การพัฒนาการตรวจจับการฉ้อโกงในระบบการเงิน โดยการใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์ข้อมูลเครือข่าย

ENHANCING FRAUD DETECTION IN FINANCIAL SYSTEMS USING MACHINE LEARNING AND NETWORK ANALYSIS TECHNIQUES

สุเชฏฐ์ เหร่บุตร[1]
เอกสิทธิ์ พัชรวงศ์ศักดา[2]

บทคัดย่อ

การเข้าถึงการทำธุรกรรมทางการเงินอิเล็กทรอนิกส์ (Online Banking) ที่ง่ายขึ้น จึงเป็นอีกสาเหตุหนึ่งของมิจฉาชีพที่ใช้ช่องทางนี้ในการฉ้อโกง ซึ่งมาในรูปแบบของการรับจ้างเปิดบัญชี (บัญชีม้า) โดยงานวิจัยนี้จะนำเสนอการตรวจจับการฉ้อโกงในระบบการเงิน โดยการใช้เทคนิคของการเรียนรู้ของเครื่องและการวิเคราะห์ข้อมูลเครือข่าย ซึ่งมีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมที่เกิดการฉ้อโกง โดยประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์เครือข่ายในการพัฒนาโมเดลสำหรับตรวจจับธุรกรรมที่เกิดการฉ้อโกง โดยจะทำการคัดเลือกตัวแปรที่มีความสำคัญและเหมาะสมที่จะนำมาใช้ในการพัฒนาแบบจำลองเพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพ งานวิจัยนี้ได้ทำการสร้างตัวแปรที่เกี่ยวข้องกับเครือข่าย 2 ค่าคือ degree centrality และ closeness centrality จากนั้นคัดเลือกคุณลักษณะของข้อมูล (Feature Selection) ด้วยการใช้โมเดล Extra Trees Classifier หา feature ที่มีความสำคัญ 9 ลำดับแรก นำไปพัฒนาโมเดลซึ่งมีวิธีจัดการข้อมูลที่ไม่สมดุลด้วยการใช้ cost sensitive และนำไปเปรียบเทียบกับโมเดลที่ไม่ได้ใช้ feature ที่เกี่ยวข้องกับการวิเคราะห์เครือข่าย ซึ่งโมเดล Extreme Gradient Boosting Tree (XGBoost) ที่ใช้มีการใช้ feature ที่เกี่ยวข้องกับการวิเคราะห์เครือข่ายและมีการคัดเลือก feature ที่สำคัญมา 9 ตัว ให้ประสิทธิภาพดีที่สุดโดยมีความ Overfitting ต่อตัวข้อมูลน้อยกว่าวิธีที่เหลือ และให้ค่า Precision อยู่ที่ 74.88%, Recall อยู่ที่ 88.47%, F-1 score อยู่ที่ 81.11% และAccuracy อยู่ที่ 99.77%

---

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์หลัก

คำสำคัญ : การตรวจจับการฉ้อโกง, การวิเคราะห์เครือข่าย, การเรียนรู้ของเครื่อง, การจัดการข้อมูลที่ไม่สมดุล

Abstract

The increasing accessibility of electronic financial transactions, or online banking, has become a catalyst for fraudsters who exploit this platform, particularly through the establishment of "mule" accounts. This research aims to address the detection of such fraudulent activities within the financial system, employing machine learning techniques and network data analysis. The objective is to identify significant and suitable variables or factors for detecting fraudulent transactions, by applying machine learning and network analysis to develop a model for this purpose. This involves selecting variables of relevance and appropriateness for developing a predictive model that delivers effective results. This study created two network-related variables: degree centrality and closeness centrality. Following this, feature selection was conducted using an Extra Trees Classifier model to identify the top 9 important features. These were then utilized to develop a model which manages imbalanced data using cost-sensitive approaches. The results were then compared with models that did not use network analysis-related features. The most effective model was the Extreme Gradient Boosting Tree (XGBoost) that incorporated network analysis-related features and selected the top 9 important features. This model demonstrated less overfitting to the data compared to other approaches and achieved a precision score of 74.88%, a recall score of 88.47%, an F-1 score of 81.11%, and an accuracy of 99.77%.

