วารสารบัณฑิตศึกษา มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีที่ 12 ฉบับที่ 3 (เมษายน 2567-กรกฎาคม 2567)

การพัฒนาระบบตรวจจับความผิดปกติของข้อมูลบัญชีการเงิน ในรูปแบบอนุกรมเวลาด้วยเทคนิคการเรียนรู้ของเครื่อง

AN ANOMALY DETECTION SYSTEM DEVELOPMENT FOR FINANCIAL TIME-SERIES DATA USING MACHINE LEARNING TECHNIQUES

อลงกรณ์ ติดทะ[1]
เอกสิทธิ์ พัชรวงศ์ศักดา[2]

บทคัดย่อ

สารนิพนธ์ฉบับนี้ศึกษาวิธีการตรวจจับความผิดปกติในข้อมูลอนุกรมเวลาทางการเงิน โดยมุ่งเน้นไปที่การตรวจจับธุรกรรมที่น่าสงสัยในข้อมูลธุรกรรมรายวัน โดยมีเป้าหมายคือการหาวิธีที่เหมาะสมที่สุดในการตรวจจับความผิดปกติเพื่อใช้ในการระบุธุรกรรมที่ผิดปกติซึ่งอาจนำไปสู่การทุจริตทางการเงิน โดยวิธีการที่ใช้ในการศึกษาครั้งนี้ใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (unsupervised learning) ด้วยโครงข่าย Autoencoder สำหรับการตรวจจับความผิดปกติ โดยประสิทธิภาพของแบบจำลองในงานวิจัยนี้ ประเมินด้วยใช้ตัวชี้วัดต่าง ๆ เช่น ความแม่นยำ, ความครอบคลุม (recall), พื้นที่ใต้กราฟ AUC-PRC, ROC, F1-Score, และความถูกต้อง ผลการศึกษาพบว่าวิธีการแต่ละวิธีมีข้อดีและข้อด้อยของแต่ละแบบจำลอง แต่แบบจำลองของ Bi-LSTM Autoencoder นั้นมีประสิทธิภาพดีที่สุดในการระบุความผิดปกติ

นอกจากนี้ยังได้มีการศึกษาการประเมินค่าเกณฑ์ (threshold) ที่เหมาะสม เพื่อปรับปรุงอัตราการตรวจจับของวิธีการเรียนรู้ของเครื่องเพิ่มเติม โดยแบบจำลองที่ได้จะถูกนำไปใช้สำหรับการตรวจสอบธุรกรรมรายวันโดยอัตโนมัติเพื่อการตรวจจับความผิดปกติ และแสดงผลลัพธ์ในรูปแบบแดชบอร์ด

คำสำคัญ : การเรียนรู้ของเครื่อง, การตรวจจับความผิดปกติ, อนุกรมเวลา

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่
[2] ที่ปรึกษาสารนิพนธ์หลัก

ABSTRACT

This independent study studies methods for detecting anomalies in financial time-series data, focusing on identifying suspicious transactions in daily transaction data. The goal is to find the most appropriate method for anomaly detection to identify abnormal transactions that may lead to financial fraud. The study employs unsupervised machine learning techniques, specifically Autoencoder networks, for anomaly detection. The model's performance in this research is evaluated using various metrics such as accuracy, recall, AUC-PRC, ROC, F1-Score, and precision. The study found that each method has its advantages and disadvantages, but the Bi-LSTM Autoencoder model performed best in identifying anomalies.

Additionally, the study explores the evaluation of appropriate thresholds to improve the detection rate of machine learning methods. The resulting model will be used for automated daily transaction monitoring to detect anomalies and display the results in a dashboard format.

Keywords: Machine Learning, Anomaly Detection, Time-Series Data

บทนำ

อุตสาหกรรมการเงินเป็นหนึ่งในเป้าหมายของการทำทุจริตอย่างต่อเนื่อง การกระทำที่ผิดกฎหมายเหล่านี้ไม่เพียงแต่นำไปสู่ความเสียหายทางการเงิน แต่ยังบั่นทอนความมั่นใจของผู้บริโภคและความมั่นคงของระบบการเงิน ด้วยการเพิ่มขึ้นของขนาดและความซับซ้อนของธุรกรรมทางการเงินเนื่องจากการเปลี่ยนแปลงสู่สังคมดิจิทัล วิธีการดั้งเดิมที่พึ่งพาการจัดการทางบัญชีด้วยตัวเองกลายเป็นวิธีที่ไม่มีประสิทธิภาพและไม่ทันกาลในการระบุธุรกรรมทุจริตในยุคปัจจุบัน

การนำระบบอัตโนมัติมาใช้วิเคราะห์ข้อมูลทางการเงินได้อย่างมีประสิทธิภาพ สามารถทำได้โดยการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งพัฒนาแบบจำลองด้วยข้อมูลที่มี และปรับปรุงระบบตามการเปลี่ยนแปลงของพฤติกรรมผู้บริโภคและการทุจริต และระบุรูปแบบและความผิดปกติที่ซับซ้อนที่ระบบพึ่งพากฎ (rule-based system) ทำได้ยาก

