การพัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน

DEVELOPMENT OF MACHINE LEARNING MODELS FOR FRAUDULENT TRANSACTION DETECTION IN FINANCIAL SYSTEMS

กีรติ ศาสนพิทักษ์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2567

การพัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน

กีรติ ศาสนพิทักษ์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2567

DEVELOPMENT OF MACHINE LEARNING MODELS FOR FRAUDULENT TRANSACTION DETECTION IN FINANCIAL SYSTEMS

KEERATI SASSANAPITAK

An Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2024

สารนิพนธ์

เรื่อง

การพัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน

ของ

กีรติ ศาสนพิทักษ์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... กรรมการ

(อาจารย์ ดร.บรรพตรี คมขำ)

ชื่อเรื่อง	การพัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน
ผู้วิจัย	กีรติ ศาสนพิทักษ์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2567
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องในการตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน โดยมุ่งเน้นการแก้ไขปัญหาข้อมูลไม่สมดุล การศึกษาใช้ชุดข้อมูลธุรกรรมทางการเงินจำนวน 594,643 รายการที่มีสัดส่วนธุรกรรมปกติ 98.79% และธุรกรรมฉ้อโกงเพียง 1.21% การวิจัยเปรียบเทียบโมเดล 5 ประเภท ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression ร่วมกับเทคนิคจัดการข้อมูลไม่สมดุล 4 วิธี ได้แก่ การเพิ่มจำนวนข้อมูลด้วย SMOTE การลดข้อมูลซ้ำซ้อนด้วย Tomek Links วิธีผสมผสานระหว่าง SMOTE และ Tomek Links และการปรับน้ำหนักคลาส นอกจากนี้ ยังมีการสร้างคุณลักษณะใหม่ที่เกี่ยวข้องกับพฤติกรรมของลูกค้า ร้านค้า และหมวดหมู่ธุรกรรม ผลการวิจัยพบว่า XGBoost ร่วมกับ Tomek Links ที่ผ่านการปรับแต่งพารามิเตอร์ให้ผลลัพธ์ดีที่สุดในภาพรวม ด้วยค่า Precision 0.88, Recall 0.82, F1-Score 0.85 และ Precision-Recall AUC 0.93 ในขณะที่ LightGBM ร่วมกับ Class Weight ให้ค่า Recall สูงสุดถึง 0.98 แม้จะมี Precision ต่ำ (0.41) การวิเคราะห์ด้วย SHAP แสดงให้เห็นว่าปัจจัยที่มีอิทธิพลสูงสุดต่อการทำนายการฉ้อโกงคือ จำนวนธุรกรรมของร้านค้า มูลค่าธุรกรรม และจำนวนธุรกรรมในแต่ละหมวดหมู่ โดยร้านค้าที่มีประวัติธุรกรรมน้อยผิดปกติหรือธุรกรรมที่มีมูลค่าสูงผิดปกติมีความเสี่ยงสูง งานวิจัยนี้แสดงให้เห็นว่าการเลือกเทคนิคจัดการข้อมูลไม่สมดุลที่เหมาะสมร่วมกับโมเดลที่เหมาะสมมีความสำคัญต่อประสิทธิภาพในการตรวจจับการฉ้อโกงทางการเงิน โดยขึ้นอยู่กับว่าองค์กรจะเน้นความสมดุลระหว่างการแจ้งเตือนผิดกับการพลาดธุรกรรมฉ้อโกง หรือเน้นการตรวจจับให้ได้มากที่สุด

คำสำคัญ : การตรวจจับการฉ้อโกง, การเรียนรู้ของเครื่อง, ข้อมูลไม่สมดุล, SMOTE, Tomek Links, XGBoost, SHAP

Title	DEVELOPMENT OF MACHINE LEARNING MODELS FOR FRAUDULENT TRANSACTION DETECTION IN FINANCIAL SYSTEMS
Author	KEERATI SASSANAPITAK
Degree	MASTER OF SCIENCE
Academic Year	2024
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

This research aims to develop and compare the effectiveness of machine learning models for fraud detection in financial systems, focusing on solving class imbalance problems. The study used a financial transaction dataset of 594,643 records with 98.79% normal transactions and only 1.21% fraudulent transactions. The research compared five models: XGBoost, Random Forest, CatBoost, LightGBM, and Logistic Regression, combined with four class imbalance handling techniques: SMOTE for oversampling, Tomek Links for undersampling, a hybrid approach combining both methods, and Class Weight adjustment. Additionally, new features related to customer behavior, merchant patterns, and transaction categories were created to enhance detection capability. Results showed that XGBoost combined with Tomek Links and optimized parameters delivered the best overall performance, with Precision of 0.88, Recall of 0.82, F1-Score of 0.85, and Precision-Recall AUC of 0.93. Meanwhile, LightGBM with Class Weight adjustment achieved the highest Recall at 0.98, though with lower Precision (0.41). SHAP analysis revealed that merchant transaction frequency, transaction amount, and transaction frequency per category were the most influential predictors, with merchants having unusually low transaction history or transactions with abnormally high values presenting higher fraud risk. This research demonstrates that selecting appropriate class imbalance techniques and suitable models is crucial for effective financial fraud detection, depending on organizational priorities—whether to balance false alerts with missed fraud detections or to maximize detection rates.

Keyword : Fraud Detection Machine Learning Class Imbalance SMOTE Tomek Links XGBoost SHAP

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ ผศ.ดร. ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษาหลัก ที่ได้ให้คำแนะนำและข้อเสนอแนะที่มีค่าตลอดกระบวนการวิจัย รวมถึงการให้คำปรึกษาเชิงเทคนิคและวิธีการในด้านวิทยาการข้อมูลอย่างต่อเนื่อง ขอขอบพระคุณคณะกรรมการสอบปากเปล่าสารนิพนธ์ทุกท่าน ที่ได้กรุณาให้คำแนะนำและข้อเสนอแนะ เพื่อให้สารนิพนธ์ฉบับนี้มีความสมบูรณ์และมีคุณภาพยิ่งขึ้น ขอขอบพระคุณครอบครัวที่ได้ให้การสนับสนุนและเป็นกำลังใจสำคัญในการศึกษาและการทำงานวิจัยให้บรรลุผลสำเร็จ

กีรติ ศาสนพิทักษ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1. ที่มาและความสำคัญของปัญหา

ในยุคดิจิทัลที่เทคโนโลยีก้าวหน้าอย่างรวดเร็ว ระบบการเงินทั่วโลกได้เปลี่ยนแปลงไปอย่างมาก โดยธุรกรรมส่วนใหญ่เกิดขึ้นผ่านระบบอิเล็กทรอนิกส์ ทำให้การดำเนินชีวิตและการทำธุรกิจสะดวกสบายขึ้น แต่ความก้าวหน้านี้ก็นำมาซึ่งความท้าทายใหม่ๆ โดยเฉพาะปัญหาการฉ้อโกงทางการเงิน ซึ่งเป็นภัยคุกคามสำคัญต่อความมั่นคงของระบบการเงินโลก ข้อมูลสถิติจากหลายแหล่งสะท้อนให้เห็นถึงความรุนแรงของปัญหานี้:

1.1 รายงานจาก Association of Certified Fraud Examiners (ACFE) ปี 2024 วิเคราะห์ข้อมูลจากคดีการฉ้อโกง 1,921 คดีใน 138 ประเทศ ระบุว่าองค์กรทั่วโลกสูญเสียรายได้เฉลี่ย 5% ต่อปีจากการฉ้อโกง คิดเป็นมูลค่ากว่า 3.1 ล้านล้านดอลลาร์สหรัฐ และ 43% ของการฉ้อโกงถูกตรวจจับผ่านการแจ้งข้อสงสัย (Association of Certified Fraud Examiners, 2024)

1.2 ข้อมูลจาก Federal Trade Commission (FTC) ของสหรัฐอเมริกาเผยว่าในปี 2023 มีการรายงานการฉ้อโกงกว่า 2.7 ล้านกรณี โดยผู้บริโภคสูญเสียเงินกว่า 8.8 พันล้านดอลลาร์ ซึ่งเพิ่มขึ้นอย่างมากจากปีก่อนหน้า (Federal Trade Commission, 2024)

1.3 ข้อมูลจากปี 2566 ชี้ว่าประเทศไทยเป็นประเทศที่ถูกหลอกลวงทางออนไลน์สูงเป็นอันดับ 4 ของโลก คิดเป็น 3.1% ของ GDP ข้อมูลนี้มาจากสถิติการรับแจ้งความออนไลน์ของคดีอาชญากรรมทางเทคโนโลยี ระหว่างวันที่ 1 มีนาคม 2565 ถึง 30 มิถุนายน 2567 ซึ่งมีการแจ้งความผ่านระบบออนไลน์ทั้งหมด 575,507 เรื่อง คิดเป็นมูลค่าความเสียหายรวมกว่า 65,715 ล้านบาท หรือเฉลี่ยวันละกว่า 80 ล้านบาท (Siamrath, 2023)

การเติบโตของธุรกรรมออนไลน์และการใช้เทคโนโลยีทางการเงิน (FinTech) ได้เปิดช่องให้อาชญากรก่อการฉ้อโกงที่ซับซ้อนขึ้น เช่น การขโมยข้อมูลบัตรเครดิต การปลอมแปลงตัวตน การฟอกเงิน และการโจมตีทางไซเบอร์ ปัญหานี้ส่งผลกระทบไม่เพียงต่อผู้บริโภคและธุรกิจ แต่ยังบั่นทอนความเชื่อมั่นในระบบการเงินโดยรวม รูปแบบการฉ้อโกงที่พบบ่อยในประเทศไทยและทั่วโลกมีหลายประการ:

1.4 การโจมตีผ่านโทรศัพท์: แก๊งคอลเซ็นเตอร์เป็นตัวการหลักที่ใช้ในการหลอกลวง โดยมักจะใช้วิธีการสร้างเรื่องราวที่น่าเชื่อถือ เช่น การอ้างเป็นเจ้าหน้าที่รัฐหรือสถาบันการเงิน เพื่อหลอกล่อให้ผู้เสียหายโอนเงินหรือเปิดเผยข้อมูลส่วนตัว

1.5 การหลอกลวงผ่านแพลตฟอร์มออนไลน์: เช่น การซื้อขายสินค้าหรือการขอสินเชื่อ โดยเฉพาะการซื้อขายสินค้าผ่านแพลตฟอร์มที่ไม่ได้มาตรฐานหรือมีการปกป้องข้อมูลไม่เพียงพอ ทำให้ผู้ใช้ถูกหลอกให้ชำระเงินก่อนโดยไม่ได้รับสินค้า

1.6 แอปพลิเคชันที่ใช้ในการขโมยข้อมูล: หรือที่เรียกว่า "แอปดูดเงิน" ซึ่งเป็นรูปแบบใหม่ของการโจมตีที่สร้างความเสียหายให้แก่ผู้ใช้โทรศัพท์มือถือจำนวนมาก แอปพลิเคชันเหล่านี้มักถูกปลอมแปลงเป็นแอปพลิเคชันที่ดูถูกกฎหมายหรือเป็นแอปพลิเคชันที่ให้บริการทั่วไป แต่แท้จริงแล้วมีจุดประสงค์ในการขโมยข้อมูลทางการเงิน

ความท้าทายหลักในการต่อสู้กับการฉ้อโกงทางการเงิน คือการตรวจจับและป้องกันให้ทันเวลา เนื่องจากธุรกรรมเกิดขึ้นอย่างรวดเร็วและต่อเนื่องตลอด 24 ชั่วโมง การใช้มนุษย์ตรวจสอบเพียงอย่างเดียวไม่สามารถรับมือกับปริมาณข้อมูลและความซับซ้อนของรูปแบบการฉ้อโกงที่เปลี่ยนแปลงอยู่ตลอดเวลาได้

ด้วยเหตุนี้ การพัฒนาระบบตรวจจับการฉ้อโกงที่มีประสิทธิภาพจึงมีความสำคัญอย่างยิ่ง Machine Learning ได้พิสูจน์ให้เห็นถึงศักยภาพในการวิเคราะห์ข้อมูลขนาดใหญ่และซับซ้อน ทำให้สามารถระบุธุรกรรมที่น่าสงสัยได้อย่างรวดเร็วและแม่นยำ ตลาดเทคโนโลยีสำหรับการตรวจจับและป้องกันการฉ้อโกงคาดว่าจะเติบโตจาก 30.5 พันล้านดอลลาร์ในปี 2023 เป็น 76.7 พันล้านดอลลาร์ภายในปี 2028 ด้วยอัตราการเติบโตเฉลี่ยต่อปี (CAGR) 20.3% สะท้อนให้เห็นถึงความสำคัญและความต้องการเทคโนโลยีนี้ที่เพิ่มขึ้นอย่างมาก

อย่างไรก็ตาม การพัฒนาโมเดล Machine Learning สำหรับการตรวจจับการฉ้อโกงยังคงเผชิญกับความท้าทายหลายประการ เช่น:

1.7 ความไม่สมดุลของข้อมูล (Data Imbalance): ธุรกรรมฉ้อโกงมีสัดส่วนน้อยเมื่อเทียบกับธุรกรรมปกติ ทำให้โมเดลมีแนวโน้มที่จะเรียนรู้การทำนายธุรกรรมปกติได้ดีกว่า

1.8 ปริมาณข้อมูลมหาศาล: ระบบการเงินประมวลผลธุรกรรมหลายล้านรายการต่อวัน ทำให้การจัดการและประมวลผลข้อมูลเป็นความท้าทายทั้งด้านเวลาและทรัพยากรคอมพิวเตอร์

1.9 รูปแบบการฉ้อโกงที่เปลี่ยนแปลง: อาชญากรมักพัฒนาวิธีการใหม่ๆ อย่างต่อเนื่อง ทำให้โมเดลต้องยืดหยุ่นและปรับตัวได้

1.10 ความปลอดภัยของข้อมูลทางการเงิน: การรักษาความเป็นส่วนตัวของข้อมูลลูกค้าเป็นสิ่งสำคัญ ซึ่งอาจจำกัดการเข้าถึงข้อมูลสำหรับการฝึกฝนโมเดล

1.11 ความต้องการการตรวจจับแบบเรียลไทม์: ระบบต้องสามารถตรวจจับและตอบสนองต่อธุรกรรมที่น่าสงสัยได้อย่างรวดเร็ว เพื่อป้องกันความเสียหาย

การวิจัยเพื่อพัฒนาโมเดล Machine Learning ที่มีประสิทธิภาพในการตรวจจับการฉ้อโกงมีความสำคัญอย่างยิ่ง โดยเฉพาะการเปรียบเทียบประสิทธิภาพของโมเดลต่างๆ และพัฒนาเทคนิคจัดการกับปัญหาความไม่สมดุลและปริมาณข้อมูลขนาดใหญ่ ผลการวิจัยนี้ไม่เพียงช่วยปกป้องผู้บริโภคและสถาบันการเงินจากการสูญเสีย แต่ยังช่วยเสริมสร้างความมั่นคงและความน่าเชื่อถือของระบบการเงิน ซึ่งเป็นรากฐานสำคัญของเศรษฐกิจดิจิทัลในปัจจุบันและอนาคต การพัฒนาเทคโนโลยีนี้จึงมีความจำเป็นอย่างเร่งด่วนเพื่อรับมือกับภัยคุกคามทางการเงินที่ทวีความรุนแรงขึ้นทุกวัน

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อพัฒนาและประเมินประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง ในการตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน โดยมุ่งเน้นที่โมเดล 5 ประเภท ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression โดยเน้นในกลุ่มของ Classification

2.2 เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องทั้ง 5 ประเภท ในแง่ของความแม่นยำของการทำนาย (Precision), ความไว (Recall), ค่า F1-Score, พื้นที่ใต้กราฟ ROC (ROC-AUC) และพื้นที่ใต้กราฟ Precision-Recall (Precision-Recall AUC) ในการตรวจจับธุรกรรมฉ้อโกง

2.3 เพื่อศึกษาและประเมินประสิทธิภาพของเทคนิคการจัดการกับปัญหาความไม่สมดุลของข้อมูล โดยเปรียบเทียบระหว่างเทคนิค SMOTE, Tomek Links, SMOTETomek และการปรับ Class Weight ในโมเดล

2.4 เพื่อพัฒนาแนวทางในการปรับปรุงและอัพเดทโมเดลอย่างต่อเนื่อง เพื่อรับมือกับรูปแบบการฉ้อโกงที่เปลี่ยนแปลงตลอดเวลา

3. สมมติฐานของการวิจัย

3.1 การจัดการข้อมูลไม่สมดุล (Class Imbalance) ด้วยเทคนิคต่างๆ (SMOTE, Tomek Links หรือแบบผสม) จะช่วยเพิ่มประสิทธิภาพการตรวจจับธุรกรรมฉ้อโกงได้อย่างมีนัยสำคัญ เมื่อเทียบกับกรณีที่ไม่มีการจัดการข้อมูลไม่สมดุล

3.2 โมเดลแบบ Gradient Boosting (เช่น XGBoost, LightGBM) จะให้ประสิทธิภาพสูงกว่าโมเดลประเภทอื่นในการตรวจจับธุรกรรมฉ้อโกง โดยวัดจากค่า AUC และ Recall

3.3 การใช้เทคนิคปรับค่าน้ำหนัก (เช่น Class Weight) และการปรับ Hyperparameters จะช่วยเพิ่มประสิทธิภาพของโมเดลได้อย่างมีนัยสำคัญ

3.4 คุณลักษณะ (features) ด้านพฤติกรรมของร้านค้าและมูลค่าธุรกรรมจะเป็นปัจจัยสำคัญที่สุดในการบ่งชี้ธุรกรรมฉ้อโกง

4. ขอบเขตของการวิจัย

การวิจัยนี้จะมุ่งเน้นไปที่การพัฒนาและเปรียบเทียบโมเดลการเรียนรู้ของเครื่องสำหรับการตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน โดยมีขอบเขตดังนี้:

4.1 แหล่งที่มาและลักษณะของข้อมูล: ใช้ข้อมูลธุรกรรมการเงินจากเว็บไซต์ Kaggle (594,643 รายการ) ที่มีความไม่สมดุลระหว่างธุรกรรมปกติ (98.79%) และธุรกรรมฉ้อโกง (1.21%)

4.2 วิธีการจัดการกับความไม่สมดุลของข้อมูล: เปรียบเทียบเทคนิค SMOTE, Tomek Links, SMOTETomek และ Class Weight เพื่อหาวิธีที่เหมาะสมที่สุด

4.3 โมเดลที่ใช้ในการวิเคราะห์: พัฒนา XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression พร้อมปรับแต่งพารามิเตอร์และศึกษาผลของ Class Weight

4.4 เกณฑ์การประเมินผล: ใช้ Recall เป็นเมตริกหลัก ร่วมกับ Precision-Recall curve และ F1-score วิเคราะห์ความสำคัญของคุณลักษณะด้วย SHAP Analysis

4.5 วิธีการทดสอบความทนทานของโมเดล: แบ่งชุดข้อมูลเป็น Training และ Test โดยรักษาสัดส่วนดั้งเดิม ใช้เทคนิค 5-Fold Cross Validation เพื่อประเมินความเสถียรของโมเดล พร้อมทดสอบความสามารถในการตรวจจับธุรกรรมฉ้อโกงที่มีรูปแบบแตกต่าง

5. นิยามศัพท์เฉพาะ

5.1 ธุรกรรมฉ้อโกง: ธุรกรรมทางการเงินที่ดำเนินการโดยไม่ได้รับอนุญาตหรือมีเจตนาหลอกลวงเพื่อผลประโยชน์ทางการเงิน

5.2 SMOTE: เทคนิคการเพิ่มจำนวนตัวอย่างของคลาสส่วนน้อยโดยการสร้างตัวอย่างสังเคราะห์

5.3 Tomek Links: เทคนิคการลดจำนวนตัวอย่างของคลาสส่วนมากโดยกำจัดตัวอย่างที่อยู่ใกล้ขอบเขตการตัดสินใจ

5.4 SMOTETomek: เทคนิคผสมระหว่าง SMOTE และ Tomek Links เพื่อปรับปรุงขอบเขตระหว่างคลาส

5.5 Class Weight: การปรับน้ำหนักของคลาสเพื่อเพิ่มความสำคัญให้คลาสที่มีจำนวนตัวอย่างน้อย

5.6 Data Imbalance: สถานการณ์ที่คลาสหนึ่งในชุดข้อมูลมีจำนวนตัวอย่างมากกว่าอีกคลาสหนึ่งอย่างมีนัยสำคัญ

5.7 Recall: อัตราส่วนของธุรกรรมฉ้อโกงที่โมเดลตรวจจับได้ถูกต้องเทียบกับจำนวนธุรกรรมฉ้อโกงทั้งหมด

5.8 Precision: อัตราส่วนของธุรกรรมที่โมเดลทำนายว่าเป็นการฉ้อโกงและเป็นการฉ้อโกงจริง

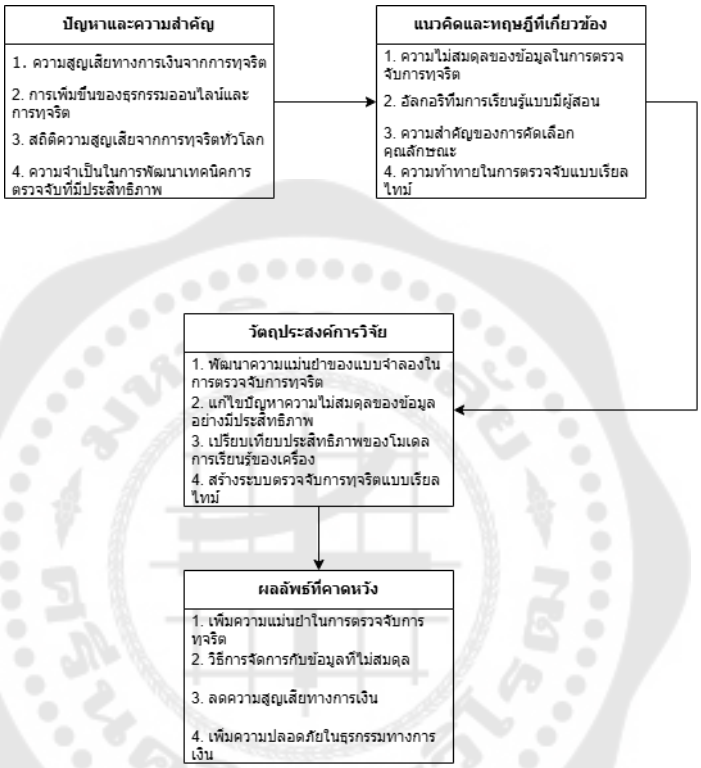
5.9 F1-Score: ค่าเฉลี่ยฮาร์โมนิกระหว่าง Precision และ Recall

5.10 Feature Importance: การวิเคราะห์ว่าแต่ละคุณลักษณะส่งผลต่อการตัดสินใจของโมเดลมากน้อยเพียงใด

5.11 Cross-Validation: เทคนิคการประเมินประสิทธิภาพของโมเดลโดยการแบ่งข้อมูลเป็นส่วนย่อยหลายส่วน

5.12 Cost Matrix: ตารางที่แสดงต้นทุนหรือผลกระทบของการทำนายผิดพลาดในแต่ละกรณี

6. กรอบแนวคิดงานวิจัย



ภาพประกอบ 1 กรอบแนวคิดการวิจัยเรื่องการตรวจจับการฉ้อโกงด้วยเทคนิคการเรียนรู้ของเครื่อง

ภาพประกอบที่ 1 แสดงกรอบแนวคิดงานวิจัยที่เชื่อมโยงระหว่างปัญหาและความสำคัญของการฉ้อโกงทางการเงิน กับแนวคิดและทฤษฎีที่จะนำมาใช้ในการแก้ปัญหา โดยมีการกำหนดวัตถุประสงค์การวิจัยที่ชัดเจน และผลลัพธ์ที่คาดว่าจะได้รับจากงานวิจัยนี้ จะเห็นได้ว่ากรอบแนวคิดนี้มุ่งเน้นการพัฒนาโมเดลการตรวจจับการฉ้อโกงให้มีประสิทธิภาพมากขึ้น โดยแก้ไขปัญหาความไม่สมดุลของข้อมูล และพัฒนาระบบตรวจจับแบบเรียลไทม์ เพื่อลดความสูญเสียทางการเงินและเพิ่มความปลอดภัยในการทำธุรกรรม

7. ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

งานวิจัยนี้จะช่วยพัฒนาโมเดลการเรียนรู้ของเครื่องที่มีประสิทธิภาพสูงสำหรับการตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน เพื่อลดความเสี่ยงจากการสูญเสียทางการเงินและเพิ่ม

ความปลอดภัยให้กับองค์กรการเงิน นอกจากนี้ งานวิจัยยังนำเสนอแนวทางที่เหมาะสมในการจัดการปัญหาความไม่สมดุลของข้อมูล ซึ่งช่วยลดต้นทุนการประมวลผลและเพิ่มความรวดเร็วในการตรวจจับธุรกรรมผิดปกติ พร้อมทั้งพัฒนาแนวทางปรับปรุงโมเดลอย่างต่อเนื่องเพื่อรองรับการฉ้อโกงในรูปแบบใหม่

บทที่ 2
วรรณกรรม และงานวิจัยที่เกี่ยวข้อง

บทนี้นำเสนอการทบทวนวรรณกรรมที่มุ่งเน้นการศึกษาและวิเคราะห์งานวิจัยเกี่ยวกับการพัฒนาโมเดลการเรียนรู้ของเครื่องเพื่อตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน ผู้วิจัยได้ทำการศึกษาอย่างละเอียดจากแหล่งข้อมูลที่หลากหลาย และได้สังเคราะห์เนื้อหาตามประเด็นสำคัญดังต่อไปนี้

1. ความหมายและประเภทของการฉ้อโกงทางการเงิน
2. ทฤษฎีเกี่ยวกับอัลกอริทึมในการจำแนกประเภท (Classification Algorithms)
3. การวัดประสิทธิภาพของแบบจำลอง (Model Performance Metrics)
4. เทคนิคการจัดการข้อมูลที่ไม่สมดุลในบริบทของการตรวจจับการฉ้อโกง
5. งานวิจัยที่เกี่ยวข้องกับการตรวจจับการฉ้อโกงในธุรกรรมการเงิน

การทบทวนวรรณกรรมนี้จะครอบคลุมงานวิจัยที่เกี่ยวข้องจากหลากหลายแหล่งข้อมูล รวมถึงบทความวิชาการ รายงานการวิจัย และกรณีศึกษาจากอุตสาหกรรมการเงิน โดยมุ่งเน้นงานวิจัยที่ตีพิมพ์ในช่วง 10 ปีที่ผ่านมา เพื่อให้ได้ข้อมูลที่ทันสมัยและสอดคล้องกับสภาพแวดล้อมทางเทคโนโลยีในปัจจุบัน ผลจากการทบทวนวรรณกรรมนี้จะนำไปสู่การระบุช่องว่างในงานวิจัยปัจจุบัน และเป็นแนวทางในการพัฒนาโมเดลการตรวจจับธุรกรรมฉ้อโกงที่มีประสิทธิภาพยิ่งขึ้นในอนาคต

1. ความหมายและประเภทของการฉ้อโกงทางการเงิน

1.1 ความหมายของการฉ้อโกงทางการเงิน

การฉ้อโกงทางการเงิน (Financial Fraud) หมายถึงการกระทำที่ไม่ถูกต้องทางกฎหมาย ซึ่งเกี่ยวข้องกับการใช้เล่ห์เหลี่ยม กลอุบาย หรือการปิดบังข้อมูลเพื่อให้ได้มาซึ่งผลประโยชน์ทางการเงิน โดยอาจทำให้บุคคลหรือองค์กรอื่น ๆ เสียเปรียบ หรือสูญเสียทรัพย์สิน ไม่ว่าจะเป็นการโกงที่ทำโดยบุคคลภายในองค์กร ผู้ประกอบการทางธุรกิจ หรือบุคคลภายนอกที่มุ่งทำให้เกิดความเสียหายต่อระบบการเงิน การฉ้อโกงทางการเงินอาจเกิดขึ้นในหลายรูปแบบและสามารถเกิดได้ในหลายช่องทาง เช่น การฉ้อโกงในการทำธุรกรรมทางการเงินผ่านอินเทอร์เน็ต การ

ปลอมแปลงข้อมูลบัตรเครดิต หรือแม้แต่การฉ้อโกงภายในองค์กร โดยการกระทำเหล่านี้มักเป็นที่มาของการสูญเสียเงินมหาศาล และอาจก่อให้เกิดความเสียหายที่ยากต่อการกู้คืนสำหรับทั้งผู้บริโภคและสถาบันการเงิน

1.2 ประเภทของการฉ้อโกงทางการเงิน

การฉ้อโกงทางการเงินมีหลายประเภทซึ่งแบ่งได้ตามลักษณะของการกระทำและช่องทางที่ใช้ โดยประเภทที่สำคัญ ได้แก่:

1.2.1 การฉ้อโกงบัตรเครดิต (Credit Card Fraud)

การใช้ข้อมูลบัตรเครดิตของผู้อื่นโดยไม่ได้รับอนุญาต อาจเกิดขึ้นได้จากการขโมยข้อมูลผ่านการแฮ็กระบบร้านค้าออนไลน์ การใช้ข้อมูลที่ได้จากการซื้อขายข้อมูลบัตรเครดิตในตลาดมืด หรือการปลอมแปลงบัตรเครดิต ซึ่งผู้กระทำการฉ้อโกงมักใช้ข้อมูลเหล่านี้ในการซื้อสินค้าหรือบริการทางออนไลน์

1.2.2 การฉ้อโกงการโอนเงิน (Wire Transfer Fraud)

การฉ้อโกงผ่านการโอนเงิน เช่น การปลอมแปลงบัญชีผู้รับเงิน หรือการส่งคำสั่งโอนเงินที่เป็นเท็จไปยังธนาคารหรือสถาบันการเงิน โดยเป้าหมายของการกระทำนี้คือการยักยอกเงินจากบัญชีที่ถูกหลอกลวงไปยังบัญชีของผู้ฉ้อโกง ซึ่งยากต่อการติดตามหรือยึดคืน

1.2.3 การฉ้อโกงออนไลน์ (Online Fraud)

การฉ้อโกงที่เกิดขึ้นในช่องทางออนไลน์ เช่น การปลอมแปลงเว็บไซต์เพื่อขโมยข้อมูลส่วนตัว (Phishing) การฉ้อโกงผ่านแพลตฟอร์มอีคอมเมิร์ซ การปลอมแปลงโปรไฟล์หรือการซื้อขายสินค้าที่ไม่ถูกต้องตามที่กล่าวอ้าง ซึ่งเป็นการใช้เทคโนโลยีเพื่อประโยชน์ส่วนตัวในลักษณะที่ผิดกฎหมาย

1.2.4 การฉ้อโกงด้านประกันภัย (Insurance Fraud)

การปลอมแปลงข้อมูลหรือสร้างเหตุการณ์เท็จเพื่อนำไปเรียกร้องค่าสินไหมทดแทนจากบริษัทประกันภัย เช่น การปลอมแปลงอุบัติเหตุ หรือการเรียกร้องค่าสินไหมทดแทนเกินความจริง ซึ่งการกระทำเช่นนี้ไม่เพียงส่งผลเสียต่อบริษัทประกัน แต่ยังส่งผลให้ผู้ถือกรมธรรม์อื่น ๆ ต้องจ่ายเบี้ยประกันที่สูงขึ้น

1.2.5 การฟอกเงิน (Money Laundering)

การฟอกเงินเป็นกระบวนการที่บุคคลหรือองค์กรแปลงเงินที่ได้จากการกระทำผิดกฎหมาย เช่น การค้ายาเสพติด หรือการคอร์รัปชัน ให้กลายเป็นเงินที่ถูกกฎหมาย ผ่านการลงทุนในธุรกิจที่ดูเหมือนถูกต้อง การโอนเงินผ่านบัญชีหลายแห่ง หรือการซื้อขายทรัพย์สินมูลค่าสูง เช่นอสังหาริมทรัพย์และงานศิลปะ เพื่อปกปิดแหล่งที่มาของเงินที่ผิดกฎหมาย

1.2.6 การฉ้อโกงทางภาษี (Tax Fraud)

การหลีกเลี่ยงการเสียภาษีโดยการปกปิดรายได้จริง หรือรายงานค่าใช้จ่ายที่ไม่ถูกต้อง เพื่อให้ได้มาซึ่งประโยชน์ทางภาษีอย่างไม่ถูกต้อง ซึ่งส่งผลให้รัฐสูญเสียรายได้และกระทบต่อการจัดการงบประมาณของประเทศ

1.2.7 การฉ้อโกงโดยใช้บัญชีธนาคารปลอม (Fake Bank Account Fraud)

การเปิดบัญชีธนาคารปลอมเพื่อวัตถุประสงค์ในการฉ้อโกง เช่น การเปิดบัญชีในชื่อผู้อื่นหรือชื่อปลอมเพื่อยักยอกเงินจากการโอนเงินหรือการฉ้อโกงผ่านธนาคารออนไลน์

1.2.8 การฉ้อโกงด้านหลักทรัพย์ (Securities Fraud)

การฉ้อโกงในตลาดหลักทรัพย์ เช่น การปลอมแปลงข้อมูลทางการเงิน การปั่นหุ้น (Insider Trading) การให้ข้อมูลที่ไม่ถูกต้องต่อสาธารณะเพื่อปั่นราคาหุ้น หรือการฉ้อโกงนักลงทุนในรูปแบบต่าง ๆ

การป้องกันและแก้ไขปัญหาการฉ้อโกงทางการเงินเพื่อลดโอกาสของการฉ้อโกงทางการเงินสถาบันการเงินต่าง ๆ ต้องพัฒนากลยุทธ์และเทคโนโลยีใหม่ ๆ ในการตรวจจับการกระทำที่น่าสงสัย เช่น การใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อวิเคราะห์พฤติกรรมการทำธุรกรรมในเวลาจริง (Real-time Transaction Monitoring) รวมถึงการใช้ระบบป้องกันข้อมูลที่ทันสมัย เช่น การเข้ารหัสข้อมูล การตรวจสอบสองขั้นตอน (Two-factor Authentication) และระบบการแจ้งเตือนการทำธุรกรรมที่ผิดปกติ

2. ทฤษฎีเกี่ยวกับอัลกอริทึมในการจำแนกประเภท (Classification Algorithms)

อัลกอริทึมการจำแนกประเภท (Classification Algorithms) เป็นเทคนิคในสาย Machine Learning ที่ทำหน้าที่ทำนายหรือจำแนกกลุ่ม (Class) ให้กับข้อมูลใหม่ โดยใช้

แบบจำลองที่เรียนรู้มาจากข้อมูลตัวอย่าง (Training Data) ที่มีการระบุคลาสไว้อย่างชัดเจน งานตรวจจับธุรกรรมฉ้อโกงส่วนใหญ่จัดอยู่ในหมวดการจำแนกประเภท เพราะโมเดลต้องทำนายว่าธุรกรรมหนึ่ง ๆ เป็น ฉ้อโกง (Fraud) หรือ ไม่ฉ้อโกง (Non-Fraud)

2.1 Random Forest

Random Forest: เป็นอัลกอริทึมการเรียนรู้แบบ ensemble ที่พัฒนาต่อยอดจากต้นไม้ตัดสินใจ (Decision Tree) เพื่อแก้ไขข้อบกพร่องของต้นไม้เดี่ยว เช่น การ overfitting และความไวต่อความผันผวนของข้อมูล Random Forest ทำงานโดยการสร้าง “ป่า” ของต้นไม้ตัดสินใจจำนวนมากที่มีความหลากหลายและไม่สัมพันธ์กันมากนัก ผ่านกลไกสองขั้นตอนหลัก ได้แก่ การสุ่มตัวอย่างข้อมูลด้วยวิธี Bootstrap Aggregating (bagging) ซึ่งเป็นการสุ่มเลือกข้อมูลจากชุดฝึกแบบมีการคืนตัวอย่าง (sampling with replacement) และการสุ่มเลือก subset ของคุณลักษณะในแต่ละ node ของต้นไม้เพื่อกำหนดเกณฑ์การแยก (split) ผลการทำนายของ Random Forest มาจากการโหวตข้างมาก (majority voting) จากทุกต้นไม้ในป่าโดยคำนวณความน่าจะเป็นของคลาส c ได้ดังสมการที่ (1)

