi ro phá sản (Z-Score). Kết quả trên cho thấy Z-Score rất hữu ích trong việc phát hiện gian lận, tương tự như nghiên cứu của [3].

Lalith P. Samarakoon & Tanweer Hasan [59] đã dự đoán cuộc khủng hoảng tài chính của 26 công ty niêm yết tại một thị trường mới nổi bằng cách sử dụng ba mô hình Z-Score của Altman. Kết quả nghiên cứu cho thấy Z-Score đã dự đoán đúng 81% các công ty gặp khó khăn về tài chính. Nghiên cứu được thực hiện tại một thị trường mới nổi này đã cung cấp thêm bằng chứng thực nghiệm rằng Z-Score rất có khả năng dự đoán tình hình tài chính của các công ty.

Dalnial và cộng sự. [22] đã nghiên cứu sự khác biệt về giá trị của các chỉ tiêu tài chính giữa báo cáo tài chính gian lận và không gian lận. Kết quả nghiên cứu cũng cho thấy chỉ số Z-Score có ảnh hưởng đáng kể đến việc phát hiện khả năng có gian lận báo cáo tài chính.

Các nghiên cứu nói trên cho thấy Z-Score không chỉ dự đoán ban đầu sự phá sản hay khó khăn về tài chính của một doanh nghiệp mà còn có khả năng dự đoán các báo cáo tài chính gian lận.

Bên cạnh đó, từ việc ứng dụng M-Score, Rasa Kanapickiene và Zivile Grundiene [37] đã thực hiện một nghiên cứu thực nghiệm để xây dựng mô hình phân loại đạt 84,8% độ chính xác đối với mẫu nghiên cứu.

Repousis [56] đã ứng dụng Beneish M-Score trên tập dữ liệu gồm hơn 25000 doanh nghiệp ở Hy Lạp và kết quả cho thấy 8.486 công ty hay 33% của toàn bộ mẫu có M-Score lớn hơn -2,2, đây là một tín hiệu cho thấy các công ty có khả năng thao túng báo cáo tài chính.

Muntari Mahama [45] đã thực hiện nghiên cứu thực nghiệm về trường hợp gian lận của Công ty Enron, trong đó tác giả sử dụng mô hình M-Score kết hợp với Z-Score để phát hiện khả năng gian lận báo cáo tài chính và khủng hoảng tài chính của công ty này. Nghiên cứu phát hiện ra rằng Enron đã rơi vào khủng hoảng tài chính từ lâu (1997) nhưng mãi đến năm 2001, hãng mới tuyên bố phá sản với thu nhập được điều chỉnh một cách có chủ ý.

Năm 2017, Maccarthy dựa trên việc phân tích vụ bê bối của Enron đã chỉ ra rằng, việc chỉ dựa vào phán đoán cũng như các kết luận chuyên môn của các kiểm toán viên, các nhà phân tích hay những đối tượng khác mà thị trường tin tưởng là không đủ để phát hiện gian lận báo cáo tài chính. Thông qua kết quả nghiên cứu của mình [44], tác giả Maccarthy khuyến khích sử dụng kết hợp Altman Z-score và Beneish M-Model để phát hiện khả năng gian lận trên báo cáo tài chính.

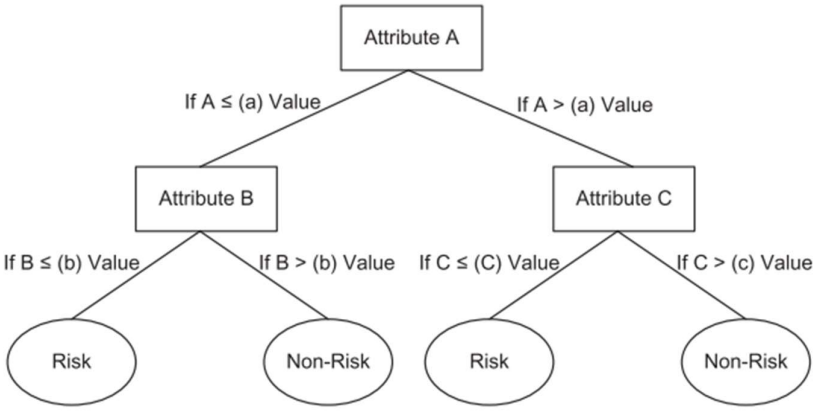
1.4. Các thuật toán học máy được sử dụng trong các nghiên cứu dự đoán gian lận báo cáo tài chính

1.4.1. Các thuật toán Bagging

Bagging là một phương pháp giảm phương sai dự đoán bằng cách tạo thêm dữ liệu để đào tạo từ tập dữ liệu bằng cách sử dụng các kết hợp có lặp lại để tạo ra nhiều bộ dữ liệu gốc [68].

1.4.1.1. Decision tree

Đây là một thuật toán học có giám sát [49]. DT có dạng cấu trúc hình cây (Hình 1). Cây bao gồm nút gốc và các nút khác được chia theo cách nhị phân hoặc đa tách thành các nút con. Sau đó, mỗi cây sử dụng thuật toán riêng của nó để thực hiện quá trình tách, cho đến khi không cần tách nữa sẽ tạo ra sự khác biệt trong mô hình của tác giả, liên kết mỗi thuộc tính với một giá trị là biến đầu vào liên quan đến phương thức đang được sử dụng như mô tả của Y. Sahin và E. Duman [58].



Hình 1. Ví dụ về DT

1.4.1.2. Random Forest

Do tính không ổn định của các cây đơn lẻ và sự nhạy cảm của chúng với một số dữ liệu huấn luyện, một mô hình mới gọi là rừng ngẫu nhiên đã được phát triển. Rừng ngẫu nhiên có hiệu quả tính toán cao hơn vì mỗi cây được xây dựng độc lập với các cây khác [32]. Về cơ bản, nó là một tập hợp các cây hồi quy và / hoặc phân loại có

được phương sai giữa các cây của chúng và do đó dễ sử dụng do chỉ sử dụng hai nguồn ngẫu nhiên hoặc các tham số được sử dụng để xây dựng cây bằng cách sử dụng dữ liệu được đào tạo riêng biệt khởi động cùng với các mẫu với chỉ một tập hợp con thuộc tính dữ liệu ngẫu nhiên để xây dựng từng cây như được chỉ định [13].

1.4.1.3. Logistic Regression

Khi biến phụ thuộc là biến nhị phân, đây là một kỹ thuật thích hợp để phân tích dự đoán [52]. Kỹ thuật này có thể được sử dụng vì biến phân loại các giao dịch là gian lận hay không là một biến nhị phân. Tập dữ liệu được sử dụng làm đầu vào cho mô hình được phân loại trước khi nó được sử dụng để huấn luyện và kiểm tra mô hình. Sau khi mô hình đã được đào tạo, nó được đưa vào thử nghiệm để xem liệu có giá trị ngưỡng giới hạn tối thiểu để dự đoán hay không. Các biến thể quan trọng nhất sau đó được chọn và mô hình được tinh chỉnh cho phù hợp. Do đó, các điểm ngoại lệ không được xử lý hiệu quả [51]. Nó tính toán xác suất và cho thấy rằng kết quả thuộc một loại cụ thể bằng cách sử dụng hàm logarit tự nhiên [13].

1.4.1.4. Support Vector Machine

Như đã nêu trong [13], SVM là bộ phân loại tuyến tính hoạt động ở độ cao nhiều chiều vì tác vụ phi tuyến tính trong đầu vào trở thành tuyến tính trong không gian đa chiều, làm cho SVM cực kỳ hữu ích để phát hiện gian lận. Nó có khả năng tổng quát hóa cao do có hai đặc điểm quan trọng nhất: một hàm nhân để biểu diễn chức năng phân loại trong sản phẩm chấm của phép chiếu điểm dữ liệu đầu vào và thực tế là nó cố gắng tìm một siêu mặt phẳng để tối đa hóa sự tách biệt giữa các lớp đồng thời giảm thiểu việc trang bị quá nhiều dữ liệu đào tạo [47], [48].