Keywords: Fraud Detection, Network analysis, Machine learning, Imbalance Data

บทนำ

การทำธุรกรรมทางการเงินออนไลน์นับเป็นเรื่องที่เป็นปกติในชีวิตประจำวัน โดยเฉพาะอย่างยิ่งการโอนเงินและการรับเงิน อีกทั้งมาตรการกระตุ้นธุรกรรมทางการเงินอิเล็กทรอนิกส์ (Online Banking) จากทางภาครัฐ เช่น พร้อมเพย์ โครงการคนละครึ่ง เป็นต้น

การเข้าถึงการทำธุรกรรมที่ง่ายขึ้น จึงเป็นอีกสาเหตุหนึ่งของมิจฉาชีพที่ใช้ช่องทางนี้ในการฟอกเงินในรูปของการรับจ้างเปิดบัญชี (บัญชีม้า) คือมิจฉาชีพที่จ้างให้ตนไปเปิดบัญชีธนาคารเป็นชื่อของตน แล้วก็เอา

สมุดคู่ฝากกับบัตรเอทีเอ็มของตนไปเป็นของตัวเอง โดยมีจุดประสงค์เพื่อนำบัญชีไปรับโอนเงินที่ได้มาแบบผิดกฎหมาย เช่น ค้ายา การพนัน (ใครมาชวน "รับจ้างเปิดบัญชี" ห้ามหลงเชื่อเด็ดขาด! ., 2563)

ถ้าหากได้นำเส้นทางการเงินของกลุ่มที่รับจ้างเปิดบัญชี มาวิเคราะห์ จะพบว่าเครือข่ายทางการเงินของบัญชีเหล่านี้ว่ามีการโอนต่อไปให้กลุ่ม/ตัวบุคคล หรือมีการรับเงินค่าจ้างจากกลุ่ม/ตัวบุคคลเหล่านั้น

ทั้งนี้ด้วยข้อจำกัดของข้อมูลจริงที่เกี่ยวกับธุรกรรมการฉ้อโกงนั้นหาได้ยาก งานวิจัยนี้จึงได้ใช้ข้อมูลที่จำลองขึ้นมาบนพื้นฐานข้อมูลจริงของการทำธุรกรรมบนมือถือของผู้ให้บริการในแถบประเทศแอฟริกา (Lopez-Rojas et al., 2016)

ซึ่งงานวิจัยนี้ได้ทำการศึกษาปัจจัยที่เกี่ยวข้องกับการตรวจจับการฉ้อโกงในธุรกรรมทางการเงิน รวมถึงมีการประยุกต์ใช้ตัวแปรที่เกี่ยวก้องกับเครือข่าย เช่น Degree centrality และ closeness centrality จากนั้นนำมาคัดเลือกตัวแปรที่มีความสำคัญ เพื่อนำไปพัฒนาโมเดลตรวจจับการฉ้อโกงในธุรกรรมทางการเงิน ส่วนสุดท้ายสร้างเว็บแอพพลิเคชั่นในการรับค่า input พร้อมทั้งแสดงผลการทำนาย

วัตถุประสงค์ของงานวิจัย

1. เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับพยากรณ์ธุรกรรมที่เกิดการฉ้อโกง
2. เพื่อประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์เครือข่ายในการพัฒนาโมเดลสำหรับตรวจจับธุรกรรมที่เกิดการฉ้อโกง
3. เพื่อเปรียบเทียบประสิทธิภาพของโมเดล และนำเสนอโมเดลที่มีประสิทธิภาพดีที่สุดและเหมาะสมกับชุดข้อมูล
4. เพื่อพัฒนา Web Application ด้วยโมเดลการตรวจจับธุรกรรมที่เกิดการฉ้อโกง

ขอบเขตงานวิจัย

1. เป็นข้อมูลการจำลอง (Pay-Sim) จาก Kaggle ที่นำมาใช้งาน
2. เป็นข้อมูลรูปแบบ structured
3. ข้อมูลรูปแบบของ transactions มีทั้งสิ้น 6,362,620 แถว

ทฤษฎี และผลงานที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

1. ธุรกรรมทางการเงินที่เกิดการฉ้อโกง (Fraud Transactions)

ธนาคารแห่งประเทศไทย (2566) ธุรกรรมทางการเงินที่นับว่าเข้าข่ายทุจริต โดยมีการใช้บัญชีเงินฝาก หรือบัญชี e-Money ที่โอนเงินได้ของบุคคลอื่นมาเป็นช่องทางในการรับเงินและถ่ายโอนเงินที่ได้มาจากการกระทำความผิด หรือ บัญชีม้า) ซึ่งธุรกรรมนั้นมีลักษณมีการโอนเงินไปยังบัญชีอื่นที่ต้องสงสัย และ บัญชีที่มีการโอนเงินเข้ามาจำนวนมากโดยโอนเข้ามาจากหลายบัญชีแต่โอนเงินออกไปยังบัญชีปลายทางเพียงไม่กี่บัญชี เป็นต้น

