รายงานวิจัยฉบับสมบูรณ์

ผลกระทบของราคาหุ้นในฐานะตัวแปรสำหรับการตรวจจับการฉ้อโกงทางการเงิน ด้วยโมเดล Random Forest และ Logistic Regression

พฤศจิ ลิ้มสุวรรณเกสร[1]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการตรวจจับการฉ้อโกงทางการเงินในบริษัทจดทะเบียนในตลาดหลักทรัพย์แห่งประเทศไทย โดยใช้กรณีศึกษาของบริษัท พลังงานบริสุทธิ์ จำกัด (มหาชน) (EA) และบริษัท สตาร์ค คอร์ปอเรชั่น จำกัด (มหาชน) (STARK) งานวิจัยได้วิเคราะห์ข้อมูลทางการเงินและราคาหุ้นย้อนหลังร่วมกับการประยุกต์ใช้โมเดล Machine Learning ได้แก่ Logistic Regression (LR) และ Random Forest (RF) เพื่อเปรียบเทียบประสิทธิภาพของโมเดลทั้งสอง ผลการศึกษาแสดงให้เห็นว่า RF มีประสิทธิภาพสูงกว่า LR ในตัวชี้วัดสำคัญ เช่น ความแม่นยำ (Accuracy), ความแม่นยำเชิงบวก (Precision) และความครอบคลุม (Recall) โดยเฉพาะเมื่อเพิ่มตัวแปรราคาหุ้น อย่างไรก็ตาม ปัญหาด้านข้อมูลรั่วไหลและความไม่สมดุลในข้อมูลยังคงเป็นความท้าทายที่ต้องได้รับการแก้ไข การศึกษานี้ชี้ให้เห็นถึงความสำคัญของการใช้ข้อมูลราคาหุ้นและ Machine Learning ในการตรวจจับการฉ้อโกง พร้อมเสนอแนวทางปรับปรุงระบบตรวจจับเพื่อเพิ่มความโปร่งใสและเสถียรภาพของตลาดทุนไทย

[1] นักศึกษาคณะเศรษฐศาสตร์ มหาวิทยาลัยธรรมศาสตร์ รหัสนักศึกษา 6404681550 อีเมล์: pruesaji.l@st.econ.tu.ac.th

GitHub: https://github.com/Pruesaji/EC439

1. บทนำ

ตลาดทุนและตลาดเงินเป็นหนึ่งในกลไกสำคัญของเศรษฐกิจไทย โดยเฉพาะอย่างยิ่งในด้านการระดมทุนที่ช่วยให้ธุรกิจขนาดเล็กและขนาดใหญ่สามารถเข้าถึงแหล่งเงินทุนเพื่อการขยายตัว การจัดสรรทรัพยากรทางเศรษฐกิจให้เกิดประสิทธิภาพสูงสุด และการสร้างโอกาสในการลงทุนสำหรับประชาชนทั่วไป รวมถึงการสนับสนุนความมั่นคงทางการเงินในระดับประเทศ (พิชิต อัคราทิตย์และคณะ, 2000) อย่างไรก็ตาม ท่ามกลางความสำคัญของระบบตลาดทุน ความเสี่ยงที่เกิดจากการฉ้อโกงทางการเงินยังคงเป็นปัญหาที่ส่งผลกระทบอย่างรุนแรงในหลากหลายมิติ ไม่ว่าจะเป็นความสูญเสียทางการเงินโดยตรง และต้นทุนทางอ้อม เช่น ความเสียหายต่อชื่อเสียงของผู้มีส่วนได้ส่วนเสีย นักลงทุน เจ้าหนี้ และพนักงาน ในกรณีที่รุนแรง การฉ้อโกงทางการเงินอาจนำไปสู่การล้มละลายขององค์กร (Dorris, 2020)

ในกรณีของประเทศไทย ตามผลสำรวจอาชญากรรมทางเศรษฐกิจและการทุจริต ประจำปี 2020 ของ PwC พบว่า 22% ขององค์กรไทยที่ตอบแบบสำรวจระบุว่า พวกเขาประสบเหตุทุจริต คอร์รัปชัน หรืออาชญากรรมทางเศรษฐกิจและการเงินในช่วง 24 เดือนที่ผ่านมา ประเภทของการทุจริตที่พบบ่อย ได้แก่ การยักยอกทรัพย์สิน การทุจริตในการจัดซื้อจัดจ้าง การติดสินบนและการคอร์รัปชัน กรณีส่วนใหญ่ของการทุจริตในประเทศไทยเกิดขึ้นจากพนักงานภายในองค์กร โดย 59% ของการทุจริตเกิดจากบุคคลภายใน และ 18% เกิดจากการสมรู้ร่วมคิดกับบุคคลภายใน (PricewaterhouseCoopers, 2020)

การวิจัยนี้จึงมุ่งเน้นไปที่รูปแบบการฉ้อโกงดังกล่าว โดยใช้กรณีศึกษาการฉ้อโกงในงบการเงินของบริษัท สตาร์ค คอร์ปอเรชั่น จำกัด (มหาชน) (STARK) ซึ่งถูกตรวจพบว่ามีการตกแต่งงบการเงินอย่างเป็นระบบเพื่อสร้างภาพลักษณ์ทางการเงินที่ดี โดยมีการบิดเบือนข้อมูลทางการเงิน เช่น การรายงานกำไรเกินจริงถึง 30% และสินทรัพย์ที่ไม่มีอยู่จริง ทั้งนี้การกระทำดังกล่าวไม่เพียงแต่ส่งผลกระทบต่อความน่าเชื่อถือของบริษัท แต่ยังทำให้นักลงทุนสูญเสียเงินลงทุนรวมกว่า 5,000 ล้านบาท และราคาหุ้นลดลงถึง 60% หลังการเปิดเผยข้อมูล และกรณีของบริษัท พลังงานบริสุทธิ์ จำกัด (มหาชน) (EA) ที่เกี่ยวข้องกับการทุจริตในกระบวนการจัดซื้ออุปกรณ์ ซึ่งก่อให้เกิดความเสียหายในด้านต้นทุนกว่า 2,000 ล้านบาท และทำให้ความไว้วางใจของผู้เกี่ยวข้องในอุตสาหกรรมลดลงอย่างมีนัยสำคัญ (Wiriyapong et al., 2024) จะเห็นได้ว่าตัวแปรที่สำคัญอย่างงบการเงินได้สะท้อนถึงความโปร่งใสและความถูกต้องในการรายงานทางการเงิน ขณะที่ราคาหุ้นสามารถบ่งชี้ถึงความเชื่อมั่นของนักลงทุนในบริษัท ทั้งสองตัวแปรนี้อาจช่วยในการประเมินผลกระทบของการทุจริตและตรวจจับความผิดปกติในการบริหารจัดการทางการเงินได้อย่างมีประสิทธิภาพ

การตรวจจับการฉ้อโกงนี้เป็นเรื่องที่ท้าทายเนื่องจากมีความซับซ้อนและมีหลายวิธีในการฉ้อโกง วิธีการตรวจจับการฉ้อโกงแบบดั้งเดิมมักจะพึ่งพาตรวจสอบจากผู้สอบบัญชีภายนอก วิธีการเหล่านี้มักมีประสิทธิภาพต่ำเนื่องจากใช้เวลานานและมีค่าใช้จ่ายสูง (Dyck et al., 2010; West et al., 2016) นอกจากนี้ยังมีปัญหาทางจริยธรรมว่าผู้สอบบัญชีบางครั้งอาจมีความสัมพันธ์ทางการเงินกับบริษัทที่พวกเขาตรวจสอบ ซึ่งอาจก่อให้เกิดความ

ขัดแย้งทางผลประโยชน์ นั่นหมายความว่าพวกเขาอาจไม่ซื่อสัตย์หรือไม่ละเอียดในการตรวจสอบ (Simnett et al., 2019) ดังนั้นการจัดการกับการฉ้อโกงทางการเงินอย่างรวดเร็วเป็นเรื่องสำคัญมาก ไม่เพียงแต่เพื่อเรียกคืนเงินที่สูญเสียไป แต่ยังเพื่อฟื้นฟูความเชื่อมั่นในระบบตลาดเงินและตลาดทุน การพัฒนาระบบที่ดีในการตรวจจับการฉ้อโกงจึงเป็นสิ่งสำคัญสำหรับทุกฝ่ายที่เกี่ยวข้อง ไม่ว่าจะเป็นนักลงทุน บริษัทตรวจสอบบัญชี และหน่วยงานกำกับดูแล (Abbasi et al., 2012; Albrecht et al., 2008)

ในการพัฒนาระบบที่สามารถตรวจจับการฉ้อโกงได้มีหลากหลายวิธี ซึ่งในปัจจุบันหนึ่งในเทคนิคที่มีความแม่นยำสูงและสามารถจัดการข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพคือการใช้ปัญญาประดิษฐ์ (Artificial Intelligence: AI) (Sadgali et al., 2019) ซึ่งมีส่วนประกอบย่อยที่ผนวกรวมไว้สองส่วนคือ การเรียนรู้ของเครื่อง (Machine Learning: ML) และการเรียนรู้เชิงลึก (Deep Learning: DL) โดย ML ใช้งานง่ายกับข้อมูลที่มีโครงสร้าง เช่น ข้อมูลตารางที่มีแถวและคอลัมน์ ตีความผลลัพธ์ได้ชัดเจน ประมวลผลได้เร็วกว่า DL และต้องการทรัพยากรคอมพิวเตอร์น้อยกว่า แต่อาจไม่เหมาะกับข้อมูลที่ซับซ้อนหรือไม่มีโครงสร้าง และประสิทธิภาพขึ้นอยู่กับการเลือกตัวแปรที่เหมาะสม (Feature Selection) ในขณะที่ DL ซึ่งเป็นสาขาหนึ่งของ ML มีศักยภาพสูงในการประมวลผลข้อมูลที่ซับซ้อนและไม่มีโครงสร้าง เช่น ข้อความ เสียง หรือภาพ สามารถจับความสัมพันธ์ที่ซับซ้อนได้และพัฒนาประสิทธิภาพสูงสุดในปัญหาที่มีข้อมูลจำนวนมาก แต่ผลลัพธ์ที่ได้ตีความยาก ใช้ทรัพยากรสูง และมีความเสี่ยงต่อ Overfitting[2] ในกรณีข้อมูลน้อย (Zohuri et al., 2020)

ด้วยเหตุนี้ การพัฒนาโมเดลที่เหมาะสมสำหรับการตรวจจับการฉ้อโกงจึงใช้เทคโนโลยี ML เพื่อปรับตัวให้เข้ากับลักษณะเฉพาะของข้อมูลและเป้าหมายในการวิจัย โดยเฉพาะเมื่อข้อมูลมีลักษณะเชิงโครงสร้าง เช่น ข้อมูลงบการเงินที่สะท้อนถึงความโปร่งใสและความถูกต้องในการรายงานทางการเงิน รวมถึงข้อมูลราคาหุ้นที่สามารถบ่งชี้ถึงความเชื่อมั่นของนักลงทุนในบริษัท การรวมตัวแปรเหล่านี้จะช่วยในการประเมินผลกระทบจากการทุจริตและตรวจจับความผิดปกติในการบริหารจัดการทางการเงินได้อย่างมีประสิทธิภาพ

งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาการใช้ ML เพื่อตรวจจับการฉ้อโกงทางการเงินในองค์กร โดยเน้นไปที่โมเดลที่เป็นประเภท Supervised Learning ที่มีข้อดีในเรื่องของการประมวณผลได้เร็วและประเมิณผลได้ง่าย โมเดลที่มีความแม่นยำในการตรวจจับมากที่สุดก็คือ Logistic Regression ในประเภทของ Regression และ Random Forest ในประเภทของ Classification โดยวิเคราะห์ผลกระทบของราคาหุ้นต่อการตรวจจับการฉ้อโกงทางการเงิน เปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression และ Random Forest เพิ่มความแม่นยำของโมเดลโดยใช้ราคาหุ้นเป็นตัวแปรเพิ่มเติม

---

[2] เกิดขึ้นเมื่อโมเดลเรียนรู้จากชุดข้อมูลฝึก (Training Set) ได้ดีมาก แต่ไม่สามารถทำงานได้ดีเมื่อใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (Test Set) ซึ่งมักจะเป็นสัญญาณว่าโมเดลกำลังจำรายละเอียดที่ไม่จำเป็นจากข้อมูลฝึก ซึ่งสาเหตุหลักมาจากปัญหาข้อมูลไม่สมดุล (Imbalanced Data) (Piyapanichayakul, R., 2023)

2. การทบทวนวรรณกรรม

การฉ้อโกงทางการเงินเป็นปัญหาสำคัญที่ส่งผลกระทบอย่างรุนแรงต่อเศรษฐกิจและตลาดทุนทั่วโลก การตรวจจับปัญหานี้เป็นความท้าทายที่นักวิจัยได้พยายามแก้ไขผ่านการพัฒนาโมเดล AI และ Machine Learning อย่างไรก็ตาม ในบริบทของประเทศไทย การวิเคราะห์ผลกระทบของราคาหุ้นยังไม่ได้รับการศึกษาอย่างลึกซึ้ง งานวิจัยของเราจึงมีความสำคัญในการเพิ่มความเข้าใจเกี่ยวกับบทบาทของราคาหุ้นในการตรวจจับการฉ้อโกง โดยเปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression และ Random Forest ซึ่งช่วยเพิ่มความแม่นยำของการวิเคราะห์ข้อมูลและลดความเสี่ยงจากการฉ้อโกงในตลาดทุนไทยได้อย่างมีนัยสำคัญ นักวิจัยหลายคนได้พยายามพัฒนาเครื่องมือและเทคนิคที่สามารถช่วยในการตรวจจับการฉ้อโกงได้อย่างมีประสิทธิภาพ โดยเฉพาะการประยุกต์ใช้เทคโนโลยี AI และ Machine Learning ซึ่งมีความสามารถในการวิเคราะห์ข้อมูลที่ซับซ้อนและมีปริมาณมากได้อย่างแม่นยำ