1.4.1.5. K Nearest Neighbors

Đây là một kỹ thuật học tập có giám sát thường cho kết quả tốt hơn các kỹ thuật phát hiện gian lận dựa trên thống kê khác [71]. Hiệu suất của nó bị ảnh hưởng nhiều bởi ba yếu tố: khoảng cách để xác định các láng giềng xa nhất, một số quy tắc để suy ra phân loại từ một láng giềng gần nhất thứ k và số lượng láng giềng để gắn nhãn mẫu mới [39]. Thuật toán này phân loại bất kỳ báo cáo tài chính nào xảy ra theo điểm gần nhất so với báo cáo tài chính cụ thể này và nếu người hàng xóm ít xa nhất này bị coi là gian lận, báo cáo tài chính mới cũng bị coi là gian lận. Khoảng cách Euclide là một lựa chọn tuyệt vời để tính toán khoảng cách trong trường hợp này. Kỹ thuật này nhanh chóng và tạo ra các cảnh báo lỗi. Tối ưu hóa chỉ số khoảng cách có thể cải thiện hiệu suất của nó [19].

1.4.2. Các thuật toán Boosting

Boosting là tập hợp một nhóm các thuật toán học yếu (weak leaners). Các bản ghi được phân loại trực tiếp thông qua thuật toán tăng cường bằng cách đánh trọng số cao hơn. Những thuật toán học yếu này sau đó được kết hợp lại để tạo thành một mô hình mạnh hơn (a single strong learner) [60].

1.4.2.1. eXtreme Gradient Boosting (XGBoost)

XGBoost cải tiến loss function và mở rộng đa thức Taylor bậc hai của nó có thể cải thiện tính gần đúng hàm mất mát của mô hình. Hơn nữa, XGBoost là một mô hình cây phân tán hỗ trợ tính toán quy mô lớn và phần lớn các đầu vào trong các ứng dụng thực tế là rất thưa thớt. Về vấn đề này, XGboost sử dụng thuật toán cảm biến thưa thớt, có thể chấp nhận một lượng lớn dữ liệu thưa thớt và thực hiện các phép tính một cách hiệu quả. Theo các nghiên cứu, thuật toán cảm biến thưa thớt nhanh hơn hàng trăm lần so với các phương pháp truyền thống. Hơn nữa, XGBoost không chỉ xử phạt số lượng lá mà còn cả trọng số của lá khi cắt tỉa. Bằng cách sử dụng tham số điều chỉnh, có thể giảm phương sai của mô hình, ngăn chặn việc overfitting, kiểm soát độ phức tạp của mô hình và đơn giản hóa mô hình đào tạo [18]. Khi so sánh với mô hình tuyến tính, XGBoost cũng có thể giải quyết các đặc điểm về kích thước khác nhau, cũng như xử lí tốt các ngoại lệ.

1.4.2.2. Light Gradient Boosted Machine (LightGBM)

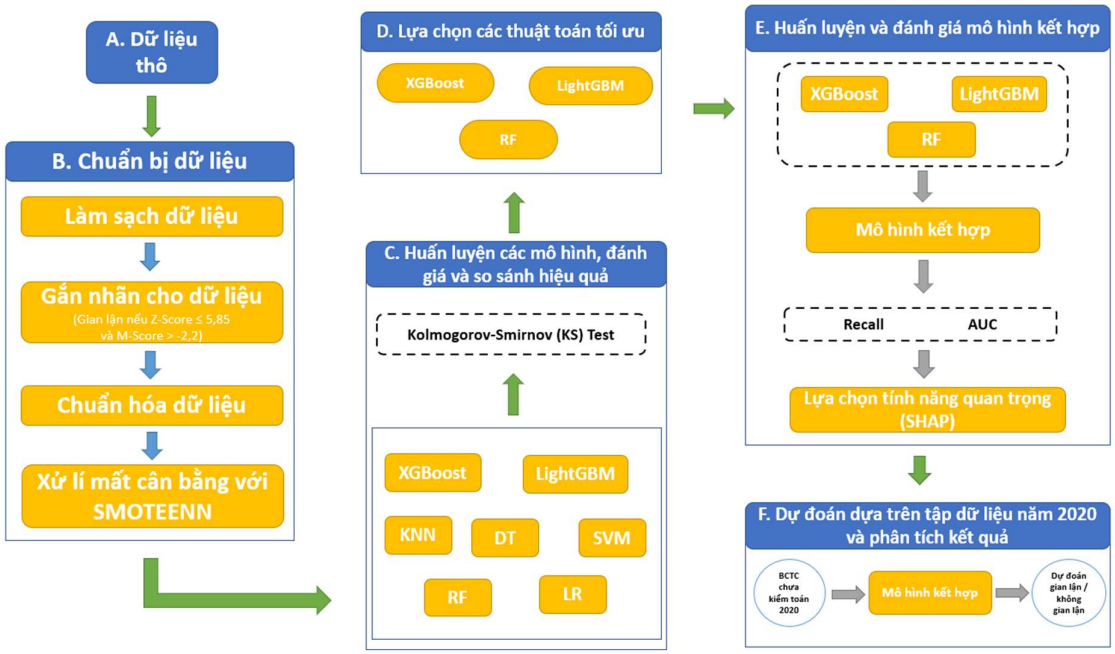
LightGBM là một FrameWork để xử lý thuật toán tăng cường độ dốc (Gradient Boosting) được phát triển bởi Microsoft. Gradient Boosting là một thuật toán xuất phát từ thuật toán DT, nó thực hiện việc xây dựng tuần tự nhiều cây quyết định và tiến hành học tập. LightGBM sử dụng công nghệ GOSS (Gradient base One-Side Sampling), để nâng cao lợi thế của nó trong việc xử lý lượng lớn dữ liệu. Trong hầu hết các trường hợp, hiệu suất của mô hình được đào tạo bởi thuật toán GOSS tốt hơn so với hiệu suất của thuật toán lấy mẫu ngẫu nhiên thông thường. Mặt khác, phương pháp GOSS cũng làm tăng tính đa dạng của các mô hình cơ sở (base learner), từ đó cải thiện khả năng tổng quát hóa của mô hình. Ngoài ra, nó sử dụng chiến lược tăng trưởng khôn ngoan của lá (leaf) để giảm số lượng tính toán và ngăn ngừa overfitting bằng cách kiểm soát độ sâu tối đa của cây (depth). Bên cạnh đó, LightGBM còn hỗ trợ tính toán song song đem lại hiệu quả cao. Những ưu điểm này giúp LightGBM giảm đáng kể thời gian tính toán mà vẫn đảm bảo hiệu suất cao, đồng thời nó cũng đặc biệt xuất sắc trong các tác vụ phân loại và dự đoán [38].

Kết luận chương 1

Trong chương 1, tác giả đã trình bày rõ về khái niệm gian lận báo cáo tài chính theo VSA240 và IAS240, đồng thời nêu ra các nghiên cứu nhằm nhận diện gian lận trên báo cáo tài chính bao gồm hai mô hình Altman's Z-Score và Beneish M-Score. Ngoài ra, nghiên cứu đồng thời tổng hợp các thuật toán học máy được sử dụng phổ biến cho bài toán phân loại nói chung hay bài toán nhận diện báo cáo tài chính gian lận nói riêng để tạo tiền đề cho việc xây dựng mô hình nhằm dự đoán báo cáo gian lận tài chính.

CHƯƠNG 2: QUY TRÌNH NGHIÊN CỨU

2.1. Sơ đồ quy trình thực hiện nghiên cứu



Hình 2. Sơ đồ mô hình nghiên cứu

Trên đây là quy trình thực hiện nghiên cứu của tác giả, theo đó chi tiết các bước sẽ được trình bày ở phần tiếp theo.