เนื่องจากข้อมูลทางการเงินมีลักษณะหลายมิติ และเป็นข้อมูลอนุกรมเวลา มีความสัมพันธ์กันซับซ้อน แบบจำลองการเรียนรู้ของเครื่องแบบทั่วไปอาจไม่สามารถประมวลผลข้อมูลลักษณะเหล่านี้ได้อย่างมีประสิทธิภาพ ดังนั้น การใช้แบบจำลองที่ซับซ้อนขึ้นเช่น LSTM (Long Short-Term Memory) ที่ออกแบบมาเพื่อจัดการกับข้อมูลอนุกรมเวลา จึงถูกนำมาใช้พัฒนาแบบจำลองข้อมูลธุรกรรมทางการเงิน และใช้วิธีการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (unsupervised learning) อย่าง Autoencoder เพื่อแยกแยะความผิดปกติระหว่างข้อมูล

วัตถุประสงค์ของงานวิจัย

1. เพื่อพัฒนาระบบตรวจจับข้อมูลการเงินในรูปแบบอนุกรมเวลาหลายตัวแปร สามารถค้นหาความผิดปกติในธุรกรรมที่เคยมีรูปแบบระบุมาแล้ว หรือที่ยังไม่เคยเกิดขึ้นมาก่อน และแสดงผลการตรวจจับในรูปแบบแดชบอร์ดที่ง่ายต่อการสำรวจข้อมูลเชิงลึก

2. เพื่อทดสอบประสิทธิภาพของระบบที่พัฒนาขึ้น โดยเปรียบเทียบกับเทคนิคการตรวจจับความผิดปกติแต่ละแบบ

3. เพื่อศึกษาเกี่ยวกับปัญหาและอุปสรรคในการพัฒนาระบบตรวจจับการทุจริตในข้อมูลการเงินแบบอนุกรมเวลาและการแก้ไข

ขอบเขตงานวิจัย

1. ข้อมูล: งานวิจัยนี้จะใช้ข้อมูลธุรกรรมทางการเงินที่เผยแพร่สาธารณะที่สามารถนำมาแปลงเป็นข้อมูลอนุกรมเวลาหลายตัวแปรได้

2. เทคนิคการเรียนรู้ของเครื่อง: งานวิจัยนี้มุ่งเน้นที่การพัฒนาระบบตรวจจับความผิดปกติด้วยการใช้ unsupervised machine learning ประกอบกับอัลกอริทึมสำหรับอนุกรมเวลา ประกอบด้วย LSTM Autoencoder และ Bi-LSTM Autoencoder

3. การประยุกต์ใช้: งานวิจัยนี้มุ่งเน้นที่การพัฒนาระบบตรวจสอบต่อเนื่องรูปแบบแดชบอร์ด สำหรับตรวจจับรายการผิดปกติจากการวิเคราะห์ข้อมูลธุรกรรมทางการเงิน

4. การประเมินผล: งานวิจัยนี้ใช้ตัวชี้วัดการประเมินผลที่เกี่ยวข้องกับการพัฒนาแบบจำลอง เช่น ความแม่นยำ (precision), ความครอบคลุม (recall), คะแนน F1 เป็นต้น ประกอบกับตัวชี้วัดอื่นๆ

ทฤษฎี และผลงานที่เกี่ยวข้อง

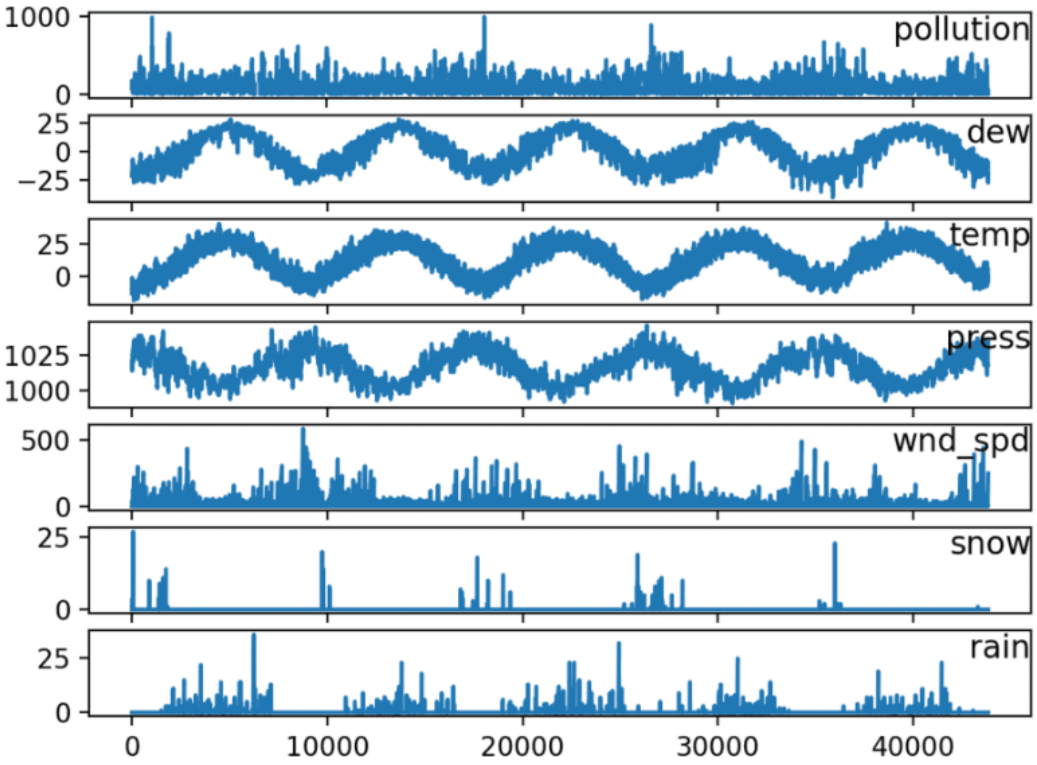
1. การตรวจจับความผิดปกติในข้อมูลทางการเงิน (Anomaly detection in financial data)