หนึ่งในปัจจัยสำคัญที่ช่วยในการตรวจจับการฉ้อโกงทางการเงิน คือการเลือกตัวแปรที่เหมาะสม ซึ่งรวมถึงข้อมูลทางการเงิน เช่น อัตราส่วนทางการเงิน (Financial Ratios) และตัวแปรที่ได้จากราคาหุ้น ได้รับการพิสูจน์ว่ามีประโยชน์ในการระบุพฤติกรรมผิดปกติในตลาด โดยเฉพาะในกรณีของบริษัทที่มีประวัติการฉ้อโกง ในการศึกษาวิจัยหลายฉบับ (Feroz et al., 2000; Stice et al., 1991; Persons, 1995; Wells, 1997; Fanning & Cogger, 1998; Beneish, 1999; Spathis et al., 2002; Lenard & Alam, 2009; Ravisankar et al., 2011) การวิเคราะห์อัตราส่วนทางการเงินได้รับการเลือกให้เป็นหนึ่งในวิธีการที่ใช้ในการตรวจจับการฉ้อโกง หลังจากการวิจัยทางทฤษฎี อัตราส่วนทางการเงินที่ใช้ในวรรณกรรมทางวิทยาศาสตร์ถูกจัดกลุ่มออกเป็น 5 กลุ่มและกลุ่มย่อยของอัตราส่วนทางการเงิน[3] ซึ่งยืนยันว่าผู้วิจัยต่างเลือกใช้อัตราส่วนทางการเงินที่แตกต่างกันในการตรวจจับการฉ้อโกง

อีกทั้งตัวแปรความยากลำบากทางการเงินอาจเป็นแรงจูงใจให้ผู้บริหารมีส่วนร่วมในกิจกรรมการฉ้อโกง ตามที่ Fanning & Cogger (1998), Kirkos et al. (2007), Ravisankar et al. (2011) ระบุว่า ระดับหนี้สินที่สูงขึ้นอาจเพิ่มโอกาสที่งบการเงินจะถูกฉ้อโกงด้วย อัตราส่วนที่ใช้บ่อยในการวิจัยที่เกี่ยวข้องกับการตรวจจับการฉ้อโกง ได้แก่ อัตราส่วนหนี้สินรวมต่อสินทรัพย์รวม (TD/TA) (Kirkos et al., 2007; Gaganis, 2009; Dalnial et al., 2014) หรือ อัตราส่วนหนี้สินรวมต่อสินทรัพย์รวม (TL/TA) (Lenard & Alam, 2009) อัตราส่วนหนี้สินรวมต่อส่วนของผู้ถือหุ้น (TD/Eq) (Spathis et al., 2002; Kirkos et al., 2007; Dalnial et al., 2014) และสภาพคล่องที่ต่ำอาจเป็นแรงจูงใจให้ผู้บริหารมีส่วนร่วมในการฉ้อโกงงบการเงิน โดยทั่วไปแล้ว สภาพคล่องจะถูกวัดโดยอัตราส่วนทุน

---

[3] 1a. อัตราผลตอบแทนจากการขาย (Return on Sales: ROS), 1b. อัตราผลตอบแทนจากการลงทุน (Return on Investment: ROI), 2. อัตราส่วนสภาพคล่อง, 3. อัตราส่วนความสามารถในการชำระหนี้ระยะยาว (Solvency Ratios), 4. อัตราส่วนกิจกรรม (Activity Ratios), 5a. อัตราส่วนโครงสร้างสินทรัพย์รวม (Total Assets Structure Ratio), 5b. อัตราส่วนโครงสร้างสินทรัพย์หมุนเวียน (Current Assets Structure Ratio), 5c. อัตราส่วนโครงสร้างสินทรัพย์ประเภทที่ดิน (Property Structure Ratio)

หมุนเวียนต่อสินทรัพย์รวม (WC/TA) หรือ อัตราส่วนสินทรัพย์หมุนเวียนต่อหนี้สินหมุนเวียน (CA/CL) (Lenard & Alam, 2009; Ravisankar et al., 2011)

จากงานวิจัย Song et al. (2014), Stice et al. (1991) ระบุว่าอีกหนึ่งแรงจูงใจในการฉ้อโกงของผู้บริหารบริษัทคือการรักษาการเติบโตของบริษัท เพื่อที่จะตรวจสอบว่าบริษัทยังคงเติบโตอยู่หรือไม่ ได้ใช้การวิเคราะห์อัตราส่วนทางการเงินที่เกี่ยวข้องกับกิจกรรม ความสามารถในการทำกำไร และการจัดการสินทรัพย์ เช่น อัตราส่วนยอดขายต่อสินทรัพย์รวม (SAL/TA), อัตราส่วนกำไรสุทธิต่อยอดขาย (NP/SAL), อัตราส่วนกำไรสุทธิต่อสินทรัพย์รวม (ROA), อัตราส่วนสินทรัพย์หมุนเวียนต่อสินทรัพย์รวม (CA/TA)

รวมถึงงานวิจัยของ Stice et al. (1991), Persons (1995), Kaminski et al. (2004), Kirkos et al. (2007), Perols (2011) ระบุว่า สินค้าคงคลังและลูกหนี้เป็นตัวแปรในงบการเงินที่สามารถประเมินได้โดยใช้วิจารณญาณ ดังนั้นอัตราส่วนที่ใช้ในการตรวจจับงบการเงินที่มีการฉ้อโกงได้แก่ อัตราส่วนสินค้าคงคลังต่อยอดขาย (INV/SAL), อัตราส่วนสินค้าคงคลังต่อสินทรัพย์รวม (INV/TA), อัตราส่วนลูกหนี้ต่อยอดขาย (REC/SAL)

การอ้างอิงในวรรณกรรมชี้ให้เห็นว่า การใช้อัตราส่วนทางการเงินในการตรวจจับการฉ้อโกงในงบการเงินเป็นวิธีที่สะดวกและตรงไปตรงมา อย่างไรก็ตาม ปัญหาที่เกิดขึ้นคือการตีความผลลัพธ์ เช่น ค่าอัตราส่วนทางการเงินใดบ้างที่บ่งชี้ว่า งบการเงินนั้นมีการฉ้อโกง

การใช้เทคโนโลยี Machine Learning และ AI ในการตรวจจับการฉ้อโกงทางการเงินได้รับการพัฒนาและปรับปรุงอย่างต่อเนื่อง โดยมีการใช้งานทั้งในรูปแบบที่มีการควบคุม (supervised learning) เช่น Logistic Regression Random Forest Decision Tree และ SVM ซึ่งมีความสามารถในการจำแนกและทำนายการฉ้อโกงในกรณีต่าง ๆ และไม่มีการควบคุม (unsupervised learning) ใช้เทคนิคการตรวจจับข้อผิดปกติ (anomaly detection) เช่น ANN (Guo et al., 2022)

อย่างไรก็ตามการเลือกตัวแปรที่เหมาะสมในการวิเคราะห์ข้อมูลทางการเงินจะเป็นกุญแจสำคัญในการตรวจจับการฉ้อโกงและใช้โมเดล Machine Learning และ AI ที่หลากหลาย ก็ยังมีความท้าทายหลายประการในเรื่องของการตีความผลลัพธ์ที่ได้จากโมเดลต่าง ๆ เช่น การเลือกอัตราส่วนทางการเงินที่เหมาะสม หรือการใช้ข้อมูลจากราคาหุ้น ซึ่งอาจจะมีความสัมพันธ์ที่ไม่แน่ชัดกับพฤติกรรมการฉ้อโกงบางประเภท นอกจากนี้ยังมีปัญหาเกี่ยวกับข้อมูลที่ไม่สมดุล (imbalanced data) จึงอาจมีปัญหาความไม่สมดุลของสัดส่วนตัวอย่างและจำนวนตัวอย่างที่น้อยเกินไป ซึ่งอาจทำให้โมเดลเกิดการ overfitting (Guo et al., 2022; Qian and Luo, 2015)

ในบริบทของประเทศไทย Sawangarreerak & Thanathamathee (2021) ได้ศึกษาและพัฒนาเทคนิคการตรวจจับการฉ้อโกงในตลาดหลักทรัพย์แห่งประเทศไทย (SET) โดยใช้ Association Rule Mining เพื่อค้นหาความผิดปกติจากข้อมูลทางการเงิน ผลการศึกษาแสดงให้เห็นว่ารายการทางการเงิน 9 รายการที่มีความสัมพันธ์กับการฉ้อโกงในงบการเงิน ซึ่งสามารถใช้ในการตรวจจับสัญญาณของการจัดการทางการเงินที่ไม่ถูกต้องหรือการฉ้อโกงได้แก่ กำไรขั้นต้น, รายได้จากธุรกิจหลัก, อัตราส่วนรายได้ธุรกิจหลักต่อสินทรัพย์รวม, อัตราส่วนทุนและเงินสำรอง

ต่อหนี้รวม, อัตราส่วนหนี้ระยะยาวต่อทุนรวมและสำรอง, อัตราส่วนลูกหนี้ต่อรายได้จากธุรกิจหลัก, อัตราส่วนกำไรขั้นต้นต่อกำไรธุรกิจหลัก, อัตราส่วนหนี้ระยะยาวต่อสินทรัพย์รวม, และสินทรัพย์รวม ตัวแปรเหล่านี้สะท้อนถึงความสามารถในการทำกำไร การใช้สินทรัพย์ การจัดการหนี้สิน และความเสี่ยงทางการเงิน หากมีการจัดการเพื่อแสดงข้อมูลไม่พึงประสงค์จะทำให้สถานะทางการเงินดูดีเกินจริง ซึ่งสามารถช่วยระบุบริษัทที่มีความเสี่ยงต่อการฉ้อโกงได้

อย่างไรก็ตามงานวิจัยนี้ไม่ได้พิจารณาถึงบทบาทของราคาหุ้นซึ่งเป็นตัวแปรสำคัญในการสะท้อนความเชื่อมั่นของตลาดและพฤติกรรมการซื้อขายของนักลงทุน ผู้วิจัยจึงให้ความสำคัญในการนำข้อมูลราคาหุ้นมาใช้ร่วมกับอัตราส่วนทางการเงิน เพื่อพัฒนาโมเดล Logistic Regression และ Random Forest ที่สามารถเพิ่มความแม่นยำในการตรวจจับการฉ้อโกงและตอบโจทย์ความท้าทายที่ยังไม่ได้รับการแก้ไขในบริบทของประเทศไทย

ความสำคัญของโมเดล Logistic Regression และ Random Forest มีในหลายงานวิจัย เช่น การศึกษาของ Gottlieb et al., (2006) ใช้เทคนิค Machine Learning เพื่อตรวจจับการฉ้อโกงทางการเงิน โดยทดสอบ Logistic Regression, Naive Bayes และ SVM บนข้อมูลจาก 9,000 บริษัทที่จดทะเบียนในตลาดหลักทรัพย์สหรัฐฯ ครอบคลุม 40 ไตรมาส รวม 360,000 จุดข้อมูล พบว่าเพียง 2,000 จุดที่เกี่ยวข้องกับการฉ้อโกง ข้อมูลที่ไม่สำคัญถูกตัดออก เหลือ 173 คุณลักษณะที่สำคัญในการทำนาย การศึกษาพบว่า Logistic Regression ทำงานได้ดี โดย 10% ของกลุ่มเสี่ยงสูงสุดมีโอกาสถูกฉ้อโกงมากขึ้น มีความแม่นยำสูงสุดอยู่ที่ 26% ทำงานได้ดีพอ ๆ กับ Gaussian SVM ในการทำนายการฉ้อโกง

งานวิจัยของ Li and Wang (2020) เปรียบเทียบโมเดลการทำนายทางการเงินของบริษัทจดทะเบียนโดยใช้ Machine Learning โดยใช้สามโมเดล ได้แก่ Logistic Regression, Support Vector Machine (SVM), และ Decision Tree เพื่อจำแนกชุดข้อมูลการฝึกอบรม พบว่าโมเดลทั้งสามมีความแม่นยำสูง โดยโมเดล Decision Tree มีความสามารถในการทำนายดีที่สุด ตามด้วย SVM และ Logistic Regression โดยเชื่อว่าโมเดล Decision Tree เป็นตัวเลือกที่ดีกว่าในการทำนายงบทางการเงินของบริษัท

ในงานวิจัยของ Ma (2019) ได้ใช้โมเดล Decision Tree, XGBoost, และ Random Forest ในการระบุการฉ้อโกงทางการเงินของบริษัท พบว่าโมเดล Random Forest มีการเรียกคืน (recall) สูงสุดในสามโมเดล ดังนั้นผู้เขียนจึงพิจารณาให้โมเดล Random Forest เป็นโมเดลที่ดีที่สุด

งานวิจัยของ Xu, H., Fan, G., & Song, Y. (2022) ได้วิเคราะห์โมเดลการทำนายและการตรวจจับการฉ้อโกงทางการเงินทั้งในและต่างประเทศ โดยเลือกโมเดลการเรียนรู้ของเครื่อง 4 แบบ ได้แก่ GBDT, Random Forest, Support Vector Machine (SVM), และ Decision Tree ในบรรดาโมเดลที่ทดสอบทั้งหมด โมเดลการตัดสินใจ (Decision Tree) พบว่ามีความสามารถในการทำนายดีที่สุด ตามด้วยเครื่องเวกเตอร์สนับสนุน (Support Vector Machine) และแบบจำลองการถดถอยโลจิสติก (Logistic Regression) ซึ่งแสดงให้เห็นว่าโมเดลการตัดสินใจเป็นตัวเลือกที่มีประสิทธิภาพสูงสุดสำหรับการทำนายการฉ้อโกงทางการเงิน

งานวิจัย Wyrobek (2020) เน้นไปที่การเปรียบเทียบประสิทธิภาพของอัลกอริธึม Machine Learning และ AI (Neural Networks) กับโมเดลแบบดั้งเดิมเช่น การถดถอยเชิงเส้น (Linear Regression) และการถดถอยโลจิสติก (Logistic Regression) เมื่อใช้ข้อมูลเดียวกัน โดยพบว่าอัลกอริธึมการเรียนรู้ของเครื่องและปัญญาประดิษฐ์มีประสิทธิภาพดีกว่า แต่โมเดลประเภท white-box ซึ่งเป็นโมเดลที่สามารถเข้าใจและตรวจสอบได้ง่าย และยังคงเป็นที่นิยมมากกว่า รวมถึงทำการศึกษาความสามารถของงบการเงินในการตรวจจับพฤติกรรมการทุจริตในบริษัท โดยใช้ข้อมูลจาก 54 บริษัทที่มีชื่อเสียงเกี่ยวกับการทุจริตและเปรียบเทียบกับ 58 บริษัทที่ไม่มีการทุจริต ส่วนใหญ่เป็นบริษัทที่จดทะเบียนใน NYSE หรือ NASDAQ ผลการฝึกอัลกอริธึมจากงบการเงิน 1,317 ฉบับ พบว่าอัลกอริธึมที่มีประสิทธิภาพสูงที่สุดคือ Gradient-boosted decision trees (XGB) และ Random Forest โดยมีความถูกต้อง 93.5% และ 94.7% ตามลำดับ อีกทั้งยังมีการทดสอบกับโมเดลแบบดั้งเดิม เช่น Linear Discriminant Analysis ซึ่งให้ผลลัพธ์ที่ดีด้วยความถูกต้อง 92.2% งานวิจัยนี้แสดงให้เห็นว่าโมเดลที่ทันสมัยสามารถตรวจจับการทุจริตได้ดีกว่า ในขณะที่โมเดลแบบดั้งเดิมก็ยังมีประสิทธิภาพในบางกรณี