2. Centrality

งานวิจัยนี้ได้ใช้ค่า Degree Centrality และ Closeness Centrality เพื่อมาช่วยคำนวณหาความเป็นศูนย์กลางของบัญชี เพื่อใช้ระบุบัญชีของผู้ทำการฉ้อโกง เนื่องจากบัญชีม้ามีลักษณะคือ บัญชีที่เจ้าของบัญชีตัวจริงไม่ได้เปิดเพื่อใช้เอง แต่ขายบัญชีให้คนร้าย ยอมให้คนร้ายเอาไปใช้หรือถูกคนร้ายขโมยข้อมูล หรือถูกสวมตัวตนมาเปิดบัญชี ใช้เป็นช่องทางในการรับ / โอนเงินที่ได้มาจากการกระทำความผิดเพื่อปกปิดไม่ให้มีหลักฐานหรือถูกเชื่อมโยงมาถึงตัวได้ (บัญชีม้า ,ม.ป.ป.)

3. Decision Tree

งานวิจัยนี้ได้เลือกใช้ต้นไม้ตัดสินใจ (Decision Tree) เนื่องมาจากสามารถตีความได้ง่าย โดย Decision Tree มีกระบวนการทำงานในการแบ่งแยกข้อมูล โดยให้ตัวแปรที่สามารถจำแนกข้อมูลได้มากที่สุดเป็น node แรก (root node) และชั้นความลึกถัดไปจะเป็นตัวแปรที่แบ่งแยกข้อมูลได้รองลงมา ทำเช่นนี้จนกระทั่งครบความลึกที่ได้ระบุไว้หรือจนไม่สามารถแบ่งแยกข้อมูลออกมาได้ (Geron, 2019)

4. Random Forest

งานวิจัยนี้ได้เลือกใช้ป่าสุ่ม (Random Forest) เนื่องมาจากมีการใช้ Decision tree หลายต้น(Ensemble of Decision Trees) มาร่วมกันตัดสินเพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพเพิ่มขึ้น โดยในการ trained โมเดลนั้นจะใช้วิธี bagging method (random sampling with replacement) ซึ่งจะนำมาสร้างแบบจำลองต้นไม้โดยแต่ละต้นจะมีชุดข้อมูลสำหรับ trained ไม่ซ้ำกัน (subsets of the training set) โดยแบบจำลองจะมีการทำนายผลออกมาซึ่งจะนำผลการทำนายที่ได้มาโหวตหาผลการทำนายที่ได้รับการโหวตมากที่สุด (Geron, 2019)

5. Extreme Gradient Boosting Tree (XGBoost)

งานวิจัยนี้ได้เลือกใช้ XGBoost เนื่องมาจากอัลกอริทึมนี้ถูกพัฒนามาจาก Gradient Boosting ซึ่งออกแบบให้มีประสิทธิภาพสูง , ยืดหยุ่น และสามารถนำไปใช้ได้กับระบบต่างๆ (XGBoost: A Scalable Tree Boosting System, n.d.) โดยการทำงานของมันจะใช้เทคนิคการนำการเรียนรู้ต้นไม้ตัดสินใจจำนวนหลายๆ โมเดลมาทำนายต่อกัน ซึ่งแต่ละต้นไม้ตัดสินใจจะได้เรียนรู้จากค่าความผิดพลาดของต้นไม้ตัดสินใจก่อนหน้า โดยนำค่าความผิดพลาดนั้นมาปรับปรุงในการสร้างโมเดลถัดไป (Geron, 2019)

6. การคัดเลือกคุณลักษณะของข้อมูล

ก่อนจะนำข้อมูลไปสร้างโมเดลจะต้องมีขั้นตอนการคัดเลือกคุณลักษณะหรือตัวแปรที่มีความสำคัญและเหมาะสมเพื่อให้ได้โมเดลที่มีประสิทธิภาพและยังช่วยให้การทำงานไวขึ้น

การคัดเลือกคุณลักษณะของข้อมูล โดยใช้ Extra trees Classifier เป็นแนวทางในการใช้ โมเดลในการหา feature ที่สำคัญ ซึ่งมีวิธีการคล้ายกับ random forest แต่ต้นไม้ในโมเดล จะมีความสัมพันธ์กันน้อยกับต้นไม้ต้นอื่นในโมเดล จึงทำให้ผลลัพธ์ที่ได้นั้น overfit กับตัว data น้อยลง (ML | Extra Tree Classifier for Feature Selection, 2023)