ในช่วงไม่กี่ปีที่ผ่านมา พบงานวิจัยจำนวนมากที่ตีพิมพ์เกี่ยวกับการตรวจจับความผิดปกติในข้อมูลทางการเงินโดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยมุ่งเน้นไปที่การตรวจสอบเพื่อประกันความมั่นใจและการปฏิบัติตามข้อกำหนด ซึ่งรวมถึงการตรวจสอบบัญชีและการบัญชีนิติวิทยา เมื่อเปรียบเทียบกับการตรวจจับความผิดปกติที่ใช้กฎเกณฑ์ (rule-based anomaly detection) ซึ่งจำกัดความสามารถอยู่ที่ตรรกะที่ผู้ใช้กำหนดขึ้นเองจากประสบการณ์ วิธีการนี้จึงมีประโยชน์จากการวิเคราะห์ข้อมูลและการตรวจจับรูปแบบที่

ซ่อนอยู่ โดยอิงจากข้อมูลที่มีการติดป้ายกำกับหรือความเบี่ยงเบนจากรูปแบบความปกติทั่วทั้งชุดข้อมูล ซึ่งเหมาะสำหรับการปรับเปลี่ยนให้เข้ากับข้อมูลรูปแบบใหม่ที่อาจเปลี่ยนแปลงอย่างรวดเร็ว

2. ข้อมูลอนุกรมเวลาแบบหลายตัวแปร (Multivariate Time-Series Data)

ข้อมูลอนุกรมเวลาแบบหลายตัวแปรเป็นข้อมูลที่ประกอบด้วยค่าตัวแปรหลายตัวที่ถูกบันทึกไว้ตามลำดับเวลา ดังนั้น ตัวแปรหนึ่งตัวอาจเปลี่ยนแปลงค่าตามเวลา มีลำดับข้อมูลสัมพันธ์กับเส้นเวลา เช่น ข้อมูลอุณหภูมิ, ความชื้น, และความดันอากาศที่ถูกบันทึกไว้ทุก ๆ ชั่วโมง ณ สถานีอุตุนิยมวิทยา ข้อมูลธุรกรรมทางการเงินที่ธนาคารเก็บบันทึกในแต่ละรายการ เป็นต้น

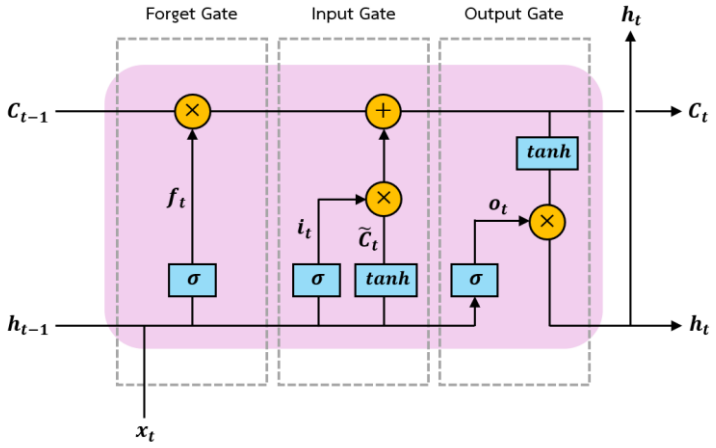


ภาพที่ 1 ข้อมูลอนุกรมเวลาแบบหลายตัวแปร

3. LSTM (Long Short-Term Memory)

เป็นโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network หรือ RNN) รูปแบบหนึ่ง ที่ถูกออกแบบมาเพื่อแก้ปัญหาการเสื่อมค่าของเกรเดียนต์ (vanishing gradient problem) ที่เกิดขึ้นใน RNN แบบดั้งเดิม ซึ่งการเรียนรู้และจดจำความสัมพันธ์ระยะยาวของข้อมูลขาดประสิทธิภาพ