2.1.1. Tập dữ liệu nghiên cứu

Tập dữ liệu được sử dụng trong nghiên cứu này là các báo cáo tài chính chưa được kiểm toán từ năm 2010 đến năm 2019, được cung cấp bởi FiinGroup. Tập dữ liệu bao gồm 1506 công ty phi tài chính trên các sàn giao dịch HNX, HoSE và UpCOM với 189 chỉ tiêu tài chính.

2.1.2. Chuẩn bị dữ liệu

2.1.2.1. Làm sạch dữ liệu

Tại bước này, tác giả tiến hành tìm kiếm và xử lí các bản ghi bị lỗi như không có dữ liệu trên một trong ba báo cáo: bảng cân đối kế toán, báo cáo kết quả hoạt động kinh doanh và báo cáo lưu chuyển tiền tệ. Sau đó tiến hành xóa các bản ghi bị thiếu dữ liệu trên ra khỏi tập dữ liệu.

2.1.2.2. Gắn nhãn cho dữ liệu

Từ các lý thuyết đã được trình bày ở chương 1, nghiên cứu quyết định sử dụng kết hợp 2 chỉ số Z-Score và M-Score để dán nhãn cho các báo cáo tài chính chưa kiểm toán là có gian lận hay không gian lận. Việc kết hợp các chỉ số lại với nhau nhằm

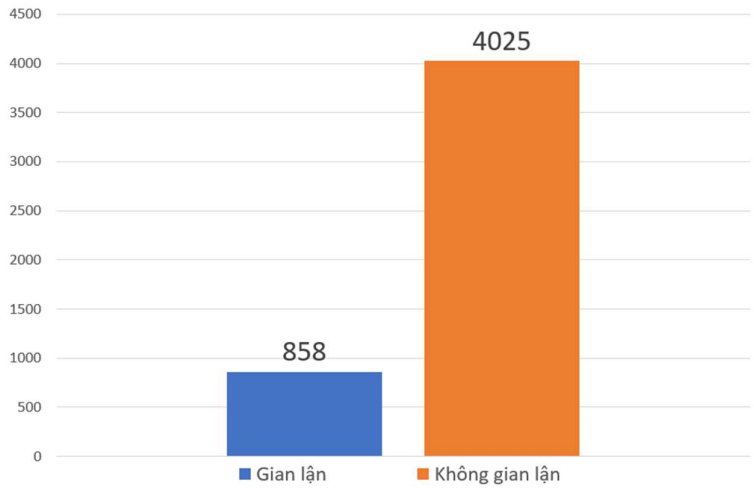
tăng khả năng phân loại báo cáo tài chính, tạo một tập dữ liệu được gắn nhãn có độ chính xác cao và đáng tin cậy.

Theo [4], Altman Z-score có độ chính xác cao trong việc dự đoán tình trạng khó khăn tài chính doanh nghiệp ở Hoa Kỳ cũng như ở các thị trường mới nổi. Đồng thời, nghiên cứu của Altman và Hotchkiss [5] áp dụng cho các thị trường mới nổi như Việt nam để đánh giá bức tranh tài chính của doanh nghiệp. Giá trị Z bé hơn hoặc bằng 5,85 chứng tỏ doanh nghiệp nằm trong vùng xám, đang gặp khó khăn về tài chính; vùng nguy hiểm với nguy cơ phá sản rất cao là các doanh nghiệp có Z-Score bé hơn hoặc bằng 4,35. Trên thực tế, nếu công ty bị áp lực về tài chính (gồm vùng xám và vùng nguy hiểm) thì khả năng gian lận trên báo cáo tài chính cũng sẽ cao hơn.

Nghiên cứu [69] chỉ ra rằng M – score (được phát triển bởi [12]) được sử dụng để đo lường mức độ thao túng báo cáo tài chính của công ty. Giá trị M dưới -2,22 cho thấy, công ty sẽ không có dấu hiệu thao túng báo cáo tài chính; giá trị M lớn hơn -2,22 báo hiệu rằng công ty có khả năng thao túng, nghĩa là có xác suất lớn về việc công ty chủ động gian lận trên báo cáo tài chính.

Đồng thời nghiên cứu [69] cũng cho thấy rằng việc công ty liên tục nằm trong vùng xám của mô hình Z-Score và mô hình M-Score chỉ ra rằng báo cáo tài chính có khả năng bị thao túng có thể dẫn đến khả năng báo cáo tài chính của công ty có gian lận. Dựa trên nền tảng này, tác giả quyết định gắn nhãn gian lận cho các công ty nằm trong vùng xám hay nhóm có nguy cơ phá sản, đạt Z-Score bé hơn hoặc bằng 5,85 [5] và M-Score lớn hơn -2,22 [69].

Sau khi loại bỏ các bản ghi có giá trị bị thiếu hoặc các bản ghi không tính đoán được chỉ số Beneish M-Score (do thiếu dữ liệu của năm liền trước), tập dữ liệu bao gồm 4883 dòng và 190 cột (bao gồm 189 chỉ tiêu tài chính và 1 cột nhãn gian lận hoặc không gian lận). Một công ty có thể gian lận nếu có Z-Score $\leqslant$ 5,85 [5] và M-Score > -2,22 [69]. Cuối cùng, tác giả có 858 báo cáo tài chính được dán nhãn gian lận và 4025 không gian lận, sự mất cân bằng được thể hiện rõ ràng như ở hình 3.



Hình 3. Biểu đồ thể hiện số lượng BCTC được gắn nhãn gian lận và không gian lận

2.1.2.3. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là một cách hiệu quả để tăng độ chính xác của các mô hình học máy và nếu không chuẩn hóa dữ liệu, một số mô hình học máy không hoạt động hiệu quả do sự chênh lệch về giá trị giữa các bản ghi và các cột [61]. Trong mô hình đề xuất này, tác giả đã sử dụng Min Max Scaler (MMS) làm mô hình Chuẩn hóa. MMS thu nhỏ dữ liệu trong phạm vi [0, 1] hoặc [-1, 1] theo phương trình [50]:

Công thức 1. Công thức tính MMS

$$x'_{i,n} = \frac{x_{i,n} - \min(x_i)}{max(x_i) - \min(x_i)}(nMax - nMin) + nMin$$

Tập dữ liệu sau khi được gắn nhãn như phương pháp đã được đề cập ở trên sẽ được chuẩn hóa về khoảng [0,1] theo phương pháp Min-max để giảm bớt chênh lệch các giá trị trên báo cáo tài chính giữa các công ty nhưng vẫn giữ được tỷ lệ và phân phối ban đầu.

Trong ngôn ngữ lập trình Python, tác giảtác giả sử dụng câu lệnh như hình để thực hiện kĩ thuật (4) Min-max scaling.

```
from sklearn.preprocessing import MinMaxScaler
cols = df1[df1.columns.difference(['Mã CK','Tên', 'Sàn', 'Year', 'fraud'], sort=False)]
pd.isnull(df1[cols.columns]).sum() > 0
df1[cols.columns] = MinMaxScaler().fit_transform(df1[cols.columns])
df1.head()
```

Hình 4. Code thực hiện bước Min-max scaling

2.1.2.4. Xử lí mất cân bằng dữ liệu

Thường có ba phương pháp chính để xử lí các tập dữu liệu mất cân bằng: (1) Oversampling, (2) Undersampling và (3) Hybrid (kết hợp (1) và (2)) [54]. Trong số đó, SMOTEENN, một phương pháp kết hợp giữa kỹ thuật lấy mẫu quá mức (SMOTE) [17] và kỹ thuật lấy mẫu dưới (ENN) [67], cho kết quả tốt hơn đối với mô hình trước hoạt động với dữ liệu mất cân bằng [54], [70].

Tiếp theo, tập dữ liệu được chia theo tỷ lệ 70:30 cho tập huấn luyện và tập thử nghiệm trước khi áp dụng kỹ thuật lấy mẫu lại.