7. การจัดการกับข้อมูลที่ไม่สมดุล

ปัญหา imbalanced classification เกิดขึ้นเมื่อจำนวนข้อมูลในแต่ละ Class คำตอบมีจำนวนข้อมูลแตกต่างกันมาก การมีข้อมูลที่ไม่สมดุลนี้ส่งผลกระทบต่อทำงานของโมเดล ทั้งนี้ในการจัดการปัญหา imbalanced data หลักๆ จะมีสองแนวทางคือ Data Sampling และ Class weighting

การให้น้ำหนักในคลาสเป็นวิธีที่นิยมในการจัดการกับปัญหาความไม่สมดุลของคลาสในโมเดลการเรียนรู้ของเครื่อง โดยการกำหนดน้ำหนักที่แตกต่างกันให้กับแต่ละคลาสในระหว่างการฝึกฝน เพื่อให้แน่ใจว่าทุกคลาสจะได้รับความสำคัญที่เท่ากัน (Kumar, 2023)

8. ตัววัดประสิทธิภาพของโมเดล

ในงานวิจัยนี้ได้ใช้ Confusion Matrix ในการเปรียบเทียบประสิทธิภาพของโมเดลภายใต้เงื่อนไขต่างๆ โดยทั่วไปจะใช้เปรียบเทียบค่าที่เกิดจากการทำนาย และค่าที่เกิดขึ้นจริง (Confusion Matrix, n.d.)

9. Web Application

Web Application คือแอปที่ถูกเขียนขึ้นมาให้สามารถเปิดใช้ในเว็บเบราว์เซอร์ได้โดยตรงทำให้โดยรวมแล้วกินทรัพยากรค่อนข้างน้อย สามารถเปิดใช้งานได้รวดเร็ว (Web application คืออะไร? ต่างจากเว็บไซต์ทั่วไปอย่างไร?, ม.ป.ป.)

10. งานวิจัยที่เกี่ยวข้อง

Petr Hajek., Mohammad Zoynul Abedin., Uthayasankar Sivarajah (2022) ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนระบบจำลองการชำระเงินผ่านมือถือ โดยได้มีการเทียบประสิทธิภาพของโมเดล ทั้งแบบใช้วิธี Supervised Learning Method และ Outlier Detection Methods ในงานวิจัยนี้พบว่า semi-supervised ensemble model integrating multiple unsupervised outlier detection และใช้ algorithms XGBoost classifier (XGBOD) ให้ผลลัพธ์ที่ดีที่สุด ให้ค่า accuracy เท่ากับ 0.9994 , F1 เท่ากับ 0.8737 , Precision เท่ากับ 0.9942 และ Recall เท่ากับ 0.7793 ขณะที่โมเดลที่ดีที่สุดในแง่ของ cost saving of fraud detection system (ค่า recall มากที่สุด) ได้แก่ combining random under-sampling and XGBoost (RUS+XGBoost) ให้ค่า accuracy เท่ากับ 0.9963 , F1 เท่ากับ 0.2812 , Precision เท่ากับ 0.1637 และ Recall เท่ากับ 0.9976

Jianping Li., Shigang Wen., Mingxi Liu ., Xiaoqian Zhu (2022) ในงานวิจัยนี้ได้นำเสนอวิธีการตรวจจับการฉ้อโกงในภาคการเงินโดยอาศัย kg (knowledge graph)ในการศึกษาความสัมพันธ์ระหว่างผู้จัดการกับสถาบันการเงินที่เกี่ยวข้อง โดยได้มีการเทียบประสิทธิภาพของโมเดลทั้งหมด 4 โมเดล และโมเดลที่มีการใช้ kg อีก 4โมเดล โดยได้ผลการทดลองว่าโมเดลที่มีประสิทธิภาพสูงสุดได้แก่ SVM ที่มีการใช้ centrality measure ให้ค่า average Accuracy เท่ากับ 0.9318 , average F1 เท่ากับ 0.5852 , average Precision เท่ากับ 0.6845 และให้ค่า average Recall เท่ากับ 0.5611