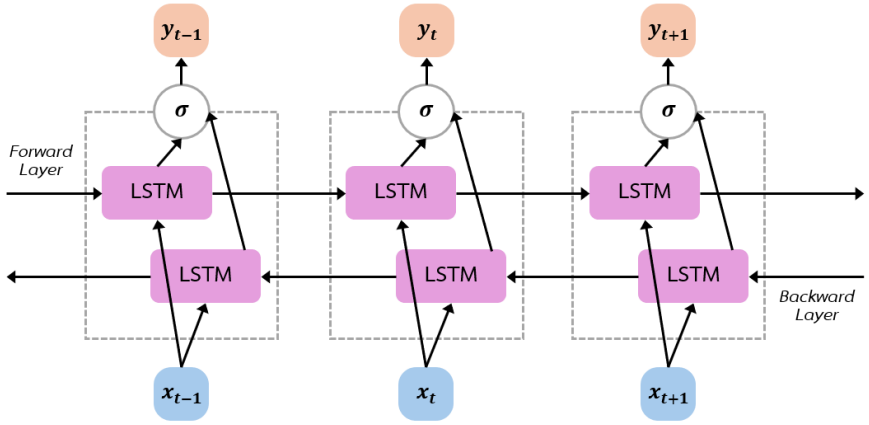
ลักษณะเด่นของ LSTM คือมีการเพิ่มเซลล์หน่วยความจำ (memory cell) และประตู (gate) เข้าไปในโครงสร้าง เพื่อช่วยในการควบคุมการไหลของข้อมูลในแต่ละเซลล์



ภาพที่ 2 โครงสร้างของ LSTM

4. Bi-LSTM

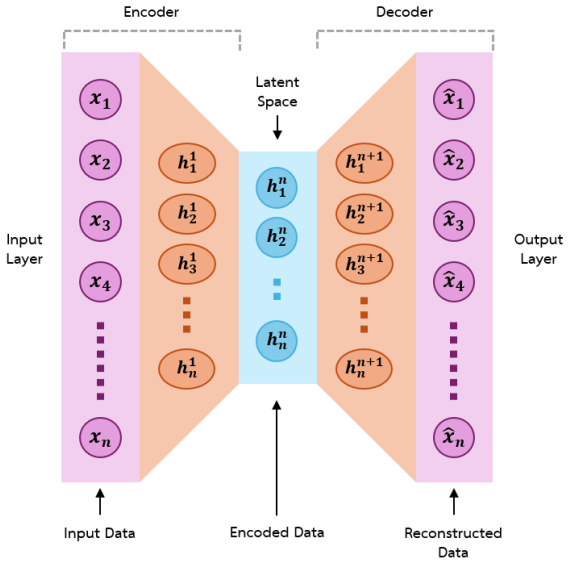
โครงข่าย Bi-LSTM ได้เพิ่มชั้นของ LSTM เพื่อประมวลผลข้อมูลในทิศทางย้อนกลับ การฝึกแบบจำลองโครงข่าย Bi-LSTM เทียบเท่าได้กับการฝึกโครงข่าย LSTM แบบทิศทางเดียวสองโครงข่ายแยกกัน โครงข่ายแรกถูกฝึกด้วยลำดับข้อมูลต้นฉบับ ในขณะที่อีกโครงข่ายถูกฝึกด้วยลำดับข้อมูลสำเนากลับด้าน (reversed copy) วิธีฝึกฝนโครงข่ายแบบนี้จะช่วยให้การเรียนรู้ของเครื่องจดจำข้อมูลสำคัญได้มากขึ้นและเร็วขึ้น



ภาพที่ 3 โครงสร้างของ Bi-LSTM

5. Autoencoder

Autoencoder เป็นโครงข่ายประสาทเทียมชนิดหนึ่งที่ใช้นพัฒนาการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (unsupervised learning) โดยมีวัตถุประสงค์หลักเพื่อบีบอัดข้อมูลนำเข้าให้อยู่ในรูปแบบที่มีขนาดเล็กกว่าเดิม (encoding) และจากนั้นทำการขยายข้อมูลที่บีบอัดนั้นกลับคืนมาเป็นรูปแบบเดิม (decoding) ที่ใกล้เคียงกับต้นฉบับ [13] โดยการฝึกฝน Autoencoder จะทำการลดความผิดพลาด (error) ระหว่างข้อมูลนำเข้าที่แท้จริงและข้อมูลที่ถูกสร้างใหม่ (reconstructed output) ให้มีค่าน้อยที่สุด