การวิจัยในปัจจุบันชี้ให้เห็นถึงความสำคัญของการเลือกตัวแปรที่เหมาะสมและการใช้เทคนิค Machine Learning ที่หลากหลายเพื่อปรับปรุงประสิทธิภาพในการตรวจจับการฉ้อโกง โดยเฉพาะการรวมตัวแปรราคาหุ้นเป็นข้อมูลเพิ่มเติม ซึ่งเป็นการเพิ่มมิติใหม่ให้กับกระบวนการตรวจจับการฉ้อโกง การศึกษานี้จึงเน้นการประเมินว่าการใช้ราคาหุ้นในโมเดล Logistic Regression และ Random Forest สามารถเพิ่มความแม่นยำและตอบโจทย์ปัญหาความซับซ้อนของข้อมูลได้อย่างไร โดยเฉพาะในกรณีที่ต้องการระบุความผิดปกติทางการเงินของบริษัทในตลาดทุนไทย อย่างไรก็ตามยังมีความท้าทายหลายประการทั้งการจัดการกับข้อมูลที่ไม่สมดุลและการเพิ่มความสามารถในการอธิบายผลลัพธ์ของโมเดลที่ยังต้องการพัฒนาเพิ่มขึ้นในอนาคต

3. วิธีการศึกษา

การวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ผลกระทบของราคาหุ้นต่อการตรวจจับการฉ้อโกงทางการเงิน เปรียบเทียบประสิทธิภาพของโมเดล LR และ RF เพิ่มความแม่นยำของโมเดลโดยใช้ราคาหุ้นเป็นตัวแปรเพิ่มเติม งานวิจัยนี้จะดำเนินการตามขั้นตอนดังต่อไปนี้

3.1 การเก็บข้อมูล (Data Collection)

ข้อมูลที่ใช้ในการวิจัยนี้ประกอบไปด้วย**งบแสดงฐานะการเงิน (Balance Sheet) งบกำไรขาดทุน (Income Statement) งบกระแสเงินสด (Cash Flow) อัตราส่วนทางการเงิน (Ratios-Key Metrics) และราคาหุ้นรายสัปดาห์ย้อนหลัง (Price History)** ตั้งแต่บริษัทเข้าสู่ตลาดหลักทรัพย์แห่งประเทศไทย สำหรับกลุ่มบริษัทที่มีรายงานข่าวเกี่ยวกับการฉ้อโกง ได้แก่ บริษัท STARK และ EA และบริษัทอื่น ๆ ที่เป็นคู่แข่งในอุตสาหกรรมเดียวกันกับ STARK หรือ EA แต่ไม่มีรายงานการทุจริต โดยรายชื่อจะแสดงใน *ภาคผนวก ก.*

ในการเลือกกลุ่มบริษัทที่ใช้ในการวิเคราะห์นั้น จะเลือกบริษัทที่อยู่ในกลุ่มอุตสาหกรรมเดียวกันการเนื่องจากการเปรียบเทียบบริษัทต่างอุตสาหกรรมอาจมีกฎระเบียบและแรงกระตุ้นทางเศรษฐกิจที่แตกต่างกัน เช่น

อุตสาหกรรมการท่องเที่ยวได้รับผลกระทบจากการระบาดของโควิด-19 มากกว่าอุตสาหกรรมอื่น ๆ (จิตเกษม พรประพันธ์ และพรชนก เทพขาม, 2020) การเลือกบริษัทในอุตสาหกรรมเดียวกันจึงช่วยลดความคลาดเคลื่อนของข้อมูลและเพิ่มความแม่นยำของการวิเคราะห์ (จันทนา วัฒนกาญจนะ, 2016)

3.2 การเตรียมข้อมูล (Data Preprocessing)

เพื่อให้โมเดลสามารถประมวลผลข้อมูลได้อย่างมีประสิทธิภาพ การเตรียมข้อมูลเป็นขั้นตอนสำคัญในการวิจัยนี้ โดยขั้นตอนจะครอบคลุมดังนี้

1. การทำความสะอาดข้อมูล (Data Cleaning) ในการจัดการกับ Missing Values, การแปลงค่า "--" เป็น np.nan ถูกใช้เพื่อแสดงข้อมูลที่ขาดหายไปในคอลัมน์ต่าง ๆ ของ Balance Sheet, Cash Flow, และ Income Statement นอกจากนี้ยังได้ใช้ ffill (Forward Fill) หรือ bfill (Backward Fill) เพื่อแทนที่ Missing Values (NaN) ในคอลัมน์นั้น ๆ ในกรณีที่ข้อมูลในบางคอลัมน์มีลักษณะเป็นข้อมูลที่ไม่สมบูรณ์มากเกินไป การลบคอลัมน์ที่มี Missing Values ด้วย dropna() อีกทั้งยังมีการทำความสะอาดข้อมูลอักขระที่ไม่จำเป็นออก เช่น การลบ "," (ลูกน้ำ) "(" (วงเล็บเปิด) ")" (วงเล็บปิด) เพื่อให้อยู่ในรูปแบบที่เหมาะสมสำหรับการประมวลผลด้วยโปรแกรมคอมพิวเตอร์

2. การแปลงข้อมูล (Data Transformation) เนื่องจากข้อมูลทางการเงินมีหน่วยที่แตกต่างกัน เช่น ข้อมูลทางการเงิน อาจอยู่ในรูปของพันล้านบาทหรือหน่วยเปอร์เซ็นต์ การปรับสเกลข้อมูลจึงจำเป็นเพื่อให้ตัวแปรทั้งหมดมีขนาด ใกล้เคียงกัน ซึ่งจะช่วยลดปัญหาความเบี่ยงเบนในการฝึกโมเดล โดยใช้เทคนิค Min-Max Scaling[4] จากไลบรารี sklearn.preprocessing เป็นวิธีหนึ่งที่ช่วยปรับขนาดข้อมูลในคอลัมน์ตัวแปรต้นของ Balance Sheet, Income Statement, Cash Flow, Ratios-Key Metrics, Price History แสดงใน*ภาคผนวก ข.* ให้มีค่าตั้งแต่ 0 ถึง 1 ซึ่งช่วยให้ข้อมูลไม่เกิดการครอบงำจากค่าที่มีขนาดใหญ่มากเกินไปเมื่อใช้ในโมเดลทางสถิติsหรือลดปัญหา Outliers จากนั้นแปลงข้อมูลตัวเลขในคอลัมน์ต่าง ๆ ถูกแปลงด้วยฟังก์ชัน pd.to_numeric() เพื่อให้ข้อมูลเป็นตัวเลขที่สามารถนำไปใช้ในโมเดลได้อย่างถูกต้อง ในขณะเดียวกันคอลัมน์ที่เป็นข้อความก็สามารถแปลงเป็นตัวเลขได้ โดยการสร้างคอลัมน์ใหม่ เช่น คอลัมน์ 'Fraud' ที่ใช้กำหนดค่า 1 สำหรับบริษัทที่ถูกระบุว่ามีการฉ้อโกง (เช่น EA และ STARK) และค่า 0 สำหรับบริษัทอื่น ๆ

3. การรวมข้อมูล (Data Integration) ข้อมูลจาก Price History, Balance Sheet, Cash Flow, Income Statement, และ Ratios-Key Metrics ถูกนำมารวมกันโดยใช้ pd.concat() ซึ่งช่วยให้สามารถนำข้อมูลจากหลายแหล่งมารวมเป็นชุดข้อมูลเดียวกันได้ เพื่อให้สามารถทำการวิเคราะห์ได้อย่างครบถ้วนและมีประสิทธิภาพ

3.3 การสร้างและฝึกฝนโมเดล (Model Building and Training)

---

[4] Min-Max Scaling เป็นการปรับค่าของ feature ให้อยู่ในช่วง 0 ถึง 1 โดยใช้สูตร $x_{scaled} = \frac{x-min}{\max - min}$ โดยที่ x คือค่าของ feature ในแต่ละตัวอย่าง, min คือค่าของ feature ที่น้อยที่สุด, และ max คือค่าของ feature ที่มากที่สุด การใช้ Min-Max Scaling จะช่วยให้ข้อมูลมีช่วงที่สอดคล้องกันและช่วยเพิ่มประสิทธิภาพในการเรียนรู้ของโมเดลในบางกรณี (กิตติมศักดิ์ ในจิต, n.d.)

การวิจัยนี้ใช้โมเดล LR และ RF ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการจำแนกประเภทข้อมูล โมเดลแต่ละแบบจะถูกสร้างและฝึกฝนด้วยข้อมูลที่เตรียมไว้เพื่อให้สามารถคาดการณ์โอกาสการเกิดการฉ้อโกงได้ โดยจะใช้ขั้นตอนการสร้างและฝึกฝนโมเดลดังนี้

Logistic Regression Model (LR)

สมการ Logistic function จะแปลงค่าให้มีค่าอยู่ในช่วง 0 ถึง 1 ตามรูปแบบความสัมพันธ์ต่อไปนี้และ เมื่อค่าอยู่ในช่วง 0 ถึง 1 จึงสามารถตีความได้ว่าค่าดังกล่าวคือความน่าจะเป็นที่สิ่งหนึ่งจะเกิดขึ้น ให้ P(y = 1) (การฉ้อโกง) (Hosmer et al., 2013)

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

โดยที่ $P(y = 1|x)$ คือความน่าจะเป็นที่ y จะเท่ากับ 1 (การฉ้อโกง) โดยมีตัวแปรอิสระ x

β_0 คือค่าคงที่ (intercept)

$\beta_1, \beta_2, \dots, \beta_n$ คือค่าสัมประสิทธิ์ของตัวแปรอิสระ

$x_1, x_2, \dots, x_n$ คือตัวแปรอิสระที่ใช้ในการพิจารณาความน่าจะเป็น เช่น ตัวชี้วัดทางการเงินหรือราคาหุ้น

สร้างและฝึกโมเดลโดยแยกข้อมูลเป็น Training Set และ Testing Set ด้วยเทคนิค Train-Test Split (80:20) จากนั้นสร้างโมเดลด้วย LogisticRegression จากไลบรารี sklearn ใช้ method fit() ในการฝึกฝนโมเดลกับ Training Set วิเคราะห์ค่าสัมประสิทธิ์ β เพื่อศึกษาความสำคัญของตัวแปร (Feature Importance) โดย feature ที่มีค่าสัมประสิทธิ์มาก (ทั้งค่าบวกและค่าลบ) บ่งบอกถึง feature ที่มีผลต่อการทำนายมาก

ประเมินผลโมเดล ใช้ชุด Testing Set เพื่อทดสอบโมเดล โดยผลลัพธ์การคาดการณ์ ($\hat{y}$) ได้จากเกณฑ์ P(y=1|X) > 0.5 ใช้ตัวชี้วัด ได้แก่ Accuracy, Precision, Recall และ F1-Score ในการวัดประสิทธิภาพของโมเดล

Random Forest Model (RF)

RF เป็นอัลกอริทึมที่ใช้การสร้างต้นไม้การตัดสินใจ (Decision Trees) จำนวนมาก และรวมผลลัพธ์จากต้นไม้แต่ละต้นเพื่อลดความแปรปรวนของโมเดล การคาดการณ์ขั้นสุดท้ายอิงจากการโหวตส่วนใหญ่ (Majority Voting) (Breiman, 2001)

$$\hat{y} = majority\ vote\ (y_1, y_2, \dots, y_m)$$

โดยที่ $\hat{y}$ คือผลลัพธ์ที่โมเดลทำนาย

$y_1, y_2, \dots, y_m$ คือผลลัพธ์ของแต่ละต้นไม้ใน Random Forest ทั้งหมด m ต้น

หากคลาสที่ได้รับคะแนนเสียงมากที่สุดคือ "1" (การฉ้อโกง) โมเดลก็จะทำนายว่าเป็นการฉ้อโกง ในขณะที่หากคะแนนเสียงมากที่สุดเป็น "0" โมเดลจะทำนายว่าไม่มีการฉ้อโกง

แยกข้อมูลเป็น Training Set และ Testing Set เหมือนกับ Logistic Regression ใช้ RandomForestClassifier จากไลบรารี sklearn ตั้งค่าพารามิเตอร์ ดังนี้

- n_estimators: จำนวนต้นไม้ = 100
- max_depth: ความลึกสูงสุดของแต่ละต้นไม้ = None (ไม่มีการจำกัดความลึก)

โมเดลถูกฝึกฝนด้วย Training Set และตรวจสอบความสำคัญของตัวแปร (Feature Importance) สำหรับต้นไม้แต่ละต้น โดยพิจารณาจากค่าการลดความไม่แน่นอน (Impurity) ของข้อมูลในแต่ละโหนด ซึ่งคำนวณโดยใช้สูตร Gini Impurity

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

โดยที่ p_i คือความน่าจะเป็นของการเกิดแต่ละคลาส i ในโหนดนั้น ๆ เช่น การฉ้อโกงหรือไม่ฉ้อโกง

C คือจำนวนของคลาสหรือกลุ่มที่มีอยู่ในข้อมูล

อีกทั้งยังปรับปรุงโมเดลด้วยการใช้ GridSearchCV เพื่อค้นหาพารามิเตอร์ที่เหมาะสมที่สุดช่วยเพิ่มประสิทธิภาพของ Random Forest Classifier โดยการทดสอบค่าพารามิเตอร์ต่าง ๆ ได้แก่

- n_estimators = range(2, 20, 2)
- max_depth = [None, 5, 10]
- min_samples_split = [2, 5, 10]

และประเมินผลด้วย Cross-Validation แบบ 5-Fold ซึ่งช่วยลดความเสี่ยงของ Overfitting และเพิ่มความสามารถในการทำนายข้อมูลใหม่ (Generalization) ได้ดียิ่งขึ้น ซึ่งจะอธิบายวิธีการในหัวข้อถัดไป

3.4 การประเมินผล (Model Evaluation)

หลังจากการฝึกฝนโมเดลแล้ว จะมีการประเมินประสิทธิภาพของโมเดลโดยใช้ตัวชี้วัดการทำนายที่สำคัญเพื่อวัดความสามารถของโมเดลในการตรวจจับการฉ้อโกง เครื่องมือที่ใช้คือ **Confusion Metrix** ใช้ในการวัดประสิทธิภาพของโมเดลในการทำงานด้าน Classification โดยแสดงให้เห็นรายละเอียดของการทำนายทั้งหมดในรูปแบบของตาราง ช่วยให้เข้าใจได้ว่าโมเดลทำนายค่าถูกต้องหรือผิดพลาดอย่างไรบ้าง สำหรับปัญหาแบบสองคลาส (Binary Classification) โดย Confusion Matrix จะมีโครงสร้างเป็น 2x2 ดังแสดงใน*ภาคผนวก ค.*