Aanchal Sahu., GM Harshvardhan., Mahendra Kumar Gourisaria (2020) ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนเครดิตการ์ด ในงานวิจัยนี้ได้เปรียบเทียบวิธีในการจัดการปัญหาความไม่สมดุลของข้อมูล 2 วิธี 1. Oversampling minority class และ 2. Cost based โดยการปรับค่าน้ำหนักให้กับ minority class เพื่อให้มีผลกระทบสูงขึ้นต่อการสร้างโมเดล โดยได้ทำการทดลองทั้งสิ้น 5 โมเดล ดังนี้ ANN, LR, SVM, DT, RF ซึ่งโมเดลที่มีประสิทธิภาพสูงสุดในแง่ของค่า F1 ได้แก่โมเดล Random forest ทั้ง 2 วิธี โดยวิธี Oversampling minority class ให้ค่า accuracy เท่ากับ 96.57, F1 เท่ากับ 95.21, precision เท่ากับ 97.59 และ recall เท่ากับ 94.95 และวิธี Cost based ให้ค่า accuracy เท่ากับ 96.48 , F1 เท่ากับ 92.03 , precision เท่ากับ 94.1 และ recall เท่ากับ 86.88

Dejan Varmedja., Mirjana Karanovic., Srdjan Sladojevic., Marko Arsenovic., Andras Anderla (2019) ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนเครดิตการ์ด ซึ่งได้ใช้เทคนิค SMOTE ในการจัดการปัญหาความไม่สมดุลของข้อมูล โดยได้ทำการทดลองทั้งสิ้น 4 โมเดล ดังนี้ Logistic Regression, Random Forest, Naïve Bayes และ Multilayer Perceptron ซึ่งโมเดลที่มีประสิทธิภาพสูงสุดได้แก่ Random forest ให้ค่า accuracy เท่ากับ 99.96 , precision เท่ากับ 96.38 และ recall เท่ากับ 81.63

ระเบียบวิธีวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยมีขั้นตอนการดำเนินการดังต่อไปนี้

1. การเก็บข้อมูล

งานวิจัยนี้ได้ใช้ข้อมูล PaySim จาก Kaggle ซึ่งเป็นจำลองการทำธุรกรรมเงินสดผ่านมือถือในประเทศแอฟริกาใดประเทศหนึ่งโดยอิงจากตัวอย่างการทำธุรกรรมจริงที่สกัดมาจาก financial logs ในระยะเวลา 1 เดือน ทั้งนี้ชุดข้อมูลจำลองนี้ถูกลดขนาดลงเป็น 1/4 จากชุดข้อมูลดั้งเดิมและสร้างขึ้นเพื่อใช้งานใน Kaggle เท่านั้น (Lopez-Rojas, E., 2017)

2. การเตรียมข้อมูล

2.1 ตัวแปรที่ใช้ในการศึกษา

นำข้อมูลตัวแปร (Field) isFlaggedFraud ออกเนื่องจากเป็นการ flagged transaction ที่มียอดมากและเป็นการป้องกันของ Business model จึงไม่นับเป็นปัจจัยในการนำมาศึกษา

2.2 ข้อมูลที่ใช้ในการศึกษา

ข้อมูล transactions ที่เลือกใช้สำหรับการศึกษานี้คือ type ของ transactions ที่เกิด Fraud เท่านั้น โดย type ของ transactions ดังกล่าวคือ TRANSFER และ CASH_OUT

3. การสร้างตัวแปร

3.1 Centrality : ทำการคำนวณ degree centrality , closeness centrality ของทุก Node (nameOrig กับ nameDest)

3.2 Cumulative sum of transactions : ทำการคำนวณ transaction ของ type Transfer และ type Cashout ในลักษณะของ Cumulative sum เพื่อนับจำนวน transactions ที่เกิดขึ้น ของทั้ง nameOrig กับ nameDest

4. การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพ

งานวิจัยนี้ได้ใช้วิธีได้ใช้วิธีการตรวจสอบความถูกต้องโดยแยกตามสัดส่วน (Split validation test) โดยแบ่งข้อมูลเป็นข้อมูลสำหรับเรียนรู้ (Training data) 70% และเป็นข้อมูลสำหรับทดสอบ (Testing data) 30%

5. การคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)

การสร้างโมเดลจะต้องมีการคัดเลือกตัวแปรที่มีความสำคัญ 9 ตัว (เท่ากับจำนวนตัวแปรก่อนมีการสร้าง feature centrality) เพื่อให้โมเดลมีประสิทธิภาพสูง ซึ่งในงานวิจัยนี้เลือกใช้เทคนิค ใช้ model Extratree Classifier หาค่า Feature importance ของชุดข้อมูล เพื่อดูความสัมพันธ์ระหว่างตัวแปรต่างๆ และตัวแปรคำตอบ