Trong ngôn ngữ lập trình Python, tác giả sử dụng câu lệnh như hình để thực hiện các kỹ thuật (5) chia dữ liệu, (6) SMOTEEN.

```
from sklearn.model_selection import train_test_split
X = df1[df1.columns.difference(['Mã CK','Tên', 'Sàn', 'Year', 'fraud'], sort=False)]
y = df1.iloc[:, df1.columns == 'fraud']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

print("Number transactions X_train dataset: ", X_train.shape)
print("Number transactions y_train dataset: ", y_train.shape)
print("Number transactions X_test dataset: ", X_test.shape)
print("Number transactions y_test dataset: ", y_test.shape)
```

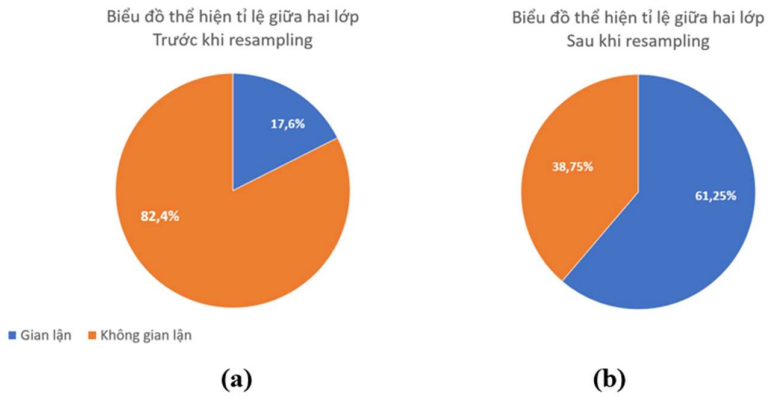
```
Number transactions X_train dataset:  (3418, 189)
Number transactions y_train dataset:  (3418, 1)
Number transactions X_test dataset:  (1465, 189)
Number transactions y_test dataset:  (1465, 1)
```

Hình 5. Code và kết quả khi thực hiện bước chia tập dữ liệu

```
from imblearn.combine import SMOTEENN
smote_enn = SMOTEENN(random_state=42)
X_train_res, y_train_res = smote_enn.fit_resample(X_train, y_train_encoded)
```

Hình 6. Code khi thực hiện kĩ thuật Resampling bằng SMOTEENN

Ngoài ra, để đảm bảo tính độc lập của tập dữ liệu thử nghiệm, tác giả chỉ áp dụng kỹ thuật SMOTEENN cho tập dữ liệu huấn luyện. Sau khi áp dụng kỹ thuật SMOTEENN, tập dữ liệu đào tạo thu được bao gồm 2575 công ty gian lận và 1629 công ty không gian lận. Sự phân bố của hai lớp trước và sau khi xử lý SMOTEENN được thể hiện trong Hình 7. Hình 7a cho thấy rằng hai nhãn của dữ liệu ban đầu là cực kỳ không cân bằng. Trong hình 7b, chúng ta có thể quan sát thấy rằng hai loại mẫu trong tập huấn luyện đã đạt đến trạng thái cân bằng. Sử dụng tập dữ liệu cân bằng sẽ có lợi cho việc đào tạo mô hình phân loại và có thể đạt được hiệu suất cao hơn.



Hình 7. Tỷ lệ giữa hai lớp (a) trước khi resampling (b) sau khi resampling

2.1.3. Huấn luyện mô hình, đánh giá và so sánh kết quả giữa các thuật toán phân loại

Sau quá trình tiền dữ liệu, tác giả tiến hành huấn luyện các mô hình. Các thuật toán ML được triển khai bao gồm các thuật toán đóng gói (ví dụ: KNN, DT, SVM, LR. Và RF) và tăng cường (ví dụ: XGBoost và LightGBM) và so sánh các mô hình này về hiệu suất của chúng.

Trong nghiên cứu này, tác giả sử dụng chỉ số KS (Thống kê Kolmogorov-Smirnov) để lựa chọn thuật toán phân loại đưa vào mô hình ensemble sau cùng.

Theo [15] trong các bài toán phân loại nhị phân, chỉ số KS đã được sử dụng làm thước đo khác biệt để đánh giá khả năng phân loại của mô hình bằng cách đo khoảng cách mà điểm của nó tạo ra giữa các hàm phân phối tích lũy (CDF) của hai lớp dữ liệu [21], còn được gọi là KS2. Từ đây KS đánh giá thuật toán nào có khả năng khái quát hóa trên tập test dựa vào phân phối giữa hai lớp. KS càng lớn, mức độ phù hợp giữa giá trị dự đoán và giá trị thực tế càng cao hay mô hình dự đoán càng chính xác và ngược lại KS thấp chứng tỏ mô hình đang ở trạng thái overfitting hoặc underfitting.

Bằng ngôn ngữ lập trình Python, tác giả tiến hành huấn luyện và đánh giá các mô hình phân loại từ các thư viện có sẵn là sklearn, lightgbm và xgboost như hình 8.

```
def ks_stat(y, yhat):
    y = y.reshape(1,-1)[0]
    yhat = yhat.reshape(1,-1)[0]
    return ks_2samp(yhat[(y==1)], yhat[(y!=1)]).statistic

def ml_model(X_train,X_test, Y_train, Y_test):
  MLA = [XGBClassifier(),lgb.LGBMClassifier(),RandomForestClassifier(), SVC(probability = True),
         LogisticRegression(), KNeighborsClassifier(), tree.DecisionTreeClassifier()]
  MLA_columns = []
  MLA_compare = pd.DataFrame(columns = MLA_columns)
  row_index = 0
  for alg in MLA:
    predicted = alg.fit(X_train, Y_train).predict(X_test)
    MLA_name = alg.__class__.__name__
    MLA_compare.loc[row_index,'Model Name'] = MLA_name
    MLA_compare.loc[row_index, 'KS'] = round(ks_stat(Y_test, predicted),3)

    row_index+=1
    MLA_compare.sort_values(by = ['KS'], ascending = False, inplace = True)


  return MLA_compare
ml_model(X_train, X_test, y_train, y_test)
```

Hình 8. Code thực hiện chức năng huấn luyện các mô hình phân loại sẵn có

Để tìm ra thuật toán phân loại có hiệu suất tốt nhất, trong nghiên cứu này, tác giả đã lựa chọn bảy thuật toán phân loại được sử dụng nhiều nhất đã được trình bày ở chương 1 là KNN, DT, SVM, LR. RF, XGBoost và LightGBM.

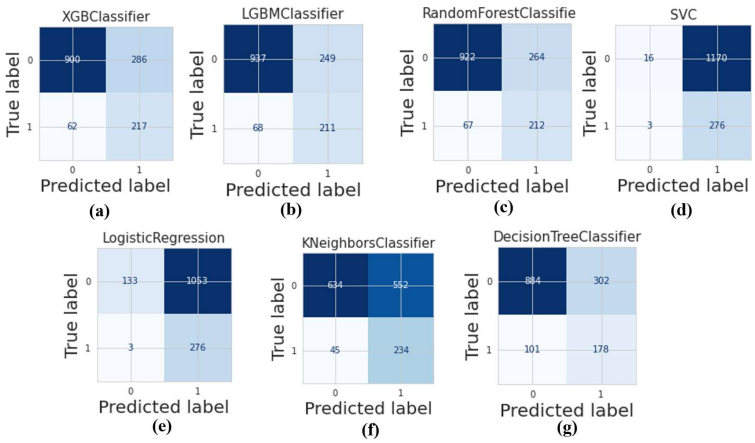
Bảng 2 cho thấy giá trị KS của bảy mô hình này. Để đơn giản, hiệu quả của các thuật toán được tóm tắt như sau.