1) Accuracy เป็นค่าที่แสดงถึงความถูกต้องของการทำนายโดยรวม โดยคำนวณจากจำนวนการทำนายที่ถูกต้องทั้งหมดหารด้วยจำนวนข้อมูลทั้งหมด แม้ว่า Accuracy จะเป็นตัวชี้วัดที่ใช้บ่อย แต่มักมีข้อจำกัดในกรณีที่ข้อมูลมีความไม่สมดุล
2) Precision เป็นค่าที่แสดงถึงความแม่นยำในการทำนายการฉ้อโกง โดยคำนวณจากจำนวนครั้งที่ทำนายว่ามีการฉ้อโกงที่ถูกต้อง หารด้วยจำนวนครั้งที่ทำนายว่ามีการฉ้อโกงทั้งหมด ตัวชี้วัดนี้ช่วยประเมินความน่าเชื่อถือของโมเดลเมื่อทำนายว่ามีการฉ้อโกงเกิดขึ้น

3) Recall หรือ Sensitivity เป็นค่าที่แสดงถึงความสามารถของโมเดลในการตรวจจับการฉ้อโกง โดยคำนวณจากจำนวนครั้งที่ทำนายว่ามีการฉ้อโกงที่ถูกต้อง หารด้วยจำนวนการฉ้อโกงที่เกิดขึ้นจริง ตัวชี้วัดนี้ช่วยให้มั่นใจว่าโมเดลสามารถตรวจจับการฉ้อโกงทั้งหมดได้ครอบคลุม
4) F1-score: เป็นการรวม Precision และ Recall เข้าไว้ด้วยกันเพื่อหาค่าที่สมดุลระหว่างความแม่นยำและความครอบคลุมของโมเดล โดยค่า F1-score ที่สูงแสดงให้เห็นถึงความสามารถของโมเดลที่แม่นยำและครอบคลุม

นอกจากนี้ กรณีมีข้อจำกัดด้านปริมาณข้อมูลและต้องการหลีกเลี่ยงปัญหา overfitting หรือ bias จะใช้ประเมินผลด้วย **Cross-Validation** โดยผลลัพธ์จากแต่ละรอบ เช่น ค่า Accuracy, Precision, Recall หรือ F1-score จะถูกนำมาคำนวณค่าเฉลี่ยเพื่อใช้เป็นตัวแทนประสิทธิภาพของโมเดล ดังนี้

$$CV_{Score} = \frac{1}{k}\sum_{i=1}^{k} S_i$$

โดยที่ S_i คือค่าประสิทธิภาพของโมเดใน fold ที่ i

k คือจำนวน fold (ในที่นี้ให้เท่ากับ 5)

ข้อดีของ Cross-Validation คือการลดความลำเอียงในการประเมินผล เนื่องจากทุกข้อมูลใน dataset ได้มีโอกาสถูกใช้งานทั้งในชุดฝึกและชุดทดสอบ และยังสามารถระบุความแปรปรวนของโมเดลได้จากการวิเคราะห์ค่าผลลัพธ์แต่ละ fold เช่น ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน อย่างไรก็ตามการประมวลผลต้องใช้เวลามากขึ้นเนื่องจากต้องสร้างและทดสอบโมเดลหลายรอบตามจำนวน K ที่เลือก โดย ค่า K ถูกตั้งไว้เป็น 5 เพื่อให้ได้ผลลัพธ์ที่มีความสมดุลระหว่างความแม่นยำและเวลาในการประมวลผล (Paul, 2023)

3.5 สมมติฐาน

- การตรวจจับการฉ้อโกงทางการเงินโดยใช้ข้อมูลทางการเงินและราคาหุ้นร่วมกันจะมีความแม่นยำสูงกว่าการใช้ข้อมูลทางการเงินเพียงอย่างเดียวในการทำนายการฉ้อโกงทางการเงิน
- การใช้โมเดล Random Forest จะมีประสิทธิภาพในการตรวจจับการฉ้อโกงทางการเงินสูงกว่าโมเดล Logistic Regression โดยวัดจาก Accuracy, Precision, Recall และค่า F1 score

4. แหล่งข้อมูลที่ใช้

การวิจัยนี้ใช้แหล่งข้อมูลสำคัญจากฐานข้อมูล Eikon-DataStream และตลาดหลักทรัพย์แห่งประเทศไทย (SET) ซึ่งเป็นแหล่งข้อมูลทางการเงินและราคาหุ้นที่มีความน่าเชื่อถือ

ฐานข้อมูล Eikon-DataStream ประกอบด้วยข้อมูลงบการเงิน เช่น งบแสดงฐานะการเงิน งบกำไรขาดทุน และอัตราส่วนทางการเงิน ซึ่งสามารถใช้วิเคราะห์สภาพคล่องและความสามารถในการบริหารจัดการของบริษัท นอกจากนี้ยังมีข้อมูลราคาหุ้นรายสัปดาห์ย้อนหลัง ที่สะท้อนถึงความเชื่อมั่นของนักลงทุนและการตอบสนองต่อข่าวสารเกี่ยวกับบริษัทซึ่งช่วยในการสร้างตัวแปรสำคัญในการตรวจจับการฉ้อโกง

ตลาดหลักทรัพย์แห่งประเทศไทย (SET) เป็นแหล่งข้อมูลที่ให้ข้อมูลกลุ่มอุตสาหกรรม รวมถึงมาตรการกำกับดูแลที่ใช้ในการควบคุมพฤติกรรมการซื้อขายที่ผิดปกติ ข้อมูลเหล่านี้ช่วยให้สามารถเลือกกลุ่มบริษัทที่มีความคล้ายคลึงกันในแง่ของหมวดธุรกิจ ทำให้การวิเคราะห์มีความครอบคลุมและสามารถเปรียบเทียบความเสี่ยงในการฉ้อโกงได้ในบริบทของประเทศไทย

5. ผลการศึกษา

จากการนำเข้าข้อมูลและสร้าง Data frame ของ งบแสดงฐานะการเงิน (Balance sheet) งบกำไรขาดทุน (Income statement) งบกระแสเงินสด (cash-flow statement) อัตราส่วนทางการเงินที่สำคัญ (Ratios-key metrics) และราคาหุ้นย้อนหลัง (Price history) พบการกระจายตัวของข้อมูลที่สามารถดูได้จาก *ภาคผนวก ง.* ซึ่งแสดงข้อมูลที่มีความผิดปกติบางประการในการกระจายตัวของข้อมูล โดยในเบื้องต้นพบความผิดปกติของข้อมูล เช่น **outliers** ที่อาจมีผลกระทบต่อการสร้างโมเดลทำนายการฉ้อโกง ข้อมูลที่เป็น outliers อาจบิดเบือนผลการวิเคราะห์และส่งผลต่อประสิทธิภาพของโมเดลในการทำนายความน่าจะเป็นของการฉ้อโกงได้

ในการตรวจจับการฉ้อโกงทางการเงิน การลดจำนวน **false positives**[5] เป็นสิ่งสำคัญ เนื่องจากการระบุบริษัทที่มีการฉ้อโกงผิดพลาดว่าไม่มีการฉ้อโกงอาจนำไปสู่ความเสียหายทางการเงินอย่างมหาศาล ซึ่งอาจสูงถึงหลายพันล้านดอลลาร์ ดังนั้น การประเมินประสิทธิภาพของโมเดลจึงควรมุ่งเน้นที่ **precision** ซึ่งวัดความแม่นยำในการระบุเหตุการณ์ที่เป็นบวกจริงเมื่อโมเดลทำนายว่ามีการฉ้อโกง

งบแสดงฐานะการเงิน (Balance sheet)

ตารางที่ 1 สรุปการทดสอบและการประเมิณผล

Logistic Regression	Random Forest
Precision: 0.8889 หมายความว่าเมื่อโมเดลทำนายว่ามีการฉ้อโกงได้ 88.89%	**Precision**: 1.0000 ชี้ว่าโมเดลสามารถทำนายการฉ้อโกงได้ถูกต้องทั้งหมด
Accuracy: 0.9474แสดงถึงความแม่นยำโดยรวมที่สูง	**Accuracy**: 0.9708 มีความแม่นยำโดยรวมสูงกว่า LR
Recall: 0.6957 ชี้ให้เห็นถึงความสามารถในการตรวจจับการฉ้อโกงที่ค่อนข้างต่ำ	**Recall**: 0.7826 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่สูงกว่า LR
F1 Score: 0.7805 แสดงถึงความสมดุลระหว่างความแม่นยำและการครอบคลุมที่ยังต้องการการปรับปรุง	**F1 Score**: 0.8780 สะท้อนถึงความสมดุลที่ดีระหว่างความแม่นยำและการครอบคลุม

Cross-Validation

จากการใช้ cross-validation แบบ 5-fold กับแบบจำลอง RF พบว่าแบบจำลองมีคะแนนความแม่นยำเฉลี่ยอยู่ที่ 0.92 ซึ่งบ่งชี้ว่าแบบจำลองสามารถทำนายผลลัพธ์ได้อย่างถูกต้องประมาณ 92% โดยแสดงรายละเอียดแต่ละ fold ใน*ภาคผนวก จ.*

[5] แสดงในภาคผนวก ค.

ตารางที่ 2 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์

ลำดับ	Logistic Regression		Random Forest	
	Feature	ความสำคัญ (Coefficient)	Feature	ความสำคัญ
1.	**Retained Earnings (Accumulated Deficit)**	3.843123	Minority Interest	0.068724
2.	Long Term Debt	2.282651	**Other Payables**	0.060696
3.	**Other Payables**	-1.878364	Receivables - Other	0.059268
4.	**Minority Interest - Non Redeemable**	-1.835439	**Minority Interest - Non Redeemable**	0.055311
5.	Curr. Port. of LT Capital Leases, Suppl.	-1.815288	**Retained Earnings (Accumulated Deficit)**	0.051349

งบกำไรขาดทุน (Income statement)

ตารางที่ 3 สรุปการทดสอบและการประเมิณผล

Logistic Regression	Random Forest
Precision: 0.8947 หมายความว่าเมื่อโมเดลทำนายว่ามีการฉ้อโกง จะถูกต้องประมาณ 89.47%	**Precision**: 1.0000 ชี้ว่าโมเดลสามารถทำนายการฉ้อโกงได้ถูกต้องทั้งหมด
Accuracy: 0.8901 แสดงถึงความแม่นยำโดยรวมที่ค่อนข้างสูง	**Accuracy**: 0.9725 มีความแม่นยำสูงกว่ามาก
Recall: 0.4857 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่ต่ำ	**Recall**: 0.8571 บ่งบอกถึงความสามารถในการตรวจจับการฉ้อโกงที่ดีกว่า LR
F1 Score: 0.6296 ต้องการการปรับปรุงในส่วนของการตรวจจับ	**F1 Score**: 0.9231แสดงถึงความสมดุลที่ดีและมีประสิทธิภาพในการตรวจจับการฉ้อโกง

Cross-Validation

จากการประเมินประสิทธิภาพของแบบจำลอง RF พบว่าแบบจำลองมีความแม่นยำเฉลี่ยอยู่ที่ 0.93 ซึ่งบ่งชี้ว่าแบบจำลองสามารถทำนายผลลัพธ์ได้อย่างถูกต้องประมาณ 93% แสดงรายละเอียดแต่ละ fold ใน*ภาคผนวก จ.*

ตารางที่ 4 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์

ลำดับ	Logistic Regression		Random Forest	
	Feature	ความสำคัญ (Coefficient)	Feature	ความสำคัญ

1.	**Other, Net**	-5.371017	**Other, Net**	0.162174
2.	Normalized EBITDA	2.706634	Normalized Income After Taxes	0.092276
3.	Inc Tax Ex Impact of Sp Items	1.663117	Normalized Inc. Avail to Com.	0.066833
4.	Total Operating Expense	-1.253544	Selling/General/Admin. Expenses, Total	0.059423
5.	Diluted EPS Excluding ExtraOrd Items	-1.138966	Income Available to Com Excl ExtraOrd	0.055324

งบกระแสเงินสด (Cash-flow statement)

ตารางที่ 5 สรุปการทดสอบและการประเมินผล

Logistic Regression	Random Forest
Precision: 1.0000 หมายความว่าโมเดลทำนายว่ามีการฉ้อโกงได้อย่างแม่นยำ	**Precision:** 1.0000 หมายความว่าโมเดลทำนายว่ามีการฉ้อโกงได้อย่างแม่นยำ
Accuracy: 0.9669 แสดงถึงความแม่นยำโดยรวมที่สูง	**Accuracy:** 0.9613 น้อยกว่า Logistic Regression เล็กน้อย แต่ยังถือว่ามีความแม่นยำสูง
Recall: 0.8125 ชี้ให้เห็นถึงความสามารถในการตรวจจับการฉ้อโกงที่แท้จริงค่อนข้างสูง	**Recall:** 0.7812 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่ต่ำกว่า LR
F1 Score: 0.8966 แสดงถึงความสมดุลระหว่างความแม่นยำและการครอบคลุมที่น่าพอใจ	**F1 Score:** 0.8772 มีความสมดุลที่ดีระหว่างความแม่นยำและการครอบคลุม

Cross-Validation

จากการประเมินประสิทธิภาพของแบบจำลอง RF พบว่าแบบจำลองนี้มีความแม่นยำเฉลี่ยอยู่ที่ 0.95 ซึ่งถือว่าสูงมาก โดยแสดงรายละเอียดแต่ละ fold ใน*ภาคผนวก จ.*

ตารางที่ 6 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์

ลำดับ	**Logistic Regression**		**Random Forest**	
	Feature	**ความสำคัญ (Coefficient)**	**Feature**	**ความสำคัญ**
1.	Net Income/Starting Line	3.048152	Issuance (Retirement) of Debt, Net	0.116953
2.	Other Assets	2.604460	Accounts Receivable	0.116451

3.	Lease liability Reduced, Supplemental	-2.603353	Changes in Working Capital	0.104605
4.	Unusual Items	2.402014	Cash Taxes Paid	0.058855
5.	Net Changes in Working Capital	2.055454	Cash from Financing Activities	0.057712

อัตราส่วนทางการเงินที่สำคัญ (Ratios-key metrics)

ตารางที่ 7 สรุปการทดสอบและการประเมินผล

Logistic Regression	Random Forest
Precision: 1.0000 หมายความว่าเมื่อโมเดลทำนายว่ามีการฉ้อโกงถูกต้องทั้งหมด	**Precision:** 1.0000 แสดงว่าโมเดลสามารถทำนายการฉ้อโกงได้อย่างถูกต้องทั้งหมด
Accuracy: 0.9615 แสดงถึงความแม่นยำโดยรวมที่สูง	**Accuracy:** 0.9890 มีความแม่นยำโดยรวมสูงมาก
Recall: 0.7083 ชี้ให้เห็นถึงความสามารถในการตรวจจับการฉ้อโกงที่สูง	**Recall:** 0.9167 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่สูงมาก
F1 Score: 0.8293 แสดงถึงความสมดุลระหว่างความแม่นยำและการครอบคลุมที่น่าพอใจ	**F1 Score:** 0.9565 สะท้อนถึงความสมดุลที่ดีเยี่ยมระหว่างความแม่นยำและการครอบคลุม

Cross-Validation

จากการประเมินประสิทธิภาพของแบบจำลอง RF พบว่าแบบจำลองนี้มีความแม่นยำเฉลี่ยอยู่ที่ 0.94 ซึ่งแสดงให้เห็นถึงความสามารถในการทำนายผลลัพธ์ได้อย่างถูกต้องในระดับสูง แสดงรายละเอียดใน*ภาคผนวก* จ.