(Label) ซึ่งตัวแปรที่มีค่าน้ำหนักมากหมายถึงมีความสำคัญมากกว่าอีกตัวแปรที่มีค่าน้ำหนักน้อยกว่า ทั้งนี้ได้มีการจัดการปัญหาข้อมูลที่ไม่สมดุลโดยมีการปรับค่า weight เพื่อเพิ่มน้ำหนักให้กับ minority class

6.การสร้างโมเดล

หลังจากทำการคัดเลือกคุณลักษณะ (Feature Selection) ที่ดีที่สุด 9 ลำดับแรก เสร็จเรียบร้อยแล้ว ก็นำตัวแปรที่ได้ไปสร้างโมเดล โดยงานวิจัยนี้ได้ทดลองใช้ทั้งหมด 3 algorithm ได้แก่ Decision Tree , Random Forest และ Extreme Gradient Boosting (XGBoost)

7. การประเมินผล

ขั้นตอนนี้เป็นการวัดประสิทธิภาพของโมเดลโดยแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนที่ทำการฝึกฝนโมเดลและส่วนที่สำหรับประเมินโมเดล โดยได้ใช้เกณฑ์ประเมินซึ่งวัดจากค่า F-1 score, Precision, Recall และ Accuracy ทั้งนี้ยังเปรียบเทียบผลโมเดลที่ใช้และไม่ใช้ตัวแปร centrality โดยคำนึงถึง Overfitting ของตัวโมเดล

ผลการศึกษา

งานวิจัยนี้เป็นการศึกษาเพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยมีรายละเอียดของผลการศึกษาดังต่อไปนี้

ตารางที่ 1 ผลของโมเดลที่มีการใช้ feature centrality และมีการทำ feature selection

Model	Precision		Recall		F-1 Score		Accuracy
	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	
Decision Tree	99.99%	8.62%	94.20%	98.32%	97.01%	15.85%	94.22%
Random Forrest	99.96%	12.24%	96.30%	93.30%	98.09%	21.65%	96.28%
XGBoost	99.94%	74.88%	99.84%	88.47%	99.89%	81.11%	99.77%

ตารางที่ 2 ผลของโมเดลที่ไม่มีการใช้ feature centrality และไม่ได้ทำ feature selection

Model	Precision		Recall		F-1 Score		Accuracy
	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	
Decision Tree	99.98%	31.62%	98.84%	96.99%	99.41%	47.69%	98.83%
Random Forrest	99.99%	22.84%	98.16%	98.06%	99.07%	37.06%	98.16%
XGBoost	99.96%	91.18%	99.95%	92.99%	99.96%	92.08%	99.91%

ตารางที่ 3 ผลของโมเดลที่ใช้เพียง feature หลังจากการ prepare data

Model	Precision		Recall		F-1 Score		Accuracy
	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	Class Legitimate	Class Fraud	
Decision Tree	99.98%	32.59%	98.88%	97.31%	99.43%	48.83%	98.88%
Random Forrest	99.99%	22.37%	98.13%	97.45%	99.05%	36.39%	98.12%
XGBoost	99.96%	91.79%	99.95%	93.03%	99.96%	92.41%	99.92%

จากตารางที่ 1 ,2 และ 3 พบว่า Extreme Gradient Boosting (XGBoost) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight ได้ผลดีที่สุดทั้ง 3 วิธี

วิธีที่ 1. มีการใช้ feature centrality ร่วมกันกับ feature selection เลือก feature 9 ตัวที่ให้ค่า importance สูงสุด

วิธีที่ 2. ไม่มีการใช้ feature centrality และไม่มีการทำ feature selection เนื่องจากมี feature อยู่ทั้งสิ้น 9 ตัว

วิธีที่ 3. ไม่มีการใช้ทั้ง feature centrality และไม่มี feature cumulative sum transactions