Bảng 2. Giá trị KS của các mô hình phân loại

Model Name	KS
LightGBM	0.546
XGBoost	0.537
Random Forest	0.523
KNeighbors	0.373
Decision Tree	0.344
Logistic Regression	0.101
SVM	0.003

Kết quả dự đoán của mô hình được trình bày trong Bảng 2. LightGBM cho thấy hiệu suất tốt nhất, với giá trị KS là 54,6% trong bộ thử nghiệm và cho thấy rằng thuật toán này có khả năng phân biệt mạnh mẽ giữa các mô hình. Hiệu suất của XGBoost và Random Forest cũng cao, với lần lượt giá trị KS là 53,7% và 51,5%. Kết quả trình bày trong Bảng 2 cho thấy mô hình tích hợp LightGBM, XGBoost và Random Forest rất hiệu quả trong nghiên cứu này. Trong khi đó thuật toán LR và SVM nhanh chóng đạt đến trạng thái overfitting khi kết quả KS rất thấp lần lượt là 10,1% và 0,3%.

Hơn nữa, để chứng minh một cách trực quan dự đoán của bộ phân loại, tác giả hiển thị ma trận phân loại nhị phân của bảy mô hình trong Hình 9. Chúng ta có thể dễ dàng thấy rằng XGBoost, Random Forest và LightGBM đưa ra dự đoán chính xác hơn các thuật toán còn lại. Trong khi đó, SVM và LR thể hiện trường hợp bị overfitting.



Hình 9. Confusion matrix của các mô hình phân loại (0 – BCTC không chứa gian lận, 1 – BCTC có gian lận, Predicted label – Nhãn mô hình phân loại dự đoán, True label – Nhãn thực tế)

2.1.4. Lựa chọn các thuật toán tối ưu

Từ những kết quả trên, tác giả đã chọn XGBoost, LightGBM và Random Forest để đưa vào mô hình ensemble cuối cùng. Sau đó đánh giá hiệu quả mô hình ensemble bằng các chỉ số là Độ nhạy và AUC.

2.1.5. Mô hình kết hợp (Mô hình Ensemble)

2.1.5.1. Lựa chọn tính năng quan trọng (Feature importance)

Trong phạm vi nghiên cứu này, để tiến hành lựa chọn các tính năng quan trọng, tác giả sử dụng SHAP. SHAP (Shapley Additive exPlanation) là một phương pháp diễn giải mô hình độc lập với mô hình và có thể định lượng đóng góp của từng đặc điểm vào các dự đoán của mô hình [43]. Kỹ thuật này tính đến tác động của một đối tượng cũng như tác động của các nhóm đối tượng và tác động tổng hợp có thể có giữa các đối tượng. Giá trị SHAP dựa trên giá trị Shapley, là một khái niệm lý thuyết trò chơi.

Sau khi giá trị SHAP được tính toán bằng thư viện Shap trong Python và 20 tính năng hàng đầu của tất cả các mẫu đã được vẽ biểu đồ. Từ đó mô hình ensemble gợi ý các chỉ tiêu tài chính quan trọng mà KTV có thể chú trọng hơn các khoản mục này.

Tác giả sử dụng chỉ số Độ nhạy (Recall) và điểm AUC (Area Under The Curve) để đánh giá hiệu suất mô hình ensemble vì mục tiêu của nghiên cứu là xác định nhóm gian lận trong mẫu và đánh giá khả năng mô hình bị overfitting.

2.1.6. Dự đoán dựa trên tập dữ liệu năm 2020 và phân tích kết quả

Sau khi có được mô hình ensemble, tác giả sử dụng mô hình để dự đoán gian lận trên báo cáo tài chính cho 941 công ty trên các sàn HNX, HoSE và UpCOM dựa trên đầu vào là các báo tài chính chưa được kiểm toán năm 2020 tương ứng.

Kết luận chương 2

Trong chương 2, tác giả đã trình bày quá trình thực hiện.nghiên cứu, từ việc xử lí dữ liệu cho đến đưa ra kết quả dự đoán báo cáo tài chính gian lận cho năm 2020.

Dựa trên cơ sở lí luận của chương 1, nghiên cứu đã chỉ ra được tính hợp lí và đáng tin cậy khi sử dụng kết hợp hai chỉ số Z-Score ($\leqslant$ 5,85) và M-Score (> -2,2) khi gắn nhãn gian lận cho các báo cáo tài chính ở thị trường mới nổi như Việt Nam.

Để tiến hành xây dựng mô hình, nghiên cứu đã thu thập dữ liệu gồm 1506 báo cáo tài chính chưa kiểm toán công ty phi tài chính trên các sàn HNX, HoSE và UpCOM được cung cấp bởi FiinGroup, trong khoản thời gian từ năm 2010 – 2019. Sau đó là quá trình xử lí dữ liệu thô và dữ liệu mất cân bằng.

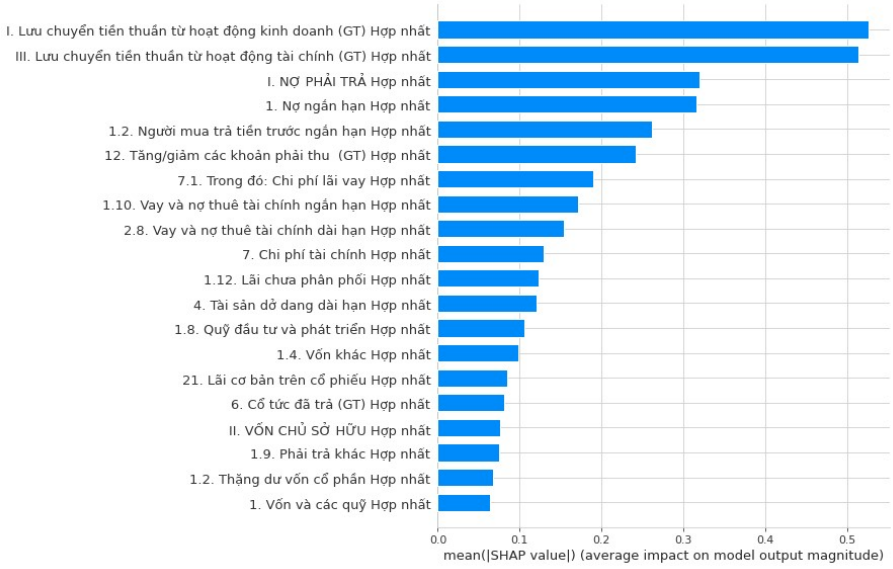
Nghiên cứu đã tiến hành so sánh 7 thuật toán phổ biến cho bài toán phân loại và nhận thấy XGBoost, LightGBM và Random Forest cho kết quả tốt và ổn định hơn các mô hình còn lại. Sau đó là kết hợp ba mô hình để xây dựng mô hình ensemble và tiến hành phân tích các đặc trưng quan trọng bằng phương pháp SHAP. Dựa trên các khoản mục quan trọng đóng góp vào khả năng dự đoán báo cáo tài chính gian lận, mô hình có thể đưa ra một số khoản mục cần lưu ý kiểm tra nhằm phát hiện gian lận cho kiểm toán viên.

CHƯƠNG 3: KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN CÁC KẾT QUẢ

3.1. Kết quả kiểm thử mô hình trên tập test và phân tích SHAP

Sau khi xây dựng được mô hình và kiểm thử trên tập test, nghiên cứu thu được kết quả mô hình ensemble đạt AUC là 76,3% và Độ nhạy đạt 74,9%. Từ đó nhận thấy rằng mô hình ensemble từ LightGBM, XGBoost và Random Forest đạt hiệu quả tốt khi tiến hành phát hiện báo cáo tài chính gian lận từ tập dữ liệu sử dụng. Độ nhạy của mô hình đạt 74,9% có nghĩa là cứ trong 100 báo cáo tài chính gian lận, mô hình ensemble có thể phát hiện chính xác đến gần 75 báo cáo.