ภาพที่ 4 โครงสร้างของ Autoencoder อย่างง่าย

6. งานวิจัยที่เกี่ยวข้อง

Detecting anomalies in financial data using Machine Learning, Alexander Bakumenko. (2022). ในงานวิจัยนี้ ผู้วิจัยได้แสดงวิธีการใช้การเรียนรู้ของเครื่องแบบมีผู้สอนและไม่มีผู้สอนเพื่อตรวจจับความผิดปกติในข้อมูลบัญชีแยกประเภททั่วไป (General Ledger) และแสดงวิธีการปรับปรุงแบบจำลองเพื่อเพิ่มประสิทธิภาพในการตรวจจับ โดยเลือกใช้การเรียนรู้ของเครื่องแบบมีผู้สอน 7 แบบ และการเรียนรู้ของเครื่องแบบไม่มีผู้สอน 2 แบบ แบบจำลองที่เลือกวิจัยได้แก่ Random Forest และ Isolation Forest ซึ่งแสดงประสิทธิภาพในการตรวจจับสูง และผู้วิจัยได้จำกัดคุณลักษณะที่เลือกใช้ให้อยู่เฉพาะกับคุณลักษณะเชิงหมวดหมู่ (categorical features) เท่านั้น โดยไม่ได้ทดสอบคุณลักษณะอื่น ๆ หรือการสร้างคุณลักษณะเพิ่มเติมที่รวมถึงคุณลักษณะด้านจำนวนเงินและคุณลักษณะด้านเวลาตามประเภทของความผิดปกติที่ต้องการศึกษา

A Financial Fraud Detection Model Based on LSTM Deep Learning Technique, Yara Alghofaili , Albatul Albattah & Murad A. Rassam (2022) งานวิจัยนี้ได้นำเสนอการใช้ LSTM ในการตรวจจับการทุจริตทางการเงิน โดยมีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพของเทคนิคการตรวจจับที่มีอยู่ และเพิ่มความแม่นยำในการตรวจจับการทุจริตในบริบทของข้อมูลขนาดใหญ่ ผลการทดลองใช้ชุดข้อมูลจริงเกี่ยวกับการทุจริตบัตรเครดิต และนำผลลัพธ์ไปเปรียบเทียบกับโมเดลการเรียนรู้เชิงลึกที่มีอยู่ เช่น Autoencoder และเทคนิคการเรียนรู้ของเครื่องอื่น ๆ ผลการศึกษาพบว่าแบบจำลอง LSTM มีความแม่นยำในการตรวจจับการทุจริตสูงถึง 99.95% ภายในเวลาน้อยกว่าหนึ่งนาที

ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการศึกษาข้อมูลอนุกรมเวลา (Time Series Data) ทางการเงินและปัจจัยของตัวแปรที่มีผลต่อการตรวจจับความผิดปกติ เพื่อสร้างแบบจำลองสำหรับพยากรณ์ความผิดปกติ และนำเสนอผลลัพธ์จากการพยากรณ์ในรูปแบบของ Dashboard

1. การทำความเข้าใจธุรกิจ (Business Understanding)

ปัญหาทางธุรกิจ: การทุจริตทางการเงินเป็นปัญหาที่มีผลกระทบอย่างมากต่อองค์กรการเงิน ไม่ว่าจะเป็นธนาคารหรือบริษัทบัตรเครดิต ความสูญเสียจากการทุจริตสามารถก่อให้เกิดผลกระทบต่อกำไรและชื่อเสียงขององค์กร การตรวจจับการทุจริตอย่างมีประสิทธิภาพจึงเป็นสิ่งสำคัญ

วัตถุประสงค์ทางธุรกิจ: พัฒนาระบบที่สามารถตรวจจับการทุจริตทางการเงินได้อย่างแม่นยำและรวดเร็ว ลดจำนวนธุรกรรมที่ทุจริตและความสูญเสียที่เกิดจากการทุจริตเพิ่มความปลอดภัยและความน่าเชื่อถือให้กับระบบการเงินขององค์กร

2. การทำความเข้าใจข้อมูล (Data Understanding)

ชุดข้อมูล PaySim เป็นชุดข้อมูลที่ได้จากโปรแกรมจำลองทางการเงินที่จำลองธุรกรรมการเงินบนมือถือโดยอิงจากชุดข้อมูลต้นฉบับ ชุดข้อมูลนี้จำลองธุรกรรมที่เกิดขึ้นใน 1 เดือน (31 วัน) จากการใช้บริการทางการเงินผ่านมือถือในประเทศในทวีปแอฟริกาประเทศหนึ่ง ชุดข้อมูลที่เลือกใช้นี้ลดขนาดลงเหลือ 1 ใน 4 ส่วนจากชุดข้อมูลต้นฉบับ โดยมีรายการทุจริตจำนวน 8,213 รายการจากรายการธุรกรรมทั้งหมด 6,362,620 รายการ ชุดข้อมูลนี้มีการกระจายตัวไม่สมดุลสูง โดยมีคลาสบวก ซึ่งเป็นการทุจริต (fraud) นับเป็นสัดส่วนเพียง 0.129% ของธุรกรรมทั้งหมด

3. การเตรียมข้อมูล (Data Preparation)