ตารางที่ 8 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์

ลำดับ	Logistic Regression		Random Forest	
	Feature	**ความสำคัญ (Absolute Coefficient)**	**Feature**	**ความสำคัญ**
1.	Pretax ROA	-4.686737	**Asset Turnover**	0.130300
2.	**ROE**	3.228335	**ROE**	0.124088
3.	**Times Interest Earned**	3.096309	**Times Interest Earned**	0.111766
4.	Pretax ROE	2.714419	EBITDA Margin	0.109864
5.	**Asset Turnover**	-2.464797	x Tax Complement	0.075747

ราคาหุ้นย้อนหลัง (Price history) และตัวแปรสำคัญ 3 อันดับแรกข้างต้นของทุกงบทั้งสองโมเดล

ตารางที่ 9 สรุปการทดสอบและการประเมินผล**ก่อน**รวมราคาหุ้นกับตัวแปรสำคัญ 3 อันดับ

Logistic Regression	Random Forest
Precision: 0.9752 หมายความว่าการทำนายว่ามีการฉ้อโกงของโมเดลถูกต้องประมาณ 97.52%	**Precision:** 1.0000 แสดงว่าโมเดลสามารถทำนายการฉ้อโกงได้ถูกต้องทั้งหมด
Accuracy: 0.9861 แสดงถึงความแม่นยำโดยรวมที่สูงมาก	**Accuracy:** 1.0000 มีความแม่นยำโดยรวมสูงที่สุด
Recall: 0.9255 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่แท้จริงสูง	**Recall**: 1.0000 ชี้ให้เห็นถึงความสามารถในการตรวจจับการฉ้อโกงสูงที่สุด
F1 Score: 0.9497 สะท้อนถึงความสมดุลที่ดีระหว่างความแม่นยำและการครอบคลุม	**F1 Score:** 1.0000 บ่งบอกถึงความสมดุลที่ยอดเยี่ยมระหว่างความแม่นยำและการครอบคลุมที่สุด

กรณีนี้แบบจำลอง RF ที่ได้คะแนนสมบูรณ์แบบอาจเกิดจากข้อมูลรั่วไหล ชุดข้อมูลไม่สมดุล หรือแบบจำลองที่ซับซ้อนเกินไป และได้ทำการแก้ไขโดยใช้ cross-validation ได้ผลดังนี้ต่อไปนี้

Cross-Validation

จากการประเมินประสิทธิภาพของแบบจำลอง RF พบว่าแบบจำลองมีความแม่นยำเฉลี่ยอยู่ที่ 0.95 ซึ่งบ่งชี้ถึงความสามารถในการทำนายผลลัพธ์ได้อย่างถูกต้องในระดับที่สูงโดยแสดงรายละเอียดแต่ละ fold ในภาคผนวก จ. และช่วยลดความเสี่ยงของ overfitting และทำให้มั่นใจได้ว่าแบบจำลองมีความสามารถในการสรุปข้อมูลใหม่ได้ดีสนับสนุนความน่าเชื่อถือของแบบจำลอง RF

ตารางที่ 10 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์**ก่อน**รวมราคาหุ้นกับตัวแปรสำคัญ 3 อันดับ

ลำดับ	**Logistic Regression**		**Random Forest**	
	Feature	**ความสำคัญ (Coefficient)**	**Feature**	**ความสำคัญ**
1.	Pretax ROA	-19.539314	Other, Net	0.183847
2.	ROE	18.693615	Other Operating Expenses, Total	0.151875
3.	Other, Net	-6.546091	Asset Turnover	0.141930
4.	Other Assets	5.167174	ROE	0.075885
5.	Changes in Working Capital	4.650499	Minority Interest - Non Redeemable	0.048680
6.	Times Interest Earned	3.636727	Times Interest Earned	0.041376
7.	Other Operating Expenses, Total	3.323447	Minority Interest	0.038468

8.	Net Changes in Working Capital	-3.318248	Other Payables	0.037701
9.	Inc Tax Ex Impact of Sp Items	3.226638	Inc Tax Ex Impact of Sp Items	0.037321
10.	Lease liability Reduced, Supplemental	-2.641305	Changes in Working Capital	0.036902
11.	Minority Interest	-2.620326	Issuance (Retirement) of Debt, Net	0.036741
12.	Long Term Debt	2.359958	Net Changes in Working Capital	0.026137
13.	Total Operating Expense	-1.947789	Long Term Debt	0.024349
14.	Normalized EBITDA	-1.907221	Other Assets	0.021880
15.	Other Payables	-1.737138	Pretax ROA	0.021009
16.	Asset Turnover	-1.631240	Total Operating Expense	0.020423
17.	Net Income/Starting Line	1.578257	Net Income/Starting Line	0.017987
18.	Minority Interest - Non Redeemable	1.333969	Retained Earnings (Accumulated Deficit)	0.014966
19.	Retained Earnings (Accumulated Deficit)	0.731331	Normalized EBITDA	0.012840
20.	Issuance (Retirement) of Debt, Net	0.277355	Lease liability Reduced, Supplemental	0.009681

ตารางที่ 11 สรุปการทดสอบและการประเมินผล**หลัง**รวมราคาหุ้นกับตัวแปรสำคัญ 3 อันดับ

Logistic Regression	Random Forest
Precision: 0.9590 หมายความว่าการทำนายว่ามีการฉ้อโกงของโมเดลถูกต้องประมาณ 95.90%	**Precision:** 1.0000 แสดงว่าโมเดลสามารถทำนายการฉ้อโกงได้ถูกต้องทั้งหมด
Accuracy: 0.9839 แสดงถึงความแม่นยำโดยรวมที่สูงมาก	**Accuracy:** 1.0000 มีความแม่นยำโดยรวมสูงที่สุด
Recall: 0.9262 แสดงถึงความสามารถในการตรวจจับการฉ้อโกงที่แท้จริงสูงถึง 92.62%	**Recall**: 1.0000 ชี้ให้เห็นถึงความสามารถในการตรวจจับการฉ้อโกงที่สูงมากที่สุด

F1 Score: 0.9423 สะท้อนถึงความสมดุลที่ดีระหว่างความแม่นยำและการครอบคลุม	**F1 Score:** 1.0000 บ่งบอกถึงความสมดุลที่ดีที่สุดระหว่างความแม่นยำและการครอบคลุม

เช่นเดียวกัน ในกรณีนี้แบบจำลอง RF ที่ได้คะแนนสมบูรณ์แบบอาจเกิดจากข้อมูลรั่วไหล ชุดข้อมูลไม่สมดุล หรือแบบจำลองที่ซับซ้อนเกินไป และได้ทำการแก้ไขโดยใช้ cross-validation ได้ผลดังนี้ต่อไปนี้

Cross-Validation

จากการประเมินประสิทธิภาพของแบบจำลอง RF พบว่าแบบจำลองมีความแม่นยำเฉลี่ยอยู่ที่ 0.95 ซึ่งบ่งชี้ถึงความสามารถในการทำนายผลลัพธ์ได้อย่างถูกต้องในระดับที่สูงถึง 95% แสดงรายละเอียด ในภาคผนวก จ. ตารางที่ 12 สรุปความสำคัญของฟีเจอร์ต่าง ๆ ตามค่าค่าสัมประสิทธิ์**หลัง**รวมราคาหุ้นกับตัวแปรสำคัญ 3 อันดับ

ลำดับ	**Logistic Regression**		**Random Forest**	
	Feature	**ความสำคัญ (Coefficient)**	**Feature**	**ความสำคัญ**
1.	Pretax ROA	-19.565968	Other, Net	0.303896
2.	ROE	18.315308	Other Operating Expenses, Total	0.163865
3.	Other, Net	-6.619273	Asset Turnover	0.099335
4.	Other Assets	5.087376	Net Changes in Working Capital	0.076027
5.	Changes in Working Capital	4.416327	Other Payables	0.065055
6.	Times Interest Earned	3.697296	Minority Interest - Non Redeemable	0.041126
7.	Other Operating Expenses, Total	3.227230	Lease liability Reduced, Supplemental	0.036629
8.	Inc Tax Ex Impact of Sp Items	3.186754	Long Term Debt	0.034441
9.	Net Changes in Working Capital	-3.126576	Times Interest Earned	0.033373
10.	Minority Interest	-3.105136	Normalized EBITDA	0.030905
11.	Lease liability Reduced, Supplemental	-2.813632	Retained Earnings (Accumulated Deficit)	0.030696
12.	Long Term Debt	2.251620	ROE	0.025559

13.	Total Operating Expense	-2.147609	Pretax ROA	0.025482
14.	Normalized EBITDA	-1.904049	Net Income/Starting Line	0.009770
15.	Minority Interest - Non Redeemable	1.627962	Close	0.005961
16.	Other Payables	-1.500919	Total Operating Expense	0.005377
17.	Asset Turnover	-1.478210	Issuance (Retirement) of Debt, Net	0.004324
18.	Net Income/Starting Line	1.249308	Inc Tax Ex Impact of Sp Items	0.004035
19.	Close	1.166638	Other Assets	0.003287
20.	Retained Earnings (Accumulated Deficit)	0.875481	Changes in Working Capital	0.000857
21.	Issuance (Retirement) of Debt, Net	0.538709	Minority Interest	0.000000

จากการทดสอบและประเมินผลการใช้โมเดล LR และ RF หลังจากรวมราคาหุ้นกับตัวแปรสำคัญ 3 อันดับแรก พบว่าโมเดลทั้งสองมีประสิทธิภาพสูงในการทำนายการฉ้อโกงทางการเงิน โดยโมเดล LR ได้คะแนน Precision ที่ 95.90% ซึ่งแสดงถึงความแม่นยำในการทำนายการฉ้อโกงที่สูง ในขณะที่โมเดล RF มี Precision, Accuracy, Recall และ F1 Score ที่เท่ากับ 1.0000 ซึ่งแสดงให้เห็นถึงความสามารถในการทำนายการฉ้อโกงได้อย่างถูกต้องทั้งหมด และสามารถตรวจจับการฉ้อโกงได้สมบูรณ์ แต่แน่นอนว่าผลลัพธ์นี้อาจเกิดจากปัญหาของข้อมูลรั่วไหลหรือชุดข้อมูลที่ไม่สมดุล จึงใช้ cross-validation ได้ช่วยลดปัญหานี้ โดยได้ความแม่นยำเฉลี่ยที่ 95% ซึ่งแสดงถึงความสามารถในการทำนายที่สูง

สำหรับการประเมินความสำคัญของฟีเจอร์ต่าง ๆ พบว่าใน **LR** ฟีเจอร์ที่มีความสำคัญสูงสุดคือ **Pretax ROA** (-19.565968) ซึ่งเป็นตัวบ่งชี้ความสามารถในการทำกำไรก่อนภาษีของบริษัท หากค่านี้ลดลงมาก อาจสะท้อนถึงความพยายามในการบิดเบือนข้อมูลเพื่อปกปิดผลประกอบการที่แท้จริง รองลงมาคือ **ROE** (18.315308) อาจบ่งบอกถึงผลตอบแทนต่อผู้ถือหุ้นที่สูงผิดปกติ ซึ่งอาจชี้ถึงการจัดการรายงานผลกำไรเกินจริง และ **Other, Net** (-6.619273) อาจสะท้อนถึงรายการที่ไม่ชัดเจนในงบการเงิน เช่น รายการที่เกิดขึ้นเพียงครั้งเดียวหรือรายการปรับปรุงที่อาจถูกใช้เพื่อซ่อนการฉ้อโกง ขณะที่ใน **RF** ฟีเจอร์ที่สำคัญที่สุดคือ **Other, Net** (0.303896) ตามด้วย **Other Operating Expenses, Total** (0.163865) เป็นตัวบ่งชี้ถึงต้นทุนที่ไม่โปร่งใสหรือการบันทึกค่าใช้จ่ายผิดปกติในงบการเงิน และ **Asset Turnover** (0.099335) แสดงถึงประสิทธิภาพในการใช้สินทรัพย์ของบริษัท หากค่าสูงผิดปกติ อาจเป็นผลจากการบันทึกสินทรัพย์หรือรายได้เกินจริง

เมื่อเปรียบเทียบผลการทดสอบทั้งก่อนและหลังรวมตัวแปรราคาหุ้น พบว่า LR ความแม่นยำลดลงเล็กน้อย หลังจากการเพิ่มตัวแปรราคาหุ้น แต่ยังคงแสดงถึงความสามารถในการทำนายการฉ้อโกงได้ดี โมเดล RF มีความสามารถในการทำนายที่ดีกว่า LR แต่ผลลัพธ์ก่อนและหลังรวมราคาหุ้นใน RF มีปัญหาข้อมูลรั่วไหล (Data leakage) และชุดข้อมูลที่ไม่สมดุล (Imbalanced Data) อีกทั้ง cross-validation มีผลลัพธ์ที่ 95% เท่ากัน จึงไม่สามารถสรุปได้แน่ชัดของอิทธิพลของตัวแปรราคาหุ้นในการเพ่มประสิทธิภาพการตรวจจับการฉ้อโกงใน RF

6. บทสรุปและนัยเชิงนโยบาย

การศึกษานี้ได้เปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression (LR) และ Random Forest (RF) ในการตรวจจับการฉ้อโกงทางการเงิน โดยใช้ข้อมูลทางการเงินและราคาหุ้นย้อนหลังจากบริษัทในตลาดหลักทรัพย์แห่งประเทศไทย ผลการศึกษาพบว่า RF มีประสิทธิภาพที่ดีกว่า LR ในตัวชี้วัดสำคัญ เช่น ความแม่นยำ (Accuracy), ความแม่นยำเชิงบวก (Precision), และความครอบคลุม (Recall) โดยเฉพาะเมื่อเพิ่มตัวแปรราคาหุ้น อย่างไรก็ตาม ผลลัพธ์ที่สมบูรณ์แบบในบางกรณีของ RF ชี้ให้เห็นถึงปัญหาข้อมูลรั่วไหล (Data Leakage) และความไม่สมดุลของข้อมูล (Imbalanced Data) ซึ่งต้องได้รับการแก้ไขผ่านการใช้วิธี Cross-Validation