Sử dụng phương pháp SHAP thu được các chỉ số tài chính quan trọng đối với mô hình ensemble được thể hiện ở hình 10 sau:



Hình 10. Top 20 tính năng quan trọng được mô hình ensemble lựa chọn

Giải thích toàn cục (Global explaination)

Từ hình 10, có thể thấy rằng mỗi chỉ tiêu tài chính đóng góp khác nhau vào kết quả dự đoán. Cụ thể, Lưu chuyển tiền thuần từ hoạt động kinh doanh có đóng góp quan trọng nhất, sau đó là Lưu chuyển tiền thuần từ hoạt động tài chính và Nợ phải trả, v.v. Khi thực hiện kiểm toán, các chỉ tiêu sẽ có này ảnh hưởng vượt trội hơn các chỉ tiêu khác đến quyết định của KTV. Từ đó đưa ra gợi ý cho các KTV cần chú trọng hơn các khoản mục được đề xuất khi thực hiện kiểm toán.

Giải thích cục bộ (Local explaination)

Để giải thích ý nghĩa cục bộ của mô hình, tác giả chọn mẫu là báo cáo tài chính chưa được kiểm toán của công ty H trong năm 2020. Mô hình ensemble gắn nhãn dự báo có dấu hiệu gian lận trên báo cáo tài chính của công ty H. Khi tính toán giá trị SHAP cho mẫu này, nghiên cứu thu được kết quả đóng góp của các chỉ tiêu tài chính vào kết quả dự đoán của mô hình được thống kê ở bảng 3 như sau, chi tiết sẽ được giải thích ở phần 3.2.

Table 3. Top 20 chỉ tiêu quan trọng do mô hình chọn lựa cho mẫu là BCTC của công ty H

Chỉ tiêu tài chính	Tỉ lệ quan trọng (%)	Xếp hạng
Lưu chuyển tiền thuần từ hoạt động kinh doanh	0.201	1
Lưu chuyển tiền thuần từ hoạt động tài chính	0.196	2
NỢ PHẢI TRẢ	0.136	3
Nợ ngắn hạn	0.124	4
Người mua trả tiền trước ngắn hạn	0.118	5
Tăng/giảm các khoản phải thu	0.112	6
Vay và nợ thuê tài chính ngắn hạn	0.101	7
Vay và nợ thuê tài chính dài hạn	0.1	8
Chi phí lãi vay	0.092	9
Lãi chưa phân phối	0.071	10
Cổ tức đã trả	0.069	11
Lãi cơ bản trên cổ phiếu	0.067	12
Quỹ đầu tư và phát triển	0.066	13
Chi phí tài chính	0.065	14
Tài sản dở dang dài hạn	0.062	15
Vốn và các quỹ	0.059	16
Vốn khác	0.056	17
Phải trả khác	0.053	18
Quỹ khen thưởng, phúc lợi	0.05	19
Nguyên giá TSCĐ vô hình	0.049	20

3.2. Ứng dụng mô hình ensemble dự đoán BCTC chưa kiểm toán năm 2020

Tác giả sử dụng mô hình đã xây dựng để đưa ra dự đoán về 941 báo cáo tài chính chưa được kiểm toán của năm 2020. Mô hình đã dự đoán 160 báo cáo tài chính gian lận trong tổng số 941 báo cáo. Kết quả thử nghiệm cho thấy mô hình có độ nhạy cao và có tiềm năng là một công cụ hỗ trợ đắc lực cho kiểm toán viên trong việc nhận diện báo cáo tài chính có khả năng gian lận hay không, từ đó có kế hoạch kiểm toán phù hợp.

Ví dụ, mô hình hoạt động tốt khi áp dụng cho báo cáo tài chính chưa kiểm toán của ba công ty vào năm 2020, bao gồm H, Y và V. Chênh lệch giữa lợi nhuận sau thuế chưa kiểm toán và lợi nhuận sau thuế đã kiểm toán của các doanh nghiệp này là 9,6%, 16% và 2%, cụ thể được trình bày ở bảng 5 (ĐVT: VNĐ).

Bảng 4. Kết quả kiểm chứng mô hình trên 3 mã chứng khoán trên sàn HoSE

Công ty	Lợi nhuận sau thuế (Chưa kiểm toán)	Lợi nhuận sau thuế (Đã kiểm toán)	Chênh lệch	Dự đoán của mô hình
H	-2.174.699.681.000	-2.383.339.850.000	9,6%	Gian lận
Y	-150.873.436.102	-179.998.905.303	16%	Gian lận
V	704.824.558.926	719.298.770.749	2%	Không gian lận

Đối với báo cáo tài chính của H [9], khi chú ý vào 3 chỉ tiêu: Lưu chuyển tiền thuần từ hoạt động kinh doanh, Lưu chuyển tiền thuần từ hoạt động tài chính và Nợ phải trả Hợp nhất, so sánh đối chiếu kết quả trước và sau kiểm toán thu được kết quả được trình bày ở bảng 5 (ĐVT: VNĐ):

Bảng 5. Phân tích các chỉ tiêu tài chính quan trọng của mô hình cho H

Công ty	Giá trị chưa kiểm toán	Giá trị đã kiểm toán	Chênh lệch
Lưu chuyển tiền thuần từ hoạt động kinh doanh	-723.360.928.000	-1.764.136.007.000	143,88%
Lưu chuyển tiền thuần từ hoạt động tài chính	3.020.579.316.000	3.074.497.367.000	1,79%
Nợ phải trả Hợp nhất	26.625.879.708.000	27.238.024.092.000	2,30%

Kết quả thu được ở bảng 5 cho thấy trong năm 2020, đối với các chỉ tiêu tài chính quan trọng được mô hình chọn lựa,công ty H đều có biến động và đặc biệt biến động lớn ở khoản mục Lưu chuyển tiền thuần từ hoạt động kinh doanh. Vào ngày 20/4/2021, HOSE đã ra Quyết định số 224/QĐ-SGDHCM chuyển cổ phiếu của công ty H từ diện cảnh báo sang diện kiểm soát. Báo cáo tài chính có ý kiến kiểm toán nhấn mạnh liên quan đến việc:

> *Khoản lỗ lũy kế của Tập đoàn là 7.371,6 tỷ đồng, điều kiện này cùng với những vấn đề khác được nêu trong Thuyết minh 2.6 cho thấy sự tồn tại của yếu tố không chắc chắn trọng yếu có thể dẫn tới nghi ngờ đáng kể về khả năng hoạt động liên tục của Tập đoàn. Tập đoàn đã điều chỉnh một số dữ liệu tương ứng trên báo cáo tài chính hợp nhất giữa niên độ để phản ánh các điều chỉnh nhằm sửa chữa các sai sót đã thực hiện trong kỳ trước.* [9]

Tiến hành phân tích tương tự cho công ty Y [10], kết quả được trình bày ở bảng 6 (ĐVT: VNĐ) như sau:

Bảng 6. Phân tích các chỉ tiêu tài chính quan trọng của mô hình cho Y

Công ty	Giá trị chưa kiểm toán	Giá trị đã kiểm toán	Chênh lệch
Lưu chuyển tiền thuần từ hoạt động kinh doanh	-108.570.994.384	-427.973.816.543	294,19%
Lưu chuyển tiền thuần từ hoạt động tài chính	-107.511.603.695	52.710.907.370	-149,03%
Nợ phải trả Hợp nhất	535.010.392.130	504.657.255.257	-5,67%

Kết quả thu được ở bảng 6 cũng cho thấy trong năm 2020, đối với các chỉ tiêu tài chính quan trọng được mô hình chọn lựa, công ty Y đều có biến động và đặc biệt biến động lớn ở khoản mục Lưu chuyển tiền thuần từ hoạt động kinh doanh và Lưu chuyển tiền thuần từ hoạt động tài chính. Vào ngày 05/04/2021, HOSE tiếp tục duy trì diện kiểm soát đối với cổ phiếu Y theo Quyết định số 186/QĐ-SGDHCM.