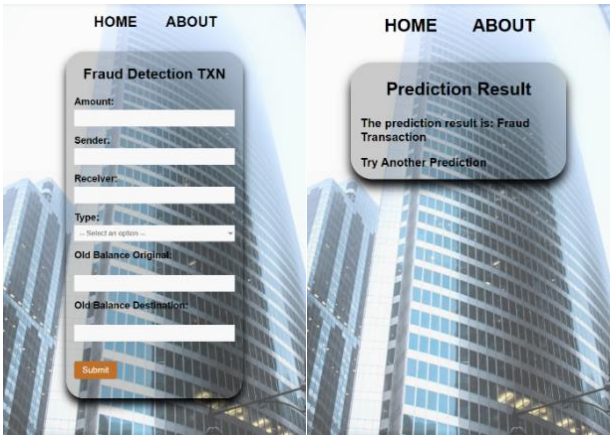
จากทั้ง 3 วิธีจึงได้ทำการเปรียบเทียบความ Overfitting ดังตารางที่ 4

ตารางที่ 4 เปรียบเทียบผลโมเดล Extreme Gradient Boosting (XGBoost) ของ 3 วิธี เทียบความ overfitting

Approach	Dataset	Precision Class Fraud	Recall Class Fraud	F-1 Score Class Fraud	Accuracy
1	Training Data	55.50%	100%	71.38%	99.85%
1	Testing Data	74.88%	88.47%	81.11%	99.77%
2	Training Data	87.25%	100%	93.19%	99.97%
2	Testing Data	91.18%	92.99%	92.08%	99.91%
3	Training Data	87.19%	100%	93.15%	99.77%
3	Testing Data	91.79%	93.03%	92.41%	99.92%

จากตารางพบว่าการใช้โมเดล Extreme Gradient Boosting (XGBoost) ในวิธีการที่ 1 ค่อนข้าง Overfitting น้อยที่สุดจากทั้ง 3 วิธีกล่าวคือ ผลของโมเดลในวิธีการที่ 2 และ 3 มีค่าตัววัดต่าง ๆ มากกว่า 85% ค่อนข้างจะไม่ generalize และหากเปรียบเทียบกับวิธีการที่ 1 ที่ ค่า precision และ recall ใน training Data น้อยกว่าวิธีการที่ 2 และ 3 อย่างมาก และในการประยุกต์ใช้จริงต้องคำนึงถึงการใช้โมเดลที่ไม่ยึดติดกับตัว data มากเกินไป

ดังนั้นจึงพิจารณาเลือกวิธีการที่ 1 ในการนำโมเดลไปใช้ในการจำลองการตรวจจับ transactions ที่เกิดการฉ้อโกง บน Web-Application



ภาพที่ 1 หน้า User interface บน Web-Application และการแสดงผลการทำนาย

สรุปและอภิปรายผลการวิจัย

1. ได้หาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับการตรวจจับธุรกรรมการฉ้อโกงโดยการประยุกต์ใช้ค่า Centrality นำมาสร้างเป็นตัวแปรในโมเดล ซึ่งพบว่าสามารถลดความ Overfitting กับตัวข้อมูลของโมเดลที่นำมาใช้ได้ เมื่อเปรียบเทียบกับวิธีการอื่นที่ไม่ได้มีการใช้ตัวแปรที่เกี่ยวข้องกับค่า Centrality

2. ได้พัฒนาโมเดล Extreme Gradient Boosting (XGBoost) (ตามวิธีการทดลองที่ 1) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight, มีการใช้ feature degree centrality และ closeness centrality และใช้ตัวแปรที่มีค่า feature importance 9 ลำดับแรกซึ่งให้ค่า Precision อยู่ที่ 74.88% ค่า Recall อยู่ที่ 88.47% ค่า F-1 Score อยู่ที่ 81.11% และ Accuracy อยู่ที่ 99.77%

3. ได้พัฒนาเครื่องมือที่สามารถใส่ค่า input พร้อมทั้งทำการตรวจสอบ transactions โดยการใช้โมเดลที่ได้ทำการพัฒนาและแสดงข้อมูล transaction ว่าเป็น Legitimate หรือ Fraud

ข้อเสนอแนะ

1. พิจารณา feature ค่า Centrality ตัวอื่นๆ เช่น Betweenness, PageRank ทั้งนี้หากใช้ Betweenness อาจต้องมี hardware ระดับสูงเพื่อลดระยะเวลาการคำนวณ

2. พัฒนาประสิทธิภาพของโมเดลหรือทดลองทำโมเดลอื่นเพิ่มเติมเพื่อให้ได้ผลการตรวจจำที่แม่นยำมากขึ้น

3. พิจารณาหรือทดลองการจัดการข้อมูลที่ไม่สมดุลแนวทางอื่นเพิ่มเติมเพื่อให้ได้ผลการตรวจจำแม่นยำมากขึ้น