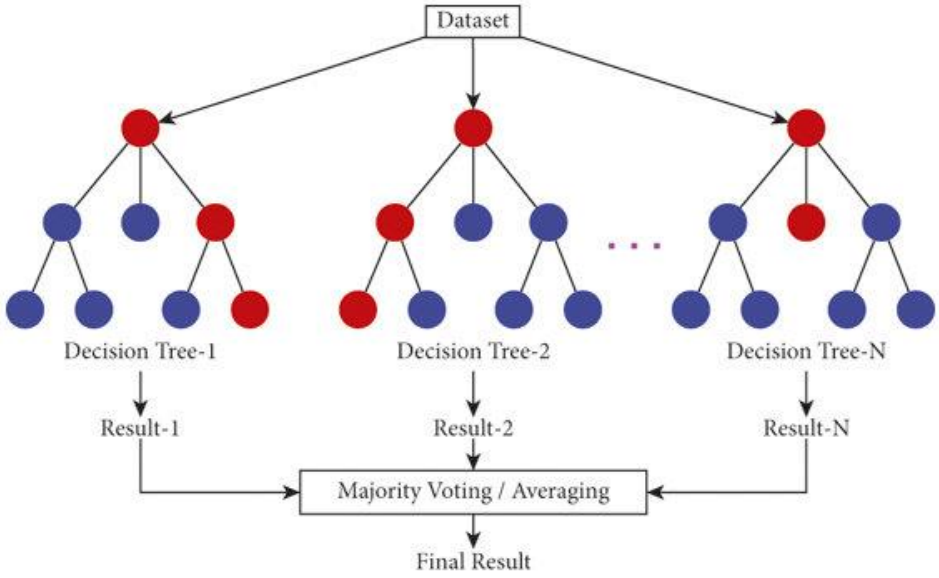
$$P(C) = \frac{1}{T}\sum_{t=1}^{T} I(h_t(x) = C) \tag{1}$$

โดยที่

$h_t(x)$ คือ ผลลัพธ์จากต้นไม้ต้นที่ t

$I(\cdot)$ คือ ฟังก์ชั่นตัวบ่งชี้ ซึ่งให้ค่า 1 หากทำนายตรงกับคลาส c และ 0 หากไม่ตรง

โมเดลนี้เหมาะกับการตรวจจับการฉ้อโกงที่ต้องพิจารณาตัวแปรหลายตัวพร้อมกัน เช่น จำนวนเงินธุรกรรมสูงผิดปกติ สถานที่ทำรายการที่อยู่นอกพฤติกรรมปกติของผู้ใช้ ช่วงเวลาที่ทำรายการที่น่าสงสัย หรือประวัติการทำรายการที่ผ่านมา โดยแต่ละต้นไม้ในป่าสามารถจับรูปแบบย่อยของข้อมูลที่แตกต่างกันได้ ตัวอย่างเช่น ต้นไม้หนึ่งอาจเน้นที่สถานที่ทำรายการที่ไม่เคยปรากฏมาก่อน ขณะที่อีกต้นหนึ่งอาจเน้นที่จำนวนเงินที่สูงเกินค่าเฉลี่ยของผู้ใช้



ภาพประกอบ 2 แสดงขั้นตอนการทำงานของ Random Forest โดยรวมผลจากหลายต้นไม้ด้วยการโหวต

ที่มา : (Khan และคณะ, 2021)

จากงานวิจัยของ (Breiman, 2001) ได้กล่าวไว้ว่า Random Forest มีข้อดีคือความสามารถในการปรับตัวกับรูปแบบข้อมูลที่ซับซ้อนและลดโอกาสการเกิด overfitting เมื่อเทียบกับต้นไม้ตัดสินใจเดี่ยว ด้วยการรวมผลจากหลายโมเดลย่อยที่ทำให้การทำนายเสถียรขึ้น ขณะที่ข้อเสีย การตีความผลลัพธ์ทำได้ยาก เนื่องจากโมเดลสุดท้ายประกอบด้วยต้นไม้จำนวนมาก

นอกจากนี้ จากงานวิจัยของ (Aslam และ Hussain, 2024) ได้กล่าวไว้ว่า Random Forest ให้ความแม่นยำสูงและทนทานต่อ noise ในข้อมูล แต่ข้อเสียคือใช้ทรัพยากรในการคำนวณมากและเวลาในการฝึกนานกว่าโมเดลที่เรียบง่าย โมเดลนี้จึงเหมาะกับงานที่ต้องการความแม่นยำสูง แต่ไม่เหมาะหากต้องการความโปร่งใสหรือการตอบสนองที่รวดเร็ว

ด้วยคุณสมบัติดังกล่าว Random Forest จึงเหมาะสมอย่างยิ่งสำหรับการประยุกต์ใช้ในบริบทของการตรวจจับการฉ้อโกงทางการเงิน ซึ่งมักมีลักษณะข้อมูลที่มีจำนวนมาก มีความซับซ้อนสูง และไม่สมดุลระหว่างกลุ่มธุรกรรมปกติกับธุรกรรมฉ้อโกง ในงานวิจัยฉบับนี้ Random Forest ถูกนำมาประเมินร่วมกับโมเดลอื่น ได้แก่ XGBoost, LightGBM และ CatBoost เพื่อเปรียบเทียบประสิทธิภาพในการตรวจจับธุรกรรมฉ้อโกงภายใต้สถานการณ์ที่มีความไม่สมดุลของข้อมูล

2.2 Logistic Regression

เป็นเทคนิคการวิเคราะห์เชิงสถิติพื้นฐานที่ใช้กันอย่างแพร่หลายสำหรับการจำแนกแบบทวิภาค (binary classification) โดยมีเป้าหมายเพื่อทำนายความน่าจะเป็นของเหตุการณ์ เช่น การจำแนกธุรกรรมว่าเป็น “ปกติ” หรือ “ฉ้อโกง” โมเดลนี้ตั้งอยู่บนสมมติฐานว่ามีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอินพุตและ log-odds ของผลลัพธ์เป้าหมาย และใช้ฟังก์ชันซิกมอยด์ (sigmoid) เพื่อแปลงผลรวมเชิงเส้นของตัวแปรอินพุตให้อยู่ในช่วง [0, 1] สมการของฟังก์ชันซิกมอยด์แสดงดังสมการที่ (2) และ (3)

$$x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

โดยที่

x คือ ผลรวมเชิงเส้นของตัวแปรอินพุต

$x_1, x_2, \ldots, x_n$ คือ ตัวแปรอินพุต เช่น จำนวนเงินของธุรกรรม สถานที่ที่ทำรายการ เวลา หรือประวัติธุรกรรมก่อนหน้า

$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ คือ ค่าน้ำหนักของแต่ละตัวแปร (สัมประสิทธิ์)

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

โดยที่

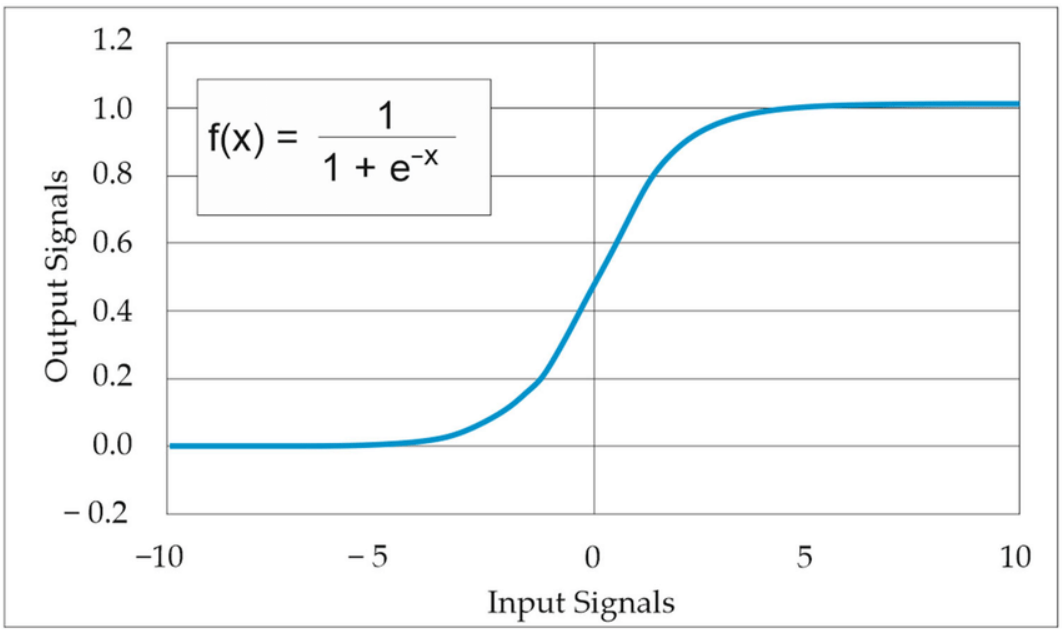
$f(x)$ คือ ความน่าจะเป็นที่มีค่าอยู่ในช่วง [0, 1]

(x) คือ ผลรวมเชิงเส้นจากสมการที่ 1

ผลลัพธ์ของ $f(x)$ ถูกตีความเป็น ความน่าจะเป็นของการเป็น Fraud ซึ่งสามารถใช้เกณฑ์ (threshold) เช่น 0.5 ในการตัดสินใจได้

หาก $f(x) > 0.5$ จัดเป็น "ฉ้อโกง" (Fraud)

หาก $f(x) \leq 0.5$ จัดเป็น "ปกติ" (Non-Fraud)



ภาพประกอบ 3 แสดงลักษณะของฟังก์ชันซิกมอยด์

ที่มา : (Gor และคณะ, 2022)

จากงานวิจัยของ (Dhawan, 2024) ได้กล่าวไว้ว่า Logistic Regression มีข้อดีคือความเรียบง่ายและตีความผลลัพธ์ได้ง่าย เนื่องจากน้ำหนักที่โมเดลเรียนรู้สามารถบ่งบอกถึงความสัมพันธ์ระหว่างตัวแปรแต่ละตัวกับโอกาสการเกิดการฉ้อโกงได้อย่างชัดเจน อีกทั้งใช้เวลาฝึกโมเดลไม่นานและไม่ต้องการทรัพยากรสูง ขณะที่ข้อเสีย โมเดลนี้มีข้อจำกัดในการจำแนกรูปแบบที่ซับซ้อน เนื่องจากถูกจำกัดด้วยสมมติฐานเชิงเส้น หากข้อมูลมีความสัมพันธ์ไม่เชิงเส้น โมเดลจะไม่มีประสิทธิภาพเพียงพอ

นอกจากนี้ จากงานวิจัยของ (Aslam และ Hussain, 2024) ได้กล่าวไว้ว่า Logistic Regression เหมาะเป็นโมเดล baseline และทำงานได้ดีในกรณีที่ข้อมูลแยกได้เชิงเส้น แต่หากไม่มีการควบคุม เช่น การใช้ regularization อาจเกิดการ overfitting ได้

โดยสรุป Logistic Regression เป็นตัวเลือกที่ดีสำหรับงานที่ต้องการความเรียบง่าย ความเร็ว และการตีความที่ชัดเจน แต่ไม่เหมาะกับข้อมูลที่ซับซ้อนหรือมีความสัมพันธ์ไม่เชิงเส้น ซึ่งจำกัดขอบเขตการใช้งานในบริบทการฉ้อโกงที่หลากหลาย

2.3 **XGBoost** (Extreme Gradient Boosting)

เป็นอัลกอริทึมในตระกูล gradient boosting ที่ออกแบบมาเพื่อเพิ่มประสิทธิภาพทั้งในด้านความเร็วและความแม่นยำ โดยสร้างโมเดลแบบอนุกรมผ่านการฝึกต้นไม้ตัดสินใจขนาดเล็ก (weak learners) จำนวนมากอย่างต่อเนื่อง แต่ละต้นไม้ถูกพัฒนาขึ้นเพื่อแก้ไขข้อผิดพลาดของการทำนายจากโมเดลก่อนหน้า โดยเริ่มจากการทำนายค่าเริ่มต้น (เช่น ค่าเฉลี่ยของผลลัพธ์) และปรับปรุงการทำนายในแต่ละรอบผ่านการคำนวณ residual ผลลัพธ์สุดท้ายของ XGBoost คือผลรวมของการทำนายจากทุกต้นไม้ใน ensemble ดังสมการที่ (4)

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{4}$$

โดยที่

$f_k(x)$ คือการทำนายจากต้นไม้ที่ k

K คือจำนวนต้นไม้ทั้งหมดในชุดเรียนรู้

XGBoost ปรับปรุงจาก gradient boosting ดั้งเดิมด้วยการเพิ่มเทคนิค regularization เช่น L1 (Lasso) และ L2 (Ridge) เพื่อควบคุมความซับซ้อนของโมเดลและป้องกันการ overfitting รวมถึงใช้ stochastic boosting โดยการสุ่มตัวอย่างข้อมูลบางส่วนในแต่ละรอบเพื่อเพิ่มความหลากหลายและลดความลำเอียง โมเดลนี้เหมาะกับข้อมูลธุรกรรมที่มี Historical ความซับซ้อนและมิติสูง เช่น การวิเคราะห์พฤติกรรมฉ้อโกงจากตัวแปรหลายตัวพร้อมกัน รวมถึงช่วงเวลาที่ทำรายการ สถานที่ที่ใช้บัตร จำนวนเงินที่ใช้จ่าย และประวัติการทำรายการของผู้ใช้ตัวอย่างการใช้งานอาจรวมถึงการตรวจจับธุรกรรมที่เกิดขึ้นในเวลาไม่ปกติ (เช่น ตี 3) และสถานที่ที่ห่างไกลจากตำแหน่งปกติของผู้ถือบัตร

จากงานวิจัยของ (Hajek และคณะ, 2022) ได้กล่าวไว้ว่า XGBoost มีข้อดีคือประสิทธิภาพในการทำงานและความแม่นยำสูง โดยเฉพาะเมื่อปรับจูนพารามิเตอร์ได้เหมาะสมทำให้เหมาะกับการตรวจจับการฉ้อโกงที่มีรูปแบบซับซ้อน ขณะที่ข้อเสีย การฝึกโมเดลใช้เวลานานและต้องใช้ทรัพยากรสูง รวมถึงการปรับพารามิเตอร์ที่ซับซ้อน

จากงานวิจัยของ Aslam และ Hussain (2024) ได้กล่าวไว้ว่า XGBoost ทนทานต่อ noise และจับความสัมพันธ์ไม่เชิงเส้นได้ดี แต่ข้อเสียคืออาจไม่เหมาะกับระบบเรียลไทม์ที่ต้องการความเร็วสูง

โดยสรุป XGBoost เป็นอัลกอริทึมที่เหมาะอย่างยิ่งกับงานที่ต้องการ ความแม่นยำสูง และสามารถจัดการกับข้อมูลที่ ซับซ้อนและมีหลายมิติ ได้ดี อย่างไรก็ตาม โมเดลนี้ต้องการทรัพยากรประมวลผลสูง และมี ความซับซ้อนในการปรับแต่งพารามิเตอร์ ซึ่งอาจเป็นข้อจำกัดในระบบที่ต้องการ ความเร็วหรือการอธิบายผลลัพธ์ได้ง่าย

2.4 LightGBM (Light Gradient Boosting Machine)

เป็นอัลกอริทึม gradient boosting ที่พัฒนาโดย Microsoft โดยมุ่งเน้นที่ความเร็วและประสิทธิภาพในการประมวลผล เพื่อรองรับชุดข้อมูลขนาดใหญ่และงานที่ต้องการการฝึกโมเดลอย่างรวดเร็ว LightGBM ใช้เทคนิค histogram-based เพื่อลดความละเอียดในการคำนวณจุดแยก (split points) โดยจัดกลุ่มข้อมูลเป็น bin แทนการพิจารณาค่าต่อเนื่องทั้งหมด และเติบโตแบบ leaf-wise โดยขยายเฉพาะใบที่มีการลดทอนความผิดพลาด (loss reduction) สูงสุดก่อน แทนการเติบโตแบบ level-wise ที่ขยายทุก node ในระดับเดียวกัน สมการการคำนวณ gain ที่ใช้เลือกการแยกแสดงดังสมการที่ (5)

$$[\text{Gain} = \frac{1}{2}\left[\frac{(\sum g_i)^2}{\sum h_i + \lambda}\right]_{\text{left}} + \left[\frac{(\sum g_i)^2}{\sum h_i + \lambda}\right]_{\text{right}} - \left[\frac{(\sum g_i)^2}{\sum h_i + \lambda}\right] \text{parent} \quad (5)$$

โดยที่

g_i และ h_i คือ gradient และ Hessian ที่วัดความผิดพลาดของข้อมูล

λ คือ พารามิเตอร์ regularization เพื่อป้องกันความซับซ้อนเกินจำเป็น

โมเดลนี้สามารถจัดการข้อมูลหมวดหมู่ได้ดีโดยไม่ต้องแปลงเป็นตัวเลขทั้งหมด และเหมาะกับระบบที่ต้องประมวลผลข้อมูลปริมาณมากแบบใกล้เรียลไทม์ เช่น การสกรีนธุรกรรมบัตรเครดิตนับล้านรายการต่อวัน โดยพิจารณาตัวแปร เช่น เวลาที่ทำรายการ ประเทศที่ใช้บัตร จำนวนเงิน หรือประเภทร้านค้า ตัวอย่างการใช้งานอาจรวมถึงการตรวจจับธุรกรรมที่มีความผิดปกติในมิติต่างๆ เช่น การใช้จ่ายสูงในประเทศที่ผู้ถือบัตรไม่เคยไปมาก่อน

จากงานวิจัยของ (Aslam และ Hussain, 2024) ได้กล่าวไว้ว่า LightGBM มีข้อดีคือความเร็วและประสิทธิภาพสูงในการฝึกโมเดล รองรับข้อมูลใหญ่และหมวดหมู่ได้ดี ทำให้เหมาะกับ

ระบบเรียลไทม์ ขณะที่ข้อเสีย มีความเสี่ยงต่อการ overfitting บนชุดข้อมูลขนาดเล็กหากไม่ควบคุมพารามิเตอร์ดีพอ

จากงานวิจัยของ (Dhawan, 2024)ได้กล่าวไว้ว่า LightGBM สเกลได้ดีและใช้หน่วยความจำน้อย แต่ข้อเสียคือการตีความผลลัพธ์ยังคงซับซ้อน โมเดลนี้จึงเหมาะกับงานที่ต้องการความเร็วและข้อมูลปริมาณมาก

โดยสรุป LightGBM เป็นอัลกอริทึม Gradient Boosting ที่โดดเด่นด้านความเร็วในการฝึกโมเดล ประสิทธิภาพในการจัดการข้อมูลขนาดใหญ่ และความสามารถในการรองรับข้อมูลหมวดหมู่โดยไม่ต้องแปลงล่วงหน้า เหมาะสำหรับงานที่ต้องการประมวลผลข้อมูลปริมาณมากแบบใกล้เรียลไทม์ เช่น การตรวจจับการฉ้อโกงในระบบการเงิน อย่างไรก็ตาม จุดอ่อนของ LightGBM คือมีความเสี่ยงต่อการเกิด overfitting เมื่อใช้กับชุดข้อมูลขนาดเล็ก และยังคงมีข้อจำกัดในด้านการตีความผลลัพธ์ที่ซับซ้อน ดังนั้น LightGBM จึงเหมาะสำหรับบริบทที่ต้องการความเร็วสูง ความแม่นยำดี และสามารถรับมือกับข้อมูลจำนวนมาก ได้อย่างมีประสิทธิภาพ

2.5 **CatBoost** (Categorical Boosting)

เป็นอัลกอริทึม gradient boosting ที่พัฒนาโดย Yandex โดยมีจุดเด่นในการจัดการข้อมูลหมวดหมู่ (categorical data) ได้อย่างมีประสิทธิภาพ และลดความยุ่งยากในการเตรียมข้อมูลก่อนฝึกโมเดล CatBoost ใช้เทคนิค ordered target encoding เพื่อเข้ารหัสตัวแปรหมวดหมู่โดยพิจารณาอินสแตนซ์ก่อนหน้าในการคำนวณค่าเฉลี่ย แทนการใช้ข้อมูลทั้งหมด ซึ่งช่วยป้องกันการรั่วไหลของข้อมูล (target leakage) และลด bias ที่อาจเกิดจากการเข้ารหัสแบบดั้งเดิม สมการการทำนายของ CatBoost คล้ายกับ XGBoost แต่มีการปรับปรุงวิธีการคำนวณ split ด้วย ordered boosting ดังสมการที่ (6)

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{6}$$

โดยที่

$f_k(x)$ คือ กระบวนการ Order boosting

โมเดลนี้เหมาะกับข้อมูลธุรกรรมที่มีตัวแปรหมวดหมู่จำนวนมาก เช่น ประเภทร้านค้าที่ใช้บัตร ประเทศที่ทำรายการ ช่องทางการชำระเงิน (ออนไลน์หรือหน้าร้าน) ประเภทบัตรหรือธนาคารผู้ออกบัตร โดยสามารถจับความสัมพันธ์ซับซ้อนระหว่างตัวแปรได้ดี ตัวอย่างเช่น การ

ตรวจจับธุรกรรมฉ้อโกงจากรูปแบบการใช้บัตรในสองประเทศที่ห่างไกลกันในช่วงเวลาใกล้เคียงกัน ซึ่งอาจบ่งบอกถึงการใช้งานที่ผิดปกติ

จากงานวิจัยของ (Dhawan, 2024) ได้กล่าวไว้ว่า CatBoost มีข้อดีคือการจัดการข้อมูลหมวดหมู่ได้ดีเยี่ยม ให้ความแม่นยำสูง และลดความจำเป็นในการเตรียมข้อมูล ขณะที่ข้อเสีย การฝึกโมเดลอาจช้ากว่า LightGBM และการตีความผลลัพธ์ยังคงยาก

จากงานวิจัยของ (Aslam และ Hussain, 2024) ได้กล่าวไว้ว่า CatBoost เหมาะกับข้อมูลที่มีตัวแปรหมวดหมู่จำนวนมากและจับความสัมพันธ์ซับซ้อนได้ดี แต่ข้อเสียคือใช้ทรัพยากรมากกว่าโมเดลที่เรียบง่าย โมเดลนี้จึงเหมาะกับงานที่เน้นความแม่นยำและมีข้อมูลหมวดหมู่เด่น

CatBoost เป็นอัลกอริทึม Gradient Boosting ที่เหมาะสำหรับข้อมูลที่มีตัวแปรหมวดหมู่จำนวนมาก โดยเฉพาะในงานตรวจจับการฉ้อโกงทางการเงิน ซึ่งข้อมูลมีลักษณะซับซ้อนและมีความหลากหลายของประเภทข้อมูล จุดแข็งของ CatBoost คือความสามารถในการจัดการข้อมูลเชิงหมวดหมู่ได้โดยไม่ต้องแปลงล่วงหน้า และการใช้เทคนิค Ordered Boosting ช่วยลดปัญหา target leakage และ bias จากการ encoding แบบทั่วไป

แม้ว่า CatBoost จะใช้ทรัพยากรมากและฝึกช้ากว่าโมเดลบางตัว เช่น LightGBM แต่ด้วยความแม่นยำสูงและความสามารถในการจับความสัมพันธ์เชิงซ้อน ทำให้เหมาะสำหรับงานที่ต้องการประสิทธิภาพสูงและมีความซับซ้อนของตัวแปร โดยเฉพาะในบริบทของการตรวจจับการฉ้อโกงในระบบการเงินที่ต้องวิเคราะห์ข้อมูลเชิงพฤติกรรมและประเภทธุรกรรมร่วมกันอย่างละเอียด

3. **การวัดประสิทธิภาพของแบบจำลอง** (Model Performance Metrics)

การวัดประสิทธิภาพของแบบจำลอง (Model Performance Metrics) เป็นกระบวนการสำคัญในการประเมินความสามารถของโมเดลในการทำงานกับข้อมูล โดยเฉพาะในงานที่เกี่ยวข้องกับการตรวจจับการฉ้อโกง ซึ่งมีลักษณะข้อมูลที่ไม่สมดุล และการประเมินผลที่แม่นยำเป็นสิ่งสำคัญในการตัดสินใจเชิงกลยุทธ์ บทนี้จะอธิบายเมตริกต่างๆ ที่ใช้ในการวัดประสิทธิภาพของโมเดลที่พัฒนาขึ้น โดยแบ่งออกเป็นกลุ่มหลัก ๆ ดังนี้:

3.1 Confusion Matrix

เป็นเครื่องมือที่ใช้ในการสรุปการทำนายผลลัพธ์ของโมเดลในแต่ละคลาส โดยจัดแสดงจำนวนการทำนายที่ถูกต้องและผิดพลาดในรูปแบบของเมทริกซ์ มีลักษณะดังนี้:

	ทำนายว่าเป็นการฉ้อโกง (Positive)	ทำนายว่าไม่เป็นการฉ้อโกง (Negative)
เป็นการฉ้อโกงจริง (Positive)	True Positive (TP)	False Negative (FN)
ไม่เป็นการฉ้อโกงจริง (Negative)	False Positive (FP)	True Negative (TN)

ภาพประกอบ 4 แสดงภาพตาราง Confusion Matrix

การวิเคราะห์ข้อมูลจาก Confusion Matrix สามารถให้ข้อมูลเชิงลึกเกี่ยวกับประสิทธิภาพของโมเดลได้หลากหลายมิติ โดยสามารถสรุปผลการประเมินได้ดังนี้

1. True Positive (TP): หมายถึงจำนวนตัวอย่างที่โมเดลทำนายว่าเป็นการฉ้อโกง (Positive) และเป็นการฉ้อโกงจริง ๆ ในข้อมูลทดสอบ ซึ่งแสดงถึงความสามารถของโมเดลในการตรวจจับการฉ้อโกงได้อย่างถูกต้อง ตัวอย่างเช่น ถ้าค่า TP สูง หมายความว่าโมเดลสามารถระบุการฉ้อโกงได้แม่นยำ

2. True Negative (TN): หมายถึงจำนวนตัวอย่างที่โมเดลทำนายว่าไม่เป็นการฉ้อโกง (Negative) และไม่เป็นการฉ้อโกงจริง ๆ ในข้อมูลทดสอบ แสดงถึงความสามารถในการปฏิเสธการฉ้อโกงที่แท้จริงได้อย่างถูกต้อง ถ้าค่า TN สูง แสดงว่าโมเดลสามารถหลีกเลี่ยงการทำนายผิดพลาดที่เกี่ยวข้องกับการระบุว่าการฉ้อโกงเกิดขึ้น ทั้งที่ในความเป็นจริงไม่มีการฉ้อโกง

3. False Positive (FP): คือจำนวนที่โมเดลทำนายว่ามีการฉ้อโกง แต่ในความเป็นจริงไม่มีการฉ้อโกง การทำนายที่ผิดพลาดในส่วนนี้จะก่อให้เกิดการแจ้งเตือนหรือการดำเนินการที่ไม่จำเป็น ค่า FP สูงหมายความว่าโมเดลมีแนวโน้มที่จะเกิดการแจ้งเตือนผิดพลาดบ่อย ซึ่งไม่ดีต่อระบบการตรวจจับการฉ้อโกง

4. False Negative (FN): คือจำนวนที่โมเดลทำนายว่าไม่มีการฉ้อโกง แต่ในความเป็นจริงมีการฉ้อโกง ข้อผิดพลาดในส่วนนี้อาจส่งผลเสียมากที่สุด เพราะหมายความว่าโมเดลไม่สามารถตรวจจับการฉ้อโกงได้ ค่า FN สูง แสดงว่าโมเดลไม่สามารถระบุการฉ้อโกงได้มากพอ ซึ่งเป็นผลเสียในกรณีการตรวจจับการฉ้อโกง

3.2 Accuracy (ความแม่นยำ)

Accuracy เป็นอัตราส่วนของจำนวนการคาดการณ์ที่ถูกต้องเมื่อเปรียบเทียบกับจำนวนการคาดการณ์ทั้งหมด แสดงได้ดังสมการที่ (7)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

โดยที่

TP คือ True Positives (ตรวจจับการฉ้อโกงได้ถูกต้อง)

TN คือ True Negatives (ไม่ตรวจจับการฉ้อโกงได้ถูกต้อง)

FP คือ False Positives (ตรวจจับการฉ้อโกงแต่ไม่ใช่การฉ้อโกง)

FN คือ False Negatives (ไม่ตรวจจับการฉ้อโกงที่เป็นการฉ้อโกงจริง)

แม้ว่า accuracy จะเป็นเมตริกที่แสดงผลลัพธ์รวมของโมเดลได้ดี แต่ในกรณีที่ข้อมูลมีความไม่สมดุล (ข้อมูลฉ้อโกงมีจำนวนน้อยกว่าข้อมูลปกติ) ค่าของ accuracy อาจทำให้เข้าใจผิด เพราะโมเดลอาจจะทำนายถูกต้องในข้อมูลส่วนใหญ่ที่ไม่ใช่ฉ้อโกง แต่ไม่สามารถตรวจจับการฉ้อโกงได้ดี

3.3 Precision (ความแม่นยำของการทำนาย)

Precision เป็นอัตราส่วนของการทำนายที่ถูกต้องในกลุ่มที่โมเดลทำนายว่าเป็นการฉ้อโกง แสดงได้ดังสมการที่ (8)

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Precision ช่วยประเมินว่าเมื่อโมเดลทำนายว่ามีการฉ้อโกง การทำนายนั้นมีความแม่นยำเพียงใด หากค่า precision สูง หมายความว่าจำนวนการทำนายผิดพลาด (False Positives) น้อย

3.4 Recall (ความไว)

Recall หรือ Sensitivity คืออัตราส่วนของการทำนายที่ถูกต้องในกลุ่มที่มีการฉ้อโกงทั้งหมด แสดงได้ดังสมการที่ (9)

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Recall แสดงให้เห็นว่าโมเดลสามารถตรวจจับการฉ้อโกงได้ครบถ้วนเพียงใด หาก recall สูง หมายความว่าโมเดลสามารถตรวจจับการฉ้อโกงได้มาก

3.5 F1 Score

F1 Score เป็นการวัดค่าเฉลี่ยเชิงกลีบระหว่าง Precision และ Recall โดยช่วยให้เห็นถึงความสมดุลระหว่างทั้งสองเมตริก ซึ่งเหมาะสำหรับกรณีที่ข้อมูลมีความไม่สมดุล แสดงได้ดังสมการที่ (10)

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

F1 Score จะมีค่ามากเมื่อทั้ง Precision และ Recall มีค่าสูง ดังนั้นจึงเป็นเมตริกที่เหมาะกับการวัดความสามารถของโมเดลที่ต้องการทั้งความแม่นยำและความครอบคลุมในการตรวจจับการฉ้อโกง

3.6 ROC Curve และ AUC

ROC Curve (Receiver Operating Characteristic Curve) เป็นกราฟที่แสดงความสัมพันธ์ระหว่าง True Positive Rate (Recall) กับ False Positive Rate โดยสามารถใช้วัดความสามารถของโมเดลในการแยกแยะระหว่างคลาสต่าง ๆ ได้อย่างชัดเจน สูตรคำนวณสำหรับค่าใน ROC Curve แสดงได้ดังสมการที่ (11) และ (12)

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

3.7 AUC (Area Under the Curve)

เป็นค่าที่วัดพื้นที่ใต้กราฟ ROC Curve โดยค่า AUC ที่ใกล้เคียง 1 แสดงว่าโมเดลมีประสิทธิภาพสูงในการจำแนกประเภทของข้อมูล ค่าที่ใกล้ 0.5 หมายถึงโมเดลทำงานได้ใกล้เคียงกับการทำนายแบบสุ่ม แสดงได้ดังสมการที่ (13)

$$\text{AUC} = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \times \frac{(TPR_{i+1} + TPR_i)}{2} \quad (13)$$

โดยที่

Auc คือ ค่าพื้นที่ใต้กราฟ ROC Curve

FPR_i,TPR_i คือ ค่าของ False Positive Rate และ True Positive Rate ที่จุดi

n คือ จำนวนจุดที่ใช้คำนวณใน ROC Curve

ในบริบทของการตรวจจับการฉ้อโกงทางการเงิน ค่า AUC ที่สูงแสดงถึงความสามารถของโมเดลในการแยกแยะธุรกรรมปกติ ออกจากธุรกรรมฉ้อโกงได้อย่างมีประสิทธิภาพ

การวัดประสิทธิภาพของแบบจำลองเป็นขั้นตอนที่สำคัญในการประเมินความสามารถของโมเดล โดยเมตริกต่าง ๆ เช่น Confusion Matrix, Accuracy, Precision, Recall, F1 Score, ROC Curve และ AUC จะช่วยให้เรามองเห็นภาพรวมและประสิทธิภาพในการทำงานของโมเดล นอกจากนี้ เมตริกเหล่านี้ยังมีบทบาทสำคัญในการเลือกโมเดลที่เหมาะสมสำหรับงานที่มีความไม่สมดุลของข้อมูล ในกรณีของการตรวจจับการฉ้อโกงทางการเงิน การเลือกเมตริกที่เหมาะสมเป็นเรื่องสำคัญ เนื่องจากธุรกรรมฉ้อโกงมักมีสัดส่วนน้อย ทำให้เมตริกบางตัว เช่น Accuracy อาจไม่เหมาะสม เพราะโมเดลที่ทำนายว่าทุกธุรกรรมเป็นธุรกรรมปกติก็จะมี Accuracy สูง แต่ไม่สามารถตรวจจับการฉ้อโกงได้เลย

สำหรับปัญหาประเภทนี้ เมตริกอย่าง Recall (ความไว) มีความสำคัญมาก เพราะวัดความสามารถในการตรวจจับธุรกรรมฉ้อโกงได้ทั้งหมด ขณะที่ Precision จะบอกว่าในจำนวนธุรกรรมที่โมเดลทำนายว่าเป็นการฉ้อโกง มีความถูกต้องมากน้อยเพียงใด F1 Score เป็นค่าเฉลี่ยฮาร์โมนิกระหว่าง Precision และ Recall จึงให้ภาพที่สมดุลกว่า

4. เทคนิคการจัดการข้อมูลที่ไม่สมดุลในบริบทของการตรวจจับการฉ้อโกง

การจัดการกับข้อมูลที่ไม่สมดุล (Imbalanced Data) เป็นประเด็นสำคัญในงานวิจัยด้านการตรวจจับการฉ้อโกง (Fraud Detection) เนื่องจากการฉ้อโกงมักจะเกิดขึ้นในอัตราที่ต่ำเมื่อเทียบกับธุรกรรมที่ปกติ ส่งผลให้การสร้างโมเดลเพื่อทำนายการฉ้อโกงกลายเป็นเรื่องท้าทาย เนื่องจากโมเดลอาจให้ความสำคัญกับกลุ่มธุรกรรมปกติ (ที่มีจำนวนมากกว่า) มากเกินไป จนทำให้ไม่สามารถตรวจจับการฉ้อโกงได้ดีพอ เทคนิคที่ใช้ในการจัดการกับข้อมูลที่ไม่สมดุลในบริบทนี้สามารถแบ่งออกเป็นหลายแนวทาง ดังนี้

เทคนิคการจัดการข้อมูลไม่สมดุลมีอยู่หลายวิธีและแบ่งออกเป็น 2 กลุ่มใหญ่ คือ การสุ่มข้อมูลแบบลดจำนวน (Undersampling) และการสุ่มข้อมูลแบบเพิ่มจำนวน (Oversampling)