เมื่อเปรียบเทียบกับวรรณกรรมที่เกี่ยวข้อง งานวิจัยเช่นของ Ma (2019) และ Wyrobek (2020) แสดงให้เห็นว่า RF มักมีประสิทธิภาพสูงในงานที่เกี่ยวข้องกับการตรวจจับการฉ้อโกง โดยมีข้อได้เปรียบในด้านการจัดการข้อมูลที่มีลักษณะซับซ้อน ในขณะที่งานวิจัยของ Beneish (1999) และ Fanning & Cogger (1998) เน้นการใช้อัตราส่วนทางการเงินในการตรวจจับพฤติกรรมฉ้อโกงเช่นเดียวกับ LR แต่พบข้อจำกัดในด้านความสามารถในการจัดการข้อมูลที่ไม่สมดุลและขนาดใหญ่ ซึ่งการศึกษานี้ช่วยยืนยันว่าการรวมตัวแปรราคาหุ้นใน RF สามารถเพิ่มความแม่นยำได้อย่างมีนัยสำคัญ อย่างไรก็ตาม งานวิจัยของ Wyrobek (2020) ชี้ให้เห็นถึงความสำคัญของการรักษาความโปร่งใสในการตีความโมเดล ซึ่ง LR ยังคงมีจุดเด่นในด้านนี้เมื่อเทียบกับ RF

นัยเชิงนโยบาย

1. การพัฒนาระบบตรวจจับการฉ้อโกง หน่วยงานที่เกี่ยวข้อง เช่น ตลาดหลักทรัพย์แห่งประเทศไทย (SET) และคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์ (ก.ล.ต.) ควรนำโมเดล RF ไปปรับใช้ในระบบตรวจจับการฉ้อโกง เนื่องจากประสิทธิภาพที่สูงกว่า และสามารถประยุกต์ใช้กับข้อมูลทางการเงินและราคาหุ้นร่วมกัน
2. การลดปัญหาข้อมูลไม่สมดุล ผู้กำหนดนโยบายควรสนับสนุนการจัดเก็บข้อมูลที่ครอบคลุมทุกกลุ่มตัวอย่าง รวมถึงการเพิ่มมาตรการสำหรับข้อมูลที่ไม่ได้รับการรายงานหรือถูกบิดเบือน
3. การส่งเสริมการใช้งาน AI และ ML จัดอบรมบุคลากรในภาคการเงินเกี่ยวกับการใช้เทคโนโลยี AI ในการตรวจจับการฉ้อโกง รวมถึงการพัฒนาขีดความสามารถในการตีความผลลัพธ์ของโมเดล RF เพื่อเพิ่มความโปร่งใสและการยอมรับ

7. บรรณานุกรม

กิตติมศักดิ์ ในจิต. (n.d.). *การ Scale ข้อมูลใน Machine Learning ด้วย Python*. Kittimasak.com. Retrieved December 20, 2024, from https://kittimasak.com/scale-machine-learning-python-standardization-min-max-scaling/

จันทนา วัฒนกาญจนะ. (2016). *A study of financial standard ratio in business sectors of listed companies in the stock exchange of Thailand 2548 to 2557 B.e*. Tci-thaijo.org. https://he02.tci-thaijo.org/index.php/Veridian-E-Journal/article/viewFile/75688/60948

จิตเกษม พรประพันธ์ และพรชนก เทพขาม. (2020). *ปรับโครงสร้างเศรษฐกิจ ทางออกของไทย*. Bot.or.th. https://www.bot.or.th/th/research-and-publications/articles-and-publications/articles/Article_10Nov2020.html

พิชิต อัคราทิตย์และคณะ. (2000). *บทบาทตลาดทุนไทยในการระดมทุนสู่ภาคธุรกิจ*. Sec.or.Th. https://www.sec.or.th/TH/Documents/Research/research-1143-roles.pdf

ปริย เตชะมวลไววิทย์. (2016). Sec.or.Th. https://www.sec.or.th/TH/Template3/Articles/2560/ac-post-25600904-act-sea-manipulation.pdf

สำนักข่าวอีไฟแนนซ์ไทย. (2024). *7 หุ้นถูก Force Sell มาร์เก็ตแคปรวมหายกว่า 1.5 แสนลบ*. Efinancethai.com. https://www.efinancethai.com/HotTopic/HotTopicMain.aspx?id=WGdtbWxWOFNuUk09

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly, 36*(4), 1293. https://doi.org/10.2307/41703508

Albrecht, W. S., Albrecht, C., & Albrecht, C. C. (2008). Current trends in fraud and its detection. *Information Security Journal: A Global Perspective, 17*(1), 2–12. https://doi.org/10.1080/19393550801934331

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., ... & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, *12*(19), 9637.

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal, 55*, 24–36.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

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.

Dorris, B. (2020). *President and CEO, Association of Certified Fraud Examiners: Foreword*. Report to the Nations. Association of Certified Fraud Examiners. https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf

Dyck, A., Morse, A., & Zingales, L. (2010). Who blows the whistle on corporate fraud? *The Journal of Finance, 65*(6), 2213–2253. https://doi.org/10.1111/j.1540-6261.2010.01614.x

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *Intelligent Systems in Accounting, Finance & Management, 7*, 21–41.

Feroz, E. H., Kwon, T. M., Pastena, V. S., & Park, K. (2000). The efficacy of red flags in predicting the SEC's targets: An artificial neural networks approach. *Intelligent Systems in Accounting, Finance & Management, 9*, 145–157.

Gaganis, C. (2009). Classification techniques for the identification of falsified financial statements: A comparative analysis. *Intelligent Systems in Accounting, Finance & Management, 16*, 207–229.

Gottlieb, O., Salisbury, C., Shek, H., & Vaidyanathan, V. (2006). Detecting corporate fraud: An application of machine learning. *A publication of the American Institute of Computing*, 100–215.

Guo, L., Song, R., Wu, J., Xu, Z., & Zhao, F. (2024). Integrating a machine learning-driven fraud detection system based on a risk management framework. *Applied and Computational Engineering, 87*, 80–86.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*, 995–1003.

Lenard, M. J., & Alam, P. (2009). An historical perspective on fraud detection: From bankruptcy models to most effective indicators of fraud in recent incidents. *Journal of Forensic & Investigative Accounting, 1*, 1–27.

Li, X., & Wang, Q. (2020). Comparative research on financial prediction models of listed companies based on machine learning. *Market Modernization, 7*, 150–152.

Ma, X. (2019). *Research on machine learning based Chinese company financial risks detection system* (Unpublished doctoral dissertation). Nanjing University, Nanjing, Jiangsu, China.

Min, M. O. M. O. (2017). Researchgate.net. https://www.researchgate.net/publication/336285132_Combined_Effect_of_Economic_Variables_on_Fraud_a_Survey_of_Developing_Countries

Paul. (2023). *What is Cross-validation (CV) and why do we need it?* KBTG Life. https://medium.com/kbtg-life/what-is-cross-validation-cv-and-why-do-we-need-it-fb4bac340991

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*, 19–50.

Persons, O. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research, 11*, 38–46.

Piyapanichayakul, R. (2023). *ปัญหาข้อมูลไม่สมดุล (imbalanced data in classification model)*. NT Cloud Solutions. https://ntcloudsolutions.ntplc.co.th/knowledge/imbalanced-data-classification/

PricewaterhouseCoopers. (2020.). Economic Crime and Fraud Survey in Thailand. PwC. Retrieved September 10, 2024, from https://www.pwc.com/th/en/consulting/forensic/economic crime-and-fraud-in-thailand.html

Qian, P., & Luo, M. (2015). Predicting accounting fraud in China. *Accounting Research, 7*, 18–25.

Ravisankar, P., Ravi, V., Raghava, R. G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*, 491–500.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57*, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

Wells, J. T. (1997). *Occupational fraud and abuse*. Obsidian Publishing.

Wiriyapong, N., Praiwan, Y., & Polkuamdee, N. (2024). Scandal stalks stocks. Bangkok Post. https://www.bangkokpost.com/business/investment/2833461/scandal-stalks-stocks

Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia computer science, 148, 45-54.

Sawangarreerak, S., & Thanathamathee, P. (2021). Detecting and analyzing fraudulent patterns of financial statement for open innovation using discretization and association rule mining. *Journal of Open Innovation: Technology, Market, and Complexity, 7*(2), 128.

Sethsathira, P. (2020). Economic Crime and Fraud Survey in Thailand. https://www.pwc.com/th/en/consulting/forensic/economic-crime-and-fraud-in-thailand.html

Simnett, R., Vanstraelen, A., & Chua, W. F. (2007). Assurance on sustainability reports: An international comparison. *Social Science Research Network*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1025467

Song, X. P., Hu, Z. H., Du, J. G., & Sheng, Z. H. (2014). Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. *Journal of Forecasting, 33*, 611–626.

Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: A comparative study using multicriteria analysis and multivariate statistical techniques. *European Accounting Review, 11*, 509–535.

Stice, J. D. (1991). Using financial and market information to identify pre-engagement factors associated with lawsuits against auditors. *The Accounting Review, 66*, 516–533.

Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science, 176*, 3037–3046.

Xu, H., Fan, G., & Song, Y. (2022). [Retracted] Application analysis of the machine learning fusion model in building a financial fraud prediction model. *Security and Communication Networks, 2022*(1), 8402329.

Zohuri, B., & Moghaddam, M. (2020). Deep learning limitations and flaws. *Mod. Approaches Mater. Sci*, *2*, 241-250.

8. ภาคผนวก

ภาคผนวก ก. รายชื่อบริษัทที่ไม่มีข่าวการทุจริตทั้งหมด 16 บริษัท และบริษัทที่มีการทุจริต 2 บริษัท รวมทั้งสิน 18 บริษัท ดังแสดงในตาราง

บริษัทที่มีการทุจริต	บริษัทที่ไม่มีข่าวการทุจริต
บริษัท พลังงานบริสุทธิ์ จำกัด (มหาชน) (EA)	บริษัท แอ๊บโซลูท คลีน เอ็นเนอร์จี้ จำกัด (มหาชน) (ACE)
	บริษัท บีบีจีไอ จำกัด (มหาชน) (BBGI)
	บริษัท บีซีพีจี จำกัด (มหาชน) (BCPG)
	บริษัท ซีเค พาวเวอร์ จำกัด (มหาชน) (CKP)
	บริษัท เอิร์ธ เท็ค เอนไวรอนเมนท์ จำกัด (มหาชน)
	บริษัท เอสพีซีจี จำกัด (มหาชน) (SPCG)
	บริษัท เสริมสร้าง พาวเวอร์ คอร์ปอเรชั่น จำกัด (มหาชน) (SSP)
	บริษัท ซุปเปอร์ เอนเนอร์ยี คอร์เปอเรชั่น จำกัด (มหาชน) (SUPER)
	บริษัท ท่าฉาง กรีน เอ็นเนอร์ยี่ จำกัด (มหาชน) (TGE)
	บริษัท ทีพีไอ โพลีน เพาเวอร์ จำกัด (มหาชน) (TPIPP)
บริษัท สตาร์ค คอร์ปอเรชั่น จำกัด (มหาชน) (STARK)	บริษัท เอกรัฐวิศวกรรม จำกัด (มหาชน) (AKR)
	บริษัท ซีพีที ไดร์ แอนด์ เพาเวอร์ จำกัด (มหาชน) (CPT)
	บริษัท ไลท์ติ้ง แอนด์ อีควิปเมนท์ จำกัด (มหาชน) (L&E)
	บริษัท เอสซีไอ อีเลคตริค จำกัด (มหาชน) (SCI)
	บริษัท ธีระมงคล อุตสาหกรรม จำกัด (มหาชน) (TMIm)
	ข้อมูล บริษัท ถิรไทย จำกัด (มหาชน) (TRTm)

ภาคผนวก ข. ตัวแปรต้นหรือ Column Name (Features) ในชุดข้อมูล Balance Sheet, Income Statement, Cash Flow, Ratios-Key Metrics, Price History

Balance Sheet	Period End Date, Total Liabilities & Shareholders' Equity, Curr. Port. of LT Capital Leases, Suppl., Total Current Assets less Inventory, Capital Lease Obligations, Other Current liabilities, Total, Total Long Term Debt, Total Current Liabilities, Accounts Receivable - Trade, Gross, Accounts Receivable - Trade, Net, Other Long Term Assets, Total, Retained Earnings (Accumulated Deficit), Total Current Assets, Net Debt Incl. Pref.Stock & Min.Interest, Cash & Equivalents, Long Term Investments, Tangible Book Value, Common Equity,

	Other Property/Plant/Equipment – Net, Common Stock, Total, Income Taxes Payable, Common Stock, Total Receivables, Net, Current Port. of LT Debt/Capital Leases, Notes Payable/Short Term Debt, Minority Interest, Total Debt, Total Liabilities, Total Common Shares Outstanding, Other Liabilities, Total, Other Long Term Assets, Total Assets Property/Plant/Equipment, Total – Net, Total Equity & Minority Interest, Shares Outs - Common Stock Primary Issue, Receivables – Other, Total Equity, Other Payables, Cash and Short Term Investments, Defered Income Tax - Long Term Asset, Treas Shares - Common Stock Prmry Issue Long Term Debt, Minority Interest - Non Redeemable, Additional Paid-In Capital, Company name, **Fraud**
Cash Flow	Period End Date, Long Term Debt, Net, Other Investing Cash Flow Items, Total, Cash Taxes Paid, Other Operating Cash Flow, Issuance (Retirement) of Debt, Net, Net Cash - Beginning Balance, Lease liability Reduced, Supplemental, Free Cash Flow, Financing Cash Flow Items, Depreciation/Depletion, Sale of Fixed Assets, Long Term Debt Reduction, Accounts Receivable, Purchase of Fixed Assets, Net Change in Cash, Unusual Items, Depreciation, Other Financing Cash Flow, Net Cash - Ending Balance, Cash from Financing Activities, Capital Expenditures, Cash from Investing Activities, Other Liabilities, Cash from Operating Activities, Sale/Maturity of Investment, Other Non-Cash Items, Other Assets, Net Income/Starting Line, Cash Interest Paid, Net Changes in Working Capital, Other Investing Cash Flow, Changes in Working Capital, Non-Cash Items, Company name, **Fraud**
Income Statement	Period End Date, Income Available to Com Incl ExtraOrd, Net Income After Taxes, Basic EPS Including Extraordinary Items, Normalized Inc. Avail to Com., Net Income, Basic Weighted Average Shares, Income Available to Com Excl ExtraOrd, Net Income Before Taxes, Diluted Normalized EPS, Inc Tax Ex Impact of Sp Items, Normalized Income After Taxes, Selling/General/Admin. Expenses, Total, Other Operating Expenses, Total, Total Operating Expense, Basic Normalized EPS, Depreciation, Supplemental, Interest Inc.(Exp.),Net-Non-Op., Total, Net Income Before Extra. Items, Cost of Revenue, Minority