Từ hai phân tích trên có thể thấy rằng mô hình ensemble hoạt động hiệu quả khi được ứng dụng vào thực tiễn. Việc đề xuất các chỉ tiêu tài chính quan trọng góp phần giúp KTV chú trọng hơn các khoản mục này khi thực hiện kiểm toán ở các công ty được mô hình dự báo là có khả năng gian lận trên báo cáo tài chính chưa kiểm toán.

Kết luận chương 3

Mô hình ensemble được xây dựng đạt hiệu quả cao với khả năng phát hiện báo cáo tài chính gian lận (độ nhạy) đạt 74,9% tức là cứ 100 báo cáo tài chính gian lận thì mô hình phát hiện được chính xác gần 75 báo cáo.

Ngoài ra, khi ứng dụng mô hình vào thực tiễn phân tích cụ thể ở hai trường hợp H và Y đã thấy được tính hợp lí của các chỉ số tài chính quan trọng do mô hình đề xuất và hiệu quả dự báo của mô hình ensemble.

CHƯƠNG 4: KẾT LUẬN

4.1. Kết luận về các kết quả đạt được

Đề tài đã xây dựng được một mô hình kết hợp dựa trên các thuật toán học máy có hiệu quả cao và phù hợp với bộ dữ liệu từ thị trường chứng khoán Việt Nam, giúp cho kiểm toán viên dự đoán báo cáo tài chính gian lận để lập kế hoạch kiểm toán phù hợp. Với báo cáo tài chính chưa kiểm toán đã được mô hình dự báo là có gian lận, kiểm toán viên sẽ được gợi ý để tập trung kiểm tra các khoản mục có khả năng gian lận cao, từ đó tăng tính hữu hiệu và hiệu quả của công việc kiểm toán. Những người sử dụng báo cáo tài chính sau kiểm toán như các nhà đầu tư, chủ nợ, các cơ quan quản lý cũng có thể sử dụng mô hình như công cụ để đánh giá báo cáo tài chính còn có khả năng gian lận hay không, từ đó có thể đưa ra quyết định đúng đắn và kịp thời.

4.2. Đóng góp của nghiên cứu

Trong đề tài nghiên cứu, tác giả đã giải quyết hợp lý vấn đề thiếu dữ liệu báo cáo tài chính gian lận của thị trường Việt Nam, và đưa ra phương pháp gắn nhãn các báo cáo tài chính gian lận một cách đáng tin cậy để làm đầu vào cho mô hình học máy. Vấn đề mất cân bằng dữ liệu đã được xử lý hiệu quả với kỹ thuật được đánh giá là tốt nhất đến thời điểm hiện tại. Tác giả cũng đã đánh giá và so sánh các thuật toán học máy phổ biến khi giải quyết bài toán phân loại dựa trên thống kê Kolmogorov-Smirnov. Từ đó, đưa ra lựa chọn các thuật toán tối ưu và phù hợp với dữ liệu của thị trường Việt Nam. Bên cạnh đó, tác giả đã sử dụng phương pháp SHAP để đề xuất các chỉ tiêu tài chính quan trọng cần được chú ý khi thực hiện kiểm toán. Từ đó, kết hợp các thuật toán đơn lẻ lại với nhau thành mô hình ensemble cuối cùng nhằm tăng khả năng nhận diện trong báo cáo tài chính tốt hơn, tạo một mô hình có độ nhạy cao (74,9%), điểm AUC cũng cao (76,3%) và đảm bảo không bị overfitting. Kết quả nghiên cứu có thể giúp các KTV dự đoán báo cáo tài chính có khả năng gian lận, cũng như gợi ý các khoản mục có khả năng gian lận cao. Đồng thời, các nhà đầu tư và cơ quan quản lý cũng có thể sử dụng công cụ này để kiểm tra lại báo cáo tìa chính sau kiểm toán xem còn có khả năng gian lận hay không để đưa ra quyết định đúng đắn và kịp thời.

4.3. Hạn chế của nghiên cứu và hướng khắc phục

Tập dữ liệu nghiên cứu thô ban đầu sau khi xử lí đã mất 73,3% dữ liệu do sự thiếu sót về dữ liệu trên các báo cáo. Hướng phát triển nghiên cứu trong tương lai là thu thập tập dữ liệu đầy đủ hơn về các báo cáo tài chính chưa được kiểm toán của các công ty trên các sàn HNX, HoSE và UpCOM trong thời gian dài hơn để tăng qui mô của tập dữ liệu đào tạo. Ngoài ra, nghiên cứu sẽ sử dụng thêm các kĩ thuật như Feature Engineering hay Hyperparameter tuning để cải thiện hiệu quả của mô hình. Mục đích cuối cùng là nâng cao độ chính xác của dự đoán với thời gian thực hiện hiệu quả, giúp cho các KTV lập kế hoạch kiểm toán phù hợp để phát hiện gian lận trong báo cáo tài chính, từ đó tăng tính minh bạch của thị trường chứng khoán và giảm thiểu các rủi ro cũng như tổn thất về mặt tài chính.

XÁC NHẬN CỦA TRƯỜNG/
ĐOÀN TRƯỜNG

TÁC GIẢ

TÀI LIỆU THAM KHẢO

1. 2004 Report to the Nations on Occupational Fraud and Abuse. Copyright 2004 by the Association of Certified Fraud Examiners, Inc.
2. 2016 Report to the Nations on Occupational Fraud and Abuse. Copyright 2016 by the Association of Certified Fraud Examiners, Inc.
3. Altman, E. I. (1968) a. Financial Ratios, Discriminant Analysis and The Prediction Of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843
4. Altman, E. I., Hartzell, J., & Peck, M. (1998). Emerging market corporate bonds — a scoring system. Emerging Market Capital Flows, 391–400. doi:10.1007/978-1-4615-6197-2_25
5. Altman, E.I., & Hotchkiss, E. (Eds.). (2006). Corporate financial distress and bankruptcy: predict and avoid bankruptcy, analyze and invest in distressed debt (3rd ed.)
6. Auditing Standard 2401: Consideration of Fraud in a Financial Statement Audit, by Public Company Accounting Oversight Board, 2002
7. Babajanian, M. (2012). Misstatement Amounts and Associated Penalties for Fraudulent Financial Reporting.
8. Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63-93.
9. Báo cáo tài chính kiểm toán hợp nhất năm 2020 của công ty H
10. Báo cáo tài chính kiểm toán hợp nhất năm 2020 của công ty Y
11. Bao, Yang; Ke, B. I.N.; Li, B. I.N.; Yu, Y. Julia; Zhang, J. I.E. (2020): Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. In Journal of Accounting Research 58 (1), pp. 199–235. DOI: 10.1111/1475-679X.12292.
12. Beneish, Messod D. (1999) "The Detection of Earnings Manipulation." Financial Analysis Journal 55 (5): 24–36.
13. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613.
14. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
15. Bolton, Richard J.; Hand, David J. (2002): Statistical Fraud Detection: A Review. In Statist. Sci. 17 (3). DOI: 10.1214/ss/1042727940.
16. Cecchini, Mark; Aytug, Haldun; Koehler, Gary J.; Pathak, Praveen (2010): Detecting Management Fraud in Public Companies. In Management Science 56 (7), pp. 1146–1160. DOI: 10.1287/mnsc.1100.1174.
17. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
18. Chen, T., & Guestrin, C. (2016): Xgboost: A scalable tree boosting system, pp. 785–794. DOI: 10.1145/2939672.2939785.
19. Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., & Kerdprasop, N. (2015, March). An empirical study of distance metrics for k-nearest neighbor algorithm. In Proceedings of the 3rd international conference on industrial application engineering (pp. 280-285).