เลือกช่วงข้อมูลที่จะนำมาพัฒนาแบบจำลองการเรียนรู้ของเครื่อง เพื่อลดเวลาในการประมวลผลจากการใช้งานข้อมูลทั้งหมด จากนั้นทำการจัดกลุ่มรายการธุรกรรมแบ่งตามรายบัญชีด้วยคุณลักษณะ nameOrig และกรองข้อมูลเฉพาะบัญชีลูกค้าที่มีการทำธุรกรรมมากกว่าหนึ่งรายการ และมีรายการทุจริตปรากฏในธุรกรรมตั้งแต่หนึ่งครั้งขึ้นไป ทำการสร้างคุณลักษณะเพิ่มเติมคือค่า “วัน” (Day) จากคุณลักษณะ step ด้วยการแบ่งวันเป็น 24 ชั่วโมง วันที่ 1 สร้างจาก step ที่ 1 – 24, วันที่ 2 สร้างจาก step ที่ 25 – 48 ถึงวันที่ 31 สร้างจาก step ที่ 721 – 743 ตามลำดับ

4. การสร้างแบบจำลอง (Modeling)

1. ข้อมูลนำเข้ารูปแบบอนุกรมเวลา: ทำการจัดกลุ่มรายการธุรกรรมแบ่งตามรายบัญชีด้วยคุณลักษณะ nameOrig จากนั้นกรองเฉพาะบัญชีลูกค้าที่มีการทำธุรกรรมมากกว่าหนึ่งรายการ และมีรายการทุจริตปรากฏในธุรกรรมตั้งแต่หนึ่งครั้งขึ้นไป จากนั้นทำการแปลงให้อยู่ในรูปแบบอาร์เรย์ 3 มิติ ที่มีขนาดคือ $samples \times lookback \times features$ โดย samples คือ จำนวนตัวอย่างที่สังเกตุการณ์ ซึ่งในงานวิจัยนี้ เลือกข้อมูลสำหรับฝึกและทดสอบแบบจำลองไว้ 744 ชั่วโมง ซึ่งในการพัฒนาแบบจำลอง LSTM มุ่งหวังไปที่การมองกลับไปที่จุดข้อมูลก่อนหน้า กล่าวคือ ณ จุดเวลา t แบบจำลอง LSTM จะประมวลผลข้อมูลย้อนไปถึงจุดข้อมูลที่ $(t - lookback)$ เพื่อทำการพยากรณ์ คุณลักษณะ (features) คือจำนวนตัวแปรที่ปรากฏในข้อมูลอนุกรมเวลา ซึ่งชุดข้อมูลที่นำมาใช้ ประกอบด้วยตัวแปรเชิงตัวเลข 5 ตัวแปร คือ amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, และ newbalanceDest หลังจากประมวลผลข้อมูลแล้ว จะได้ข้อมูลอนุกรมเวลา $[X_1, X_2, X_3, \ldots, X_n]$ จำนวน $n = 9298$ ตัวอย่าง ในแต่ละตัวอย่าง X_i ประกอบด้วยอาร์เรย์สองมิติขนาด $lookback \times features$ สร้างจากคาบเวลาถอย

หลังไปในช่วงเวลา T และลำดับข้อมูล $[x_1, x_2, x_3, \ldots, x_t]$ โดย $x_t \in \mathbb{R}^m$ คือข้อมูลนำเข้าที่มี m คุณลักษณะ ณ จุดเวลา t