	Interest, Supplemental, Provision for Income Taxes, Operating Income, Minority Interest, Normalized EBIT, Other, Net, Cost of Revenue, Total, Basic EPS Excluding Extraordinary Items, Revenue, Diluted EPS Including ExtraOrd Items, Selling/General/Administrative Expense, Diluted Weighted Average Shares, Total Revenue, Diluted EPS Excluding ExtraOrd Items, Diluted Net Income, Normalized Income Before Taxes, DPS - Common Stock Primary Issue, Normalized EBITDA, Company name, **Fraud**
Ratios-Key Metrics	Period End Date, Net Margin, x Tax Complement, Pretax Margin, Current Ratio, Effective Tax Rate, x Earnings Retention, Cash Cycle (Days), Reinvestment Rate, x Pretax Margin, Pretax ROE, Assets/Equity, Pretax ROA, Asset Turnover, EBITDA, Margin Operating Margin, Quick Ratio, Gross Margin, (Total Debt - Cash) / EBITDA, ROE, Debt/Equity, Times Interest Earned, x Leverage (Assets/Equity), % LT Debt to Total Capital, Company name, **Fraud**
Price History	Exchange Date, Close, Company name, **Fraud**

ภาคผนวก ค. แสดง Confusion Matrix 2x2 ดังนี้

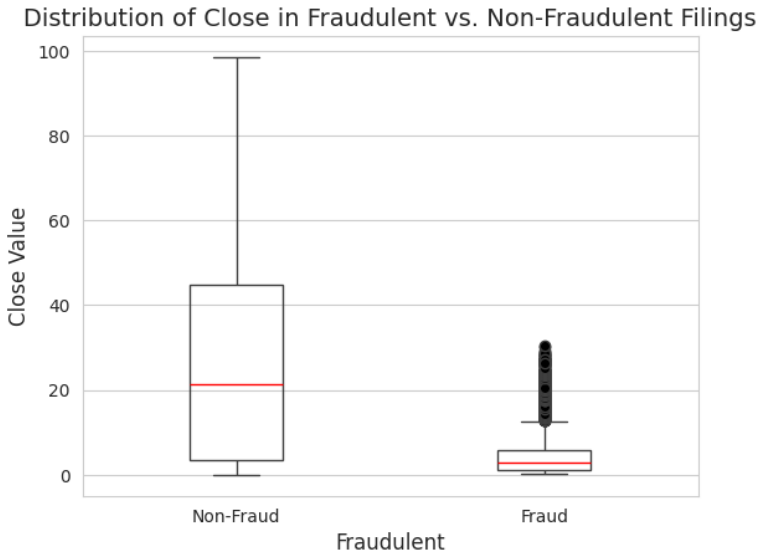
		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive (P)	True Positive (TP)	False Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

- **True Positive (TP)** สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่าจริง และสิ่งที่เกิดขึ้น คือ จริง
- **True Negative (TN)** สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น คือ ไม่จริง
- **False Positive (FP)** สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง False
- **Negative (FN)** สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

โดย TP TN FP และ FN ในตารางจะแทนด้วยค่าความถี่ ซึ่ง Confusion Matrix สามารถใช้คำนวณตัววัดผลต่าง ๆ เช่น Accuracy Precision Recall และ F1 Score เพื่อประเมินประสิทธิภาพของโมเดลได้อย่างละเอียดและครอบคลุม

ภาคผนวก ง. แสดงการกระจายตัวของข้อมูล Balance sheet, Income statement, Cash-flow statement, Ratios-key metrics และ Price history โดยแบ่งเป็นกลุ่ม Fraud และ Non-Fraud

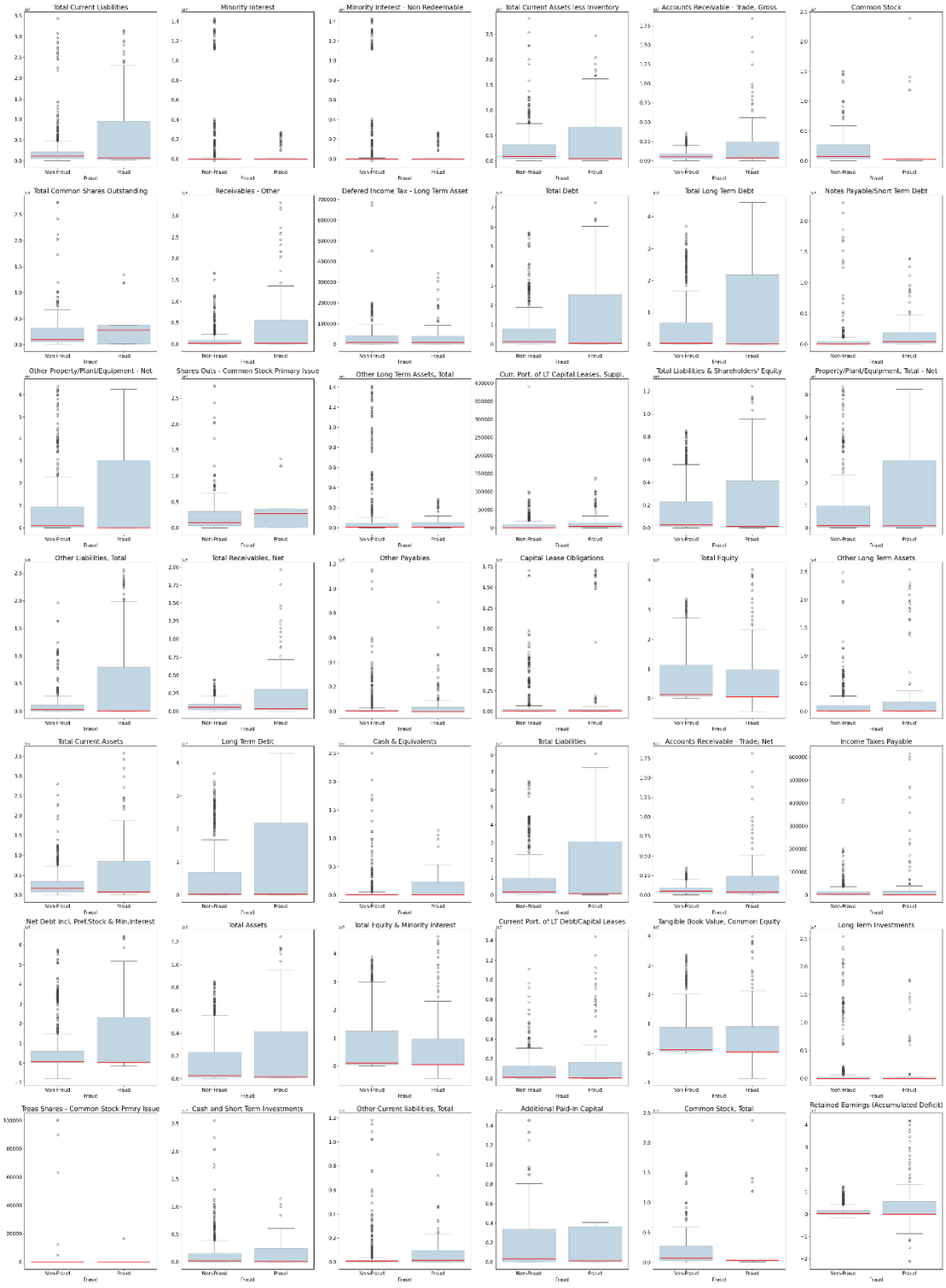
การกระจายตัวของข้อมูลราคาหุ้น (Price History)



ราคาหุ้นใน Fraud พบว่า outliers มีจำนวนมาก ซึ่งสามารถสังเกตได้จากรูปการกระจายตัวของข้อมูลที่มีค่าผิดปกติหรือแปลกปลอมออกไป โดย outliers เหล่านี้อาจเกิดจากเหตุการณ์ที่ไม่ปกติหรือการทำรายการที่มีลักษณะเฉพาะในกระบวนการทางการเงิน ซึ่งอาจเป็นตัวบ่งชี้ที่สำคัญในการตรวจจับการฉ้อโกงได้

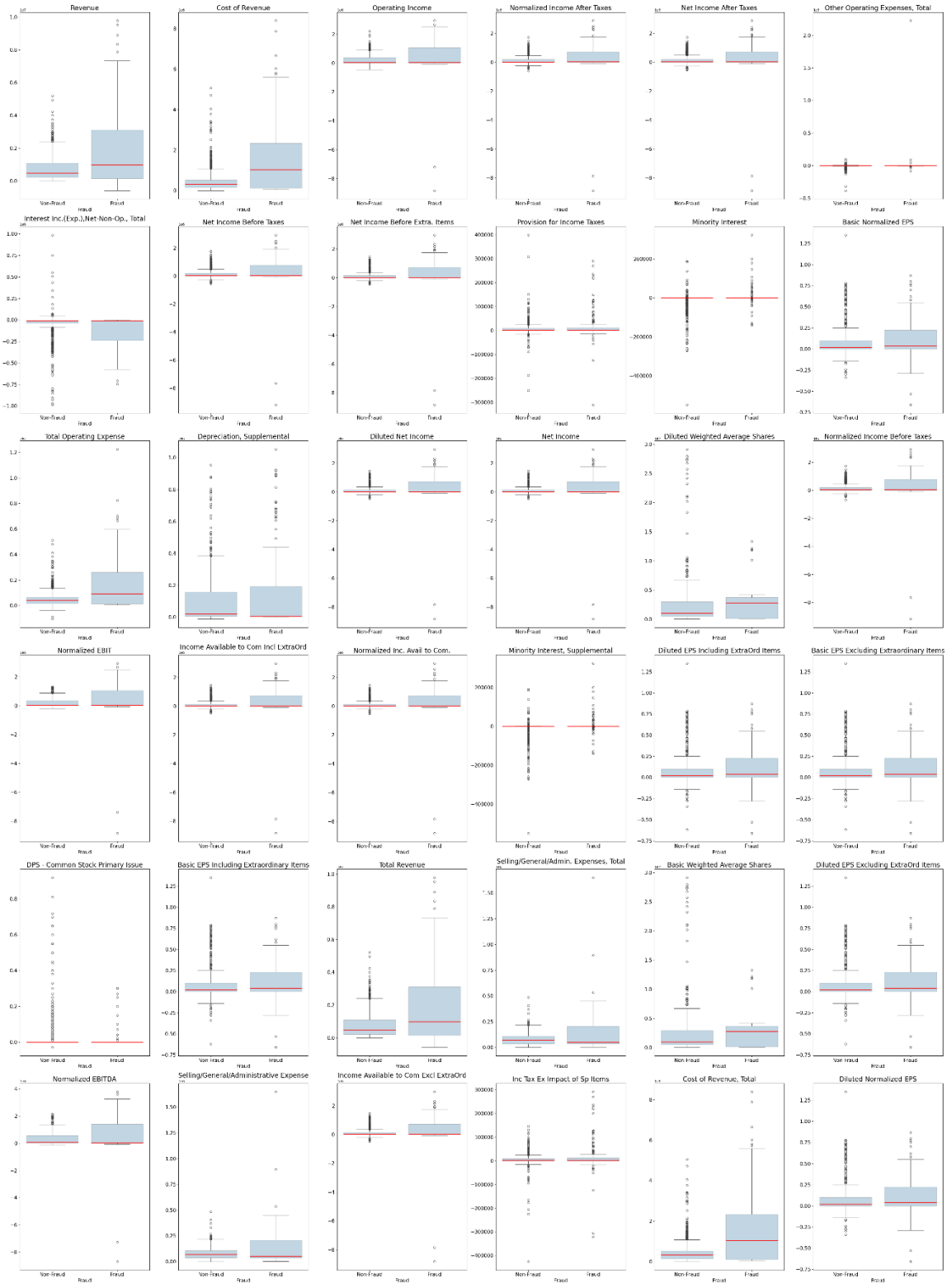
การกระจายตัวของข้อมูลงบแสดงสถานะการเงิน (Balance Sheet)

Balance Sheet's Feature Distributions by Fraud

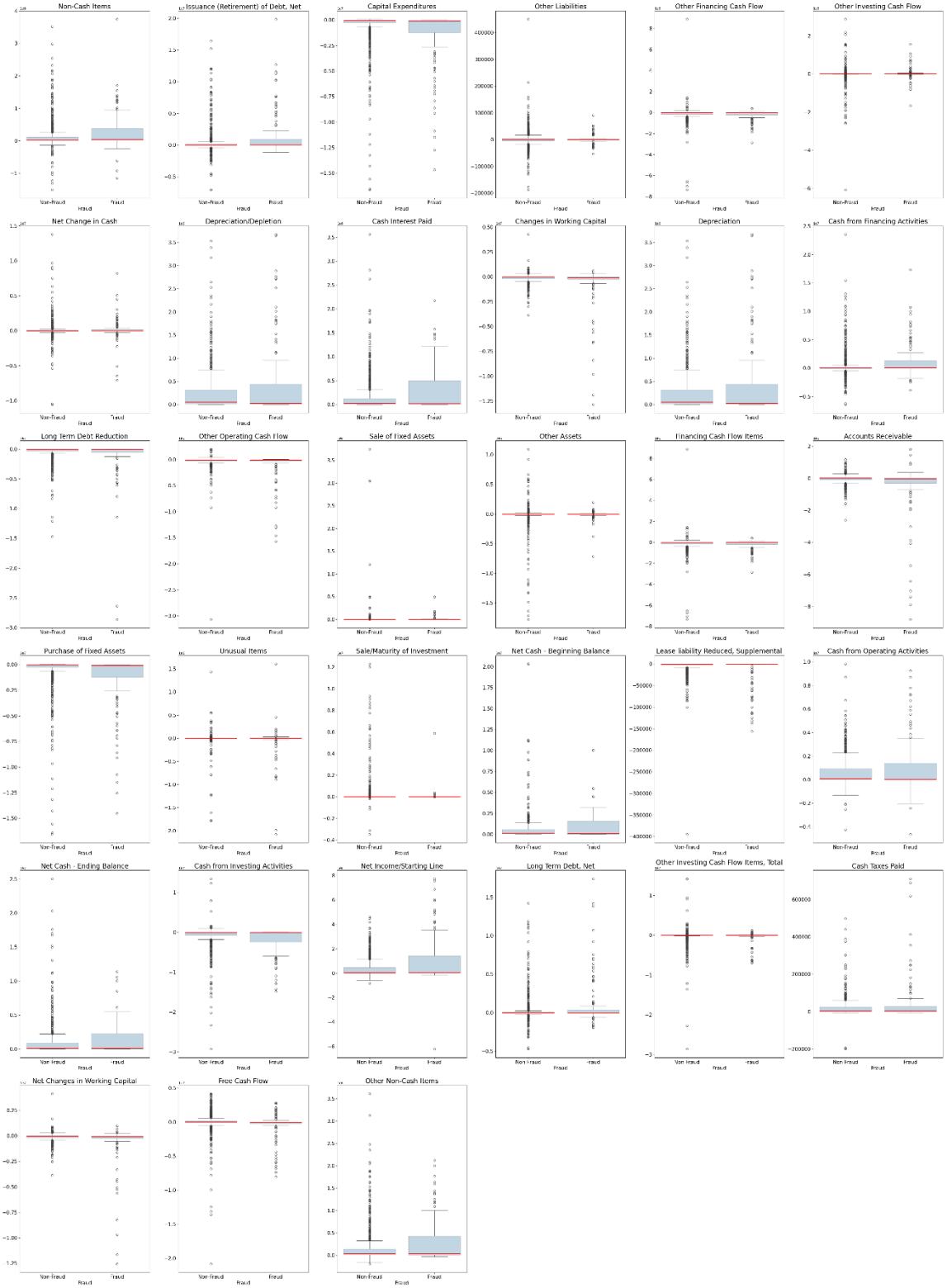


การกระจายตัวของงบกำไรขาดทุน (Income statement)