4.1 การใช้โมเดลพื้นฐานโดยไม่มีการปรับแต่งพารามิเตอร์ (Baseline model)

โมเดลพื้นฐาน (Baseline model) เป็นโมเดลที่ใช้การตั้งค่าพารามิเตอร์เริ่มต้น (default parameters) โดยไม่มีการปรับแต่งใดๆ เพิ่มเติม โมเดลนี้ถูกใช้เป็นจุดอ้างอิงเริ่มต้นในการวัดประสิทธิภาพ ก่อนที่จะมีการปรับปรุงหรือพัฒนาโมเดลให้ซับซ้อนมากขึ้น

ในบริบทของการตรวจจับธุรกรรมฉ้อโกง การสร้างโมเดลพื้นฐานมีความสำคัญ เพราะช่วยให้เข้าใจว่าโมเดลที่ไม่ได้รับการปรับแต่งมีประสิทธิภาพในการจำแนกธุรกรรมปกติและธุรกรรมฉ้อโกงอย่างไร ซึ่งให้ข้อมูลพื้นฐานที่สำคัญเกี่ยวกับลักษณะของข้อมูลและความยากง่ายในการจำแนกคลาส

การสร้างโมเดลพื้นฐานมักจะเริ่มจากการนำข้อมูลที่ผ่านการเตรียมพร้อมแล้วมาใช้ฝึกโมเดลด้วยอัลกอริทึมการเรียนรู้ของเครื่อง เช่น Random Forest, Logistic Regression หรือ Gradient Boosting โดยใช้ค่าพารามิเตอร์เริ่มต้นที่กำหนดไว้ในไลบรารี จากนั้นจึงประเมินประสิทธิภาพของโมเดลด้วยเมทริกต่างๆ เช่น ความแม่นยำ (Accuracy), ความไว (Recall), ความเฉพาะเจาะจง (Precision), F1-Score หรือ AUC-ROC

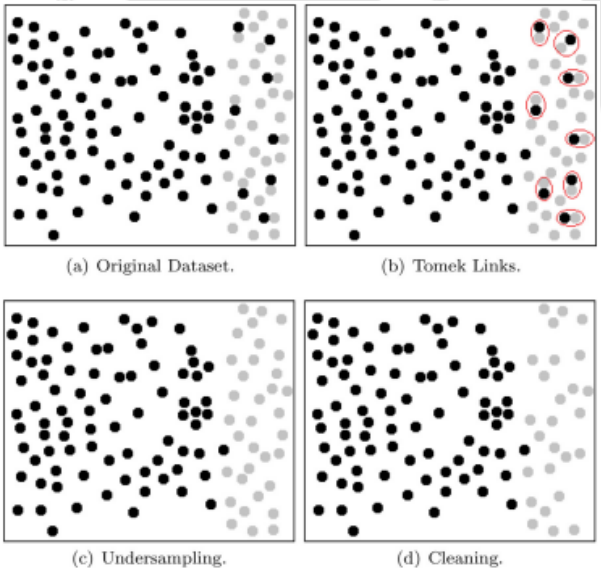
ข้อดีของการสร้างโมเดลพื้นฐานคือ ทำให้เห็นประสิทธิภาพขั้นต่ำของโมเดลที่สามารถปรับปรุงต่อไปได้ ใช้เป็นเกณฑ์เปรียบเทียบกับโมเดลที่ซับซ้อนกว่าเพื่อให้แน่ใจว่าการเพิ่มความซับซ้อนนั้นคุ้มค่า ช่วยระบุปัญหาในข้อมูลตั้งแต่เริ่มต้น และประหยัดเวลาในการทดลองเมื่อเทียบกับการเริ่มด้วยโมเดลที่ซับซ้อน

หลังจากได้ผลลัพธ์จากโมเดลพื้นฐานแล้ว จึงค่อยพิจารณาการปรับปรุงโมเดลด้วยวิธีต่างๆ เช่น การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning), การใช้เทคนิคการจัดการข้อมูลไม่สมดุล หรือการเลือกฟีเจอร์ที่สำคัญ เพื่อพัฒนาประสิทธิภาพของโมเดลให้ดียิ่งขึ้น

4.2 **การสุ่มข้อมูลแบบลดจำนวน** (Under sampling)

Tomek Links เป็นเทคนิค Undersampling ที่ใช้จัดการกับข้อมูลที่ไม่สมดุลในการตรวจจับธุรกรรมฉ้อโกง โดยมีหลักการคือการลดจำนวนตัวอย่างจากคลาสที่มีมาก (ธุรกรรมปกติ) เพื่อให้สมดุลกับคลาสที่มีน้อย (ธุรกรรมฉ้อโกง) เทคนิคนี้ทำงานโดยค้นหาคู่ข้อมูลที่อยู่ใกล้กันที่สุดแต่อยู่คนละคลาส เรียกว่า Tomek Link แล้วลบข้อมูลจากคลาสที่มีมากกว่า

กระบวนการนี้ช่วยปรับปรุงเส้นแบ่ง (boundary) ระหว่างคลาส ทำให้โมเดลแยกแยะธุรกรรมฉ้อโกงได้แม่นยำขึ้น และยังช่วยลดสัญญาณรบกวน (noise) ที่อาจทำให้โมเดลสับสน อย่างไรก็ตาม เทคนิคนี้มีข้อเสียคือใช้เวลาในการคำนวณมากกว่าการทำ Undersampling แบบสุ่ม และอาจทำให้สูญเสียข้อมูลสำคัญในคลาสที่มีจำนวนมากได้ ทั้งนี้ การลดจำนวนข้อมูลแม้จะช่วยปรับสมดุล แต่ก็อาจส่งผลให้สูญเสียข้อมูลที่อาจเป็นประโยชน์ต่อการเรียนรู้ของโมเดล



ภาพประกอบ 5 ตัวอย่างการทำงานของ Tomek Links ในการลดข้อมูลที่ไม่สมดุล

ที่มา:(Pereira และคณะ, 2020)

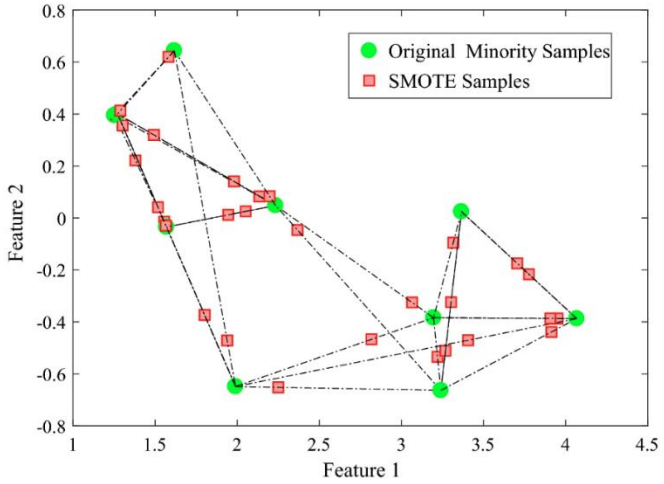
4.3 **การสุ่มข้อมูลแบบเพิ่มจำนวน** (Over sampling)

SMOTE (Synthetic Minority Over-sampling Technique) เป็นเทคนิค Oversampling ที่ใช้เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล โดยการเพิ่มจำนวนตัวอย่างในคลาสที่มีจำนวนน้อย (ธุรกรรมฉ้อโกง) เพื่อให้สมดุลกับคลาสที่มีจำนวนมาก (ธุรกรรมปกติ) แทนที่จะทำการคัดลอกข้อมูลซ้ำแบบง่ายๆ เทคนิคนี้สร้างข้อมูลสังเคราะห์ใหม่โดยใช้การคำนวณเชิงเวกเตอร์ระหว่างตัวอย่างจริงในคลาสที่มีจำนวนน้อย

กระบวนการทำงานเริ่มจากการเลือกตัวอย่างจากคลาสที่มีจำนวนน้อย แล้วหาเพื่อนบ้านที่ใกล้ที่สุด (k-nearest neighbors) จากนั้นสร้างข้อมูลใหม่โดยการคำนวณระหว่างตัวอย่างจริงและเพื่อนบ้านที่ใกล้เคียง ทำให้ได้ข้อมูลที่มีความหลากหลายมากกว่าการคัดลอกข้อมูลเดิม

ข้อดีของเทคนิคนี้คือช่วยเพิ่มความหลากหลายของข้อมูล ลดโอกาสการเกิด overfitting ที่มักพบในการทำ Random Oversampling และเพิ่มความสมดุลของข้อมูลทำให้โมเดลเรียนรู้ได้ดีขึ้น อย่างไรก็ตาม SMOTE อาจทำให้เส้นแบ่ง (boundary) ระหว่างคลาสซับซ้อนขึ้น และเพิ่มความซับซ้อนในการคำนวณ

เมื่อใช้ Tomek Links ร่วมกับ SMOTE จะช่วยลดปัญหาความไม่สมดุลของข้อมูลได้อย่างมีประสิทธิภาพ โดย Tomek Links ช่วยทำให้เส้นแบ่งระหว่างคลาสชัดเจนขึ้น ในขณะที่ SMOTE ช่วยเพิ่มความหลากหลายของข้อมูลในคลาสที่มีจำนวนน้อย



ภาพประกอบ 6 ตัวอย่างการทำงานของ SMOTE ในการเพิ่มข้อมูลของกลุ่ม Minority

ที่มา: (Elreedy และคณะ, 2023)

4.4 การใช้เทคนิคผสมผสาน (Hybrid sampling)

SMOTETomek เป็นเทคนิคผสมผสาน (Hybrid sampling) ที่รวมข้อดีของทั้ง SMOTE (Synthetic Minority Over-sampling Technique) และ Tomek Links เข้าด้วยกันเพื่อจัดการกับข้อมูลที่ไม่สมดุลในการตรวจจับธุรกรรมฉ้อโกง

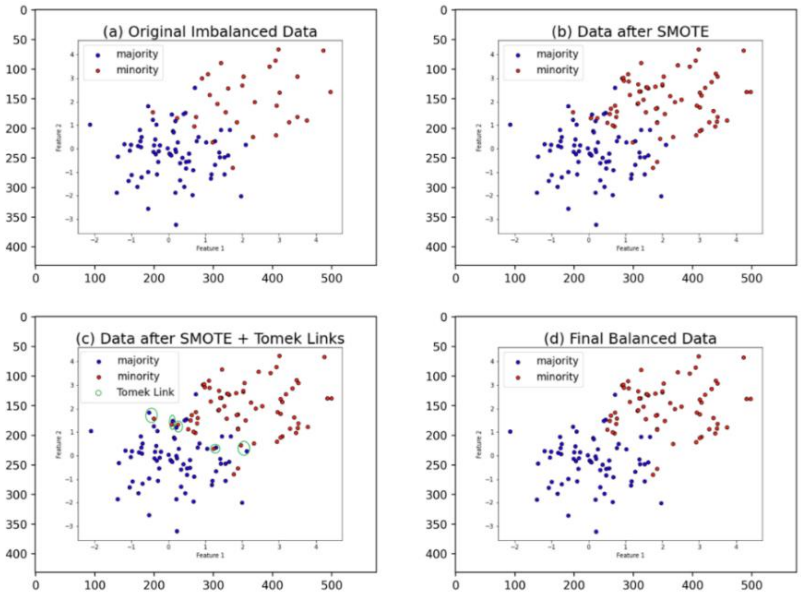
กระบวนการทำงานของ SMOTETomek แบ่งเป็นสองขั้นตอนหลัก ได้แก่:

1. ขั้นตอนการทำ SMOTE: เริ่มด้วยการสร้างข้อมูลสังเคราะห์ใหม่ในคลาสที่มีจำนวนน้อย (ธุรกรรมฉ้อโกง) โดยการคำนวณเชิงเวกเตอร์ระหว่างข้อมูลที่มีอยู่ในคลาสนั้น ทำให้เพิ่มจำนวนตัวอย่างในคลาสที่มีน้อยและสร้างความหลากหลายของข้อมูล

2. ขั้นตอนการทำ Tomek Links: หลังจากเพิ่มข้อมูลด้วย SMOTE แล้ว จะใช้ Tomek Links เพื่อระบุและกำจัดข้อมูลที่อาจทำให้เกิดความสับสนบริเวณเส้นแบ่งระหว่างคลาส โดยค้นหาคู่ข้อมูลที่อยู่ใกล้กันที่สุดแต่อยู่คนละคลาสแล้วลบข้อมูลจากคลาสที่มีจำนวนมากกว่า (ธุรกรรมปกติ)

การใช้ SMOTETomek ในการตรวจจับธุรกรรมฉ้อโกงมีข้อดีคือช่วยปรับปรุงเส้นแบ่งระหว่างคลาสให้ชัดเจนขึ้น ลดความเสี่ยงของการ overfitting ที่อาจเกิดจากการใช้ SMOTE เพียงอย่างเดียว และยังช่วยเพิ่มประสิทธิภาพของโมเดลในการตรวจจับธุรกรรมฉ้อโกง โดยเฉพาะในสถานการณ์ที่ข้อมูลมีความไม่สมดุลอย่างมาก

ด้วยการผสมผสานทั้งสองเทคนิคเข้าด้วยกัน SMOTETomek จึงเป็นทางเลือกที่มีประสิทธิภาพในการเตรียมข้อมูลก่อนนำไปฝึกโมเดลเพื่อตรวจจับธุรกรรมฉ้อโกง



ภาพประกอบ 7 ตัวอย่างกระบวนการปรับสมดุลข้อมูลโดยใช้ SMOTE และ Tomek Links

ที่มา: (Wang และคณะ, 2024)

4.5 **การปรับน้ำหนักของคลาส** (Adjust Class-weight)

การปรับน้ำหนักของคลาส (Class-Weight) เป็นเทคนิคสำคัญในการจัดการข้อมูลที่ไม่สมดุล โดยเฉพาะในกรณีการตรวจจับธุรกรรมฉ้อโกง วิธีนี้ทำงานโดยการให้ความสำคัญกับคลาสที่มีจำนวนน้อยกว่า (ธุรกรรมฉ้อโกง) มากขึ้นในการคำนวณฟังก์ชันการสูญเสีย (Loss Function) ขณะที่ลดน้ำหนักของคลาสที่มีจำนวนมากกว่า (ธุรกรรมปกติ)

เมื่อข้อมูลมีความไม่สมดุล โมเดลการเรียนรู้ของเครื่องจะเอนเอียง (Bias) ไปทางคลาสที่มีจำนวนมากกว่า เนื่องจากการทำนายคลาสนั้นบ่อยๆ จะช่วยลดค่าความผิดพลาดโดยรวมได้ง่ายกว่า ส่งผลให้โมเดลมักละเลยคลาสส่วนน้อย การปรับน้ำหนักของคลาสจึงเข้ามาแก้ไขปัญหานี้โดยเพิ่มบทลงโทษ (Penalty) สำหรับการทำนายผิดในคลาสส่วนน้อย

ข้อดีของเทคนิคนี้คือไม่ต้องเปลี่ยนแปลงข้อมูลต้นฉบับ เพียงแค่ปรับวิธีการเรียนรู้ของโมเดล ทำให้รักษาข้อมูลทั้งหมดไว้โดยไม่ต้องเพิ่มหรือลดข้อมูล นอกจากนี้ ยังสามารถปรับน้ำหนักได้อย่างยืดหยุ่นตามความสำคัญของแต่ละคลาส และสามารถใช้ร่วมกับเทคนิคอื่นๆ เช่น Undersampling หรือ Oversampling ได้

อย่างไรก็ตาม การปรับน้ำหนักของคลาสต้องระมัดระวังการให้น้ำหนักที่มากเกินไปกับคลาสส่วนน้อย เพราะอาจทำให้โมเดลมีแนวโน้มทำนายคลาสส่วนน้อยมากเกินไป (Over-prediction) ซึ่งอาจส่งผลให้เกิดจำนวนการเตือนผิดพลาด (False Alarm) สูงในกรณีตรวจจับธุรกรรมฉ้อโกง

5. งานวิจัยที่เกี่ยวข้องกับการตรวจจับการฉ้อโกงในธุรกรรมการเงิน

5.1 บทความวิจัยเรื่อง Credit Card Fraud Detection using Machine Learning Algorithms

งานวิจัยเรื่อง Credit Card Fraud Detection using Machine Learning Algorithms โดย Vaishnavi Nath Dornadula และ Geetha S.(Dornadula และ Geetha, 2019) ได้นำเสนอวิธีการตรวจจับการฉ้อโกงในธุรกรรมบัตรเครดิตโดยใช้เทคนิคการเรียนรู้ของเครื่องหลายประเภท รวมถึงการใช้ Model Random Forest, Decision Trees, Naive Bayes, Logistic Regression, และ SVM (Support Vector Machine) เพื่อเปรียบเทียบประสิทธิภาพของแต่ละโมเดล งานวิจัยนี้มีวัตถุประสงค์ในการออกแบบและพัฒนาวิธีการตรวจจับการฉ้อโกงที่มีประสิทธิภาพสำหรับข้อมูลธุรกรรมแบบสตรีมมิ่ง โดยมุ่งเน้นการวิเคราะห์รายละเอียดธุรกรรมในอดีตของลูกค้าเพื่อนำมาสกัดรูปแบบพฤติกรรมที่เป็นเอกลักษณ์ของผู้ถือบัตร โดยใช้เทคนิค Sliding Window เพื่อรวบรวมข้อมูลธุรกรรมและ SMOTE (Synthetic Minority Over-sampling Technique) เพื่อจัดการกับปัญหาความไม่สมดุลในข้อมูล

ผลการวิจัยพบว่า Random Forest และ Decision Trees ให้ผลลัพธ์ที่ดีที่สุด โดยเฉพาะหลังจากใช้ SMOTE ในการแก้ปัญหาความไม่สมดุลของข้อมูล โดยมีค่า MCC (Matthews Correlation Coefficient) สูงถึง 0.9996 เทคนิคที่ใช้ Sliding Window Technique: ใช้ในการรวบรวมข้อมูลธุรกรรมและวิเคราะห์พฤติกรรมของผู้ถือบัตร SMOTE: ใช้เพื่อเพิ่มจำนวนตัวอย่างของธุรกรรมที่เป็นการฉ้อโกงในชุดข้อมูล Feedback Mechanism: ใช้ในการจัดการกับปัญหา concept drift ซึ่งหมายถึงการเปลี่ยนแปลงของพฤติกรรมการใช้บัตรเครดิตในระยะยาว โดยรวมแล้ว งานวิจัยนี้แสดงให้เห็นถึงความสำคัญของการใช้เทคนิคการเรียนรู้ของเครื่องในการตรวจจับการฉ้อโกงในธุรกรรมบัตรเครดิต และชี้ให้เห็นถึงความท้าทายในการจัดการกับข้อมูลที่ไม่สมดุลและการเปลี่ยนแปลงของพฤติกรรมการใช้บัตรเครดิตในระยะยาว

5.2 บทความวิจัยเรื่อง Credit Card Fraud Detection using Machine Learning Techniques: A Comparative Analysis

งานวิจัยเรื่อง Credit Card Fraud Detection using Machine Learning Techniques: A Comparative Analysis โดย Awoyemi, Adetunmbi, และ Oluwadare (Awoyemi และคณะ, 2017)ศึกษาประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่องสามแบบ ได้แก่ Naïve Bayes, k-Nearest Neighbor (kNN), และ Logistic Regression สำหรับการตรวจจับธุรกรรมฉ้อโกงบัตรเครดิต

วัตถุประสงค์ งานวิจัยนี้มีเป้าหมายเพื่อเปรียบเทียบประสิทธิภาพของโมเดลทั้งสามบนข้อมูลธุรกรรมบัตรเครดิตที่ไม่สมดุลอย่างมาก และตรวจสอบผลกระทบของการสุ่มตัวอย่างแบบผสมต่อประสิทธิภาพการตรวจจับข้อมูลและวิธีการ ใช้ข้อมูลจริงจากธุรกรรมบัตรเครดิตของชาวยุโรป จำนวน 284,807 รายการ โดยมีธุรกรรมฉ้อโกงเพียง 0.172% ใช้ PCA ในการคัดเลือกคุณลักษณะ ใช้การสุ่มตัวอย่างแบบผสม (Hybrid Sampling) โดยการ Oversampling คลาสส่วนน้อย (ธุรกรรมฉ้อโกง) และ Undersampling คลาสส่วนมาก (ธุรกรรมปกติ) เพื่อสร้างชุดข้อมูลสองชุดคือ 10:90 และ 34:66

ผลลัพธ์ kNN แสดงผลลัพธ์ที่ดีที่สุด โดยมี Accuracy 97.92%, Precision และ Specificity 1.0 สำหรับทั้งสองชุดข้อมูล Naïve Bayes มีประสิทธิภาพดีรองลงมา โดยมี Accuracy 97.69% สำหรับชุดข้อมูล 34:66 Logistic Regression มีประสิทธิภาพน้อยที่สุด โดยมี Accuracy เพียง 54.86% สำหรับชุดข้อมูล 34:66

การประเมินผล ใช้หลายเมตริกในการประเมิน ได้แก่ Accuracy, Sensitivity, Specificity, Precision, Matthews Correlation Coefficient (MCC) และ Balanced Classification Rate โดย kNN มี MCC สูงสุดที่ +0.9535 สำหรับชุดข้อมูล 34:66

โดยสรุป kNN เป็นโมเดลที่มีประสิทธิภาพที่สุดในการตรวจจับการฉ้อโกงในชุดข้อมูลที่มีความไม่สมดุล การสุ่มตัวอย่างแบบผสมช่วยปรับปรุงประสิทธิภาพของโมเดลอย่างมีนัยสำคัญ โดยเฉพาะสำหรับชุดข้อมูล 34:66 Logistic Regression มีประสิทธิภาพน้อยที่สุดในการตรวจจับธุรกรรมฉ้อโกงงานวิจัยในอนาคต ผู้วิจัยเสนอแนะให้ศึกษาเพิ่มเติมเกี่ยวกับ meta-classifiers และ meta-learning approaches สำหรับการจัดการกับข้อมูลที่ไม่สมดุลอย่างมาก รวมถึงการตรวจสอบผลกระทบของวิธีการสุ่มตัวอย่างแบบอื่นๆ

5.3 บทความวิจัยเรื่อง Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets

งานวิจัยเรื่อง Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets โดย Vahid Sinap (Sinap, 2024) ได้นำเสนอการเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องหลายประเภทในการตรวจจับการฉ้อโกงบัตรเครดิต รวมถึง Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors (KNN), และ Support Vector Machine (SVM)

โดยงานวิจัยนี้มุ่งเน้นไปที่ปัญหาข้อมูลที่ไม่สมดุลซึ่งเกิดจากการที่ธุรกรรมฉ้อโกงมีจำนวนน้อยกว่าธุรกรรมปกติมาก วัตถุประสงค์ของการศึกษา งานวิจัยนี้มีเป้าหมายเพื่อประเมินและเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องในการตรวจจับการฉ้อโกง โดยวัดผลผ่านตัวชี้วัดต่าง ๆ เช่น Accuracy, Precision, Recall, F1-Score, AUC, และ AUPRC รวมถึงการใช้ ROC curves และ confusion matrices เพื่อประเมินความสามารถของโมเดล

ผลลัพธ์ที่ได้ โมเดลที่ดีที่สุด: Random Forest (RF) และ K-Nearest Neighbors (KNN) เป็นโมเดลที่มีประสิทธิภาพดีที่สุด โดยมีค่า Accuracy สูงถึง 97% คะแนน AUC-ROC และ AUPRC: RF แสดงผลลัพธ์ที่ดีที่สุดในแง่ของ AUC-ROC และ AUPRC ด้วยคะแนน 0.97 และ 0.98 ตามลำดับ ในขณะที่ KNN ก็มีคะแนนที่ใกล้เคียงกัน ปัญหาที่พบ Naive Bayes มีประสิทธิภาพที่ต่ำที่สุด โดยมี Accuracy เพียง 94% และมีข้อจำกัดในการจัดการกับข้อมูลที่ไม่สมดุล เทคนิคที่ใช้ Random Under-Sampling: ใช้เพื่อปรับสมดุลข้อมูลธุรกรรมที่ไม่สมดุล Dimensionality Reduction และ Clustering: ใช้เทคนิค t-SNE เพื่อลดมิติของข้อมูลและทำให้เห็นกลุ่มของธุรกรรมฉ้อโกงและปกติได้ชัดเจนยิ่งขึ้น

โดยสรุป งานวิจัยนี้แสดงให้เห็นว่า Random Forest และ KNN เป็นโมเดลที่มีประสิทธิภาพสูงที่สุดในการตรวจจับการฉ้อโกงบัตรเครดิตในชุดข้อมูลที่ไม่สมดุล ในขณะที่ Naive Bayes มีประสิทธิภาพน้อยกว่า

5.4 บทความวิจัยเรื่อง Credit card fraud detection using the brown bear optimization algorithm

งานวิจัยเรื่อง Credit card fraud detection using the brown bear optimization algorithm โดย Shaymaa E. Sorour และคณะ (Sorour และคณะ, 2024) ได้เสนอวิธีการตรวจจับ

การฉ้อโกงบัตรเครดิตโดยใช้ Brown Bear Optimization (BBO) ซึ่งเป็นอัลกอริทึมที่ได้รับแรงบันดาลใจจากพฤติกรรมของหมีสีน้ำตาล ผสมผสานกับการเลือกคุณลักษณะ (Feature Selection, FS) เพื่อเพิ่มประสิทธิภาพในการตรวจจับการฉ้อโกง โดยใช้ชุดข้อมูลเครดิตการ์ดจากประเทศออสเตรเลีย วัตถุประสงค์ของการศึกษา งานวิจัยนี้มุ่งเน้นการพัฒนาอัลกอริทึม BBO และสร้างเป็นเวอร์ชันไบนารีที่เรียกว่า Binary BBO (BBBO) เพื่อใช้ในการเลือกคุณลักษณะ (Feature Selection) ที่เกี่ยวข้องกับการตรวจจับการฉ้อโกง และเปรียบเทียบประสิทธิภาพของ BBBO กับอัลกอริทึมการเพิ่มประสิทธิภาพอื่น ๆ เช่น Binary Particle Swarm Optimization (BPSO) และ Binary Grasshopper Optimization Algorithm (BGOA)

ผลลัพธ์ที่ได้ โมเดล BBBO สามารถลดจำนวนคุณลักษณะที่ต้องใช้ได้ถึง 67% ในขณะที่ยังคงความแม่นยำสูงถึง 91% BBBO ได้รับการพิสูจน์ว่าเหนือกว่าอัลกอริทึมการเพิ่มประสิทธิภาพอื่น ๆ ในหลายตัวชี้วัด เช่น ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าการเรียกกลับ (Recall), ค่าความเที่ยงตรง (F1-Score), พื้นที่ใต้โค้ง ROC (AUC), และ Kappa เทคนิคที่ใช้ Feature Selection (FS) ใช้ FS เพื่อลดมิติของข้อมูลและกำจัดคุณลักษณะที่ไม่จำเป็น Meta-heuristic algorithms: เปรียบเทียบกับอัลกอริทึมอื่น ๆ รวมถึง BPSO, BGOA, Binary Harris Hawks Optimization (BHHO), และ Binary African Vultures Optimization (BAVO) การเรียนรู้ของเครื่อง (Machine Learning) ใช้ SVM, k-NN และ Xgb-tree เป็นตัวจำแนก (classifiers) ในการตรวจจับการฉ้อโกง ข้อสรุป BBBO แสดงให้เห็นถึงความสามารถในการลดจำนวนคุณลักษณะโดยไม่ลดประสิทธิภาพในการตรวจจับการฉ้อโกง โดยมีผลลัพธ์ที่ดีกว่าอัลกอริทึมการเพิ่มประสิทธิภาพอื่น ๆ ที่ใช้อยู่

5.5 **บทความวิจัยเรื่อง** Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique

งานวิจัยเรื่อง Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique โดย Huajie Huang และคณะ (Huang และคณะ, 2024) ได้นำเสนอวิธีการใหม่ในการตรวจจับการฉ้อโกงบัตรเครดิตโดยใช้เทคนิคการเรียนรู้ของเครื่องและการจัดการกับปัญหาข้อมูลไม่สมดุล

วัตถุประสงค์ของการศึกษา เพื่อพัฒนาวิธีการตรวจจับการฉ้อโกงบัตรเครดิตที่มีประสิทธิภาพโดยใช้ข้อมูลธุรกรรมและข้อมูลส่วนตัวของผู้ถือบัตร แก้ปัญหาข้อมูลไม่สมดุล

ระหว่างธุรกรรมปกติและธุรกรรมฉ้อโกง ทดสอบประสิทธิภาพของวิธีการที่นำเสนอโดยใช้ข้อมูลจริงจากธนาคารในช่วงการระบาดของ COVID-19 วิธีการที่นำเสนอ Hybrid Neural Network (HNN): ใช้ Convolutional Neural Network (CNN) สำหรับข้อมูลธุรกรรม และ Back Propagation Neural Network (BPNN) สำหรับข้อมูลส่วนตัวของผู้ถือบัตร Clustering-based Undersampling (CBU): ใช้ K-means clustering เพื่อลดจำนวนข้อมูลธุรกรรมปกติให้สมดุลกับข้อมูลธุรกรรมฉ้อโกง ผลลัพธ์ที่ได้ HNN-CUHIT มีประสิทธิภาพดีที่สุดในการตรวจจับการฉ้อโกง โดยมีค่า F1-score สูงสุดที่ 0.0572 อัตราส่วนที่เหมาะสมระหว่างธุรกรรมปกติและธุรกรรมฉ้อโกง คือ 1:1 การเพิ่มคุณลักษณะใหม่ 5 ประการและข้อมูลส่วนตัวของผู้ถือบัตรช่วยเพิ่มประสิทธิภาพของโมเดล

โดยสรุป งานวิจัยนี้แสดงให้เห็นว่า HNN-CUHIT มีประสิทธิภาพสูงในการตรวจจับการฉ้อโกงบัตรเครดิตและสามารถจัดการกับปัญหาข้อมูลไม่สมดุลได้ดี โดยการใช้ข้อมูลทั้งจากธุรกรรมและข้อมูลส่วนตัวของผู้ถือบัตร ร่วมกับเทคนิค clustering-based undersampling ช่วยเพิ่มประสิทธิภาพในการตรวจจับการฉ้อโกงได้อย่างมีนัยสำคัญ

5.6 บทความวิจัยเรื่อง Digital payment fraud detection methods in digital ages and Industry 4.0

งานวิจัยเรื่อง Digital payment fraud detection methods in digital ages and Industry 4.0 โดย Güner, S. S., Çelik, Y., & Sönmez, Y. (2022) (Chang และคณะ, 2022) ศึกษาการตรวจจับการฉ้อโกงในระบบชำระเงินดิจิทัลภายใต้บริบทของยุค Industry 4.0 ซึ่งแม้ระบบการเงินจะพัฒนาอย่างรวดเร็ว แต่กลับมีความเสี่ยงสูงต่อการถูกโจมตีทางไซเบอร์และการฉ้อโกงทางธุรกรรม

วัตถุประสงค์หลักของงานคือการเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง ได้แก่ Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree (DT) และ Random Forest (RF) บนชุดข้อมูลธุรกรรมจริงที่มีความไม่สมดุลสูง (ธุรกรรมฉ้อโกงเพียง 1.21%) โดยนำเทคนิคจัดการข้อมูลไม่สมดุล เช่น SMOTE (Oversampling), NearMiss (Undersampling) และ PCA (ลดมิติข้อมูล) มาใช้ร่วมกับการสร้างโมเดล

ผลการทดลองพบว่า Random Forest ให้ผลลัพธ์ดีที่สุด โดยเฉพาะเมื่อใช้ร่วมกับ SMOTE และ PCA ซึ่งให้ค่า AUROC สูงถึง 0.976 และ Precision 0.928 รองลงมาคือ Logistic

Regression ซึ่งให้ค่า AUROC 0.971 และ Precision 0.927 เมื่อใช้ร่วมกับ PCA ส่วน KNN และ Decision Tree มีผลลัพธ์ต่ำกว่า โดย DT ให้ค่า AUROC ต่ำสุดที่ 0.902

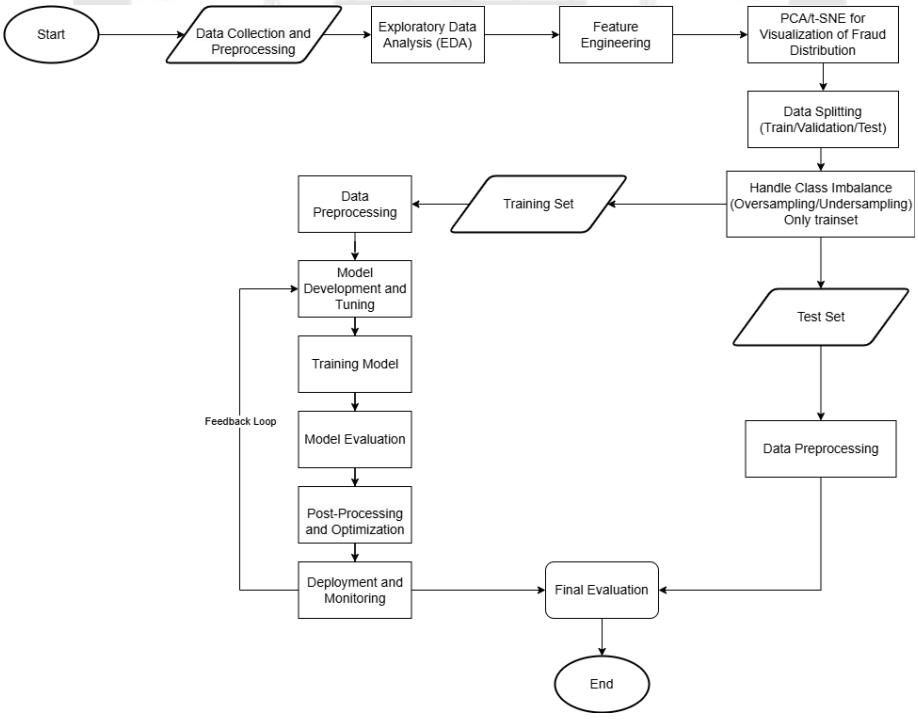
การประเมินประสิทธิภาพโมเดลใช้เมตริกหลัก ได้แก่ AUROC, Precision, Recall และ F1 Score โดยพบว่าการใช้ PCA ช่วยเพิ่มประสิทธิภาพของทุกโมเดลได้อย่างชัดเจน ทั้งนี้ งานวิจัยเสนอแนะให้ศึกษาแนวทางขั้นสูงเพิ่มเติม เช่น Meta-Learning ในอนาคต เพื่อรับมือกับปัญหาข้อมูลไม่สมดุลอย่างมีประสิทธิภาพยิ่งขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลการตรวจจับธุรกรรมที่มีความเสี่ยงต่อการฉ้อโกงในระบบการเงิน โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) กระบวนการวิจัยได้รับการออกแบบอย่างเป็นระบบเพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพสูงสุดในการตรวจจับการฉ้อโกง โดยมีขั้นตอนหลักดังนี้:

1. ภาพรวมกระบวนการวิจัย
2. การเก็บรวบรวมข้อมูล (Data Collection)
3. การสำรวจข้อมูล (Exploratory Data Analysis: EDA)
4. การเตรียมข้อมูล (Data Preparation)
5. การสร้างแบบจำลองพยากรณ์ (Model Development)

1. ภาพรวมของกระบวนการวิจัย



ภาพประกอบ 8 แสดงกระบวนการทำงานของแบบจำลอง

ภาพนี้แสดงให้เห็นถึงขั้นตอนการทำงานสำหรับการพัฒนาระบบตรวจจับการฉ้อโกงทางการเงินโดยใช้การเรียนรู้ของเครื่อง (Machine Learning) โดยสามารถอธิบายแต่ละขั้นตอนอย่างละเอียดได้ดังนี้:

1.1 **การเก็บรวบรวมและเตรียมข้อมูล** (Data Collection and Preprocessing)

ขั้นตอนแรกในงานวิจัยนี้เริ่มจากการเก็บรวบรวมข้อมูลจากชุดข้อมูลการทำธุรกรรมทางการเงินที่ได้มาจาก Kaggle ซึ่งประกอบไปด้วยข้อมูลธุรกรรมจำนวน 594,643 รายการ โดยในจำนวนนั้น 98.79% เป็นธุรกรรมปกติและ 1.21% เป็นธุรกรรมที่เป็นการฉ้อโกง เนื่องจากข้อมูลที่ได้รับอาจมีข้อมูลสูญหายหรือข้อมูลผิดพลาด จึงต้องทำการเตรียมข้อมูล (Preprocessing) เช่น การจัดการข้อมูลที่หายไป (Missing Values) การแก้ไขข้อมูลที่ผิดพลาด (Data Cleaning) และการเข้ารหัสข้อมูล (Encoding) เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับการเรียนรู้ของเครื่อง

1.2 **การวิเคราะห์ข้อมูลเบื้องต้น** (Exploratory Data Analysis: EDA)

หลังจากการเตรียมข้อมูลเบื้องต้น ผู้วิจัยได้ทำการวิเคราะห์เชิงสำรวจเพื่อทำความเข้าใจลักษณะการกระจายตัวของข้อมูล เช่น การสำรวจค่าเฉลี่ย ความแปรปรวน และค่าผิดปกติของธุรกรรม รวมทั้งวิเคราะห์ความสัมพันธ์ของคุณลักษณะต่าง ๆ เพื่อใช้เป็นข้อมูลสนับสนุนในการตัดสินใจสร้างฟีเจอร์ใหม่ (Feature Engineering)

1.3 **การสร้างคุณลักษณะใหม่** (Feature Engineering)

ผู้วิจัยดำเนินการสร้างตัวแปรเพิ่มเติม เช่น จำนวนธุรกรรมรวมที่เกี่ยวข้องกับลูกค้า ร้านค้า หมวดหมู่สินค้า และช่วงของจำนวนเงิน (customer_total_trans, merchant_total_trans, category_total_trans, และ amount_thresh_total_trans) และสร้างตัวแปรเชิงกลุ่ม (Categorical Variables) เพื่อช่วยให้โมเดลสามารถระบุและแยกแยะรูปแบบธุรกรรมฉ้อโกงได้ดียิ่งขึ้น

1.4 **การลดมิติข้อมูล** (Dimensionality Reduction) **ด้วย** PCA **และ** t-SNE

ในขั้นตอนนี้ ผู้วิจัยใช้เทคนิคการลดมิติข้อมูล ได้แก่ Principal Component Analysis (PCA) และ t-distributed Stochastic Neighbor Embedding (t-SNE) เพื่อวิเคราะห์เชิงสำรวจลักษณะการกระจายของข้อมูล โดยมีวัตถุประสงค์หลักเพื่อทำความเข้าใจ pattern ของข้อมูล โดยเฉพาะการแยกกลุ่มระหว่างธุรกรรมปกติ (non-fraud) และธุรกรรมฉ้อโกง (fraud)

เทคนิค PCA ช่วยลดจำนวนมิติของข้อมูลลง โดยยังคงรักษาโครงสร้างสำคัญไว้ได้ในรูปขององค์ประกอบหลัก (Principal Components) ที่อธิบายความแปรปรวนของข้อมูล ส่วน t-SNE เป็นเทคนิคแบบไม่เชิงเส้นที่เน้นการคงไว้ซึ่งความสัมพันธ์ของจุดข้อมูลที่อยู่ใกล้กันในมิติสูงให้อยู่ใกล้กันในมิติต่ำ ซึ่งเหมาะสำหรับการแสดงผลลักษณะกลุ่ม (clusters) ที่ซับซ้อน การใช้ PCA และ t-SNE ในการวิจัยนี้มีวัตถุประสงค์เพื่อ แสดงการกระจายของข้อมูลในเชิงภาพ (visualization) โดยไม่ได้ถูกนำไปใช้โดยตรงในกระบวนการฝึกโมเดล ทั้งนี้เพื่อสนับสนุนการทำ Feature Engineering และการวางแนวทางในการเลือกโมเดลและวิธีการจัดการข้อมูลในขั้นตอนต่อไป

1.5 **การแบ่งข้อมูล** (Data Splitting)

ในขั้นตอนนี้ ผู้วิจัยแบ่งชุดข้อมูลออกเป็นสองส่วนหลักอย่างชัดเจน ได้แก่ ชุดข้อมูลสำหรับการฝึกโมเดล (Training Set) และชุดข้อมูลสำหรับการทดสอบประสิทธิภาพ (Test Set) โดยมีสัดส่วนของข้อมูลชุดฝึกเท่ากับ 70% และชุดทดสอบเท่ากับ 30% นอกจากนี้ ผู้วิจัยยังใช้วิธี Stratified Sampling เพื่อให้แน่ใจว่าสัดส่วนระหว่างธุรกรรมปกติและธุรกรรมฉ้อโกงในแต่ละชุดข้อมูลมีความสมดุลเหมือนข้อมูลต้นฉบับ ทั้งนี้ เพื่อเพิ่มความน่าเชื่อถือในการประเมินประสิทธิภาพโมเดล และป้องกันปัญหาการ Overfitting ผู้วิจัยได้นำเทคนิค Cross Validation แบบ 5-Fold มาใช้กับชุดข้อมูลสำหรับการฝึกโมเดล (Training Set) โดยวิธีนี้จะช่วยประเมินประสิทธิภาพโมเดลอย่างรอบด้านและมีความน่าเชื่อถือมากยิ่งขึ้น

1.6 **การจัดการข้อมูลไม่สมดุล** (Handling Class Imbalance)

ผู้วิจัยได้จัดการปัญหาข้อมูลที่มีความไม่สมดุลด้วยเทคนิคต่าง ๆ ได้แก่ การใช้ Oversampling (SMOTE) เพิ่มจำนวนธุรกรรมฉ้อโกง, การใช้ Undersampling (Tomek Links) เพื่อลดข้อมูลธุรกรรมปกติที่ซ้ำซ้อน รวมถึงการปรับ Class Weight ของโมเดล เพื่อให้การเรียนรู้มีความสมดุลระหว่างกลุ่มข้อมูลมากขึ้น

1.7 **การพัฒนาและปรับแต่งแบบจำลอง** (Model Development and Tuning)

หลังจากเตรียมข้อมูลเรียบร้อยแล้ว ผู้วิจัยได้สร้างโมเดลการเรียนรู้ของเครื่อง 5 วิธี ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression โดยแต่ละโมเดลถูกพัฒนาด้วยวิธีการที่แตกต่างกัน ได้แก่: Baseline Model ใช้ข้อมูลต้นฉบับโดยไม่มีการจัดการความไม่สมดุลของข้อมูล Adjust Class Weight ปรับน้ำหนักของคลาสในโมเดล (Class Weight) Undersampling ใช้วิธี Tomek Links เพื่อลดข้อมูลส่วนใหญ่ Oversampling: ใช้วิธี

SMOTE เพื่อเพิ่มข้อมูลส่วนน้อย Hybrid Sampling: ผสมระหว่าง Oversampling และ Undersampling นอกจากนี้ ผู้วิจัยยังทำการปรับแต่งพารามิเตอร์ของแต่ละโมเดลผ่านวิธี Grid Search เพื่อค้นหาพารามิเตอร์ที่ดีที่สุด

1.8 การประเมินประสิทธิภาพของโมเดล (Model Evaluation)

ผู้วิจัยใช้วิธีการ Cross Validation (5-Fold) ในขั้นตอนการฝึกโมเดลเพื่อประเมินประสิทธิภาพของโมเดลโดยใช้เกณฑ์ AUC-ROC, Precision, Recall, และ F1-Score เพื่อวัดประสิทธิภาพในการตรวจจับธุรกรรมฉ้อโกงในสถานการณ์ที่สมจริงที่สุด

1.9 การวิเคราะห์ความสำคัญของตัวแปรด้วย SHAP Analysis

หลังจากที่ได้โมเดลที่ดีที่สุดแล้ว ผู้วิจัยได้นำเทคนิค SHAP (SHapley Additive exPlanations) มาใช้เพื่อวิเคราะห์และอธิบายว่าแต่ละตัวแปรส่งผลต่อการทำนายของโมเดลอย่างไร ซึ่งช่วยเพิ่มความเข้าใจเชิงลึกของการทำงานของโมเดล และความสำคัญสัมพัทธ์ของแต่ละฟีเจอร์ที่มีต่อการตรวจจับธุรกรรมฉ้อโกง

1.10 การปรับปรุงและเพิ่มประสิทธิภาพ (Post-Processing and Optimization)

โมเดลที่ได้รับการประเมินแล้วจะถูกปรับแต่งเพิ่มเติม เช่น การปรับ Threshold ในการจำแนกธุรกรรมให้เหมาะสมที่สุด เพื่อเพิ่มประสิทธิภาพการใช้งานจริง โดย Threshold คือค่าขีดแบ่งความน่าจะเป็นที่ใช้ตัดสินว่าธุรกรรมเป็นฉ้อโกงหรือไม่ ซึ่งการปรับค่านี้จะช่วยให้สามารถควบคุมสมดุลระหว่างการตรวจจับการฉ้อโกงและการลดการแจ้งเตือนผิดพลาดได้

1.11 การนำไปใช้และการติดตามผล (Deployment and Monitoring)

โมเดลที่มีประสิทธิภาพสูงสุดจะถูกนำไปใช้งานในสภาพแวดล้อมจริง และผู้วิจัยจะดำเนินการติดตามและประเมินผลระยะยาว เพื่อให้แน่ใจว่าโมเดลยังสามารถทำงานได้อย่างมีประสิทธิภาพเมื่อเจอสถานการณ์ที่เปลี่ยนแปลงไปในอนาคต

1.12 การประเมินผลขั้นสุดท้าย (Final Evaluation)

สุดท้าย ผู้วิจัยจะสรุปผลประสิทธิภาพของโมเดลหลังจากที่ได้นำไปใช้งานจริง เพื่อให้มั่นใจว่าโมเดลมีความแม่นยำในการตรวจจับธุรกรรมฉ้อโกงและเป็นประโยชน์ต่อการใช้งานจริงในระบบต่อไป

แผนภาพประกอบที่ 7 แสดงกระบวนการพัฒนาโมเดลการเรียนรู้ของเครื่องสำหรับการตรวจจับการฉ้อโกง ครอบคลุมทุกขั้นตอนตั้งแต่การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล

เบื้องต้น การจัดการกับความไม่สมดุลของข้อมูล การพัฒนาโมเดล ไปจนถึงการนำไปประยุกต์ใช้ในสภาพแวดล้อมจริง

2. **การเก็บรวบรวมข้อมูล** (Data Collection)

ขั้นตอนแรกของการวิจัยคือการรวบรวมข้อมูลธุรกรรมทางการเงินจากแหล่งที่เชื่อถือได้ โดยได้นำชุดข้อมูลจาก Kaggle (https://www.kaggle.com/datasets/ealaxi/banksim1) มาใช้ในการพัฒนาระบบตรวจจับการฉ้อโกง ชุดข้อมูลนี้ประกอบด้วยธุรกรรมทั้งสิ้น 594,643 รายการที่มีคุณลักษณะสำคัญ 10 ประการ ลักษณะเด่นของข้อมูลคือมีความไม่สมดุลสูง โดยธุรกรรมปกติมีจำนวน 587,443 รายการ (98.79%) ขณะที่ธุรกรรมฉ้อโกงมีเพียง 7,200 รายการ (1.21%)

ตาราง 1 แสดงรายละเอียดตัวแปรของข้อมูลธุรกรรมสำหรับการวิเคราะห์การฉ้อโกง

ชื่อฟีเจอร์	ประเภทข้อมูล	คำอธิบายข้อมูล
step	int64	วันที่เกิดธุรกรรม มีทั้งหมด 180 วัน
customer	object	รหัสลูกค้า: ขึ้นต้นด้วย "C" ตามด้วยเลข 10 หลัก รวม 4,109 รายใน dataset
age	int64	กลุ่มอายุของลูกค้าแบ่งออกเป็น 7 ประเภท ได้แก่ 0: อายุน้อยกว่า 18 ปี, 1: อายุระหว่าง 19 ถึง 25 ปี, 2: อายุระหว่าง 26 ถึง 35 ปี, 3: อายุระหว่าง 36 ถึง 45 ปี, 4: อายุระหว่าง 46 ถึง 55 ปี, 5: อายุระหว่าง 56 ถึง 65 ปี, 6: อายุมากกว่า 65 ปี, U: ไม่ทราบอายุ
gender	object	เพศของลูกค้า F: เพศหญิง M: เพศชาย E: Enterprise U: ไม่ทราบเพศ
zipcodeOri	string	รหัสไปรษณีย์ของต้นทาง (ที่อยู่ของลูกค้า)

ชื่อฟีเจอร์	ประเภทข้อมูล	คำอธิบายข้อมูล
merchant	object	รหัสของร้านค้า (ID)
zipMerchant	string	รหัสไปรษณีย์ของร้านค้า
category	object	หมวดหมู่ของสินค้า
amount	float64	จำนวนเงินในการทำธุรกรรม
fraud	int64	ตัวแปรเป้าหมายที่แสดงว่าธุรกรรมเป็นการฉ้อโกง (1) หรือไม่ฉ้อโกง (0)

3. การสำรวจข้อมูล (Exploratory Data Analysis: EDA)

ขั้นตอนถัดมาคือการทำการสำรวจและวิเคราะห์ข้อมูลเบื้องต้น (Exploratory Data Analysis: EDA) เพื่อทำความเข้าใจโครงสร้างของข้อมูล การกระจายตัวของธุรกรรมในแต่ละประเภท และตรวจสอบคุณลักษณะใดที่อาจเกี่ยวข้องกับการฉ้อโกง เช่น ความสัมพันธ์ระหว่าง amount และ fraud หรือการวิเคราะห์ค่าผิดปกติ (Outliers) ในข้อมูล การตรวจพบความไม่สมดุลของข้อมูล (Imbalanced Data) ซึ่งธุรกรรมปกติมีจำนวนมากกว่าธุรกรรมฉ้อโกงอย่างชัดเจน เป็นปัจจัยสำคัญที่ต้องนำมาพิจารณา เพื่อเลือกใช้เทคนิคที่เหมาะสมในการแก้ปัญหา เช่น การปรับสมดุลข้อมูล และเตรียมพร้อมสำหรับการพัฒนาโมเดลในขั้นตอนถัดไป

3.1 การตรวจสอบคุณภาพข้อมูล

การตรวจสอบคุณภาพข้อมูลเป็นกระบวนการสำคัญในการเตรียมข้อมูล (Data Preparation) เพื่อให้มั่นใจว่าข้อมูลมีความถูกต้อง สมบูรณ์ และเหมาะสมต่อการนำไปใช้งานในการวิเคราะห์หรือสร้างโมเดล โดยในหัวข้อนี้จะครอบคลุมการตรวจสอบในหลายมิติ เช่น ความครบถ้วน ความสม่ำเสมอ ความสมเหตุสมผล และความถูกต้องของข้อมูล

ตาราง 3 สรุปการตรวจสอบคุณภาพของข้อมูล

หัวข้อการตรวจสอบ	รายละเอียดการตรวจสอบ	ผลลัพท์
Missing Data	ตรวจสอบว่ามีคอลัมน์ใดที่มีค่าหายไปหรือไม่	ไม่มีค่าว่าง (Missing Data = 0%)
Duplicate Data	ตรวจสอบว่ามีแถวที่ซ้ำซ้อนหรือไม่	ไม่มีข้อมูลที่ซ้ำกัน

หัวข้อการตรวจสอบ	รายละเอียดการตรวจสอบ	ผลลัพท์
		(Duplicates = 0 แถว)
Outliers	คำนวณค่าผิดปกติในคอลัมน์ amount โดยใช้ IQR (Interquartile Range)	พบ Outliers จำนวน 25,781 รายการ (ค่าต่ำกว่า -29.46 และสูงกว่า 85.74)
ความสมเหตุสมผลของข้อมูล (Consistency)	ตรวจสอบอายุ (age) ที่ควรเป็นตัวเลขและต้องไม่ติดลบ	พบว่า age มีค่าที่เป็นข้อความ และบางรายการอาจไม่สมเหตุสมผล เช่น อายุที่ติดลบ
ความสม่ำเสมอของข้อมูล (Uniformity)	ตรวจสอบคอลัมน์ข้อความ เช่น customer, gender และ category ให้มีรูปแบบเดียวกัน (ตัวพิมพ์เล็ก)	ได้แปลงข้อความทั้งหมดให้เป็น ตัวพิมพ์เล็ก (lowercase) เรียบร้อยแล้ว
ค่าที่ไม่มีความหมาย (Irrelevant Data)	ตรวจสอบค่าที่ไม่มีความหมาย เช่น "unknown" หรือ "-"	ไม่พบค่าที่ไม่มีความหมายในชุดข้อมูล
การกระจายตัวของข้อมูล (Distribution)	ตรวจสอบการกระจายตัวของคอลัมน์ตัวเลข เช่น amount และ step โดยใช้ Histogram และ Boxplot	amount มีการกระจายตัวไม่สมดุล โดยพบค่าผิดปกติสูงสุดที่ 8,329.96

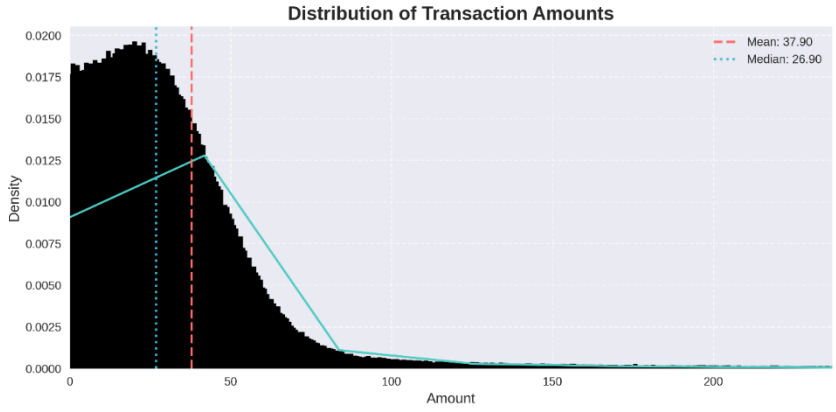
การตรวจสอบคุณภาพข้อมูลเป็นขั้นตอนสำคัญในกระบวนการวิจัย เพื่อให้มั่นใจว่าข้อมูลที่ใช้มีความถูกต้องและสมบูรณ์ ลดความคลาดเคลื่อนที่อาจเกิดขึ้น และเพิ่มความน่าเชื่อถือของผลการวิเคราะห์และการสร้างโมเดล

3.2 การวิเคราะห์ข้อมูลเชิงสำรวจและความสัมพันธ์ของตัวแปรกับการฉ้อโกง

การวิเคราะห์ข้อมูลเชิงสำรวจเป็นขั้นตอนสำคัญในการทำความเข้าใจรูปแบบข้อมูลและปัจจัยที่สัมพันธ์กับการฉ้อโกงทางการเงิน ผู้วิจัยได้ทำการวิเคราะห์ความสัมพันธ์ระหว่างมูลค่าธุรกรรม ลักษณะของลูกค้า และประเภทของธุรกรรม กับความเสี่ยงต่อการเกิดการฉ้อโกง เพื่อนำไปสู่การพัฒนาโมเดลที่มีประสิทธิภาพต่อไป

3.2.1 การวิเคราะห์การกระจายตัวของจำนวนเงิน

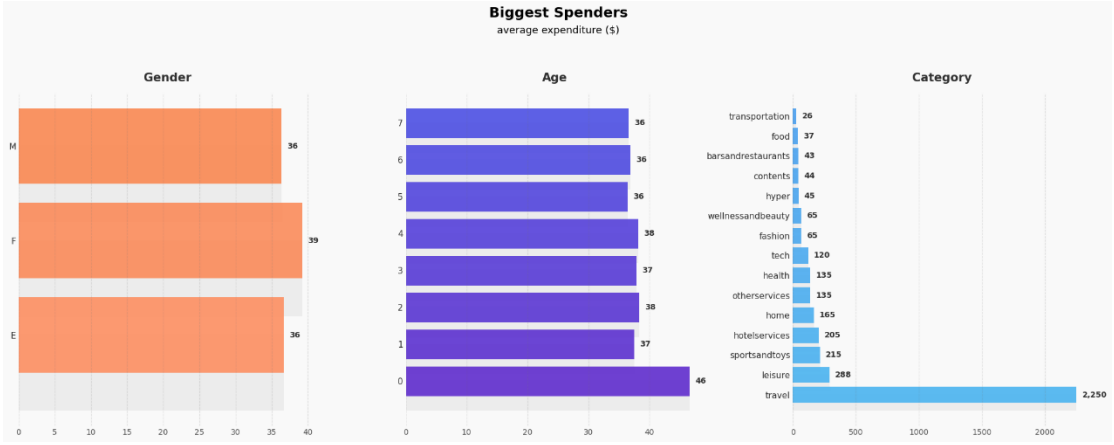
จาก Histogram แสดงให้เห็นว่าจำนวนเงินส่วนใหญ่อยู่ในช่วงต่ำกว่า 1,000 หน่วย ในขณะที่ค่าเฉลี่ย (Mean: 37.90) และมัธยฐาน (Median: 26.90) สะท้อนถึงการกระจายที่เบ้ขวา (Right-skewed) อย่างชัดเจน นอกจากนี้ยังพบ Outliers ในช่วงที่สูงกว่า 10,000 หน่วย ซึ่งอาจบ่งบอกถึงความผิดปกติของธุรกรรมบางรายการ



ภาพประกอบ 9 แสดงการกระจายตัวของจำนวนเงินในการทำธุรกรรม

3.2.2 การวิเคราะห์พฤติกรรมการใช้จ่ายตามปัจจัยทางประชากรศาสตร์

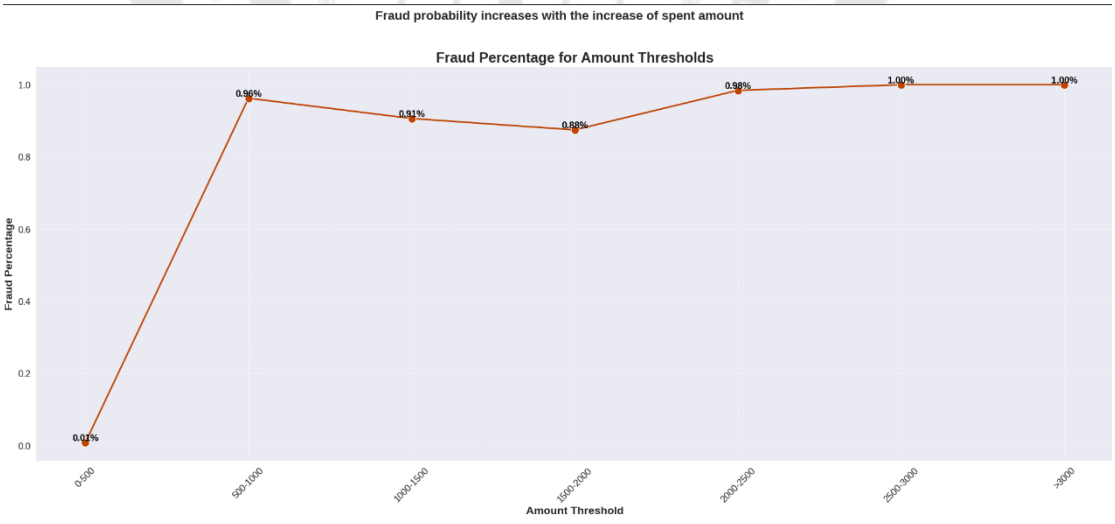
แผนภาพแสดงค่าใช้จ่ายเฉลี่ย (average expenditure) จำแนกตาม 3 ปัจจัย พบว่าเพศหญิงมีค่าใช้จ่ายสูงสุด (39 หน่วย), กลุ่มอายุน้อยกว่า 18 ปี มีการใช้จ่ายมากที่สุด (46 หน่วย), และหมวดท่องเที่ยว (travel) มีค่าใช้จ่ายสูงที่สุดอย่างเห็นได้ชัด (2,250) ตามด้วยสันทนาการ (leisure) และกีฬา (sports&toys) ภาพนี้ใช้แสดงกลุ่มที่มีการใช้จ่ายสูงสุด (Biggest Spenders) โดยเปรียบเทียบค่าใช้จ่ายเฉลี่ยตามเพศ อายุ และหมวดหมู่สินค้า/บริการ เพื่อวิเคราะห์พฤติกรรมการใช้จ่ายในแต่ละกลุ่ม ซึ่งช่วยในการระบุกลุ่มที่มีความเสี่ยงต่อการฉ้อโกงได้



ภาพประกอบ 10 แสดงการวิเคราะห์การใช้จ่ายโดยเฉลี่ย

3.2.3 ความสัมพันธ์ระหว่างมูลค่าธุรกรรมกับโอกาสการเกิดฉ้อโกง

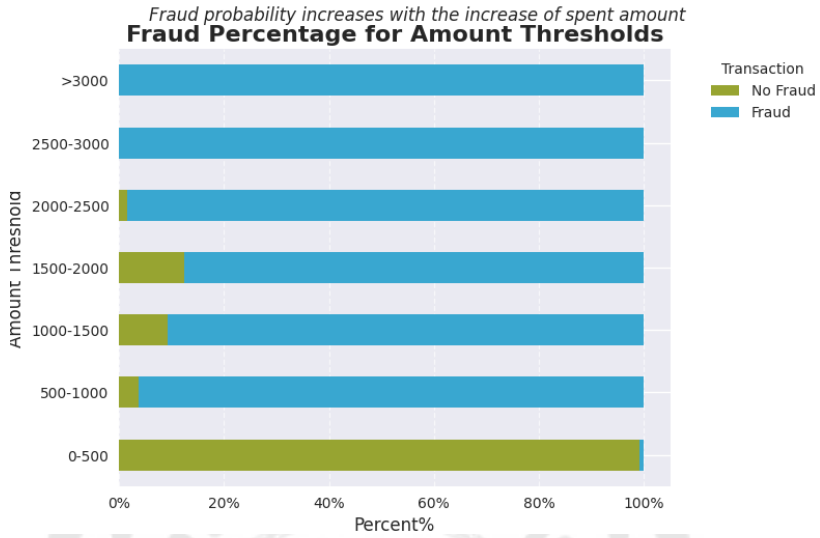
โอกาสการเกิด fraud เพิ่มขึ้นอย่างชัดเจนตามจำนวนเงินที่ใช้จ่ายสูงขึ้น โดยเฉพาะในช่วง 500-1,000 เป็นต้นไป ซึ่งมีอัตราการฉ้อโกงเกือบถึง 1% และคงที่เมื่อจำนวนเงินเกิน 2,000 ขึ้นไป



ภาพประกอบ 11 แสดงความสัมพันธ์ระหว่าง ช่วงจำนวนเงินที่ใช้จ่าย (Amount Threshold) กับเปอร์เซ็นต์การฉ้อโกง (Fraud Percentage)

ในแต่ละช่วงจำนวนเงิน (Amount Threshold) โดยพบว่าเมื่อจำนวนเงินที่ใช้จ่ายเพิ่มขึ้น สัดส่วนของ ธุรกรรม Fraud (สีน้ำเงิน) มีแนวโน้มเพิ่มขึ้นอย่างชัดเจน โดยเฉพาะ

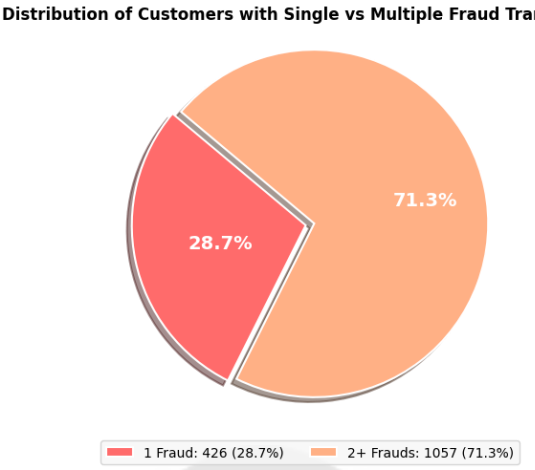
ในช่วง 500-1,000 เป็นต้นไป ซึ่งแสดงถึงความสัมพันธ์ระหว่างจำนวนเงินที่ใช้จ่ายกับโอกาสการเกิด Fraud



ภาพประกอบ 12 แสดงสัดส่วนของ ธุรกรรมที่เป็น Fraud และ Non-Fraud

3.2.4 การวิเคราะห์ลักษณะลูกค้าที่มีธุรกรรมฉ้อโกง

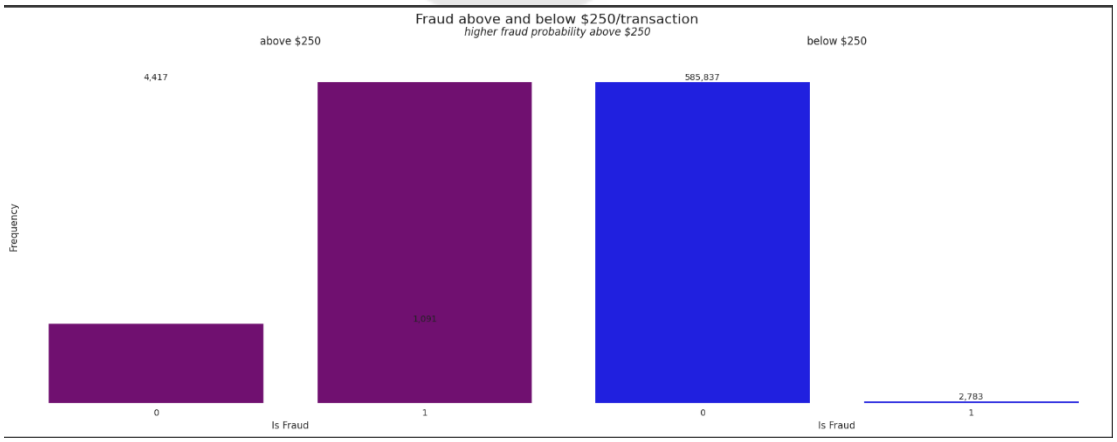
พบว่าจากข้อมูลลูกค้าทั้งหมดที่มีธุรกรรมฉ้อโกง (1,483 ราย) มีถึง 71.3% (1,057 ราย) ที่มีธุรกรรมฉ้อโกงมากกว่า 1 รายการ และมีเพียง 28.7% (426 ราย) ที่มีธุรกรรมฉ้อโกงเพียงครั้งเดียว สะท้อนให้เห็นว่าเมื่อมีการฉ้อโกงเกิดขึ้น ผู้กระทำมักจะกระทำซ้ำ ซึ่งเป็นข้อมูลสำคัญที่ช่วยในการออกแบบระบบตรวจจับและป้องกันการฉ้อโกงให้มีประสิทธิภาพยิ่งขึ้น



ภาพประกอบ 12 แสดงการกระจายตัวของลูกค้าที่มีธุรกรรมฉ้อโกง โดยจำแนกเป็นลูกค้าที่มีธุรกรรมฉ้อโกงเพียงครั้งเดียว และลูกค้าที่มีธุรกรรมฉ้อโกงหลายครั้ง

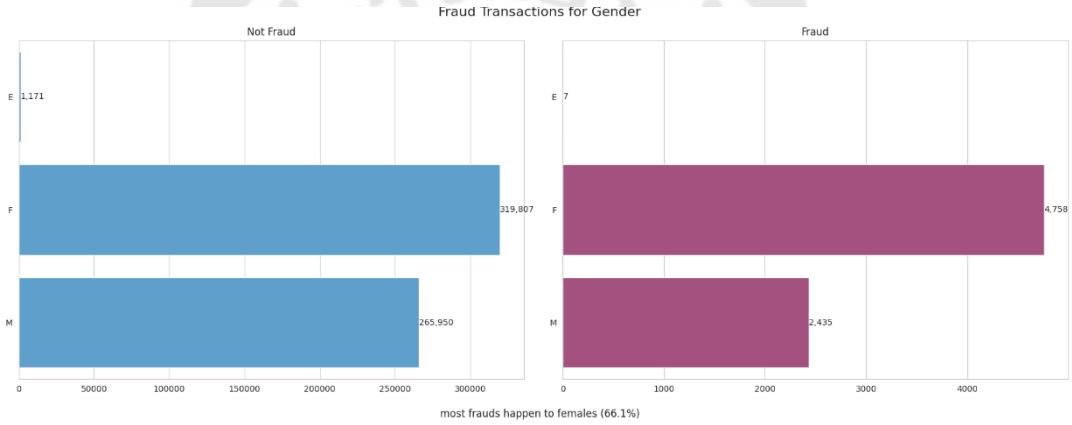
กราฟแสดงความถี่ของธุรกรรมที่เกิดการฉ้อโกง (Fraud) และไม่ฉ้อโกง (Non-Fraud) โดยจำแนกตามยอดใช้จ่ายเป็นสองกลุ่ม:ต่ำกว่า $250 (สีฟ้า): ธุรกรรมไม่ฉ้อโกงมีจำนวนมากถึง 585,837 รายการ ในขณะที่ธุรกรรมฉ้อโกงมีเพียง 2,783 รายการ สูงกว่า $250 (สีม่วง): สัดส่วนของธุรกรรมฉ้อโกงเพิ่มขึ้นชัดเจน โดยมีธุรกรรมฉ้อโกง 1,091 รายการ เทียบกับธุรกรรมไม่ฉ้อโกง 4,417 รายการ จากกราฟสามารถสรุปได้ว่า ความน่าจะเป็นในการเกิด Fraud เพิ่มขึ้นเมื่อยอดใช้จ่ายเกิน $250

3.2.5 ความสัมพันธ์ระหว่างประเภทธุรกรรมกับอัตราการฉ้อโกง



ภาพประกอบ 13 แสดงความถี่ของ ธุรกรรม Fraud และ Non-Fraud ที่จำแนกตาม ยอดใช้จ่ายต่ำกว่าและสูงกว่า $250

พบว่าเพศ F หรือผู้หญิง มีจำนวนธุรกรรมไม่ฉ้อโกงสูงที่สุดถึง 319,807 รายการ รองลงมาคือเพศ M หรือผู้ชายที่มี 265,950 รายการ ส่วนเพศ E หรือกลุ่มที่ไม่ระบุเพศมีเพียง 1,171 รายการ สำหรับกราฟทางขวาที่แสดงธุรกรรมฉ้อโกง พบว่าเพศ F มีจำนวนสูงที่สุดถึง 4,758 รายการ หรือคิดเป็น 66.1% ของธุรกรรมฉ้อโกงทั้งหมด ขณะที่เพศ M มี 2,435 รายการ และเพศ E มีจำนวนต่ำสุดเพียง 7 รายการ จากข้อมูลดังกล่าวสามารถสรุปได้ว่าธุรกรรมฉ้อโกงเกิดขึ้นมากในกลุ่มเพศ F และ M โดยเฉพาะกลุ่มเพศ F ที่มีจำนวนธุรกรรมฉ้อโกงสูงที่สุดอย่างชัดเจน



ภาพประกอบ 14 แสดงความถี่ของธุรกรรมฉ้อโกง (Fraud) และไม่ฉ้อโกง (Non-Fraud) โดยจำแนกตามเพศ (Gender)