20. Chuẩn mực kiểm toán số 240: Trách nhiệm của kiểm toán viên liên quan đến gian lận trong quá trình kiểm toán báo cáo tài chính. Ban hành kèm theo Thông tư số 214/2012/TT-BTC ngày 06 tháng 12 năm 2012 của Bộ Tài chính.
21. Conover, W. J. (1999): Practical Nonparametric Statistics: John Wiley & Sons.
22. Dalnial, H., Kamaluddin, A., Sanusi, Z. M., & Khairuddin, K. S. (2014). Accountability in financial reporting: detecting fraudulent firms. Procedia-Social and Behavioral Sciences, 145, 61-69.
23. Dbouk, B., & Zaarour, I. (2017). Towards a machine learning approach for earnings manipulation detection. AJBA, 10(2), 215-251.
24. Dichev, Ilia, John Graham, Campbell Harvey, and Shiva Rajgopal, (2013), Earnings quality: Evidence from the field, Journal of Accounting and Economics 56, 1-33
25. Ettredge, Michael L. and Sun, Lili and Lee, Picheng and Anandarajan, Asokan, Is Earnings Fraud Associated with High Deferred Tax and/or Book Minus Tax Levels?. Auditing: A Journal of Practice & Theory, Vol. 27, May 2008, Pace University Accounting Research Paper No. 2005/05, Available at SSRN: https://ssrn.com/abstract=826587
26. Fanning, Kurt M., and Kenneth O. Cogger. (1998) "Neural network detection of management fraud using published financial data." Intelligent Systems in Accounting, Finance, and Management: An International Journal 7 (1): 21–41.
27. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
28. Feroz, Ehsan H., Taek Mu Kwon, Victor Pastena, and Kyung Joo Park. (2000) "The efficacy of red flags in predicting the SEC's targets: an artificial neural networks approach." Intelligent Systems in Accounting Finance & Management 29 (3): 145–157.
29. Galindo, J., & Tamayo, P. (2000): Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications, pp. 107–143.
30. Hajek, Petr; Henriques, Roberto (2017): Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods.
31. Hancox, D. R. (2014). Internal & Performance Auditing. Traducción propia. Página Web: davehancox. com/developing-audit-findings.
32. Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.
33. Hogan, Chris E.; Rezaee, Zabihollah; Riley, Richard A.; Velury, Uma K. (2008): Financial Statement Fraud: Insights from the Academic Literature. In AUDITING: A Journal of Practice & Theory 27 (2), pp. 231–252. DOI: 10.2308/aud.2008.27.2.231.
34. Hoogs, Bethany; Kiehl, Thomas; Lacomb, Christina; Senturk, Deniz (2007): A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. In Intell. Syst. Acc. Fin. Mgmt. 15 (1-2), pp. 41–56. DOI: 10.1002/isaf.284.
35. Humpherys, Sean L., Kevin C. Moffitt, Mary B. Burns, Judee K. Burgoon, and William F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," Decision Support Systems 50 (3): 585–594.

36. Jianrong Yao; Jie Zhang; Lu Wang (2018): 2018 International Conference on Artificial Intelligence and Big Data. ICAIBD 2018, May 26-28, 2018, Chengdu, China.
37. Kanapickienė, R., & Grundienė, Ž. (2015). The model of fraud detection in financial statements by means of financial ratios. Procedia-Social and Behavioral Sciences, 213, 321-327.
38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; … & Liu, T. Y. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree, pp. 3146–3154.
39. Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, (4), 580-585.
40. Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. (2007) "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32 (4): 995–1003.
41. Kotsiantis, Sotiris, E. Koumanakos, D. Tzelepis, and V. Tampakas. (2006) "Forecasting fraudulent financial statements using data mining." International Journal of Computational Intelligence 3 (2): 104–110.
42. Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. Lecture Notes in Computer Science, 246–252. doi:10.1007/978-3-642-34062-8_32.
43. Lundberg, S. M.; Lee, S. I. (2017): A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777.
44. MacCarthy, J. (2017). Using Altman Z-score and Beneish M-score models to detect financial fraud and corporate failure: A case study of Enron Corporation. International Journal of Finance and Accounting, 6(6), 159-166.
45. Mahama, M. (2015). Detecting corporate fraud and financial distress using the Altman and Beneish models. International Journal of Economics, Commerce and Management, 3(1), 1-18.
46. Manju, B. R., & Nair, A. R. (2019, December). Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms. In 2019 9th International Symposium on Embedded Computing and System Design (ISED) (pp. 1-7). IEEE.
47. Mathur, A., & Foody, G. M. (2008). Multiclass and binary SVM classification: Implications for training and classification users. IEEE Geoscience and remote sensing letters, 5(2), 241-245.
48. Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to support vector machine (SVM) classification. IEEE transactions on neural networks, 17(3), 671-682.
49. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.
50. Pan, J.; Zhuang, Y.; & Fong, S. (2016): The impact of data normalization on stock market prediction: using SVM and technical indicators.
51. Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. Procedia computer science, 132, 385-395.

52. Pearce, J., & Ferrier, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological modelling, 128(2-3), 127-147.
53. Perols, Johan (2011): Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. In AUDITING: A Journal of Practice & Theory 30 (2), pp. 19–50. DOI: 10.2308/ajpt-50009.
54. Qian, Y., Liang, Y., Li, M., Feng, G., & Shi, X. (2014). A resampling ensemble algorithm for classification of imbalance problems. Neurocomputing, 143, 57-67.
55. Ravisankar, Pediredla, Vadlamani Ravi, Gundumalla Raghava Rao, and Indranil Bose. (2011) "Detection of financial statement fraud and feature selection using data mining techniques." Decision Support Systems 50 (2): 491–500.
56. Repousis, S. (2016). Using Beneish model to detect corporate financial statement fraud in Greece. Journal of Financial Crime.
57. Richardson, S., Sloan, R., Soliman, M., Tuna, _ I., (2005). Accrual reliability, earnings persistence and stock prices. Journal of Accounting and Economics 39, 437–485.
58. Şahin, Y. G., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines.
59. Samarakoon, L. P., & Hasan, T. (2003). Altman's Z-Score models of predicting corporate distress: Evidence from the emerging Sri Lankan stock market. Journal of the Academy of Finance, 1, 119-125.
60. Silhavy, R., Silhavy, P., & Prokopova, Z. (Eds.). (2019). Intelligent Systems Applications in Software Engineering: Proceedings of 3rd Computational Methods in Systems and Software 2019, Vol. 1 (Vol. 1046). Springer Nature.
61. Singh, D.; & Singh, B. (2020): Investigating the impact of data normalization on classification performance. In Applied Soft Computing 97, p. 105524. DOI: 10.1016/j.asoc.2019.105524.
62. Song, Xin-Ping; Hu, Zhi-Hua; Du, Jian-Guo; Sheng, Zhao-Han (2014): Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China. In J. Forecast. 33 (8), pp. 611–626. DOI: 10.1002/for.2294.
63. Sorkun, M. C., & Toraman, T. (2017). Fraud detection on financial statements using data mining techniques. Intelligent Systems and Applications in Engineering, 5(3), 132-134.
64. Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. Managerial Auditing Journal.
65. The State Securities Commission of Vietnam (2008): Công bố thông tin xử phạt vi phạm hành chính đối với Công ty Cổ phần Bông Bạch Tuyết.
66. Throckmorton, Chandra S.; Mayew, William J.; Venkatachalam, Mohan; Collins, Leslie M. (2015): Financial fraud detection using vocal, linguistic and financial cues. In Decision Support Systems 74, pp. 78–87. DOI: 10.1016/j.dss.2015.04.006.
67. Wilson Dennis L. Asymptotic properties of nearest neighbor rules using edited data // IEEE Transactions on Systems, Man, and Cybernetics. 1972. 2, 3. 408–421.

68. Wang, G., & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. Expert Systems with Applications, 39(5), 5325-5331.
69. Warshavsky, M. (2012). Analyzing earnings quality as a financial forensic tool. Financial Valuation and Litigation Expert Journal, 39(39), 16-20.
70. Xu, Zhaozhao; Shen, Derong; Nie, Tiezheng; Kou, Yue (2020): A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. In Journal of biomedical informatics 107, p. 103465. DOI: 10.1016/j.jbi.2020.103465.
71. Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.