Income Sheet's Feature Distributions by Fraud

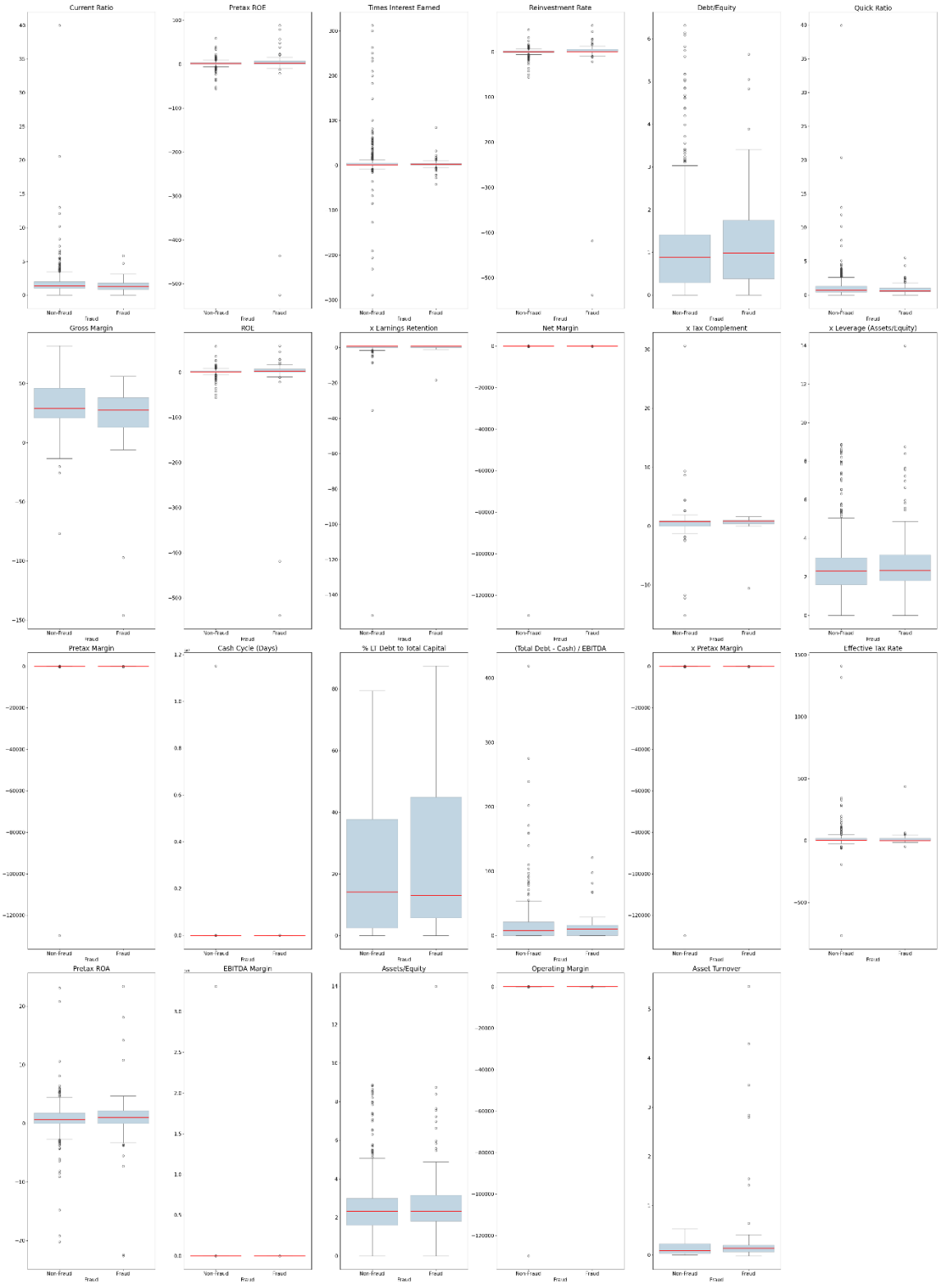


การกระจายตัวของข้อมูลงบกระแสเงินสด (Cash-flow statement)



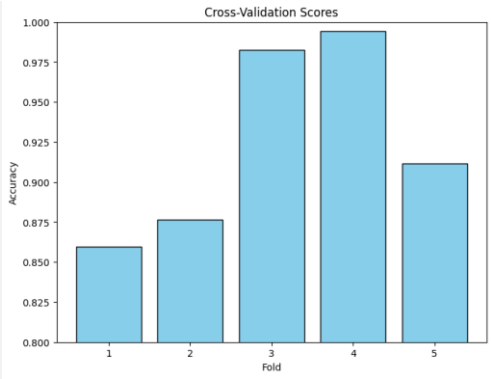
การกระจายตัวของอัตราส่วนทางการเงิน (Ratios-key Metrics)

Ratios-key's Feature Distributions by Fraud

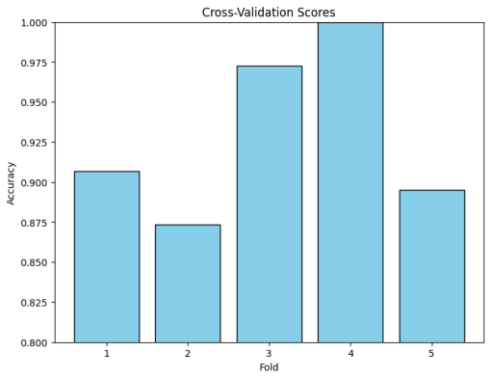


ภาคผนวก จ. กราฟแท่งแสดงผลการประเมิน Cross-validation แต่ละ fold ของแต่ละชุดข้อมูลดังนี้

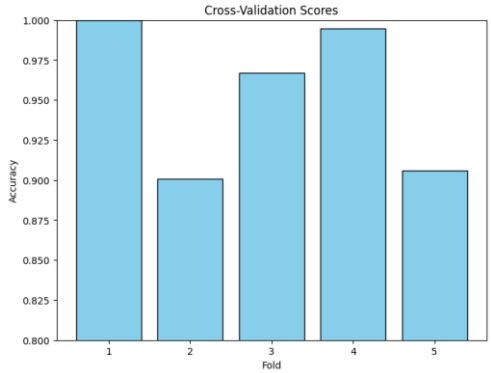
งบแสดงสถานะการเงิน (Balance Sheet)



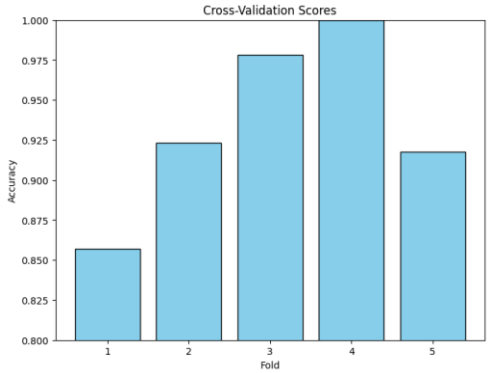
งบกำไรขาดทุน (Income statement)



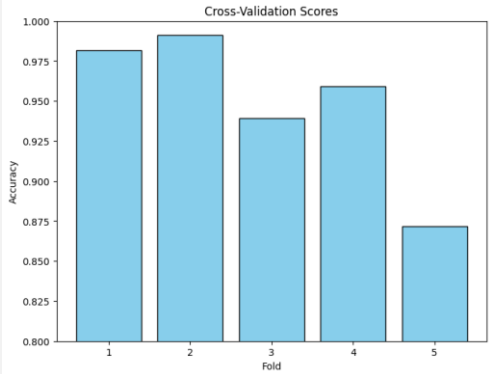
งบกระแสเงินสด (Cash-flow statement)



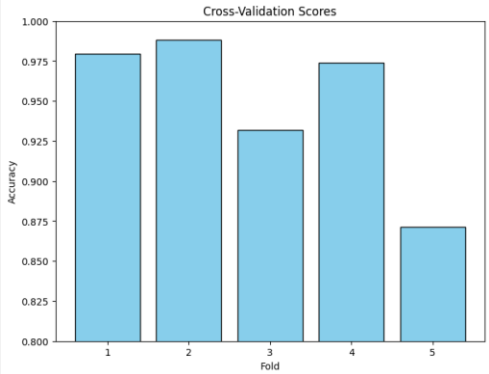
อัตราส่วนทางการเงิน (Ratios-key Metrics)



ตัวแปร 3 อันดับแรก

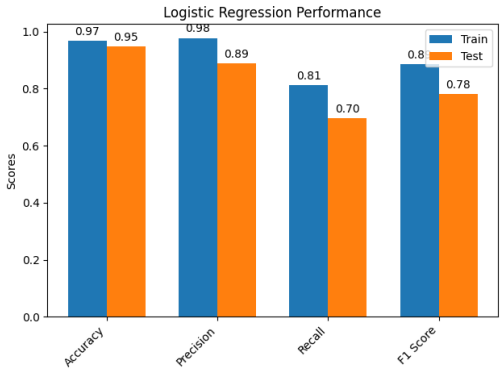


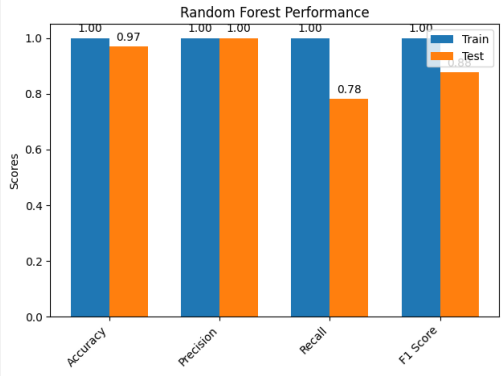
ตัวแปร 3 อันดับแรกกับราคาหุ้น



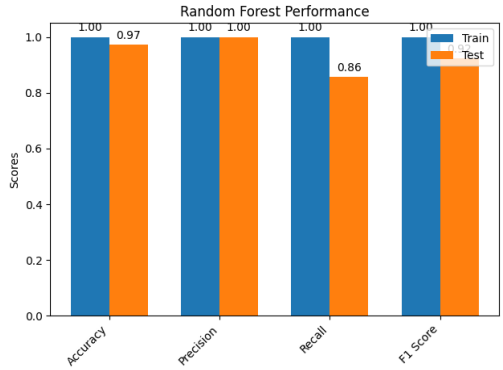
ภาคผนวก ฉ. กราฟแท่งแสดงผลการประเมินโมเดลในแง่ของค่า Accuracy, Precision, Recall และ F-1 Score สำหรับชุดข้อมูล Train และ Test

งบแสดงสถานะการเงิน (Balance Sheet)

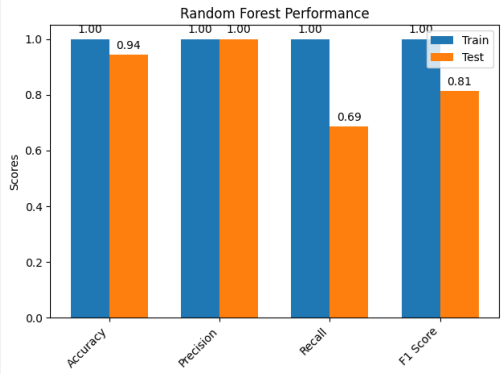




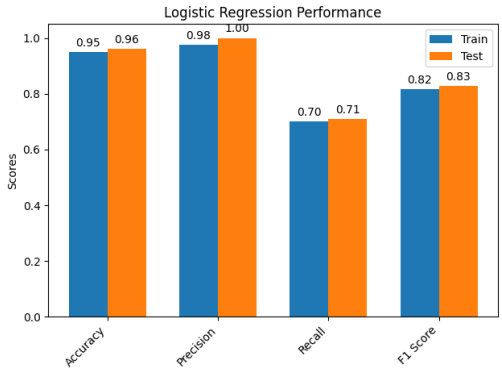
งบกำไรขาดทุน (Income statement)

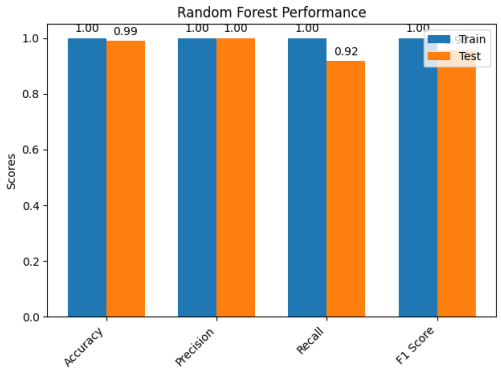


งบกระแสเงินสด (Cash-flow statement)

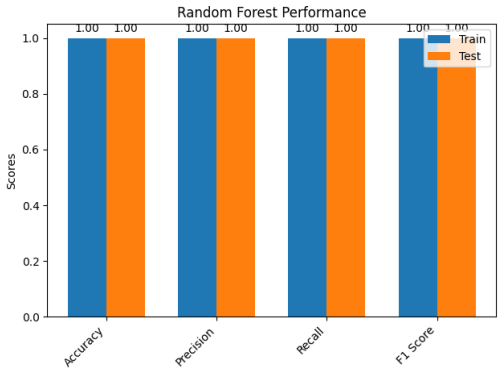


อัตราส่วนทางการเงิน (Ratios-key Metrics)





ตัวแปร 3 อันดับแรก



ตัวแปร 3 อันดับแรกกับราคาหุ้น