พบว่ากลุ่มอายุ 26 - 45 ปี (ช่วงอายุ 2 - 4) เป็นกลุ่มที่มีธุรกรรมสูงที่สุดทั้งแบบฉ้อโกงและไม่ฉ้อโกง โดยเฉพาะช่วงอายุ 2 ซึ่งมีธุรกรรมฉ้อโกงสูงที่สุดที่ 2,344 รายการ และธุรกรรมไม่ฉ้อโกงสูงถึง 184,802 รายการ ในทางตรงกันข้าม กลุ่มอายุ 0 และ 7 มีจำนวนธุรกรรมทั้งสองประเภทต่ำที่สุด จากข้อมูลนี้ชี้ให้เห็นว่ากลุ่มอายุ 26 - 45 ปีเป็นกลุ่มที่มีความเสี่ยงต่อการฉ้อโกงมากที่สุด

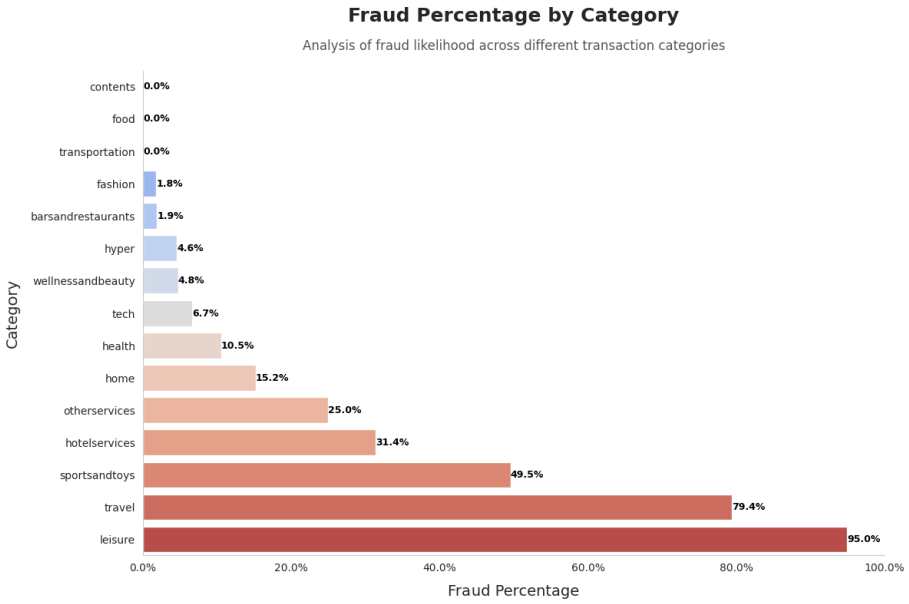
ภาพประกอบ 15 แสดงความถี่ของธุรกรรมที่เกิดการฉ้อโกง (Fraud) และไม่ฉ้อโกง (Not Fraud) โดยจำแนกตามกลุ่มอายุ (Age)

กราฟนี้แสดงเปอร์เซ็นต์ของการฉ้อโกง (Fraud Percentage) ในแต่ละประเภทของธุรกรรม (Category) เพื่อวิเคราะห์แนวโน้มและความเสี่ยงในหมวดหมู่ต่าง ๆ โดยมีข้อค้นพบสำคัญดังนี้

1. ประเภทที่มีการฉ้อโกงสูงที่สุด ได้แก่ Leisure มีเปอร์เซ็นต์การฉ้อโกงสูงถึง 95.0% และ Travel อยู่ที่ 79.4% ซึ่งเป็นหมวดหมู่ที่ต้องเฝ้าระวังเป็นพิเศษ

2. หมวดหมู่ที่มีการฉ้อโกงในระดับปานกลาง ได้แก่ Sports and Toys (49.5%), Hotel Services (31.4%) และ Other Services (25.0%)

3. หมวดหมู่ที่มีการฉ้อโกงต่ำมากหรือไม่มีเลย ได้แก่ Contents, Food, และ Transportation ที่ไม่มีการฉ้อโกงเลย (0.0%) รวมถึง Fashion (1.8%) และ Bars and Restaurants (1.9%)



ภาพประกอบ 16 แสดงเปอร์เซ็นต์ของการฉ้อโกง (Fraud Percentage) แยกตามประเภทของธุรกรรม (Category)

จากข้อมูลนี้สามารถสรุปได้ว่าหมวดหมู่ที่เกี่ยวข้องกับการพักผ่อน เช่น Leisure และ Travel มีความเสี่ยงต่อการฉ้อโกงสูงที่สุด ในขณะที่หมวดหมู่พื้นฐาน เช่น Food และ Transportation มีความเสี่ยงต่ำหรือไม่มีเลย

4. การเตรียมข้อมูล (Data Preparation)

กระบวนการเตรียมข้อมูลในงานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงคุณภาพของข้อมูลและทำให้ข้อมูลเหมาะสมสำหรับการสร้างโมเดลการเรียนรู้ของเครื่อง (Machine Learning) โดยแบ่งออกเป็นขั้นตอนสำคัญดังนี้

4.1 การตรวจสอบและทำความสะอาดข้อมูล (Data Cleaning)

1. การลบคอลัมน์ที่ไม่จำเป็น: คอลัมน์ zipcodeOri และ zipMerchant ถูกลบเนื่องจากมีข้อมูลที่ซ้ำซ้อนและไม่เพิ่มคุณค่าในการวิเคราะห์

2. จัดการข้อมูลที่ผิดปกติ: ลบอักขระพิเศษ เช่น ' จากฟีเจอร์ customer, age, gender และ category เพื่อเพิ่มความสม่ำเสมอของข้อมูล

3. ปรับค่าที่ไม่สมเหตุสมผล: ค่าที่เป็น U ในฟีเจอร์ gender ถูกลบ และค่าที่หายไปในฟีเจอร์ age ถูกแทนที่ด้วยค่า 7 ซึ่งหมายถึง "ไม่ทราบอายุ"

4. การทำความสะอาดฟีเจอร์ category: ลบคำนำหน้า es_ ในชื่อหมวดหมู่ เพื่อให้ข้อมูลมีความชัดเจนและง่ายต่อการวิเคราะห์

4.2 การตรวจสอบ Missing Values

หลังจากการตรวจสอบพบว่าไม่มีค่าที่หายไป (Missing Values = 0%) ในชุดข้อมูล ทำให้ไม่จำเป็นต้องเติมค่าหรือปรับแก้ในส่วนนี้

4.3 การจัดการค่าผิดปกติ (Outliers)

1. การใช้ IQR Method: ฟีเจอร์ amount ถูกตรวจสอบค่าผิดปกติ โดยใช้ Q1 = 13.74, Q3 = 42.54 และ IQR = 28.80 ค่าที่อยู่นอกช่วง -29.46 ถึง 85.74 ถูกจัดว่าเป็น Outliers

2. พบ Outliers จำนวน 25,781 รายการ ค่าผิดปกติในฟีเจอร์ amount ถูกเก็บไว้เนื่องจากอาจเป็นลักษณะสำคัญที่เกี่ยวข้องกับการฉ้อโกง

4.4 การสร้างฟีเจอร์ใหม่ (Feature Engineering)

1. สร้างฟีเจอร์ customer_total_trans, merchant_total_trans, category_total_trans, และ amount_thresh_total_trans เพื่อแสดงจำนวนธุรกรรมที่เกี่ยวข้องกับลูกค้า ร้านค้า หมวดหมู่สินค้า และช่วงของจำนวนเงิน

2. เพิ่ม ID สำหรับกลุ่มต่าง ๆ เช่น customer_ID, merchant_ID, category_ID, และ amount_thresh_ID เพื่อช่วยในการเชื่อมโยงข้อมูลและลดความซับซ้อน

3. สร้างฟีเจอร์ age_total_trans และ gender_total_trans เพื่อแสดงจำนวนธุรกรรมรวมตามกลุ่มอายุและเพศ

4. แปลงฟีเจอร์จำนวนธุรกรรม เช่น merchant_total_trans, age_total_trans เป็นค่าร้อยละ (% ของจำนวนธุรกรรมทั้งหมด) เพื่อเพิ่มความสามารถในการเปรียบเทียบฟีเจอร์ และช่วยให้โมเดลเข้าใจความสำคัญสัมพัทธ์ของแต่ละกลุ่ม เช่น ลูกค้าที่มียอดธุรกรรมสูงอาจมีความเสี่ยงต่อการฉ้อโกงมากขึ้น

5. การสร้างตัวแปร amount_thresh: แบ่งจำนวนเงิน (amount) ออกเป็นช่วง เช่น 0-500, 500-1000, >3000 เพื่อช่วยให้โมเดลสามารถจับลักษณะของธุรกรรมได้ดีขึ้น ช่วยลดความซับซ้อนของข้อมูลเชิงตัวเลข และช่วยให้โมเดลเรียนรู้พฤติกรรมในช่วงมูลค่าต่าง ๆ ได้ดีขึ้น

ตาราง 4 รายละเอียดของฟีเจอร์ในขั้นตอนการสร้างฟีเจอร์ (Feature Engineering)

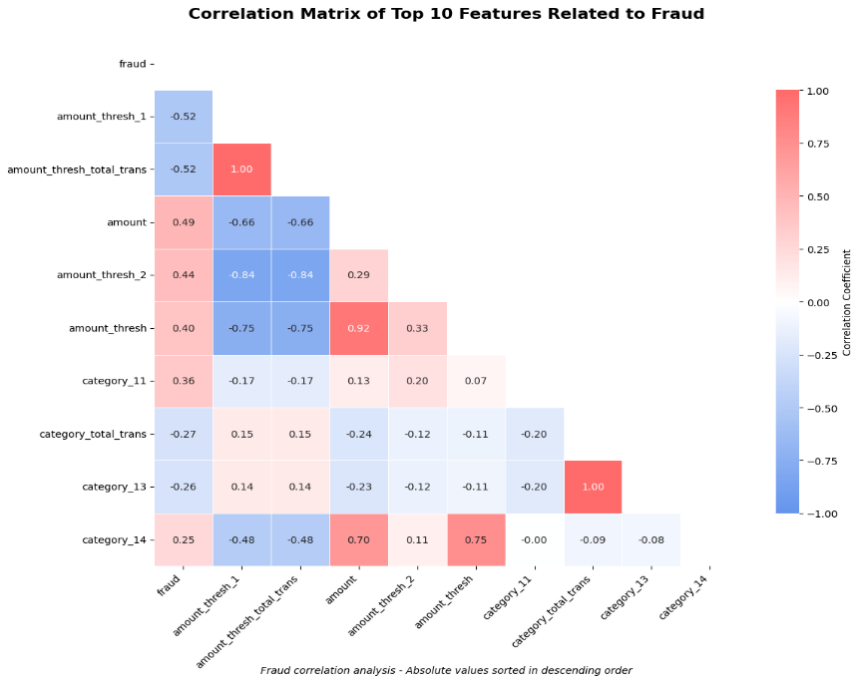
ชื่อฟีเจอร์	รายละเอียด	ประเภทข้อมูล	คำอธิบาย
customer_total_trans	จำนวนธุรกรรมรวมของลูกค้าแต่ละราย	int	คำนวณจากการนับจำนวนธุรกรรมทั้งหมดที่ลูกค้าทำ
customer_ID	รหัสของลูกค้าแต่ละราย	int	รหัสลูกค้าที่ถูกสร้างขึ้นใหม่เพื่อระบุแต่ละลูกค้า
merchant_total_trans	จำนวนธุรกรรมรวมของร้านค้าแต่ละแห่ง	int	คำนวณจากการนับจำนวนธุรกรรมที่แต่ละร้านค้าทำ
merchant_ID	รหัสของร้านค้าแต่ละแห่ง	int	รหัสร้านค้าที่ถูกสร้างขึ้นใหม่เพื่อระบุแต่ละร้านค้า
category_total_trans	จำนวนธุรกรรมรวมของแต่ละหมวดหมู่สินค้า	int	คำนวณจากการนับจำนวนธุรกรรมในแต่ละหมวดหมู่สินค้า
category_ID	รหัสของหมวดหมู่สินค้า	int	รหัสหมวดหมู่ที่ถูกสร้างขึ้นใหม่
amount_thresh_total_trans	จำนวนธุรกรรมรวมตามช่วงของจำนวนเงิน	int	คำนวณจากการนับจำนวนธุรกรรมในแต่ละช่วงของจำนวนเงิน
amount_thresh_ID	รหัสของช่วงจำนวนเงิน	int	รหัสที่ถูกสร้างขึ้นใหม่เพื่อแทนแต่ละช่วงของจำนวนเงิน
age_total_trans	จำนวนธุรกรรมรวมตามกลุ่มอายุ	int	คำนวณจากการนับจำนวนธุรกรรมในแต่ละกลุ่มอายุ
gender_total_trans	จำนวนธุรกรรมรวมตามเพศ	int	คำนวณจากการนับจำนวนธุรกรรมในแต่ละเพศ

ชื่อฟีเจอร์	รายละเอียด	ประเภทข้อมูล	คำอธิบาย
merchant_total_trans (weight)	สัดส่วนของจำนวนธุรกรรมของร้านค้าต่อจำนวนธุรกรรมทั้งหมด (%)	float	คำนวณจาก (merchant_total_trans / จำนวนธุรกรรมทั้งหมด) * 100
category_total_trans (weight)	สัดส่วนของจำนวนธุรกรรมของหมวดหมู่สินค้าต่อจำนวนธุรกรรมทั้งหมด (%)	float	คำนวณจาก (category_total_trans / จำนวนธุรกรรมทั้งหมด) * 100
amount_thresh_total_trans (weight)	สัดส่วนของจำนวนธุรกรรมตามช่วงจำนวนเงินต่อจำนวนธุรกรรมทั้งหมด (%)	float	คำนวณจาก (amount_thresh_total_trans / จำนวนธุรกรรมทั้งหมด) * 100
age_total_trans (weight)	สัดส่วนของจำนวนธุรกรรมในแต่ละกลุ่มอายุต่อจำนวนธุรกรรมทั้งหมด (%)	float	คำนวณจาก (age_total_trans / จำนวนธุรกรรมทั้งหมด) * 100
gender_total_trans (weight)	สัดส่วนของจำนวนธุรกรรมในแต่ละเพศต่อจำนวนธุรกรรมทั้งหมด (%)	float	คำนวณจาก (gender_total_trans / จำนวนธุรกรรมทั้งหมด) * 100
age	รหัสกลุ่มอายุ	int	ถูกแปลงจากค่าอายุเดิม
gender	รหัสเพศ	int	แปลงค่าจากเพศเดิมเป็นตัวเลข: M = 1, F = 2, U = 3
category	รหัสหมวดหมู่สินค้า	int	ถูกแปลงจากหมวดหมู่สินค้าเดิม

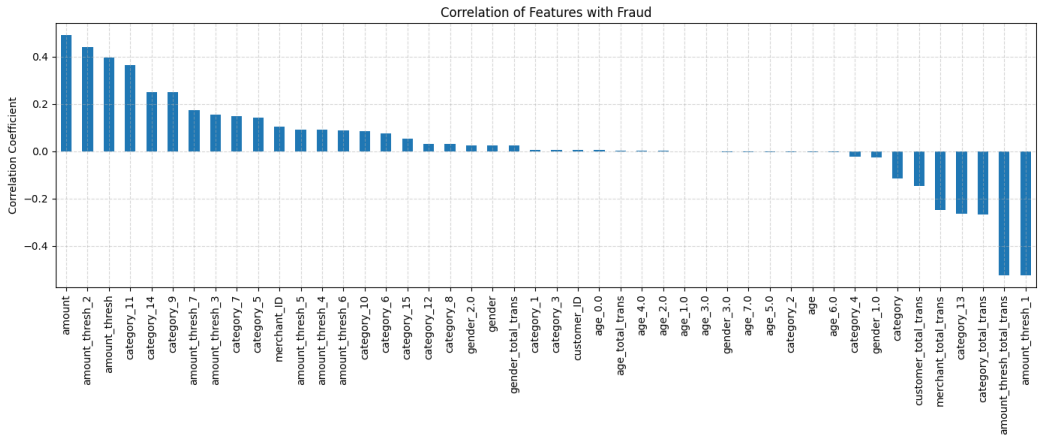
ชื่อฟีเจอร์	รายละเอียด	ประเภทข้อมูล	คำอธิบาย
amount_thresh	รหัสของช่วงจำนวนเงิน	int	ถูกแปลงจากช่วงจำนวนเงินเดิม
fraud	สถานะธุรกรรม	int	ธุรกรรม: 0=ปกติ, 1=ฉ้อโกง
age_0, age_1, ...	One-Hot Encoding ของกลุ่มอายุ	int	ใช้ One-Hot Encoding เพื่อแทนค่ากลุ่มอายุ
gender_0, gender_1, ...	One-Hot Encoding ของเพศ	int	ใช้ One-Hot Encoding เพื่อแทนค่าเพศ
category_0, category_1, ...	One-Hot Encoding ของหมวดหมู่สินค้า	int	ใช้ One-Hot Encoding เพื่อแทนค่าหมวดหมู่สินค้า
amount_thresh_0, amount_thresh_1, ...	One-Hot Encoding ของช่วงจำนวนเงิน	int	ใช้ One-Hot Encoding เพื่อแทนค่าช่วงจำนวนเงิน

4.5 **การวิเคราะห์ความสัมพันธ์ของฟีเจอร์** (Correlation Analysis)

หลังจากที่ทำ Feature engineering เสร็จเรียบร้อยแล้ว ผู้วิจัยจะทำการวิเคราะห์ความสัมพันธ์ของแต่ละ Feature โดยเลือก 10 ฟีเจอร์ที่มีความสัมพันธ์กับตัวแปรเป้าหมาย (fraud) มากที่สุดออกมาทำการวิเคราะห์เพื่อศึกษารูปแบบความสัมพันธ์ระหว่างตัวแปรและใช้เป็นข้อมูลในการพัฒนาโมเดล



ภาพประกอบ 17 Heatmap แสดงค่า Pearson Correlation



ภาพประกอบ 18 กราฟแท่งแสดงค่า Pearson Correlation ระหว่างฟีเจอร์ทั้งหมดกับตัวแปรเป้าหมาย (fraud)

ภาพประกอบที่ 17 และ 18 แสดงผลการวิเคราะห์ความสัมพันธ์ระหว่างฟีเจอร์กับตัวแปรเป้าหมาย (fraud) โดยใช้ค่า Pearson Correlation Coefficient โดยพบว่า

1. ฟีเจอร์ที่มีความสัมพันธ์เชิงบวกกับการฉ้อโกง: ตัวแปรที่มีความสัมพันธ์เชิงบวกสูงสุดได้แก่ amount (0.49), amount_thresh_2 (0.44), และ amount_thresh (0.40) ซึ่งล้วนเกี่ยวข้องกับจำนวนเงินในการทำธุรกรรม แสดงให้เห็นว่าธุรกรรมที่มีมูลค่าสูงอาจมีความเสี่ยงต่อการฉ้อโกงมากกว่า

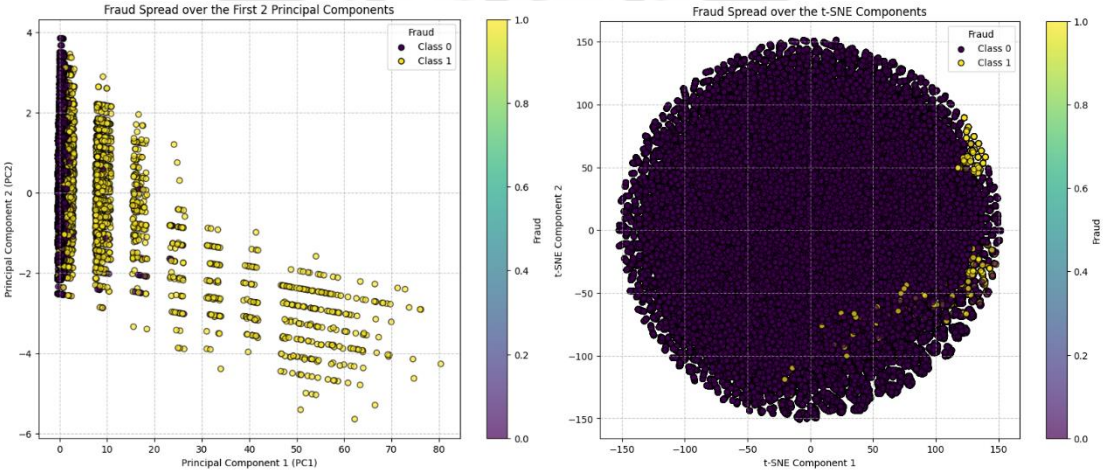
2. ฟีเจอร์ที่มีความสัมพันธ์เชิงลบกับการฉ้อโกง: ตัวแปร amount_thresh_1 และ amount_thresh_total_trans (ทั้งคู่มีค่า -0.52) มีความสัมพันธ์เชิงลบสูงสุด ซึ่งอาจบ่งชี้ว่าธุรกรรมที่มีลักษณะเหล่านี้มักเป็นธุรกรรมปกติ

3. Multicollinearity: มีความสัมพันธ์สูงระหว่างตัวแปรบางคู่ เช่น amount_thresh_1 กับ amount_thresh_total_trans (1.00) และ category_total_trans กับ category_13 (1.00) ซึ่งอาจส่งผลต่อประสิทธิภาพของโมเดลเชิงเส้น

4. ความสำคัญของประเภทธุรกรรม: ตัวแปรประเภท category มีความสัมพันธ์ระดับปานกลางกับการฉ้อโกง เช่น category_11 (0.36) และ category_14 (0.25) ซึ่งแสดงให้เห็นว่าประเภทธุรกรรมบางประเภทมีความเสี่ยงต่อการฉ้อโกงมากกว่าประเภทอื่น

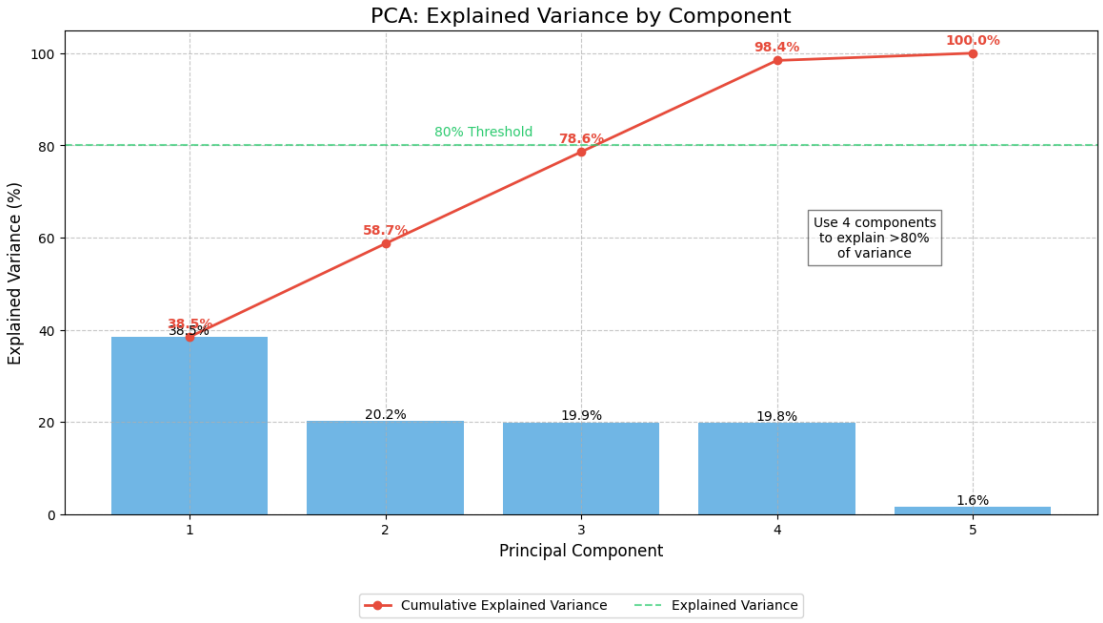
ผลการวิเคราะห์ความสัมพันธ์ระหว่างฟีเจอร์แสดงให้เห็นปัญหาเรื่อง Multicollinearity ดังนั้นผู้วิจัยจึงเลือกใช้โมเดล Tree-based เป็นหลักเพื่อจัดการกับปัญหานี้อย่างเหมาะสม

4.6 การวิเคราะห์การกระจายของ Fraud ด้วย PCA และ t-SNE



ภาพประกอบ 19 การเปรียบเทียบการกระจายของ Fraud ด้วย PCA และ t-SNE

ผู้วิจัยเปรียบเทียบการกระจายตัวของธุรกรรม fraud (สีเหลือง) และ non-fraud (สีม่วง) ด้วยเทคนิค PCA (ซ้าย) และ t-SNE (ขวา) จากผลการวิเคราะห์พบว่า t-SNE สามถแยกกลุ่มข้อมูล fraud ออกจาก non-fraud ได้ชัดเจนกว่า PCA อย่างเห็นได้ชัด ทำให้ผู้วิจัยสามารถเข้าใจลักษณะการกระจายตัวของข้อมูลและวางแผนการพัฒนาโมเดลตรวจจับการฉ้อโกงได้อย่างมีประสิทธิภาพมากขึ้น



ภาพประกอบ 20 แสดงความแปรปรวนที่อธิบายได้ (Explained Variance)

จากการวิเคราะห์ด้วยเทคนิค PCA โดยองค์ประกอบหลัก (Principal Component) ที่ 1 อธิบายความแปรปรวนได้สูงสุดที่ 38.5% ตามด้วยองค์ประกอบที่ 2 - 5 ซึ่งอธิบายได้ 20.2%, 19.9%, 19.8% และ 1.6% ตามลำดับ เมื่อพิจารณาความแปรปรวนสะสม (เส้นสีแดง) พบว่าต้องใช้ 4 องค์ประกอบแรกจึงจะอธิบายความแปรปรวนได้เกิน 80% (98.4%) ซึ่งเป็นค่าเกณฑ์มาตรฐาน (เส้นประสีเขียว) ข้อมูลนี้ช่วยสนับสนุนผลการวิเคราะห์ PCA ในการจำแนกกลุ่ม fraud ว่ามีข้อจำกัดเมื่อใช้เพียง 2 องค์ประกอบแรกซึ่งอธิบายความแปรปรวนได้เพียง 58.7% ซึ่งอาจเป็นเหตุผลที่ทำให้การแยกกลุ่มข้อมูล fraud และ non-fraud ด้วย PCA ไม่ชัดเจนเทียบกับ t-SNE

4.7 การแปลงข้อมูลหมวดหมู่ (Categorical Transformation)

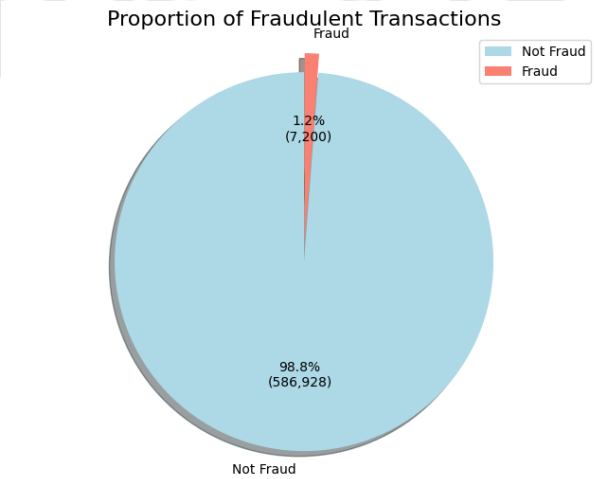
1. ฟีเจอร์หมวดหมู่ เช่น gender, age, category, และ amount_thresh ถูกแปลงให้อยู่ในรูปแบบที่โมเดลสามารถประมวลผลได้ง่าย

2. การเปลี่ยนประเภทข้อมูล ฟีเจอร์ age ถูกแปลงจากประเภท object เป็น float ฟีเจอร์ gender ถูกเข้ารหัสใหม่เป็นค่าตัวเลข (1: Male, 2: Female, 3: Missing)

4.8 การจัดการข้อมูลที่ไม่สมดุล (Handling Imbalanced Data)

ปัญหาสำคัญอย่างหนึ่งที่ผู้วิจัยพบในการตรวจจับการฉ้อโกงทางการเงินคือเรื่องข้อมูลไม่สมดุล นั่นคือจำนวนธุรกรรมปกติมีมากกว่าธุรกรรมที่เป็นการฉ้อโกงมาก ความไม่สมดุลนี้ทำให้โมเดลที่เราสร้างมักจะเอนเอียงไปทางการทำนายว่าธุรกรรมส่วนใหญ่เป็นธุรกรรมปกติ ซึ่งทำให้อาจพลาดการตรวจจับการฉ้อโกงที่แท้จริงได้

จากข้อมูลที่ศึกษา เห็นได้ชัดว่ามีความไม่สมดุลอย่างมาก ดังที่จะเห็นในภาพประกอบ 20 ที่แสดงให้เห็นว่าธุรกรรมส่วนใหญ่เป็นธุรกรรมปกติ ในขณะที่มีธุรกรรมฉ้อโกงเพียงส่วนน้อยเท่านั้น สถานการณ์แบบนี้เป็นเรื่องปกติในโลกของการเงินจริง และเป็นเหตุผลที่เราจำเป็นต้องหาวิธีพิเศษในการจัดการกับข้อมูลที่ไม่สมดุลเพื่อให้โมเดลของเราสามารถตรวจจับการฉ้อโกงได้อย่างมีประสิทธิภาพ



ภาพประกอบ 21 แสดงแผนภูมิวงกลมสัดส่วนธุรกรรมที่เป็นการฉ้อโกงและไม่เป็นการฉ้อโกง

จากแผนภูมิจะเห็นได้ว่าธุรกรรมส่วนใหญ่เป็นธุรกรรมปกติที่ไม่มีการฉ้อโกง คิดเป็น 98.8% หรือจำนวน 536,928 รายการ ในขณะที่มีเพียง 1.2% หรือเพียง 7,200 รายการเท่านั้นที่เป็นธุรกรรมฉ้อโกง แสดงให้เห็นถึงปัญหาความไม่สมดุลของข้อมูล (Data Imbalance) อย่างชัดเจน ซึ่งเป็นความท้าทายสำคัญในการพัฒนาโมเดลการตรวจจับการฉ้อโกงทางการเงิน เนื่องจากมีตัวอย่างกรณีการฉ้อโกงน้อยมากเมื่อเทียบกับธุรกรรมปกติ

1. การใช้ SMOTE และ Tomek Links: ใช้ SMOTE สร้างข้อมูลเพิ่มในกลุ่มธุรกรรมฉ้อโกงจนทั้งสองคลาสมีจำนวนเท่ากัน (410,849 รายการ) และการใช้ Tomek Links กำจัดข้อมูลซ้ำซ้อนในกลุ่มธุรกรรมปกติ ทำให้เหลือข้อมูล Non-fraud 410,240 รายการ และ Fraud 5,040 รายการ

2. การใช้ SmoteTomek รวม SMOTE และ Tomek Links ในขั้นตอนเดียวเพื่อเพิ่มความสมดุลของข้อมูล

3. การใช้ Class Weight เพื่อปรับน้ำหนักของธุรกรรมฉ้อโกงในโมเดลเพื่อให้โมเดลเรียนรู้ได้ดีขึ้น

4.9 **การสเกลข้อมูล** (Feature Scaling)

ใช้ StandardScaler เพื่อปรับสเกลข้อมูลเชิงตัวเลขให้มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ช่วยลดผลกระทบจากค่าที่มีช่วงกว้างในฟีเจอร์ต่างๆ ได้แก่ Amount (จำนวนเงิน), Merchant_total_trans (ธุรกรรมรวมของร้านค้า), Customer_total_trans (ธุรกรรมรวมของลูกค้า), Category_total_trans (ธุรกรรมรวมของหมวดหมู่สินค้า), Gender_total_trans (ธุรกรรมรวมตามเพศ), Age_total_trans (ธุรกรรมรวมตามกลุ่มอายุ) และ Amount_thresh_total_trans (ธุรกรรมรวมตามช่วงจำนวนเงิน)

4.10 **การแยกชุดข้อมูลและ** Cross Validation (Data Splitting & Cross Validation)

1. การแบ่งข้อมูล ข้อมูลถูกแบ่งเป็น Training Set (70%) และ Test Set (30%) โดยใช้คำสั่ง train_test_split พร้อมพารามิเตอร์ stratify=y เพื่อรักษาสัดส่วนของธุรกรรมปกติและฉ้อโกงในชุดข้อมูล

2. Cross Validation (5-Fold) ใช้ 5-Fold Cross Validation ในชุด Training เพื่อเพิ่มความน่าเชื่อถือของการประเมินผลโมเดล การแบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน โดยในแต่

ละรอบ จะใช้ 4 ส่วนในการฝึกโมเดล และอีก 1 ส่วนสำหรับการประเมินผล การใช้ 5-Fold Cross Validation ช่วยลด Bias และป้องกัน Overfitting โดยเฉพาะในกรณีที่ข้อมูลมีความไม่สมดุลระหว่างธุรกรรมปกติและฉ้อโกง

กระบวนการเตรียมข้อมูลในงานนี้ช่วยปรับปรุงข้อมูลให้เหมาะสมสำหรับการสร้างโมเดล โดยการจัดการค่าผิดปกติ การสร้างฟีเจอร์ใหม่ และการปรับสมดุลข้อมูล ช่วยเพิ่มศักยภาพในการตรวจจับธุรกรรมฉ้อโกง ข้อมูลที่ผ่านการปรับปรุงนี้พร้อมสำหรับการพัฒนาโมเดลในขั้นตอนถัดไป ด้วย Cross Validation ช่วยเพิ่มความน่าเชื่อถือในการประเมินผลลัพธ์ของโมเดล

5. การสร้างแบบจำลองพยากรณ์ (Model Development)

ในการศึกษานี้ ผู้วิจัยได้ออกแบบการทดลองโดยแบ่งออกเป็น 5 แนวทางหลัก เพื่อประเมินผลกระทบของเทคนิคการจัดการข้อมูลที่ไม่สมดุลต่อประสิทธิภาพของแบบจำลองในการตรวจจับธุรกรรมที่เป็นการฉ้อโกง โดยแต่ละแนวทางมีวัตถุประสงค์และกระบวนการที่แตกต่างกัน โดยดำเนินการทดลองเปรียบเทียบด้วยวิธีการต่าง ๆ ดังนี้

5.1 โมเดลพื้นฐาน (Baseline Model)

ขั้นแรกเป็นการสร้างโมเดลพื้นฐานหรือ Baseline Model เพื่อใช้เป็นเกณฑ์เปรียบเทียบผลลัพธ์ โดยใช้โมเดล Logistic Regression, Random Forest, XGBoost, LightGBM และ CatBoost โดยไม่มีการปรับแต่งข้อมูลหรือพารามิเตอร์ใด ๆ เพิ่มเติม โมเดลนี้ถูกประเมินด้วยเมตริกสำคัญ เช่น Accuracy, Precision, Recall, F1-Score และ ROC-AUC เพื่อใช้เป็นพื้นฐานสำหรับการเปรียบเทียบกับโมเดลที่ผ่านการปรับแต่งหรือใช้เทคนิคอื่นๆ ในขั้นตอนต่อไป

5.2 การปรับน้ำหนักคลาส (Adjust Class Weight)

ในขั้นตอนนี้จะเป็นการปรับน้ำหนักของแต่ละคลาสให้เหมาะสมกับความไม่สมดุลของข้อมูล โดยให้ความสำคัญกับคลาสที่เป็นธุรกรรมฉ้อโกง (Minority Class) มากขึ้น วิธีนี้ดำเนินการโดยใช้พารามิเตอร์ class_weight='balanced' สำหรับ Logistic Regression และ Random Forest และใช้ scale_pos_weight สำหรับโมเดลประเภท XGBoost และ LightGBM การปรับน้ำหนักนี้ช่วยให้โมเดลสามารถเรียนรู้จากกลุ่มธุรกรรมฉ้อโกงได้ดีขึ้น ลดปัญหาการที่โมเดลถูกดึงไปเรียนรู้เฉพาะกลุ่มที่มีจำนวนตัวอย่างมากกว่า (Majority Class)

5.3 การใช้ Undersampling (Tomek Links)