4. ข้อเสนอแนะสำหรับการวิจัยครั้งต่อไป ควรเลือกใช้ dataset ที่ไม่ถูก scaled ลงมาเพื่อยืนยันผลของ feature centrality ที่ได้สร้างขึ้น เพราะข้อมูลที่ถูก scaled ลงมาจะทำให้เกิดความยากต่อการใช้งานในเรื่องของตัวแปรที่เกี่ยวเนื่องกันแบบ consequence

บรรณานุกรม

ธนาคารแห่งประเทศไทย. (2566). *นำส่งแนวนโยบายการบริหารจัดการภัยทุจริตจากการทำธุรกรรมทางการเงิน*. https://www.bot.or.th/content/dam/bot/fipcs/documents/FOG/2566/ThaiPDF/25660066.pdf

ธนาคารกสิกรไทย. (ม.ป.ป.). *บัญชีม้า*. https://www.kasikornbank.com/th/personal/digital-banking/kbankcyberrisk/pages/sellingbankaccount.html

ออมตังค์ ความรู้ทางการเงินออนไลน์ (2563, 12 มิถุนายน). *ใครมาชวน "รับจ้างเปิดบัญชี" ห้ามหลงเชื่อ*

เด็ดขาด!.

https://oomtang.gsb.or.th/kms/kms_view/52#:~:text=%22%E0%B8%82%E0%B8%9A%E0%B8%A7%E0%B8%99%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%88%E0%B9%89%E0%B8%B2%E0%B8%87%E0%B9%80%E0%B8%9B%E0%B8%B4%E0%B8%94%E0%B8%9A%E0%B8%B1%E0%B8%8D%E0%B8%8A%E0%B8%B5%22,%E0%B9%80%E0%B8%87%E0%B8%B4%E0%B8%99%E0%B8%84%E0%B9%88%E0%B8%B2%E0%B8%88%E0%B9%89%E0%B8%B2%E0%B8%87%E0%B9%83%E0%B8%AB%E0%B9%89%E0%B9%80%E0%B8%A3%E0%B8%B2

WEBSITE CRAFTER (ม.ป.ป.). *Web application คืออะไร? ต่างจากเว็บไซต์ทั่วไปอย่างไร?*.

https://1stcraft.com/website-application-vs-general-website

Brownie, J. (2020). *Imbalanced Classification with Python Choose Better Metrics. Balance Skewed* Classes, and Apply Cost-Sensitive Learning. https://machinelearningmastery.com/imbalanced-classification-with-python

DMLC UW. (n.d.) *XGBoost: A Scalable Tree Boosting System*. https://dmlc.cs.washington.edu/xgboost.html

Geeksforgeeks. (2023, May 18). *ML | Extra Tree Classifier for Feature Selection.* https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection

Geron, A. (2019*). Hands-On Machine Learning with Scikit-Learn, Keras, and TernsorFlow, (2nd ed.)* O'Reilly Media.

Hajak, P., Abedin, M. & Sivarajah (2022). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. *SPRINGER.* https://doi.org/10.1007/s10796-022-10346-6

Kumar, A. (2023). *Dealing with Class Imbalance in Python: Techniques*. https://vitalflux.com/class-imbalance-class-weight-python-sklearn

Lopez-Rojas, E. (2017). *Synthetic Financial Datasets For Fraud Detection*. https://www.kaggle.com/datasets/ealaxi/paysim1

Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). Paysim: A financial mobile money simulator for fraud detection. In 28th European modeling and simulation symposium, EMSS 2016, Dime University of Genoa, Larnaca (pp. 249–255).

Sahu, A., GM, H. & Gourisaria, (2020). A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques. *IEEE 17th India Council International Conference (INDICON)*. https://doi.org/10.1109/INDICON49873.2020.9342462

The Science of Machine Learning. (n.d.) *Confusion Matrix*. https://www.ml-science.com/confusion-matrix

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M. & Anderla, A. (2019) .Credit Card Fraud Detection - Machine Learning methods. *IEEE 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. https://doi.org/10.1109/INFOTEH.2019.8717766

Wen, S., Li, J., Zhu, X. & Liu M. (2022). Analysis of financial fraud based on manager knowledge graph. *Procedia Computer Science, 199, 773-779.* https://doi.org/10.1016/j.procs.2022.01.096

Zafarani, R., ABBASI, M. & LIU, H. (2014). *Social Media Mining: An Introduction.* http://www.socialmediamining.info/SMM.pdf