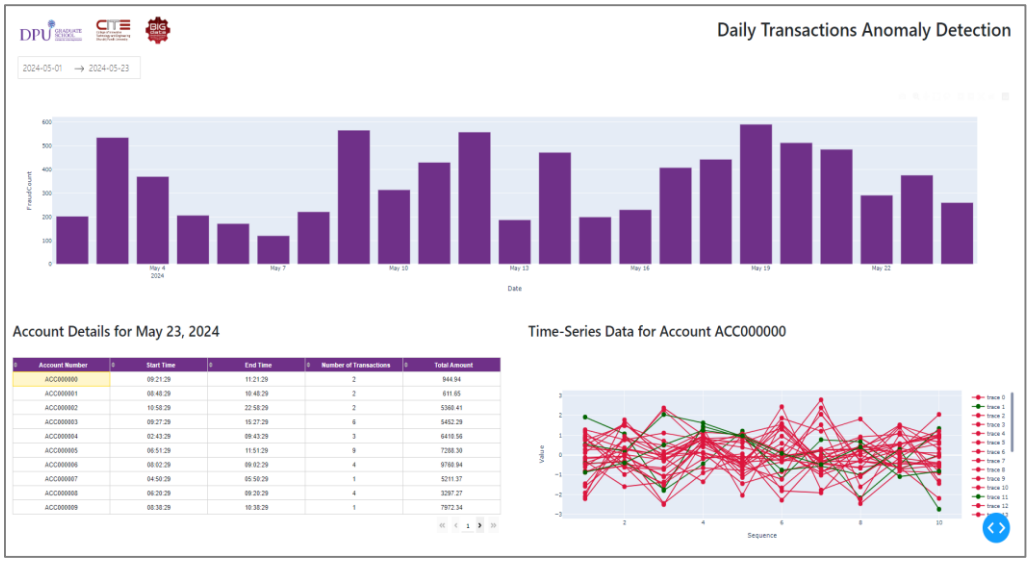
2. Bi-LSTM Encoder: Encoder หรือชุดเข้ารหัสจะรับลำดับข้อมูลนำเข้ามิติสูงและสร้างเวกเตอร์ขนาดเฉพาะขึ้นมา ด้วยการใช้เซลล์หน่วยความจำของ Bi-LSTM ชุดเข้ารหัสจะรักษาความเกี่ยวโยงของจุดข้อมูลหลาย ๆ จุดของชุดลำดับของอนุกรมเวลา พร้อมกับการลดตัวแทนเวกเตอร์นำเข้ามิติสูงให้เป็นตัวแทนมิติต่ำ ลง ในการวิจัยนี้กำหนดชั้นแรกที่ขนาด 32 เซลล์ และชั้นที่สองที่ขนาด 16 เซลล์
3. Bi-LSTM Decoder: Decoder หรือชุดถอดรหัสจะสร้างลำดับข้อมูลขึ้นใหม่จากตัวแทนของข้อมูลนำเข้าใน latent space โดยขณะที่ฟื้นฟูข้อมูลขึ้น แบบจำลองจะพยายามลดความแตกต่างระหว่างข้อมูลต้นฉบับและข้อมูลที่สร้างขึ้นใหม่ให้น้อยที่สุด เรียกว่าค่าความผิดพลาดในการฟื้นฟู (reconstruction error) เมื่อรวบรวมค่าความผิดพลาดในการฟื้นฟูจากการฝึกฝนแบบจำลองนี้แล้ว จะสามารถนำมากำหนดค่าเกณฑ์ (threshold) เพื่อนำไประบุความผิดปกติต่อไป
4. การตรวจจับความผิดปกติ (Anomaly Detection): ชุดสำรวจหนึ่งที่แตกต่างจากข้อมูลกลุ่มมากอาจหมายถึงความผิดปกติหนึ่งได้ การหาว่าข้อมูลที่สนใจแตกต่างมากเพียงใด จำเป็นต้องกำหนดค่าเกณฑ์ขึ้นมาค่าหนึ่ง หากข้อมูลที่สนใจใดมีค่าความแตกต่างเกินกว่าค่าเกณฑ์ที่กำหนดนี้จะถูกระบุว่าเป็นข้อมูลผิดปกติ ในการวิจัยนี้ได้ทำการศึกษาการกระจายตัวของค่าความผิดพลาดในการฟื้นฟูเพื่อคำนวนค่าเกณฑ์ (threshold) η ที่เหมาะสมขึ้นมาค่าหนึ่งเพื่อตรวจจับความผิดปกติ

4. การประเมินผล (Evaluation)

การประเมินผลเป็นขั้นตอนสำคัญในการพิจารณาประสิทธิภาพของแบบจำลองการตรวจจับความผิดปกติที่พัฒนาขึ้น ในการวิจัยนี้ เราใช้ตัวชี้วัดหลายประการในการประเมินผลแบบจำลอง ได้แก่ ความแม่นยำ (Accuracy), ความสามารถในการเรียกคืน (Recall), ความแม่นยำ (Precision), คะแนน F1 (F1 Score), พื้นที่ใต้กราฟ ROC (ROC AUC) และพื้นที่ใต้กราฟ Precision-Recall (AUC-PRC)

6. การนำไปใช้ (Deployment)

การนำแบบจำลองการตรวจจับความผิดปกติไปใช้งานในระบบจริงเป็นขั้นตอนสุดท้ายของงานวิจัยนี้ ซึ่งประกอบด้วยการติดตั้งแบบจำลองในสภาพแวดล้อมการผลิต การตั้งค่าแดชบอร์ดเพื่อแสดงผลลัพธ์ และการตั้งเวลาสำหรับการทำงานอัตโนมัติ



ภาพที่ 5 แดชบอร์ดแสดงการตรวจจับธุรกรรมผิดปกติ

ผลการศึกษา

จากการประเมินประสิทธิภาพของแบบจำลองในการวิจัยนี้ สามารถสรุปผลค่าชี้วัดต่าง ๆ ของแบบจำลองการตรวจจับความผิดปกติ (anomaly detection) ด้วยเทคนิค LSTM Autoencoder และ Bi-LSTM Autoencoder ดังตารางที่ 1

ตารางที่ 1 ผลการประเมินประสิทธิภาพของแบบจำลอง

แบบจำลอง	Precision	Recall	F1-Score	AUC-ROC	AUC-PRC	Accuracy
Bi-LSTM Autoencoder	0.4985	0.7466	0.5978	0.8607	0.6266	0.9611
LSTM Autoencoder	0.4893	0.5924	0.5359	0.7858	0.5474	0.9667