เป็นการจัดการปัญหาความไม่สมดุลของข้อมูลโดยใช้เทคนิค Undersampling ด้วยวิธี Tomek Links ซึ่งจะทำการลบตัวอย่างในกลุ่มธุรกรรมปกติ (Majority Class) ที่มีลักษณะใกล้เคียงหรือซ้ำซ้อนกับกลุ่มธุรกรรมฉ้อโกง (Minority Class) ออกไป วิธีนี้ช่วยลดความสับสนระหว่างคลาสต่างๆ ในการเรียนรู้ของโมเดล ซึ่งส่งผลให้โมเดลมีความสามารถในการแยกแยะธุรกรรฉ้อโกงได้ดียิ่งขึ้น และลด False Positive ที่อาจเกิดขึ้นในโมเดล

5.4 การใช้ Oversampling ด้วย SMOTE

ขั้นตอนนี้ใช้เทคนิค Oversampling โดยการสร้างตัวอย่างข้อมูลสังเคราะห์ขึ้นมาใหม่ด้วยเทคนิค SMOTE (Synthetic Minority Oversampling Technique) เพื่อเพิ่มจำนวนตัวอย่างในกลุ่มธุรกรรมฉ้อโกง (Minority Class) ให้มีจำนวนใกล้เคียงกับกลุ่มปกติ (Majority Class) ส่งผลให้โมเดลสามารถเรียนรู้รูปแบบและลักษณะของธุรกรรมฉ้อโกงได้อย่างมีประสิทธิภาพมากขึ้น วิธีนี้มีข้อดีคือช่วยเพิ่ม Recall ในการตรวจจับธุรกรรมฉ้อโกง และช่วยให้โมเดลมีความแม่นยำและเสถียรมากยิ่งขึ้น

5.5 การใช้เทคนิคแบบผสม (Hybrid Sampling)

ในขั้นตอนนี้ เป็นการนำ Oversampling ด้วย SMOTE และ Undersampling ด้วย Tomek Links มาใช้ร่วมกัน (Hybrid Sampling) เพื่อแก้ปัญหาข้อมูลไม่สมดุล โดยเริ่มจากการใช้ SMOTE สร้างข้อมูลในคลาสที่มีจำนวนน้อยก่อน หลังจากนั้นนำข้อมูลที่ได้ไปจัดการด้วย Tomek Links อีกครั้งเพื่อลบตัวอย่างที่ใกล้เคียงหรือซ้ำซ้อนออก เทคนิคนี้ช่วยให้ข้อมูลที่ได้มีความสมดุลมากขึ้นทั้งจำนวนและคุณภาพ ทำให้โมเดลมีประสิทธิภาพที่ดีขึ้นในการตรวจจับธุรกรรมฉ้อโกง

ตาราง 2 สรุปรายละเอียดการพัฒนาโมเดลโดยใช้เทคนิคต่างๆ ที่ได้อธิบายไว้ข้างต้น

No.	แนวทางการทดลอง	หลักการและวิธีการ	ข้อดี-ข้อจำกัด
1	Baseline Model	ใช้โมเดล Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost โดยไม่มีการปรับแต่งข้อมูลหรือพารามิเตอร์ใด ๆ	ใช้เป็นโมเดลพื้นฐานสำหรับเปรียบเทียบ
2	Adjust Class Weight	เพิ่มน้ำหนักคลาสของธุรกรรมฉ้อโกงโดยใช้ class weight='balanced' ใน Logistic Regression, Random Forest และใช้ scale_pos_weight ใน XGBoost, LightGBM	ช่วยเพิ่ม Recall โดยไม่เปลี่ยนแปลงข้อมูลดั้งเดิม แต่ต้องระวัง False Positives
3	Undersampling (Tomek Links)	ใช้ Tomek Links ลบข้อมูลที่ซ้ำซ้อนระหว่างธุรกรรมปกติกับธุรกรรมฉ้อโกง เพื่อลดจำนวนข้อมูลส่วนใหญ่ลง	ลด False Positives ได้ดี แต่มีโอกาสเสียข้อมูลที่สำคัญออกไปบางส่วน
4	SMOTE (Oversampling)	ใช้ SMOTE เพิ่มจำนวนข้อมูลในกลุ่มธุรกรรมฉ้อโกง โดยการสร้างข้อมูลสังเคราะห์ขึ้นมาใหม่	ช่วยปรับปรุง Recall และ F1-score แต่อาจเกิด Overfitting ได้
5	Hybrid Sampling (SMOTE + Tomek Links)	รวมเทคนิค Oversampling ด้วย SMOTE และ Undersampling ด้วย Tomek Links เพื่อให้ข้อมูลมีความสมดุลมากขึ้น	ช่วยเพิ่มความสมดุลของข้อมูลอย่างมีประสิทธิภาพ และลด Overfitting

5.6 การประเมินผลและเปรียบเทียบผลลัพธ์

ทุกโมเดลที่สร้างขึ้นในขั้นตอนต่าง ๆ ถูกประเมินด้วยเมตริกที่เหมาะสมกับการวิเคราะห์ข้อมูลที่ไม่สมดุล ได้แก่ Accuracy, Precision, Recall, F1-Score และ ROC-AUC รวมถึงวิเคราะห์ Confusion Matrix เพื่อดูการทำนายที่ถูกต้องและผิดพลาด (TP, TN, FP, FN) ผลลัพธ์ของแต่ละวิธีการถูกเปรียบเทียบ เพื่อเลือกโมเดลที่มีสมรรถนะดีที่สุดในการตรวจจับธุรกรรมฉ้อโกง

5.7 การวิเคราะห์ความสำคัญของฟีเจอร์ด้วย SHAP Analysis

ขั้นตอนสุดท้ายคือการนำเทคนิค SHAP (SHapley Additive exPlanations) มาวิเคราะห์เพื่อทำความเข้าใจการทำงานของโมเดล โดย SHAP จะช่วยอธิบายว่าแต่ละฟีเจอร์มีผลกระทบต่อการตัดสินใจของโมเดลอย่างไร ผ่านกราฟสรุปความสำคัญของฟีเจอร์ (Summary Plot) และกราฟแสดงผลกระทบของฟีเจอร์ (Dependence Plot) เพื่อให้เห็นภาพชัดเจนยิ่งขึ้นเกี่ยวกับความสัมพันธ์ของแต่ละฟีเจอร์ที่มีผลต่อการตรวจจับธุรกรรมฉ้อโกง ผลจากการวิเคราะห์นี้จะช่วยในการพัฒนาหรือเลือกฟีเจอร์ที่มีประสิทธิภาพสูงสุดต่อไป

บทที่ 4
ผลการดำเนินการวิจัย

บทนี้นำเสนอผลการวิเคราะห์และเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องในการตรวจจับธุรกรรมฉ้อโกง โดยแสดงผลการทดสอบโมเดลทั้ง 5 ประเภท ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression ผลการวิจัยถูกนำเสนอตามลำดับขั้นตอนการพัฒนา โดยมีรายละเอียดดังนี้

1. ผลการจัดการข้อมูลไม่สมดุล
2. ผลการปรับแต่งโมเดล (Hyperparameter Tuning)
3. การเปรียบเทียบประสิทธิภาพโมเดล
4. การวิเคราะห์ความสำคัญของฟีเจอร์

1. ผลการจัดการข้อมูลไม่สมดุล

ในขั้นตอนนี้ได้มีการทดสอบประสิทธิภาพของโมเดลพื้นฐานจำนวน 5 ประเภท ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression โดยใช้พารามิเตอร์เริ่มต้น (Default Parameters) เพื่อเปนเกณฑ์อ้างอิงสำหรับการเปรียบเทียบกับผลลัพธ์หลังจากการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) และการจัดการข้อมูลไม่สมดุล ในการประเมินประสิทธิภาพของโมเดลใช้ตัวชี้วัด ได้แก่ Precision, Recall, F1-Score, ROC-AUC และ Precision-Recall AUC

ตาราง 3 ผลการทดสอบโมเดลพื้นฐาน (Baseline)

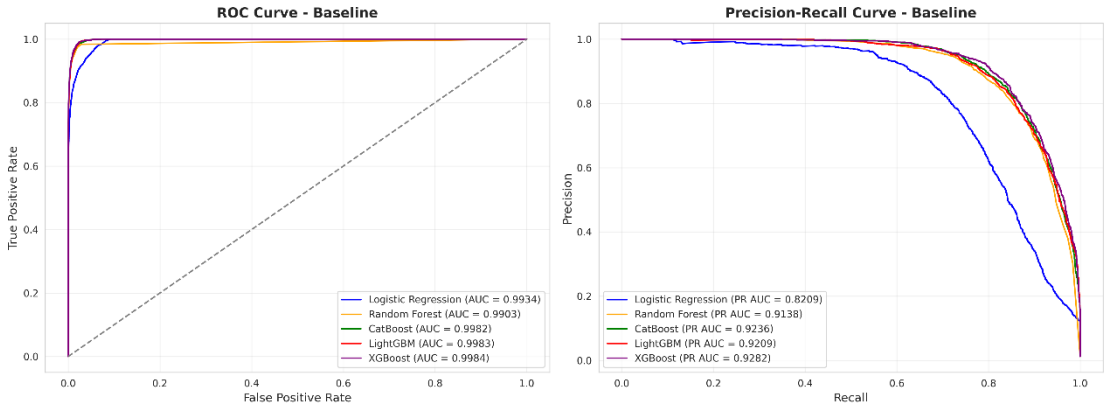
Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.88	0.66	0.75	0.993	0.821
Random Forest	0.88	0.80	0.83	0.990	0.914
CatBoost	0.88	0.80	0.84	0.998	0.924
LightGBM	0.89	0.80	0.84	0.998	0.921
XGBoost	0.89	0.81	0.85	0.998	0.928

จากตาราง 6 ผู้วิจัยพบว่า โมเดลประเภท Gradient Boosting (XGBoost, CatBoost และ LightGBM) แสดงประสิทธิภาพที่โดดเด่น โดย XGBoost ให้ผลลัพธ์ดีที่สุดด้วยค่า Precision 0.89, Recall 0.81 และ F1-Score 0.85

Random Forest ให้ผลการทดสอบอยู่ในระดับกลาง (F1-Score 0.83) ขณะที่ Logistic Regression มีประสิทธิภาพต่ำที่สุด โดยเฉพาะค่า Recall (0.66) ซึ่งสะท้อนถึงข้อจำกัดของโมเดลเชิงเส้นในการตรวจจับการฉ้อโกงที่มีความซับซ้อน

แม้ว่าทุกโมเดลมีค่า ROC-AUC สูงมาก (0.990 – 0.998) แต่ในกรณีข้อมูลไม่สมดุล ค่า Precision-Recall AUC เป็นตัวชี้วัดที่เหมาะสมกว่า ซึ่ง XGBoost ยังคงให้ค่าสูงสุดที่ 0.928

จากผลการทดสอบนี้ ผู้วิจัยเลือก XGBoost เป็นโมเดลที่มีศักยภาพมากที่สุด และจะดำเนินการปรับแต่งพารามิเตอร์เพื่อเพิ่มประสิทธิภาพต่อไป โดยเฉพาะการปรับปรุงค่า Recall เนื่องจากในบริบทการตรวจจับการฉ้อโกง การพลาดตรวจจับกรณีฉ้อโกงจริงมีผลกระทบทางธุรกิจสูง



ภาพประกอบ 21 แสดง ROC Curve และ Precision-Recall Curve สำหรับผลการทดสอบโมเดลพื้นฐาน (Baseline Model)

- ROC Curve แสดงความสามารถในการแยกแยะธุรกรรมฉ้อโกง โมเดล XGBoost, LightGBM และ CatBoost มีค่า AUC สูงสุดที่ประมาณ 0.9984, 0.9983 และ 0.9982 ตามลำดับ
- Precision-Recall Curve แสดงความสมดุลระหว่าง Precision และ Recall โมเดล XGBoost มีค่า PR AUC สูงสุดที่ 0.9282 รองลงมาคือ CatBoost และ LightGBM ที่มีค่าเท่ากันที่ 0.9209 โดยรวมแล้ว XGBoost และ CatBoost ให้ผลลัพธ์ดีที่สุดในแง่ของการตรวจจับธุรกรรมฉ้อโกง

ตาราง 4 ผลการทดสอบโมเดลหลังจากจัดการข้อมูลด้วยวิธี SMOTE

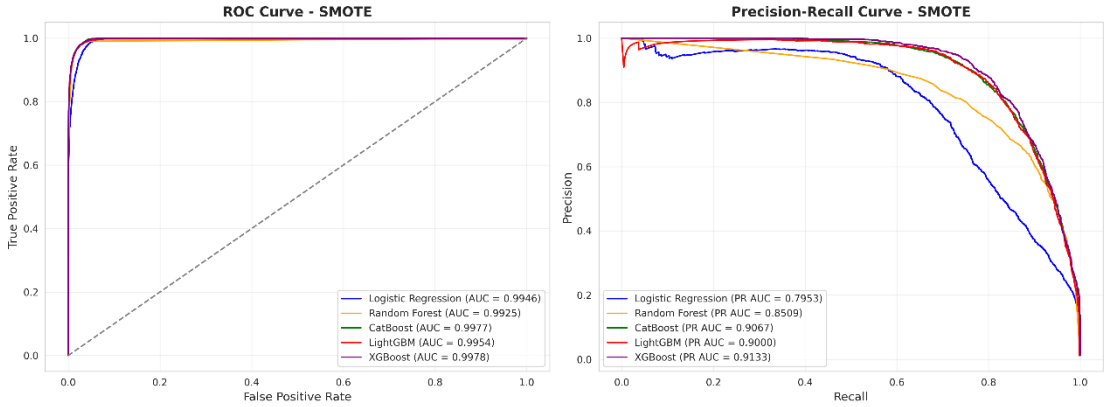
Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.21	0.98	0.35	0.995	0.795
Random Forest	0.63	0.89	0.74	0.992	0.851
CatBoost	0.54	0.93	0.69	0.998	0.907
LightGBM	0.43	0.96	0.59	0.995	0.900
XGBoost	0.49	0.95	0.65	0.998	0.913

จากตาราง 7 ผู้วิจัยพบว่า การจัดการกับข้อมูลไม่สมดุลด้วยเทคนิค SMOTE ส่งผลให้ค่า Recall ของทุกโมเดลเพิ่มขึ้นอย่างมีนัยสำคัญ แต่มีผลกระทบในเชิงลบต่อค่า Precision

Logistic Regression มีค่า Recall สูงสุด (0.98) แต่ Precision ลดลงอย่างมาก (0.21) ส่งผลให้ F1-Score ต่ำสุด (0.35) สะท้อนถึงการเพิ่มขึ้นของ false positive จำนวนมาก

Random Forest สามารถรักษาสมดุลระหว่าง Precision และ Recall ได้ดีที่สุดหลังใช้ SMOTE โดยมี F1-Score สูงสุด (0.74) แม้จะต่ำกว่าค่าที่ได้จากโมเดลพื้นฐาน XGBoost ในตารางที่ 6 โมเดล Gradient Boosting ทั้งหมดมีค่า Recall สูง (0.93 – 0.96) แต่ Precision ลดลงอย่างมาก (0.43 – 0.54) ส่งผลให้ F1-Score ลดลงเมื่อเทียบกับโมเดลพื้นฐาน โดย CatBoost ให้ผลดีที่สุดในกลุ่มนี้ (F1-Score 0.69) ค่า Precision-Recall AUC ยังคงสูงในกลุ่ม Gradient Boosting โดย XGBoost มีค่าสูงสุด (0.913) ตามด้วย CatBoost (0.907)

ผลการทดลองนี้ชี้ให้เห็นว่าการใช้ SMOTE อาจไม่เหมาะสมกับชุดข้อมูลนี้หากพิจารณาโดยรวม เนื่องจากประสิทธิภาพด้าน F1-Score ลดลงในทุกโมเดลเมื่อเทียบกับโมเดลพื้นฐาน อย่างไรก็ตาม หากภารกิจหลักคือการเพิ่ม Recall ในการตรวจจับการฉ้อโกง การใช้ SMOTE อาจเป็นทางเลือกที่เหมาะสม โดยต้องยอมรับการเพิ่มขึ้นของ false positive



ภาพประกอบ 22 แสดง ROC Curve และ Precision-Recall Curve สำหรับผลการทดสอบ SMOTE

- ROC Curve: โมเดลทั้งหมดมีค่า ROC-AUC สูงใกล้เคียงกันมาก โดยเฉพาะ XGBoost และ CatBoost ซึ่งอยู่ใกล้เคียงกับเส้น True Positive Rate = 1 อย่างชัดเจน แสดงว่าโมเดลสามารถแยกแยะธุรกรรมฉ้อโกงจากธุรกรรมปกติได้ดีมาก
- Precision-Recall Curve: CatBoost และ XGBoost มีค่า PR-AUC สูงที่สุดอยู่ที่ประมาณ 0.91 – 0.92 แสดงให้เห็นความสามารถในการตรวจจับกลุ่ม fraud ได้ดีกว่าโมเดลอื่นๆ

ตาราง 5 ผลการทดสอบโมเดลหลังจากจัดการข้อมูลด้วยวิธี Tomek Links

Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.87	0.67	0.76	0.993	0.820
Random Forest	0.86	0.81	0.83	0.990	0.913
CatBoost	0.87	0.83	0.85	0.998	0.923
LightGBM	0.88	0.81	0.84	0.998	0.921
XGBoost	0.87	0.83	0.85	0.998	0.928

จากตาราง 8 ผู้วิจัยพบว่า เทคนิค Tomek Links ช่วยปรับปรุงประสิทธิภาพของโมเดลโดยรักษาสมดุลที่ดีระหว่าง Precision และ Recall เมื่อเทียบกับเทคนิค SMOTE

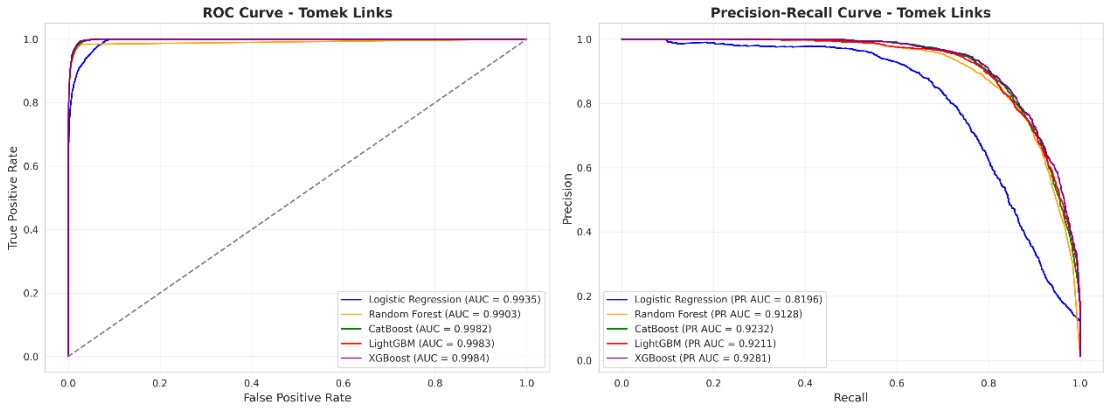
XGBoost ให้ผลลัพธ์ดีที่สุดโดยรวม ด้วยค่า F1-Score สูงสุด (0.85) และค่า Precision-Recall AUC ที่ 0.928 โดยมีค่า Precision (0.87) และ Recall (0.83) ที่สมดุล แสดงถึงความสามารถในการตรวจจับการฉ้อโกงได้อย่างแม่นยำ

CatBoost และ LightGBM มีประสิทธิภาพรองลงมา โดย CatBoost มีค่า F1-Score (0.85) ใกล้เคียงกับ XGBoost และมีค่า Recall เท่ากับ XGBoost (0.83) ขณะที่ LightGBM มีค่า Precision สูงกว่าเล็กน้อย (0.88)

Random Forest แสดงประสิทธิภาพที่สม่ำเสมอระหว่างโมเดลพื้นฐานและการใช้ Tomek Links โดยค่า F1-Score คงที่ที่ 0.83 ในทั้งสองวิธี ซึ่งบ่งชี้ว่าการใช้ Tomek Links ไม่ส่งผลกระทบต่อประสิทธิภาพโดยรวมของ Random Forest ในชุดข้อมูลนี้

Logistic Regression ยังคงมีประสิทธิภาพต่ำที่สุดในกลุ่ม โดยเฉพาะค่า Recall (0.67) แม้จะมี Precision ที่ดี (0.87) แต่ความสามารถในการตรวจจับธุรกรรมฉ้อโกงยังมีข้อจำกัด

เมื่อเปรียบเทียบกับวิธี SMOTE ผู้วิจัยพบว่า Tomek Links ให้ผลลัพธ์ที่ดีกว่าโดยรวม โดยเฉพาะค่า Precision ที่สูงกว่ามาก ซึ่งหมายถึงการลดลงของ false positive ขณะที่ยังคงรักษาความสามารถในการตรวจจับการฉ้อโกงได้ดี เหมาะสมกับการนำไปใช้ในสถานการณ์จริงที่ต้องการความสมดุลระหว่างการตรวจจับการฉ้อโกงและการลดการแจ้งเตือนที่ผิดพลาด



ภาพประกอบ 23 แสดง ROC Curve และ Precision-Recall Curve สำหรับผลการทดสอบ Tomek Links

- ROC Curve: โมเดล XGBoost แสดงประสิทธิภาพที่ดีที่สุดด้วยค่า AUC สูงสุดที่ 0.9985 ซึ่งสูงกว่าโมเดลอื่นๆ เล็กน้อย โดยที่ Random Forest มีค่า ROC-AUC ต่ำสุด (0.9903)

- Precision-Recall Curve: โมเดล XGBoost มีประสิทธิภาพสูงสุดที่ค่า PR-AUC ประมาณ 0.93 ตามมาด้วย CatBoost และ LightGBM ที่ใกล้เคียงกัน แสดงให้เห็นถึงความสมดุลที่ดีระหว่าง Precision และ Recall เมื่อใช้เทคนิค Tomek Links ซึ่งให้ผลลัพธ์ที่ดีกว่าเทคนิค SMOTE อย่างชัดเจน

ตาราง 6 ผลการทดสอบโมเดลหลังจัดการข้อมูลด้วยวิธี Hybrid (SMOTE + Tomek Links)

Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.21	0.98	0.35	0.995	0.793
Random Forest	0.63	0.89	0.74	0.994	0.852
CatBoost	0.52	0.94	0.67	0.998	0.910
LightGBM	0.43	0.96	0.60	0.997	0.904
XGBoost	0.49	0.95	0.65	0.998	0.915

จากตาราง 9 ผู้วิจัยพบว่าการใช้เทคนิค Hybrid ที่ผสมผสานระหว่าง SMOTE และ Tomek Links ส่งผลให้โมเดลทุกประเภทมีค่า Recall สูงขึ้นอย่างมาก เมื่อเทียบกับการใช้ Tomek Links เพียงอย่างเดียว

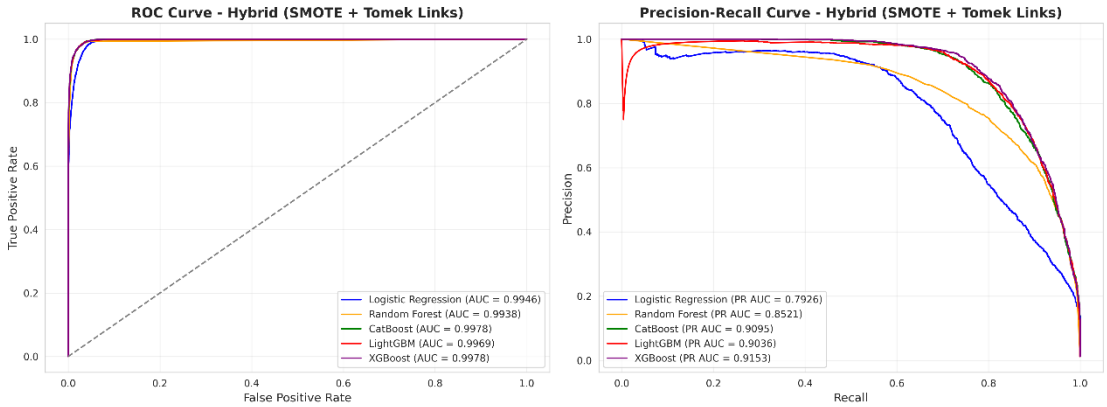
Logistic Regression มีค่า Recall สูงที่สุด (0.98) แต่มี Precision ต่ำมาก (0.21) ส่งผลให้ F1-Score ต่ำที่สุด (0.35) ซึ่งแสดงถึงจำนวนการแจ้งเตือนผิดพลาด (False Positive) ที่มากเกินไป ทำให้ไม่เหมาะสมในการนำไปใช้งานจริง

CatBoost และ XGBoost ยังคงแสดงประสิทธิภาพโดยรวมดีที่สุด โดย XGBoost มีค่า Precision-Recall AUC สูงสุด (0.915) ตามด้วย CatBoost (0.910) โดยทั้งสองโมเดลมีค่า Recall สูง (0.94 – 0.95) แต่มีค่า Precision ค่อนข้างต่ำ (0.49 – 0.52)

Random Forest มีความสมดุลระหว่าง Precision (0.63) และ Recall (0.89) ที่ดีกว่าโมเดลอื่น ส่งผลให้มีค่า F1-Score สูงสุด (0.74) แม้จะมีค่า Precision-Recall AUC ต่ำกว่ากลุ่ม Gradient Boosting

เมื่อเปรียบเทียบกับผลการทดสอบด้วย SMOTE เพียงอย่างเดียว พบว่าวิธี Hybrid ให้ผลลัพธ์ที่ใกล้เคียงกัน โดยมีค่า Precision, Recall, และ F1-Score ที่ไม่แตกต่างกันมาก แสดงให้เห็นว่าเทคนิค SMOTE มีอิทธิพลมากกว่าในวิธี Hybrid

โดยสรุป วิธี Hybrid (SMOTE + Tomek Links) มีประสิทธิภาพในการเพิ่ม Recall แต่มีข้อเสียคือทำให้ Precision ลดลงมาก ซึ่งอาจไม่เหมาะกับสถานการณ์ที่ต้องการความสมดุลระหว่างการตรวจจับการฉ้อโกงและการลดการแจ้งเตือนที่ผิดพลาด



ภาพประกอบ 24 แสดง ROC Curve และ Precision-Recall Curve สำหรับผลการทดสอบ SmoteTomek

- ROC Curve: โมเดล CatBoost และ XGBoost แสดงประสิทธิภาพสูงสุดด้วยค่า ROC-AUC ประมาณ 0.998 ตามด้วย LightGBM (0.997) ขณะที่ Logistic Regression (0.995) และ Random Forest (0.994) มีค่าต่ำกว่าเล็กน้อย อย่างไรก็ตาม ทุกโมเดลยังคงมีค่า ROC-AUC ที่สูงมาก (>0.99) แสดงถึงความสามารถในการแยกแยะที่ดี
- Precision-Recall Curve: โมเดล XGBoost และ CatBoost ให้ประสิทธิภาพดีมาก โดยเฉพาะค่า PR-AUC ที่สูงประมาณ 0.91-0.92 บ่งบอกว่าเทคนิคแบบ Hybrid ช่วยรักษาสมดุลระหว่าง Precision และ Recall ได้ดี

ตาราง 7 ผลการทดสอบโมเดลหลังจัดการข้อมูลด้วยวิธี Class Weight

Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Logistic Regression	0.19	0.99	0.32	0.995	0.798
Random Forest	0.88	0.78	0.83	0.989	0.906
CatBoost	0.47	0.96	0.63	0.998	0.919
LightGBM	0.45	0.97	0.61	0.998	0.920
XGBoost	0.52	0.93	0.67	0.998	0.916

จากตาราง 10 ผู้วิจัยพบว่าเทคนิคการถ่วงน้ำหนักคลาส (Class Weight) ส่งผลต่อประสิทธิภาพของโมเดลแตกต่างกันไป

Random Forest แสดงความสมดุลที่ดีที่สุดระหว่าง Precision (0.88) และ Recall (0.76) ส่งผลให้มี F1-Score สูงที่สุด (0.83) ในกลุ่มทั้งหมด วิธีนี้ช่วยให้ Random Forest รักษาความแม่นยำสูงไว้ได้แม้จะเพิ่มความสามารถในการตรวจจับการฉ้อโกง

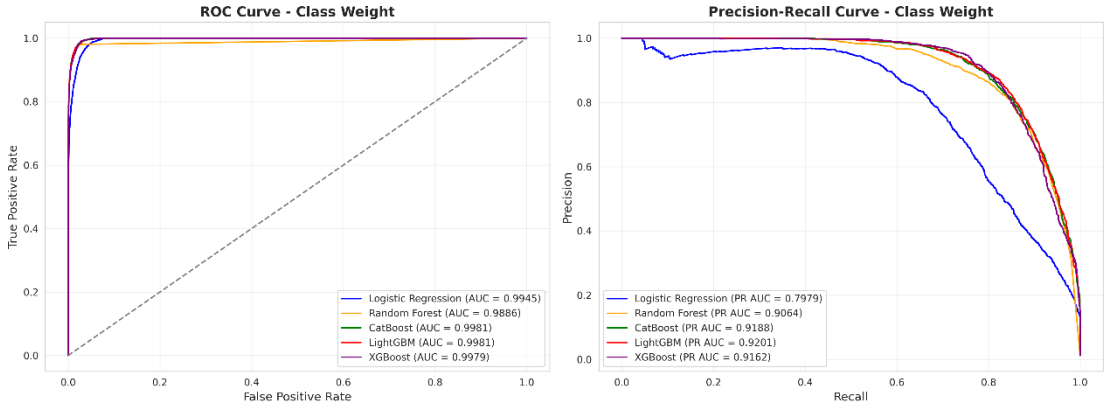
XGBoost, CatBoost และ LightGBM ได้รับผลกระทบจากการถ่วงน้ำหนักคลาสมากกว่า โดยค่า Recall เพิ่มขึ้นอย่างมาก (0.93 – 0.97) แต่ Precision ลดลงอย่างมีนัยสำคัญ (0.45 – 0.52) ทำให้ F1-Score ลดลงเมื่อเทียบกับโมเดลพื้นฐานและการใช้ Tomek Links

Logistic Regression ได้รับผลกระทบมากที่สุด มี Precision ต่ำมาก (0.19) แม้จะมี Recall สูง (0.99) ส่งผลให้มี F1-Score ต่ำที่สุด (0.32)

ด้านค่า Precision-Recall AUC พบว่าโมเดล Gradient Boosting ทั้งหมดให้ผลใกล้เคียงกัน (0.916 – 0.920) โดย LightGBM มีค่าสูงสุด (0.920) ตามมาด้วย CatBoost (0.919) และ XGBoost (0.916)

เทคนิค Class Weight ให้ผลลัพธ์คล้ายกับวิธี SMOTE โดยช่วยเพิ่ม Recall แต่มีผลทำให้ Precision ลดลง อย่างไรก็ตาม วิธีนี้มีข้อดีคือไม่จำเป็นต้องสร้างหรือลดข้อมูลจริง ทำให้การประมวลผลเร็วกว่าและไม่เสี่ยงต่อการสร้างข้อมูลสังเคราะห์ที่ไม่สมจริง ซึ่งเป็นประโยชน์ในกรณีที่มีข้อมูลขนาดใหญ่

Random Forest ดูเหมาะสมที่สุดเมื่อใช้เทคนิค Class Weight เนื่องจากสามารถรักษาความสมดุลระหว่าง Precision และ Recall ได้ดี เหมาะกับสถานการณ์ที่ต้องการทั้งความแม่นยำและความไวในการตรวจจับการฉ้อโกง



ภาพประกอบ 25 แสดง ROC Curve และ Precision-Recall Curve สำหรับผลการทดสอบ Adjust classweight

- ROC Curve: โมเดลประเภท Gradient Boosting มีค่า ROC-AUC สูงสุด (0.998) ตามด้วย Logistic Regression (0.995) และ Random Forest (0.989)
- Precision-Recall Curve: CatBoost และ LightGBM มีค่า PR-AUC สูงสุด (~0.92) ตามด้วย XGBoost (0.916) และ Random Forest (0.906) ขณะที่ Logistic Regression มีค่าต่ำสุด (0.798)
- Random Forest แสดงสมดุลที่ดีระหว่าง Precision และ Recall เมื่อใช้เทคนิค Class Weight สอดคล้องกับค่า F1-Score ที่สูง

2. **ผลการปรับแต่งโมเดล** (Hyperparameter Tuning)

การปรับแต่งค่าพารามิเตอร์ของโมเดล (Hyperparameter Tuning) เป็นขั้นตอนสำคัญในการเพิ่มประสิทธิภาพของแบบจำลองให้เหมาะสมกับข้อมูลมากที่สุด โดยในงานวิจัยนี้ ผู้วิจัยได้ดำเนินการปรับค่าพารามิเตอร์ของโมเดลด้วยเทคนิค Grid Search ร่วมกับการทำ Cross-Validation แบบ Stratified 5-Fold เพื่อให้การประเมินผลมีความเที่ยงตรงและรองรับความไม่สมดุลของข้อมูล (Class Imbalance) ได้อย่างเหมาะสม

ค่าพารามิเตอร์ที่พิจารณา ได้แก่ จำนวนต้นไม้ (n_estimators), ความลึกของต้นไม้ (max_depth), อัตราการเรียนรู้ (learning_rate), และพารามิเตอร์อื่นที่เกี่ยวข้องในแต่ละอัลกอริธึม เช่น scale_pos_weight สำหรับ XGBoost หรือ depth สำหรับ CatBoost โดยมีการประเมินผลลัพธ์จากค่า Recall เป็นหลัก เนื่องจากงานวิจัยนี้มุ่งเน้นไปที่การตรวจจับธุรกรรมที่เป็นการฉ้อโกง (fraud detection) ซึ่งเป็นกลุ่มข้อมูลส่วนน้อย (minority class) ที่ต้องให้ความสำคัญเป็นพิเศษ

ผลจากการปรับแต่งแสดงให้เห็นว่าโมเดลประเภท Boosting เช่น XGBoost, LightGBM และ CatBoost มีแนวโน้มให้ผลลัพธ์ที่แม่นยำมากขึ้น โดยเฉพาะเมื่อใช้ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุล เช่น SMOTE, SMOTE-Tomek และการปรับ class weight ซึ่งมีผลให้ค่า Recall, Precision, และ AUC เพิ่มขึ้นอย่างมีนัยสำคัญ เมื่อเปรียบเทียบกับค่าพารามิเตอร์ตั้งต้น

ตาราง 8 แสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Oversampling ด้วย SMOTE ก่อนและหลังการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning)

Technique	Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Baseline (Smote)	Logistic Regression	0.21	0.98	0.35	0.995	0.795
	Random Forest	0.63	0.89	0.74	0.992	0.851
	CatBoost	0.54	0.93	0.69	0.998	0.907
	LightGBM	0.43	0.96	0.59	0.995	0.900
	XGBoost	0.49	0.95	0.65	0.998	0.913
Tuned (Smote)	Logistic Regression	0.19	0.98	0.32	0.994	0.805
	Random Forest	0.63	0.89	0.74	0.992	0.850
	CatBoost	0.53	0.94	0.68	0.997	0.910
	LightGBM	0.47	0.95	0.62	0.994	0.896
	XGBoost	0.56	0.93	0.69	0.997	0.913

ผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค SMOTE ก่อนและหลังการปรับแต่งพารามิเตอร์

จากตาราง 11 ผู้วิจัยพบว่าการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) ไม่ได้ส่งผลให้ประสิทธิภาพของโมเดลเพิ่มขึ้นอย่างมีนัยสำคัญเมื่อใช้ร่วมกับเทคนิค SMOTE

CatBoost เป็นโมเดลที่ให้ผลดีที่สุดหลังการปรับแต่ง โดยมีค่า Recall สูง (0.94) และ PR-AUC (0.910) ที่ดี เหมาะสำหรับการตรวจจับธุรกรรมฉ้อโกงได้อย่างครอบคลุม

XGBoost แสดงการปรับปรุงที่ชัดเจนที่สุดหลังการปรับแต่ง โดย Precision เพิ่มขึ้นจาก 0.49 เป็น 0.56 และ F1-Score เพิ่มขึ้นจาก 0.65 เป็น 0.69 ขณะที่ยังรักษา Recall ในระดับสูง (0.93)

LightGBM และ Random Forest มีการเปลี่ยนแปลงเพียงเล็กน้อย ส่วน Logistic Regression กลับมีประสิทธิภาพลดลงเล็กน้อยหลังการปรับแต่ง

โดยสรุป โมเดลประเภท Gradient Boosting (CatBoost, XGBoost, LightGBM) ยังคงเป็นตัวเลือกที่เหมาะสมที่สุดสำหรับการตรวจจับการฉ้อโกงเมื่อใช้ร่วมกับเทคนิค SMOTE

ตาราง 9 แสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Undersampling ด้วย TomekLinks ก่อนและหลังการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning)

Technique	Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Baseline (TomekLinks)	Logistic Regression	0.87	0.67	0.76	0.993	0.820
	Random Forest	0.86	0.81	0.83	0.990	0.913
	CatBoost	0.87	0.83	0.85	0.998	0.923
	LightGBM	0.88	0.81	0.84	0.998	0.921
	XGBoost	0.87	0.83	0.85	0.998	0.928
Tuned (TomekLinks)	Logistic Regression	0.87	0.67	0.75	0.993	0.819
	Random Forest	0.86	0.82	0.84	0.992	0.913
	CatBoost	0.88	0.82	0.85	0.998	0.925
	LightGBM	0.88	0.82	0.85	0.998	0.922
	XGBoost	0.88	0.82	0.85	0.998	0.929

ผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค TomekLinks ก่อนและหลังการปรับแต่งพารามิเตอร์

จากตารางที่ 12 ซึ่งแสดงผลการทดลองโดยใช้เทคนิค Undersampling ด้วย Tomek Links พบว่าโมเดลที่มีประสิทธิภาพโดดเด่นที่สุดในการตรวจจับธุรกรรมฉ้อโกงคือ XGBoost หลังการปรับแต่งพารามิเตอร์ โดยให้ค่า Recall เท่ากับ 0.82 และมีค่า Precision-Recall AUC สูงถึง 0.93 ซึ่งแสดงถึงความสามารถในการตรวจจับ Fraud ได้อย่างแม่นยำและครอบคลุม

ในขณะที่ CatBoost และ LightGBM ก็ให้ผลลัพธ์ที่ใกล้เคียงกัน โดยมีค่า Recall อยู่ที่ 0.82 เช่นเดียวกัน แต่ค่า Precision-Recall AUC ต่ำกว่าเล็กน้อย ทำให้ XGBoost ยังคงเป็นตัวเลือกที่เหมาะสมที่สุดสำหรับกรณีที่ต้องการสมดุลระหว่างการตรวจจับและลดการแจ้งเตือนผิดพลาด (False Positive)

ส่วน Logistic Regression แม้จะมีค่า Precision สูง (0.87) แต่ค่า Recall ต่ำที่สุดในกลุ่ม (0.67) แสดงให้เห็นว่ามีโอกาสพลาดธุรกรรมฉ้อโกงจำนวนมาก จึงอาจไม่เหมาะในกรณีที่ต้องการลดความเสี่ยงจากการตรวจจับไม่เจอ (False Negative) ในระดับสูง

ตาราง 10 แสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Hybrid sampling ก่อนและหลังการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning)

Technique	Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Baseline (Hybrid)	Logistic Regression	0.21	0.98	0.35	0.995	0.793
	Random Forest	0.63	0.89	0.74	0.994	0.852
	CatBoost	0.52	0.94	0.67	0.998	0.910
	LightGBM	0.43	0.96	0.60	0.997	0.904
	XGBoost	0.49	0.95	0.65	0.998	0.915
Tuned (Hybrid)	Logistic Regression	0.20	0.98	0.33	0.994	0.803
	Random Forest	0.63	0.89	0.74	0.995	0.855
	CatBoost	0.53	0.94	0.68	0.997	0.910
	LightGBM	0.48	0.94	0.64	0.992	0.900
	XGBoost	0.55	0.93	0.69	0.997	0.914

ผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Hybrid sampling ก่อนและหลังการปรับแต่งพารามิเตอร์

จากตาราง 13 ผู้วิจัยพบว่าการปรับแต่งพารามิเตอร์สำหรับเทคนิค Hybrid (SMOTE-Tomek) ส่งผลให้มีการปรับปรุงประสิทธิภาพในโมเดลบางประเภท: XGBoost แสดงการปรับปรุงที่ชัดเจนที่สุด โดย Precision เพิ่มขึ้นจาก 0.49 เป็น 0.55 และ F1-Score เพิ่มขึ้นจาก 0.65 เป็น 0.69 ขณะที่ยังรักษา PR-AUC ในระดับสูง (0.914) LightGBM มีการปรับปรุงด้าน Precision จาก 0.43 เป็น 0.48 และ F1-Score จาก 0.60 เป็น 0.64 แม้ว่า Recall จะลดลงเล็กน้อยจาก 0.96 เป็น 0.94 CatBoost และ Random Forest มีการเปลี่ยนแปลงเพียงเล็กน้อยหลังการปรับแต่ง ขณะที่ Logistic Regression มีประสิทธิภาพลดลงเล็กน้อย

โดยสรุป XGBoost เป็นโมเดลที่ได้ประโยชน์มากที่สุดจากการปรับแต่งพารามิเตอร์เมื่อใช้ร่วมกับเทคนิค Hybrid sampling โดยให้ความสมดุลที่ดีระหว่างความแม่นยำและความสามารถในการตรวจจับการฉ้อโกง

ตาราง 11 แสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Class weight ก่อนและหลังการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning)

Technique	Model	Precision	Recall	F1-Score	ROC-AUC	Precision-Recall AUC
Baseline (Class weight)	Logistic Regression	0.19	0.99	0.32	0.995	0.798
	Random Forest	0.88	0.78	0.83	0.989	0.906
	CatBoost	0.47	0.96	0.63	0.998	0.919
	LightGBM	0.45	0.97	0.61	0.998	0.920
	XGBoost	0.52	0.93	0.67	0.998	0.916
Tuned (Class weight)	Logistic Regression	0.18	0.98	0.31	0.994	0.802
	Random Forest	0.67	0.89	0.76	0.997	0.894
	CatBoost	0.36	0.98	0.52	0.998	0.905
	LightGBM	0.41	0.98	0.58	0.998	0.920
	XGBoost	0.37	0.98	0.53	0.998	0.916

ผลการเปรียบเทียบประสิทธิภาพของโมเดลในเทคนิค Class weight ก่อนและหลังการปรับแต่งพารามิเตอร์

จากตารางที่ 14 ซึ่งแสดงผลการทดลองโดยใช้เทคนิค Class Weight เพื่อจัดการกับปัญหา Class Imbalance พบว่าในขั้นตอน Baseline โมเดลที่ให้ผลลัพธ์ที่ดีที่สุดคือ LightGBM โดยให้ค่า Recall สูงถึง 0.97 และมีค่า Precision-Recall AUC เท่ากับ 0.920 ซึ่งเหมาะสำหรับกรณีที่ต้องการลดโอกาสพลาดการตรวจจับธุรกรรมฉ้อโกง (False Negative) ได้อย่างมีประสิทธิภาพ

เมื่อเปรียบเทียบกับผลหลังการปรับแต่งพารามิเตอร์ (Tuned) พบว่า XGBoost ยังคงเป็นโมเดลที่มีสมรรถนะโดดเด่น โดยให้ค่า Recall เท่ากับ 0.98 และค่า Precision-Recall AUC เท่ากับ 0.916 แม้ว่า Precision จะลดลงเล็กน้อยก็ตาม

จากผลการทดลองนี้สามารถสรุปได้ว่า เทคนิค Class Weight ร่วมกับโมเดล XGBoost หรือ LightGBM เป็นทางเลือกที่ดีในการตรวจจับธุรกรรมฉ้อโกง โดยเฉพาะในกรณีที่ต้องการเน้นไปที่การไม่พลาดการตรวจจับ (Recall สูง) มากกว่าการลดการแจ้งเตือนผิดพลาด

สรุปผลการเปรียบเทียบประสิทธิภาพของโมเดลตรวจจับการฉ้อโกง

ในการศึกษานี้ ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของโมเดลตรวจจับการฉ้อโกงทางการเงินภายใต้เทคนิคจัดการปัญหา Class Imbalance 4 รูปแบบ ได้แก่ SMOTE, SMOTE-Tomek Links (Hybrid), Tomek Links (Undersampling) และ Class Weight โดยทดสอบกับโมเดล 5 ประเภท ทั้งแบบพื้นฐานและแบบปรับแต่งพารามิเตอร์

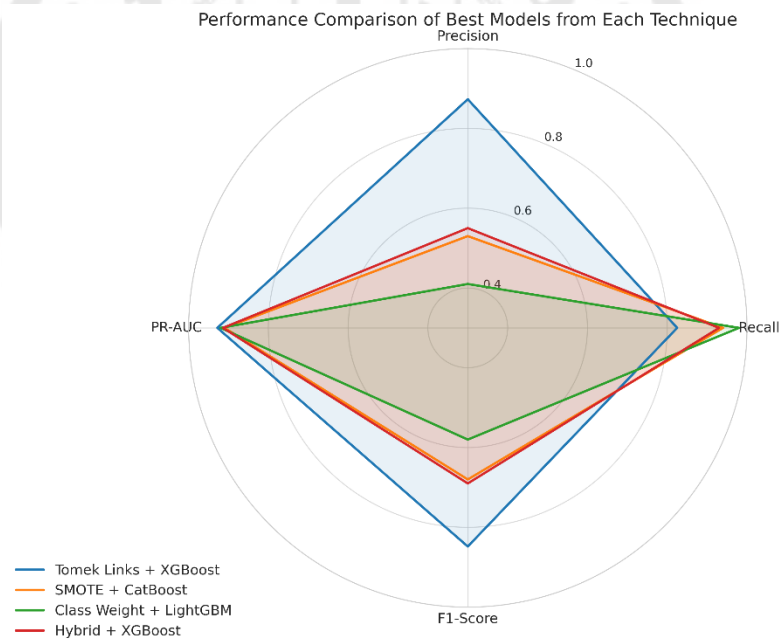
- Tomek Links + XGBoost (Tuned) ให้ผลลัพธ์ดีที่สุดโดยรวม ด้วยค่า ROC-AUC 0.998, PR-AUC 0.928 และค่า Recall 0.825 แสดงถึงความสมดุลระหว่างความแม่นยำและความไวในการตรวจจับ
- SMOTE + CatBoost (Tuned) เหมาะสำหรับกรณีที่ต้องการค่า Recall สูง (0.940) พร้อมด้วยค่า PR-AUC ที่ดี (0.910) แม้จะแลกมาด้วยค่า Precision ที่ต่ำ (0.530)
- Class Weight + LightGBM (Tuned) โดดเด่นด้านค่า Recall สูง (0.980) และ PR-AUC (0.920) เหมาะกับงานที่ต้องการตรวจจับธุรกรรมฉ้อโกงให้ได้มากที่สุด
- Hybrid (SMOTE-Tomek) + XGBoost (Tuned) มีประสิทธิภาพดีด้วยค่า Recall 0.930 และ PR-AUC 0.914 โดยมี Precision (0.550) สูงกว่าการใช้ SMOTE เพียงอย่างเดียว

กรณีต้องการความสมดุล ใช้ XGBoost ร่วมกับ Tomek Links ที่ผ่านการปรับพารามิเตอร์ เนื่องจากให้ค่า Precision และ Recall ที่สมดุล

กรณีต้องการตรวจจับครบถ้วน ใช้ LightGBM (Tuned) ร่วมกับ Class Weight หรือ CatBoost ร่วมกับ SMOTE ซึ่งให้ค่า Recall สูงกว่า 0.940

กรณีต้องการประมวลผลเร็ว วิธี Class Weight มีข้อได้เปรียบเรื่องความเร็วและไม่ต้องสร้างข้อมูลสังเคราะห์ โดย Random Forest (Baseline) ให้ความสมดุลที่ดีด้วย Precision 0.883 และ Recall 0.775

การศึกษานี้แสดงให้เห็นว่าการเลือกเทคนิคจัดการ Class Imbalance ที่เหมาะสมร่วมกับโมเดลที่เหมาะสม มีความสำคัญอย่างยิ่งต่อประสิทธิภาพในการตรวจจับการฉ้อโกงทางการเงิน โดยเฉพาะในสถานการณ์ที่มีข้อจำกัดหรือเป้าหมายเฉพาะทาง



ภาพประกอบ 26 แสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลที่ดีที่สุดหลังจากผ่านการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning)

จากภาพประกอบ 26 ผู้วิจัยพบว่าแต่ละเทคนิคจัดการ Class Imbalance มีจุดแข็งที่แตกต่างกัน

Tomek Links + XGBoost (สีน้ำเงิน) มีความสมดุลดีที่สุด ด้วยค่า Precision สูงสุด (0.87) และค่า PR-AUC สูงสุด (0.928) Class Weight + LightGBM (สีเขียว) เน้นการตรวจจับสูงสุด ด้วย Recall สูงถึง 0.98 แต่แลกด้วย Precision ต่ำ (0.41) SMOTE + CatBoost (สีส้ม) และ Hybrid + XGBoost (สีแดง) ให้ผลลัพธ์คล้ายกัน โดยมี Recall สูง (0.93 – 0.94) และ Precision ปานกลาง (0.53 – 0.55) กราฟนี้เป็นประโยชน์อย่างยิ่งในการเลือกโมเดลให้เหมาะกับความต้องการเฉพาะ เช่น หากต้องการความสมดุลที่ดีควรเลือก Tomek Links + XGBoost แต่หากต้องการเน้นตรวจจับให้ได้มากที่สุดควรเลือก Class Weight + LightGBM

ตารางนี้แสดง พารามิเตอร์ที่ดีที่สุด (Best Parameters) สำหรับแต่ละเทคนิคในการปรับแต่งโมเดล (Hyperparameter Tuning) โดยใช้วิธี Grid Search ร่วมกับ 5-Fold Cross-Validation โดยสามารถอธิบายรายละเอียดได้ดังนี้

ตาราง 12 Best Parameters สำหรับแต่ละเทคนิค

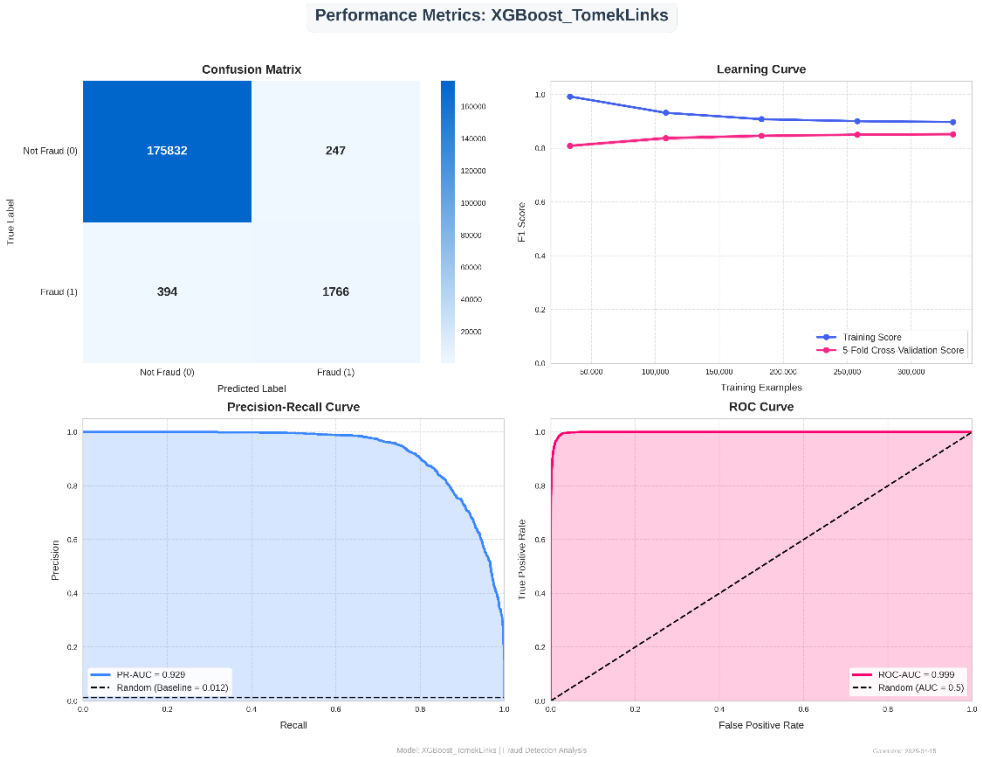
Technique	Best Model	Best Parameters	อธิบายเพิ่มเติม
SMOTE	Catboost	depth=10, iterations=200, learning_rate=0.1	เทคนิค SMOTE ช่วยเพิ่มจำนวนข้อมูลในคลาส Fraud เพื่อทำให้โมเดลเรียนรู้ได้ดีขึ้น ส่งผลให้ค่า Recall ของ CatBoost สูงถึง 0.94 โดยที่ยังรักษา PR-AUC ที่ 0.91
Hybrid	XGBoost	learning_rate=0.1, max_depth=10, n_estimators=200	เทคนิค Hybrid ผสมผสานระหว่างการเพิ่มข้อมูลในคลาส Fraud และลบข้อมูลซ้ำในคลาส Non-Fraud ส่งผลให้ค่า Recall สูง (0.93) และได้ค่า PR-AUC ที่ดี (0.914)
ClassWeight	LightGBM	learning_rate=0.1, max_depth=10, n_estimators=100, class_weight= {0:1, 1:81}	การกำหนด Class Weight ทำให้โมเดลให้ความสำคัญกับคลาส Fraud มากขึ้น ส่งผลให้ Recall สูงสุดถึง 0.98 แม้ว่าค่า Precision จะต่ำ (0.41) ซึ่งเหมาะกับกรณีที่ต้องการตรวจจับธุรกรรมฉ้อโกงให้ได้มากที่สุด

Technique	Best Model	Best Parameters	อธิบายเพิ่มเติม
Tomek Links	XGBoost	learning_rate=0.1, max_depth=6, n_estimators=200	เทคนิค Tomek Links ช่วยลดปัญหาข้อมูลซ้ำซ้อนในคลาส Non-Fraud ทำให้ค่า Precision สูงถึง 0.873 และค่า PR-AUC สูงสุดที่ 0.928 แสดงถึงความสมดุลที่ดีระหว่างความแม่นยำและความไวในการตรวจจับ

3. การเปรียบเทียบประสิทธิภาพโมเดล

ในงานวิจัยนี้ ผู้วิจัยมุ่งเน้นการเปรียบเทียบประสิทธิภาพของโมเดลที่ได้รับการปรับแต่งพารามิเตอร์แล้ว เพื่อให้เห็นความแตกต่างและจุดเด่นของแต่ละโมเดลที่ใช้เทคนิคจัดการข้อมูลไม่สมดุลต่างกัน เนื่องจากปัญหาการฉ้อโกงทางการเงินมักมีอัตราส่วนของธุรกรรมฉ้อโกงต่ำมาก ผู้วิจัยจึงคัดเลือกโมเดลที่โดดเด่นในการแก้ปัญหาข้อมูลไม่สมดุลในหลายรูปแบบ ทั้งการเพิ่มข้อมูล (SMOTE) การลดข้อมูลซ้ำซ้อน (Tomek Links) การผสมผสานสองเทคนิค (SMOTE + Tomek Links) และการกำหนดน้ำหนักให้คลาสฉ้อโกง (Class Weight)

3.1 กรณีที่ต้องการความสมดุลระหว่าง Recall และ Precision โมเดลที่เหมาะสมได้แก่ XGBoost ร่วมกับ Tomek Links ที่ผ่านการปรับพารามิเตอร์ เนื่องจากให้ค่า Precision และ Recall ที่สมดุล

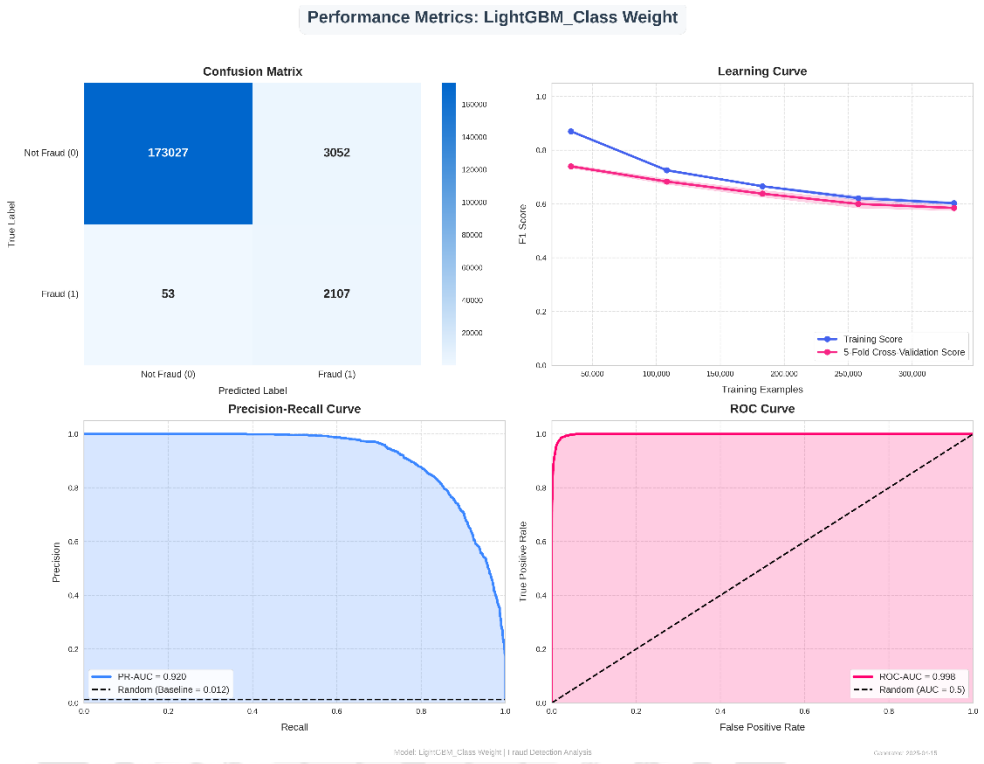


ภาพประกอบ 27 แสดงประสิทธิภาพของโมเดล XGBoost ร่วมกับ Tomek Links

ผู้วิจัยพบว่าโมเดลนี้มีความสมดุลระหว่าง Precision และ Recall ดีที่สุด โดยมี Precision ที่ 0.877 และ Recall ที่ 0.818 (1,766 จาก 2,160 กรณี) ส่งผลให้มี F1-Score ที่ดีกว่าโมเดลอื่น แม้จะพลาดการตรวจจับธุรกรรมฉ้อโกงมากกว่า แต่มี False Positive เพียง 247 กรณีเท่านั้น

Learning Curve มีช่องว่างระหว่างผลการเรียนรู้บนชุดฝึกสอนและทดสอบพอสมควร แต่มีแนวโน้มเข้าใกล้กันเมื่อข้อมูลเพิ่มมากขึ้น ผู้วิจัยวัดค่า PR-AUC ได้สูงสุดที่ 0.929 และ ROC-AUC สูงสุดที่ 0.9

3.2 กรณีต้องการตรวจจับครบถ้วน ใช้ LightGBM ร่วมกับ Class Weight ที่ผ่านการปรับพารามิเตอร์ เนื่องจากให้ค่า Recall สูงที่สุด

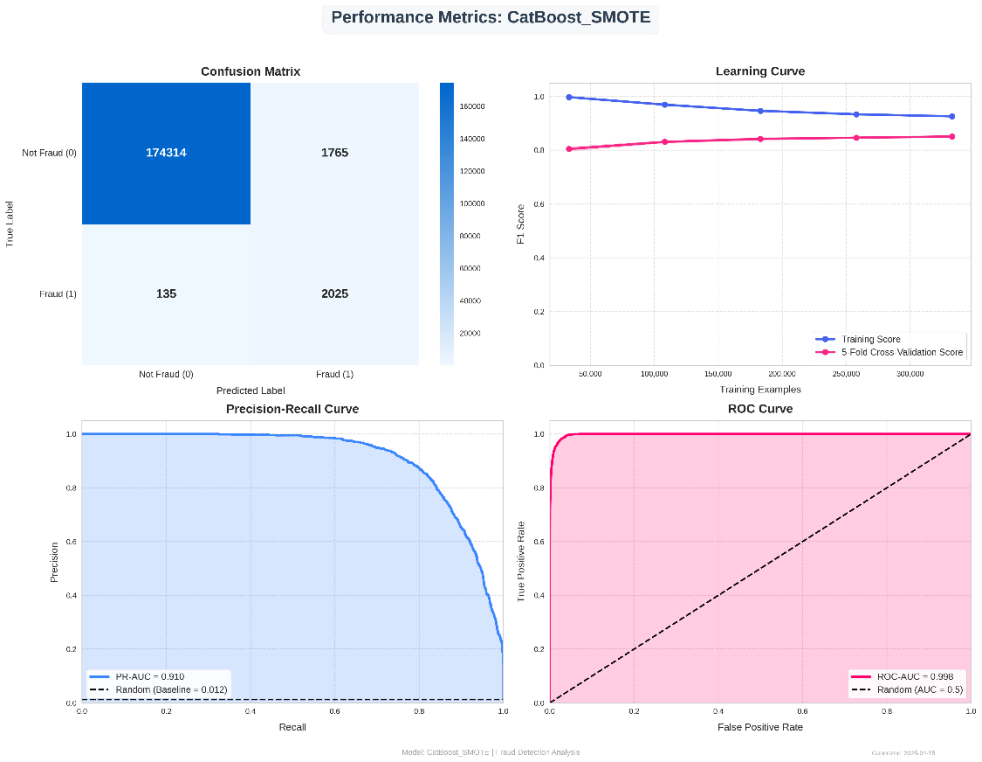


ภาพประกอบ 28 แสดงประสิทธิภาพของโมเดล LightGBM ร่วมกับ Class weight

ผู้วิจัยพบว่าโมเดลนี้มีค่า Recall สูงที่สุดในทุกการทดลองที่ 0.975 (2,107 จาก 2,160 กรณี) แต่แลกมาด้วย Precision ที่ต่ำมาก (0.408) เนื่องจากมี False Positive สูงถึง 3,052 กรณี

Learning Curve มีช่องว่างระหว่างผลการเรียนรู้บนชุดฝึกสอนและทดสอบค่อนข้างกว้างและมีแนวโน้มลดลงเมื่อข้อมูลเพิ่มขึ้น ผู้วิจัยสังเกตว่ามีการ overfitting เล็กน้อย แต่ยังมี PR-AUC ที่ดี (0.920) และ ROC-AUC สูง (0.998)

3.3 กรณีต้องการความสมดุลระหว่างประสิทธิภาพโดยรวมและการตรวจจับที่ครอบคลุม ใช้ CatBoost ร่วมกับ SMOTE ที่ผ่านการปรับพารามิเตอร์ เนื่องจากให้ค่า Recall สูง

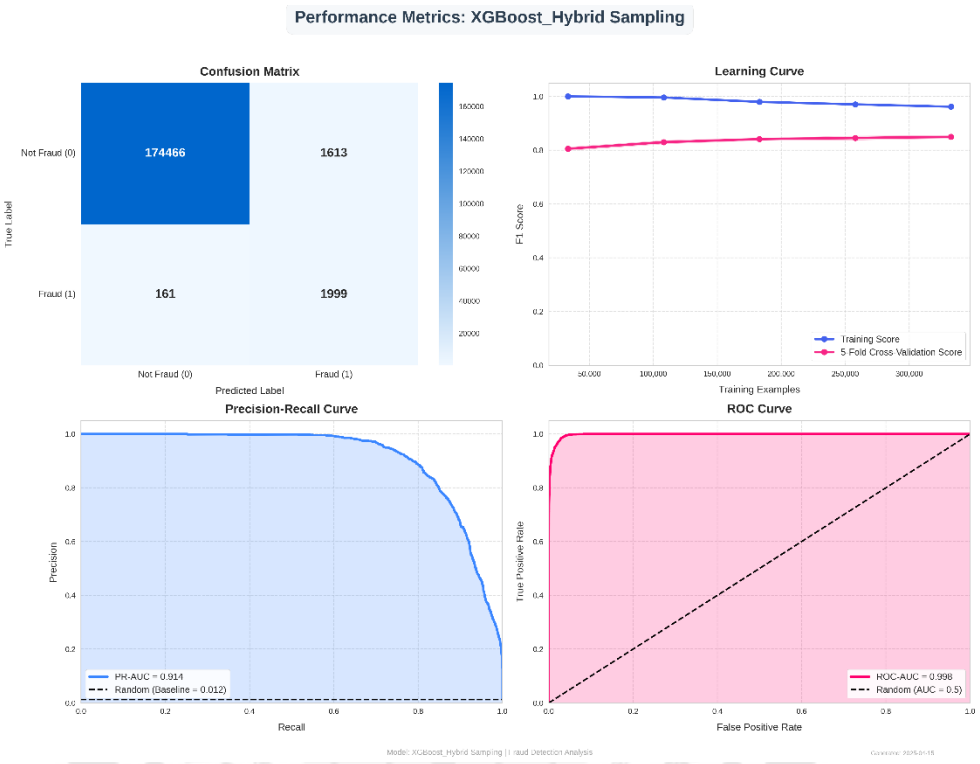


ภาพประกอบ 29 แสดงประสิทธิภาพของโมเดล Catboost ร่วมกับ SMOTE

ผู้วิจัยพบว่าจาก Confusion Matrix โมเดลสามารถตรวจจับธุรกรรมฉ้อโกงได้ถึง 2,025 จาก 2,160 กรณี คิดเป็น Recall 0.938 แต่มี False Positive จำนวน 1,765 กรณี ส่งผลให้มี Precision เพียง 0.534

Learning Curve แสดงให้เห็นว่าโมเดลมีเสถียรภาพดี โดยมีค่า F1-Score ทั้งบนชุดข้อมูลฝึกสอนและทดสอบใกล้เคียงกันและคงที่เมื่อข้อมูลเพิ่มขึ้น ผู้วิจัยวัดค่า PR-AUC ได้ 0.910 และ ROC-AUC สูงถึง 0.998 แสดงถึงความสามารถในการแยกแยะที่ดีมาก

3.4 กรณีต้องการประสิทธิภาพโดยรวมที่สูงในการตรวจจับการฉ้อโกง ใช้ XGBoost ร่วมกับ Hybrid Sampling ที่ผ่านการปรับพารามิเตอร์ เนื่องจากให้ค่า PR-AUC สูงสุด



ภาพประกอบ 30 แสดง Confusion Matrix ของโมเดล XGBoost ที่ใช้เทคนิค Hybrid sampling

ผู้วิจัยพบว่าโมเดลมีค่า Recall ดีมาก (0.925) สามารถตรวจจับธุรกรรมฉ้อโกงได้ 1,999 จาก 2,160 กรณี โดยมี Precision ที่ 0.553 ซึ่งสูงกว่า LightGBM แต่ยังไม่ดีนัก

Learning Curve มีความเสถียรดีมาก โดยทั้งค่าบนชุดฝึกสอนและทดสอบมีค่าใกล้เคียงกันและคงที่ แสดงว่าโมเดลไม่มีปัญหา overfitting หรือ underfitting ผู้วิจัยวัดค่า PR-AUC ได้ 0.914 และ ROC-AUC ที่ 0.998

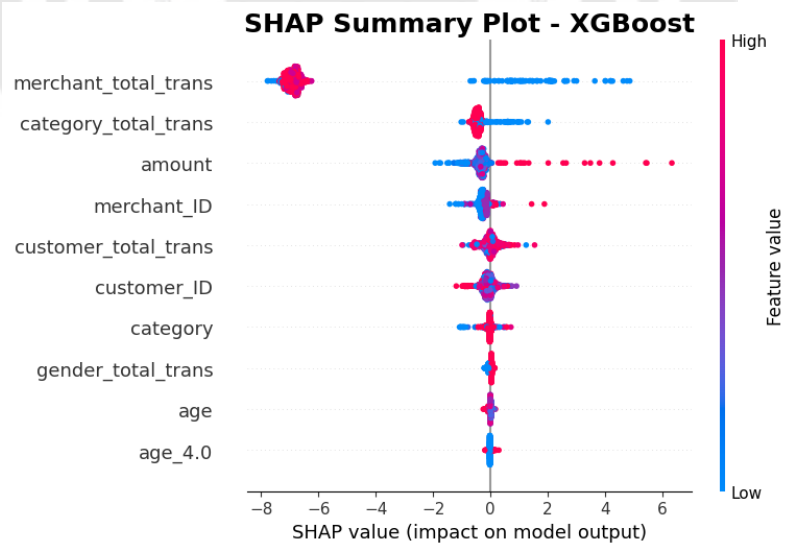
ผู้วิจัยจึงมองว่าแต่ละโมเดลมีจุดแข็งแตกต่างกัน โดยขึ้นอยู่กับเป้าหมายขององค์กรว่าจะเน้นจับธุรกรรมฉ้อโกงให้ได้มากที่สุด (Recall สูง) หรือเน้นสมดุลระหว่างการแจ้งเตือนผิดกับการพลาดธุรกรรมฉ้อโกง (Precision-Recall AUC สูง) ซึ่ง XGBoost (Tomek Links) ดูจะเป็นตัวเลือกที่ลงตัวที่สุดในภาพรวม หากให้น้ำหนักเรื่อง False Positive และ False Negative ในระดับที่สมดุล

4. การวิเคราะห์ความสำคัญของฟีเจอร์

การวิเคราะห์ความสำคัญของฟีเจอร์เป็นขั้นตอนสำคัญในการเข้าใจว่าฟีเจอร์ใดส่งผลต่อการทำนายของโมเดลมากที่สุด สำหรับโมเดลประเภท Boosting ผู้วิจัยสามารถคำนวณได้จากค่าในกระบวนการแยกต้นไม้ ในการทดลองนี้ ผู้วิจัยเลือกใช้ XGBoost ร่วมกับ Tomek Links เป็นหลัก เนื่องจากให้ค่า ROC-AUC และ F1-Score สูงสุด และได้นำ SHAP Values มาวิเคราะห์เพื่อดูผลกระทบของแต่ละฟีเจอร์ที่มีต่อการทำนาย

อย่างไรก็ตาม หากต้องการลดความเสี่ยงจาก False Negative ให้เหลือน้อยที่สุด LightGBM ร่วมกับ Class Weight จะเหมาะสมกว่าเนื่องจากให้ค่า Recall สูงกว่า แม้จะแลกมาด้วย False Positive ที่มากขึ้น ผลลัพธ์จากการวิเคราะห์ความสำคัญของฟีเจอร์จึงช่วยให้สามารถเลือกโมเดลและวิธีจัดการข้อมูลไม่สมดุลที่สอดคล้องกับเป้าหมายทางธุรกิจได้อย่างเหมาะสม