สรุปและอภิปรายผลการวิจัย

จากการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งสองพบว่า แบบจำลอง Bi-LSTM Autoencoder มีความสามารถที่เหนือกว่าในหลายๆ มิติ โดยเฉพาะอย่างยิ่งในด้าน Recall, F1-Score, AUC-ROC และ AUC-

PRC ซึ่งเป็นตัวชี้วัดที่สำคัญในการตรวจจับธุรกรรมทุจริตที่เกิดขึ้นจริง อย่างไรก็ตาม แม้ว่าแบบจำลอง LSTM Autoencoder จะมีค่า Accuracy ที่สูงกว่าเล็กน้อย แต่ในภาพรวมแล้ว Bi-LSTM Autoencoder ยังคงเป็นตัวเลือกที่ดีกว่าสำหรับการนำไปใช้ในงานตรวจจับความผิดปกติในธุรกรรมทางการเงิน เนื่องจากความสามารถในการตรวจจับและแยกแยะธุรกรรมทุจริตได้อย่างมีประสิทธิภาพมากกว่า

ข้อเสนอแนะ

1. เนื่องจากแบบจำลอง Bi-LSTM Autoencoder มีประสิทธิภาพดีกว่าในด้านการตรวจจับธุรกรรมทุจริต จึงแนะนำให้ใช้แบบจำลองนี้ในการตรวจสอบธุรกรรมทางการเงิน
2. แม้ว่า Bi-LSTM Autoencoder จะมีประสิทธิภาพที่ดี แต่ยังมีช่องว่างในการปรับปรุง เช่น การปรับค่า threshold เพื่อเพิ่มความแม่นยำในการตรวจจับ หรือการใช้เทคนิคการเรียนรู้เชิงลึกอื่น ๆ
3. เนื่องจากข้อมูลธุรกรรมทุจริตมีจำนวนที่น้อยกว่าอย่างมากเมื่อเทียบกับธุรกรรมปกติ จึงควรใช้เทคนิคการสุ่มตัวอย่าง (sampling techniques) เพื่อจัดการกับปัญหาความไม่สมดุลของข้อมูล
4. ควรมีการศึกษาวิจัยเพิ่มเติมเพื่อทดสอบแบบจำลองกับข้อมูลที่แตกต่างกันหรือในสภาพแวดล้อมที่มีความซับซ้อนมากขึ้น เพื่อประเมินความสามารถในการปรับตัวของแบบจำลองในสถานการณ์ที่หลากหลาย
5. การนำแบบจำลอง Bi-LSTM Autoencoder ไปใช้ในระบบการตรวจจับการทุจริตในธุรกรรมทางการเงินจริง ควรพิจารณาด้านความปลอดภัย ความเป็นส่วนตัวของข้อมูล และการปรับแต่งแบบจำลองให้เหมาะสมกับสภาพแวดล้อมจริง

บรรณานุกรม

ภาษาไทย

ธนาคารแห่งประเทศไทย. (2564). ธปท. ออกมาตรการจัดการภัยทุจริตทางการเงิน," ธนาคารแห่งประเทศไทย [ออนไลน์]. Retrieved from https://www.bot.or.th/th/news-and-media/news/news-20230309.html.

ภาษาต่างประเทศ

A. Bakumenko and A. Elragal. (2022). Detecting anomalies in financial data using machine learning algorithms. *Systems*.

E. A. Lopez-Rojas, S. Axelsson, and D. Baca. (2017). Applications of the PaySim simulator for fraud detection research in a financial mobile money service. *International Journal of Simulation and Process Modelling, vol. 12, no. 4,* pp. 289-298.

E. A. Lopez-Rojas. (2024). Synthetic Financial Datasets for Fraud Detection," Kaggle [Online]. Retrieved from https://www.kaggle.com/datasets/ealaxi/paysim1.

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science, vol. 313, no. 5786*, pp. 504-507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

J. P. C. Pereira. (2020). Unsupervised anomaly detection in time series data using deep learning.M. A. Rassam, Y. Alghofaili, and A. Albattah. (2020). A financial fraud detection model based on LSTM deep learning technique," Journal of Applied Security Research.

M. Sakurada and T. Yairi. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction," in Proc. of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, p. 4.

M. Schuster and K. K. Paliwal. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, vol. 45, no. 11*, pp. 2673-2681.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*, SPSS.

P. Malhotra, L. Vig, G. Shroff, and P. Agarwal. (2015). Long short-term memory networks for anomaly detection in time series," in Proc. of the Presses Universitaires de Louvain.

P. K. Kamuangu. (2014). A review on financial fraud detection using AI and machine learning," Journal of Economics, Finance and Accounting Studies.

S. Crépey, L. Nicolle, and J. Wakim. (2022). Anomaly detection on financial time series by principal component analysis and neural networks. *HAL Open Science*.

S. Karimi-Bidhendi, F. Munshi and A. Munshi. (2018). Scalable Classification of Univariate and Multivariate Time Series. 2018 IEEE International Conference on Big Data (Big Data), pp. 1598-1605.

T. K. K. Ho and A. Kharbouch. (2015). Graph anomaly detection in time series: A survey.