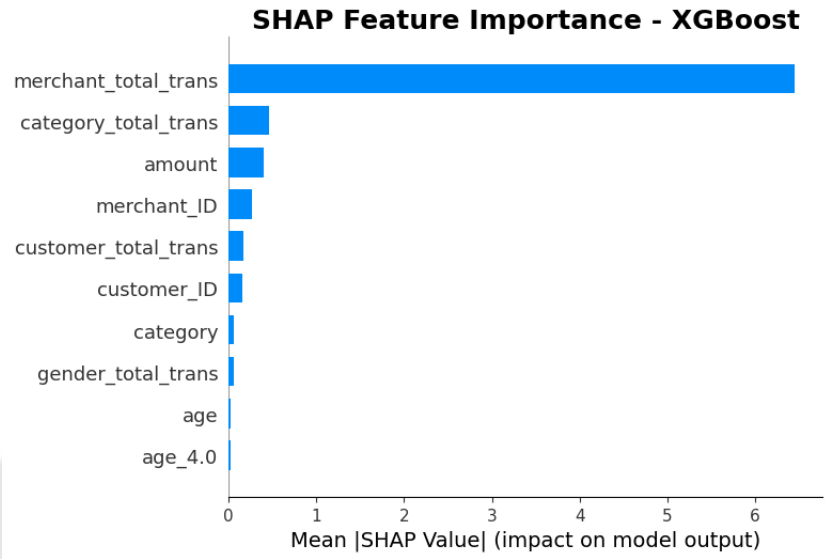
4.1 ผลการทดสอบ Undersampling ร่วมกับ Tomek Links ในโมเดล XGBoost โดยใช้ SHAP analysis



ภาพประกอบ 31 แสดง SHAP Summary Plot ของโมเดล XGBoost

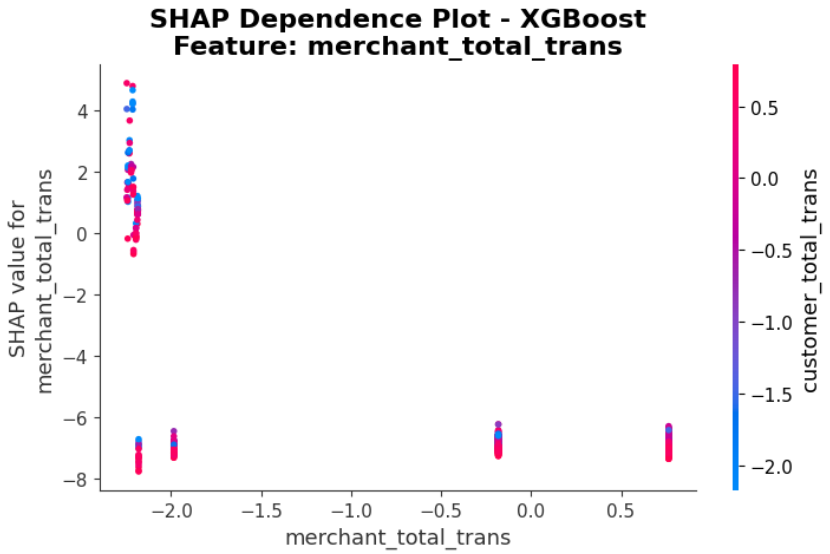
- ฟีเจอร์ merchant_total_trans มีผลกระทบต่อโมเดลสูงที่สุด โดยค่า SHAP ที่สูงทำให้โมเดลมีแนวโน้มคาดการณ์ว่าเป็นธุรกรรมที่มีโอกาสเป็น Fraud

- ฟีเจอร์ category_total_trans และ amount มีผลกระทบรองลงมา โดยเฉพาะเมื่อมีค่าสูง ส่งผลให้โมเดลมีแนวโน้มคาดการณ์ว่าเป็น Fraud
- ฟีเจอร์ age และ age_4.0 มีค่า SHAP Value ต่ำมาก ซึ่งแสดงให้เห็นว่าฟีเจอร์เหล่านี้มีผลต่อการคาดการณ์น้อย



ภาพประกอบ 32 แสดงค่าเฉลี่ยของ SHAP Value สำหรับแต่ละฟีเจอร์ในโมเดล XGBoost

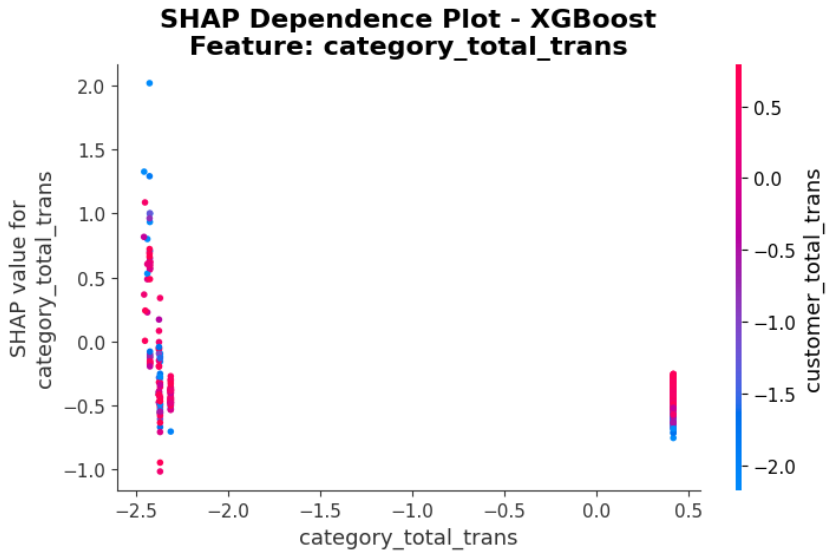
- ฟีเจอร์ merchant_total_trans มีค่าเฉลี่ย SHAP Value สูงที่สุด (~6.5) เป็นฟีเจอร์ที่มีผลกระทบมากที่สุดต่อการตัดสินใจของโมเดล
- ฟีเจอร์ category_total_trans และ amount มีผลกระทบระดับรองลงมา (~1.0 – 1.5)
- ฟีเจอร์ merchant_ID และ customer_total_trans มีผลกระทบระดับปานกลาง (~ 0.5)
- ฟีเจอร์ที่มีผลน้อย เช่น category, gender_total_trans, age, และ age_4.0 มีค่าเฉลี่ย SHAP Value ใกล้ศูนย์ แสดงว่ามีผลต่อการคาดการณ์น้อยมาก



ภาพประกอบ 33 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ merchant_total_trans กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ merchant_total_trans บนโมเดล XGBoost พบว่า

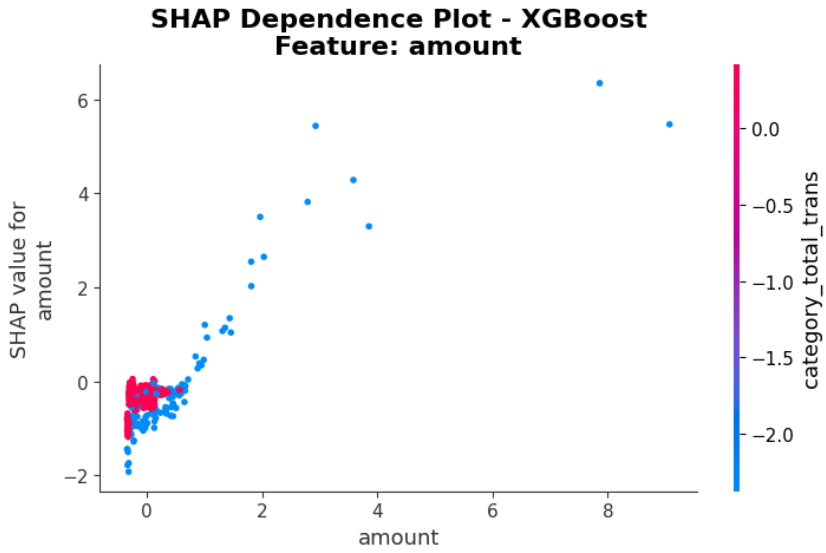
- เมื่อ merchant_total_trans มีค่าสูง (ร้านค้ามีธุรกรรมจำนวนมาก) ค่า SHAP Value จะเป็น ค่าลบ มากขึ้น โมเดลจะ ลด โอกาสทำนายว่าเป็น Fraud
- เมื่อ merchant_total_trans มีค่าต่ำ (ร้านค้ามีธุรกรรมน้อย) ค่า SHAP Value จะเป็นค่าบวก โมเดลจะ เพิ่ม โอกาสทำนายว่าเป็น Fraud
- หาก customer_total_trans สูง (จุดสีแดงเข้ม) ในช่วงที่ merchant_total_trans ต่ำ จะยิ่งผลักค่า SHAP Value ให้เป็นบวกมากขึ้น แสดงว่า ร้านค้าที่มีธุรกรรมน้อยแต่ลูกค้ามีประวัติทำธุรกรรมเยอะ จะถูกมองว่ามีความเสี่ยง Fraud มากขึ้น



ภาพประกอบ 34 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ category_total_trans กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ category_total_trans บนโมเดล XGBoost พบว่า

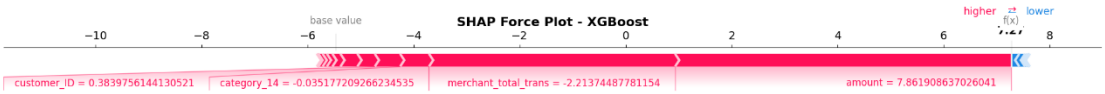
- เมื่อ category_total_trans มีค่าสูง (ธุรกรรมในหมวดหมู่นั้นมาก) ค่า SHAP Value จะเป็นค่าลบ โมเดลจะลดโอกาสทำนายว่าเป็น Fraud
- เมื่อ category_total_trans มีค่าต่ำ (ธุรกรรมน้อย) ค่า SHAP Value จะเป็นค่าบวก โมเดลจะเพิ่มโอกาสทำนายว่าเป็น Fraud
- หาก customer_total_trans สูง (จุดสีแดง) ในช่วงที่ category_total_trans ต่ำ จะยิ่งผลัก SHAP Value ให้เป็นบวกมากขึ้น แสดงว่า หมวดหมู่นั้นมีธุรกรรมน้อย แต่ลูกค้ารายนี้มีประวัติทำธุรกรรมเยอะ จึงเพิ่มความเสี่ยงการทำนายเป็น Fraud



ภาพประกอบ 35 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ amount กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ amount บนโมเดล XGBoost พบว่า

- เมื่อ amount มีค่าสูง (ยอดเงินในแต่ละธุรกรรมมาก) ค่า SHAP Value จะเป็น ค่าบวก มากขึ้น โมเดลจะ เพิ่ม โอกาสทำนายว่าเป็น Fraud
- เมื่อ amount มีค่าต่ำ (ยอดเงินน้อยหรือใกล้ศูนย์) ค่า SHAP Value จะใกล้ ศูนย์หรือลบ โมเดลจะ ลด โอกาสทำนายว่าเป็น Fraud
- ยิ่งยอดเงินสูง โมเดลยิ่งมองว่ามีโอกาสเป็น Fraud มากขึ้น และถ้าอยู่ในหมวดหมู่ที่คนใช้บ่อย (สีแดง) ก็จะยิ่งเพิ่มน้ำหนักความน่าสงสัย

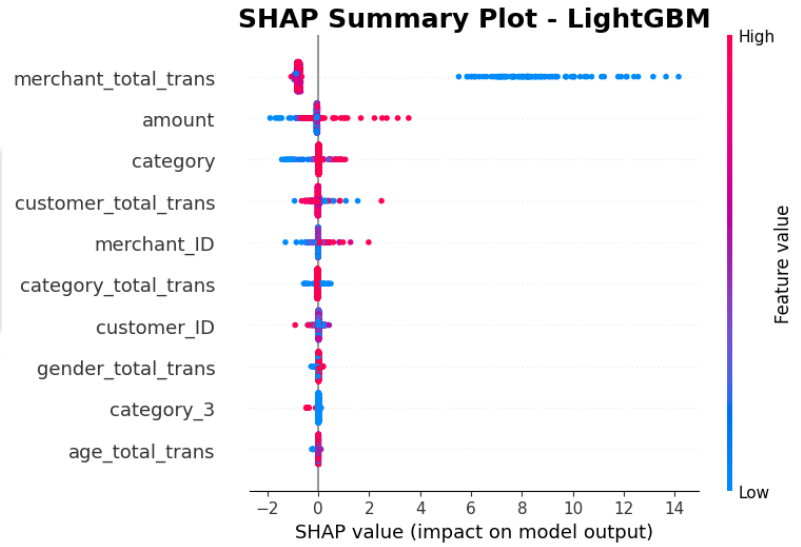


ภาพประกอบ 36 แสดง SHAP Force Plot จากโมเดล XGBoost

จาก SHAP Force Plot บนโมเดล XGBoost พบว่า Base Value (~ –3.0): จุดเริ่มต้นก่อนพิจารณาฟีเจอร์ใดๆ

- แรงดันบวก (pink bars) amount (+7.86): ปัจจัยหลักที่ผลักคะแนนให้สูง จนโมเดลคาดว่าเป็น Fraud , customer_ID (+0.38): เสริมแนวโน้ม Fraud เล็กน้อย
- แรงดันลบ (blue bars) merchant_total_trans (–2.21): ร้านค้ามีธุรกรรมเยอะ ลดโอกาสFraud , category_14 (–0.04): หมวดหมู่ 14 ลดโอกาส Fraud เล็กน้อย ผลรวมแรงดันทั้งหมดยังคงเป็นบวกมากกว่า Base Value จึงทำให้โมเดลตัดสินว่า ธุรกรรมนี้เป็น Fraud

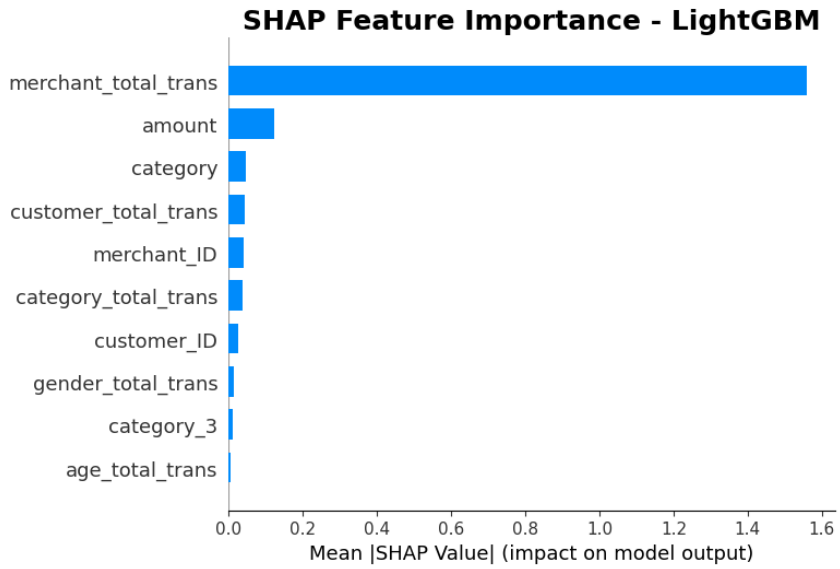
4.2 ผลการทดสอบ Class-weight ร่วมกับ LightGBM ในโมเดล XGBoost โดยใช้ SHAP analysis



ภาพประกอบ 37 แสดง SHAP Summary Plot ของโมเดล LightGBM

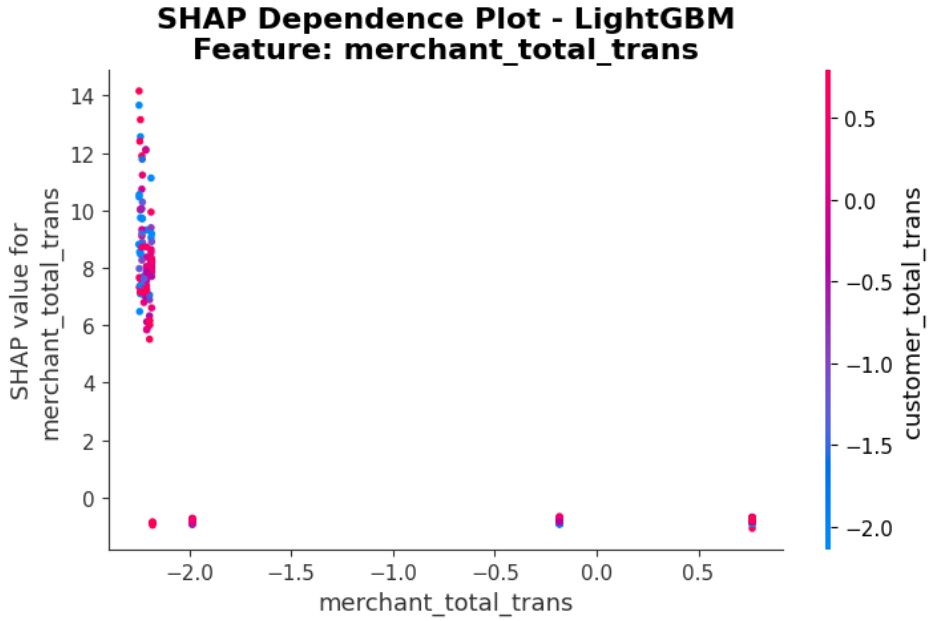
- ฟีเจอร์ merchant_total_trans มีผลกระทบต่อโมเดลสูงที่สุด โดยค่าฟีเจอร์สูง (จุดสีแดง) มักให้ SHAP Value เป็นบวกมาก โมเดลจะเพิ่มโอกาสคาดว่าเป็น Fraud
- ฟีเจอร์ amount และ category อยู่ในลำดับถัดมา เมื่อค่าฟีเจอร์สูง SHAP Value จะเป็นบวก โมเดลมีแนวโน้มคาดการณ์ว่าเป็น Fraud
- ฟีเจอร์ customer_total_trans และ merchant_ID มีผลปานกลาง จุดกระจาย SHAP Value ร่วมกับสีฟีเจอร์บ่งชี้ทิศทางบ้างเล็กน้อย

- ฟีเจอร์ category_total_trans, customer_ID, gender_total_trans, category_3 และ age_total_trans มีค่า SHAP Value กระจุกตัวใกล้ศูนย์มาก ⟶ แสดงว่าฟีเจอร์เหล่านี้มีอิทธิพลต่อการคาดการณ์ Fraud น้อยที่สุด



ภาพประกอบ 38 แสดงค่าเฉลี่ยของ SHAP Value สำหรับแต่ละฟีเจอร์ในโมเดล LightGBM

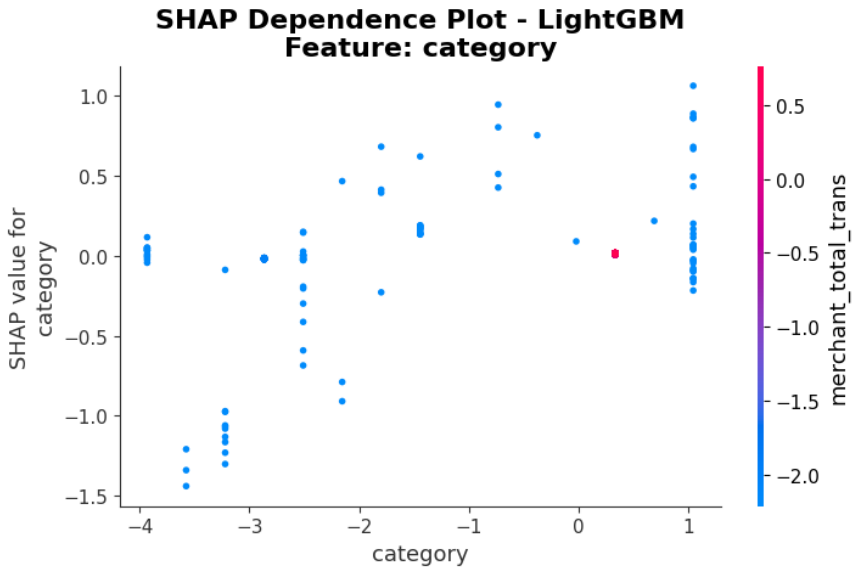
แผนภูมิแท่งสรุป Mean Absolute SHAP Value ของแต่ละฟีเจอร์ โดย merchant_total_trans และ amount ยังครองอันดับต้น ยืนยันว่าฟีเจอร์สองตัวนี้มีอิทธิพลสูงสุดต่อการตัดสินใจของโมเดล



ภาพประกอบ 39 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ merchant_total_trans กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ merchant_total_trans บนโมเดล LightGBM พบว่า

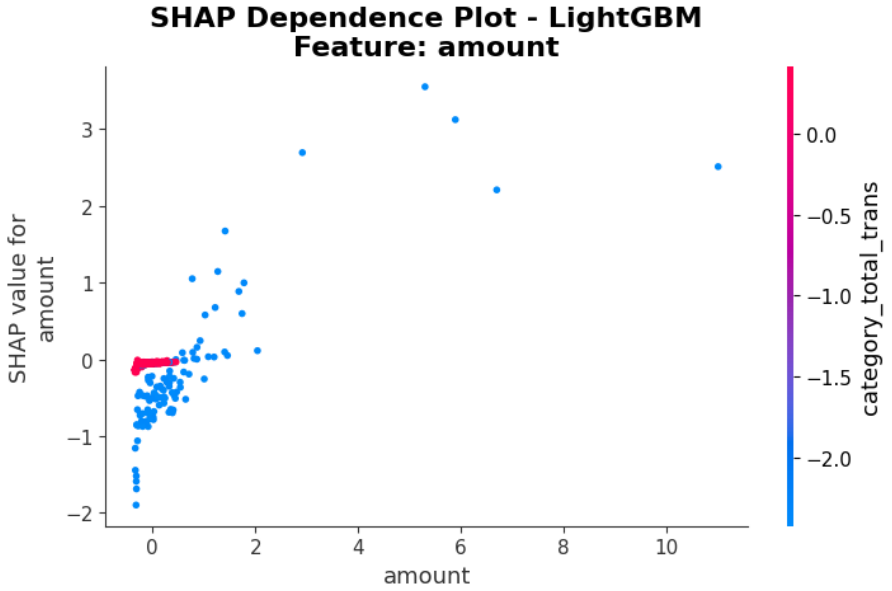
- เมื่อ merchant_total_trans มีค่าสูง (ร้านค้ามีธุรกรรมจำนวนมาก) ค่า SHAP Value จะเป็น ค่าลบ มากขึ้น โมเดลจะ ลด โอกาสทำนายว่าเป็น Fraud
- เมื่อ merchant_total_trans มีค่าต่ำ (ร้านค้ามีธุรกรรมน้อย) ค่า SHAP Value จะเป็นค่าบวก โมเดลจะ เพิ่ม โอกาสทำนายว่าเป็น Fraud
- หาก customer_total_trans สูง (จุดสีแดง) ในช่วงที่ merchant_total_trans ต่ำ จะยิ่งผลัก SHAP Value ให้เป็นบวกมากขึ้น แสดงว่า ร้านค้าที่มีธุรกรรมน้อย แต่ลูกค้ารายนี้มีประวัติทำธุรกรรมเยอะ จะถูกมองว่าสงสัย Fraud มากขึ้น



ภาพประกอบ 40 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ category กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ category บนโมเดล LightGBM พบว่า

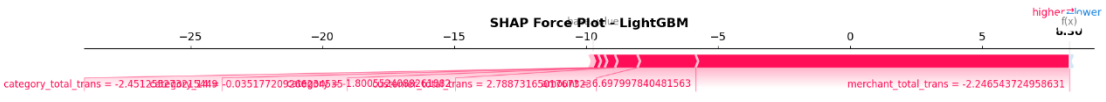
- เมื่อ category มีค่าต่ำ (เช่น –4 ถึง –3) ค่า SHAP Value จะเป็น ค่าลบ ⟶ โมเดลจะ ลด โอกาสทำนายว่าเป็น Fraud
- เมื่อ category มีค่ากลางถึงสูง (ประมาณ –2 ถึง +1) ค่า SHAP Value จะเป็น ค่าบวก โมเดลจะ เพิ่ม โอกาสทำนายว่าเป็น Fraud
- ยิ่ง merchant_total_trans สูง (จุดสีแดง) ในช่วงค่า category เดียวกัน จะยิ่งผลักSHAP Value ให้เป็นบวกสูงขึ้นเล็กน้อยแสดงว่า หมวดหมู่ธุรกรรมนั้น เมื่อเกิดกับร้านค้าที่มีประวัติทำธุรกรรมหนาแน่น จะถูกมองว่ามีความเสี่ยง Fraud มากขึ้น



ภาพประกอบ 41 แสดงความสัมพันธ์ ระหว่างฟีเจอร์ amount กับผลลัพธ์ของโมเดลผ่านค่า SHAP Value

จาก SHAP Dependence Plot ของฟีเจอร์ amount บนโมเดล LightGBM พบว่า

- เมื่อ amount มีค่าสูง (ยอดเงินในแต่ละธุรกรรมมาก) ค่า SHAP Value จะเป็น ค่าบวก มากขึ้น โมเดลจะ เพิ่ม โอกาสทำนายว่าเป็น Fraud
- เมื่อ amount มีค่าต่ำ (ยอดเงินน้อยหรือใกล้ศูนย์) ค่า SHAP Value จะเป็น ค่าลบหรือใกล้ศูนย์ โมเดลจะ ลด โอกาสทำนายว่าเป็น Fraud สีของจุดแสดงค่า category_total_trans (แดง = หมวดหมู่มีธุรกรรมเยอะ, น้ำเงิน = หมวดหมู่น้อย)
- ในระดับ amount เดียวกัน จุดสีแดงมักให้ SHAP Value บวกสูงกว่าเล็กน้อย ยอดเงินสูงในหมวดหมู่ที่คนใช้บ่อย ยิ่งถูกมองว่าน่าสงสัย Fraud มากขึ้น จุดสีน้ำเงินจะให้ SHAP Value บวกน้อยกว่า หรือแม้แต่ลบเมื่อ amount ไม่สูง



ภาพประกอบ 42 แสดง SHAP Force Plot จากโมเดล Light GBM

จาก SHAP Force Plot บนโมเดล LightGBM Base Value ≈ –10: จุดเริ่มต้นก่อนพิจารณาฟีเจอร์ใดๆ

- amount (+6.30): แรงดันโมเดลให้คาดว่าเป็น Fraud
- merchant_total_trans (–2.25) และ category_total_trans (–2.45) แรงดันลบดึงคะแนนไปฝั่ง Non-Fraud ผลรวมแรงดันทั้งหมด > 0 โมเดลตัดสินว่า ธุรกรรมนี้เป็น Fraud

จากการวิเคราะห์ความสำคัญของคุณลักษณะ (Feature Importance) โดยใช้โมเดลที่มีประสิทธิภาพสูงสุดร่วมกับเทคนิค SHAP (SHapley Additive exPlanations) พบข้อมูลเชิงลึกที่น่าสนใจดังนี้

คุณลักษณะที่มีอิทธิพลสูงสุดต่อการตัดสินใจของโมเดลคือข้อมูลที่เกี่ยวข้องกับพฤติกรรมของร้านค้าและมูลค่าธุรกรรม โดยเฉพาะ amount และ merchant_total_trans สำหรับ amount พบความสัมพันธ์ที่ชัดเจน เมื่อยอดเงินธุรกรรมสูงผิดปกติ ค่า SHAP จะเป็นบวก ส่งผลให้โมเดลมีแนวโน้มทำนายว่าเป็นธุรกรรมฉ้อโกง (Fraud) มากขึ้น ในขณะที่ merchant_total_trans แสดงความสัมพันธ์ที่ซับซ้อนกว่า ร้านค้าที่มีประวัติธุรกรรมน้อยผิดปกติ (ค่า merchant_total_trans ต่ำ) จะมีค่า SHAP เป็นบวก ทำให้โมเดลประเมินความเสี่ยงสูงขึ้น ตรงกันข้าม ร้านค้าที่มีประวัติธุรกรรมปกติหรือสูง มักมีค่า SHAP เป็นลบ ลดความเสี่ยงในการถูกประเมินว่าเป็น Fraud นอกจากนี้ คุณลักษณะรองอย่าง category_total_trans และ customer_total_trans มีบทบาทเสริมในการตัดสินใจ โดยจะเพิ่มน้ำหนักผลกระทบเมื่อพบค่าผิดปกติร่วมกับคุณลักษณะหลัก

การใช้ SHAP Plot ทั้งแบบ Dependence Plot และ Force Plot ช่วยให้เห็นภาพความสัมพันธ์ได้ทั้งในระดับภาพรวม (Global) และระดับรายธุรกรรม (Local) ทำให้เข้าใจกลไกการตัดสินใจของโมเดลได้ละเอียดยิ่งขึ้น

จากผลการวิเคราะห์นี้ สรุปได้ว่าโมเดลตรวจจับการฉ้อโกงใช้ข้อมูลพฤติกรรมร้านค้าและมูลค่าธุรกรรมเป็นปัจจัยหลักในการตัดสินใจ โดยเฉพาะเมื่อพบรูปแบบที่ผิดปกติ เช่น ร้านค้าที่มีประวัติธุรกรรมน้อยผิดปกติหรือธุรกรรมที่มีมูลค่าสูงผิดปกติ ข้อมูลเชิงลึกเหล่านี้สามารถนำไปพัฒนาระบบตรวจจับและแจ้งเตือนการฉ้อโกงให้มีประสิทธิภาพยิ่งขึ้นต่อไป

บทที่ 5
สรุปผลการวิจัย อภิปราย และข้อเสนอแนะ

บทนี้นำเสนอสรุปผลการวิจัยเกี่ยวกับการพัฒนาและเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องในการตรวจจับธุรกรรมฉ้อโกงในระบบการเงิน โดยผู้วิจัยได้ดำเนินการทดลองเปรียบเทียบโมเดล 5 ประเภท ได้แก่ XGBoost, Random Forest, CatBoost, LightGBM และ Logistic Regression ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุล 4 รูปแบบ ได้แก่ SMOTE (Oversampling), Tomek Links (Undersampling), SMOTE-Tomek Links (Hybrid) และ Class Weight การศึกษานี้มุ่งเน้นการวิเคราะห์เปรียบเทียบประสิทธิภาพของแต่ละวิธี เพื่อนำเสนอแนวทางที่เหมาะสมในการตรวจจับธุรกรรมฉ้อโกงอย่างมีประสิทธิภาพ

1. สรุปผลการวิจัย

1.1 สรุปผลการเปรียบเทียบประสิทธิภาพของโมเดลและเทคนิคการจัดการข้อมูลไม่สมดุล

1. ผลการวิจัยพบว่าโมเดลประเภท Gradient Boosting (XGBoost, CatBoost และ LightGBM) ให้ประสิทธิภาพในการตรวจจับธุรกรรมฉ้อโกงสูงกว่าโมเดลประเภทอื่น โดยเฉพาะเมื่อใช้ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุลที่เหมาะสม ผลการเปรียบเทียบโมเดลที่ดีที่สุดจากแต่ละเทคนิคแสดงให้เห็นว่า

2. XGBoost + Tomek Links ให้ประสิทธิภาพโดยรวมดีที่สุด ด้วยสมดุลระหว่าง Precision (0.87) และ Recall (0.82) ส่งผลให้มีค่า F1-Score สูงถึง 0.85 และค่า PR-AUC สูงสุดที่ 0.929 โมเดลนี้เหมาะสำหรับกรณีที่ต้องการความสมดุลระหว่างการแจ้งเตือนที่ถูกต้องและการตรวจจับครบถ้วน

3. LightGBM + Class Weight ให้ค่า Recall สูงที่สุดถึง 0.98 แต่แลกมาด้วย Precision ที่ต่ำลง (0.41) เหมาะสำหรับกรณีที่ต้องการตรวจจับธุรกรรมฉ้อโกงให้ได้มากที่สุด และยอมรับการแจ้งเตือนผิดพลาด (False Positive) ได้

4. CatBoost + SMOTE ให้ความสมดุลระหว่างประสิทธิภาพโดยรวมและการตรวจจับด้วยค่า Recall สูง (0.94) และ PR-AUC ที่ดี (0.91) โดยมี Precision ปานกลาง (0.53)

5. XGBoost + Hybrid (SMOTE + Tomek Links) ให้ประสิทธิภาพดีในด้านการตรวจจับ (Recall 0.93) และความแม่นยำที่พอใช้ได้ (Precision 0.55) โดยมีค่า PR-AUC สูง (0.914)

6. Logistic Regression มีประสิทธิภาพต่ำที่สุดในทุกเทคนิค โดยเฉพาะเมื่อใช้ร่วมกับ SMOTE หรือ Class Weight ซึ่งส่งผลให้ค่า Precision ต่ำมาก (0.18-0.21) แม้จะมีค่า Recall สูง (0.98-0.99) ก็ตาม

ผลการวิจัยนี้สอดคล้องกับสมมติฐานที่ว่า โมเดลแบบ Gradient Boosting จะให้ประสิทธิภาพสูงกว่าโมเดลประเภทอื่น และการใช้เทคนิคจัดการข้อมูลไม่สมดุลที่เหมาะสมจะช่วยเพิ่มประสิทธิภาพการตรวจจับธุรกรรมฉ้อโกงได้อย่างมีนัยสำคัญ

1.2 สรุปผลการวิเคราะห์ความสำคัญของคุณลักษณะ (Feature Importance)

จากการวิเคราะห์ความสำคัญของคุณลักษณะโดยใช้ SHAP (SHapley Additive exPlanations) พบว่า

1. merchant_total_trans มีความสำคัญสูงสุดต่อการตัดสินใจของโมเดล โดยค่า SHAP Value เป็นลบเมื่อร้านค้ามีธุรกรรมจำนวนมาก (ลดโอกาสการทำนายเป็น Fraud) และเป็นบวกเมื่อร้านค้ามีธุรกรรมน้อย (เพิ่มโอกาสการทำนายเป็น Fraud)

2. amount เป็นคุณลักษณะสำคัญอันดับสอง โดยมูลค่าธุรกรรมที่สูงมักมี SHAP Value เป็นบวก (เพิ่มโอกาสการทำนายเป็น Fraud) ขณะที่มูลค่าต่ำมักมี SHAP Value เป็นลบหรือใกล้ศูนย์

3. category_total_trans และ customer_total_trans มีความสำคัญรองลงมา โดยหมวดหมู่สินค้าที่มีธุรกรรมน้อยแต่ลูกค้ามีประวัติทำธุรกรรมเยอะจะเพิ่มโอกาสการทำนายเป็น Fraud

4. คุณลักษณะที่เกี่ยวกับข้อมูลส่วนบุคคล เช่น age และ gender มีความสำคัญน้อยกว่าคุณลักษณะอื่น ๆ อย่างมีนัยสำคัญ

ผลการวิเคราะห์นี้แสดงให้เห็นว่าพฤติกรรมของร้านค้าและมูลค่าธุรกรรมเป็นปัจจัยหลักในการบ่งชี้ธุรกรรมฉ้อโกง โดยรูปแบบที่น่าสงสัย ได้แก่ ร้านค้าที่มีประวัติธุรกรรมน้อยแต่มีธุรกรรมมูลค่าสูง หรือลูกค้าที่มีประวัติทำธุรกรรมเยอะในหมวดหมู่ที่มีธุรกรรมน้อย

2. อภิปรายผลการวิจัย

2.1 ประสิทธิภาพของโมเดลและเทคนิคการจัดการข้อมูลไม่สมดุล

ผลการวิจัยแสดงให้เห็นว่าโมเดลประเภท Gradient Boosting มีประสิทธิภาพสูงกว่าโมเดลอื่น ๆ ในการตรวจจับธุรกรรมฉ้อโกง ซึ่งสอดคล้องกับผลการวิจัยของ Chang และคณะ (2022) ที่พบว่า Random Forest และ Gradient Boosting ให้ประสิทธิภาพสูงในการตรวจจับการฉ้อโกงในระบบการชำระเงินดิจิทัล โดยเฉพาะเมื่อใช้ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุล

ผลการทดลองยังพบว่าเทคนิคการจัดการข้อมูลไม่สมดุลแต่ละวิธีมีจุดแข็งและข้อจำกัดที่แตกต่างกัน สอดคล้องกับงานวิจัยของ Dornadula และ Geetha (2019) ที่แสดงให้เห็นถึงความสำคัญของการใช้ SMOTE ในการปรับสมดุลข้อมูลเพื่อเพิ่มประสิทธิภาพการตรวจจับการฉ้อโกง

อย่างไรก็ตาม ผลการวิจัยพบประเด็นที่น่าสนใจ คือ การใช้ SMOTE เพียงอย่างเดียวมักทำให้ค่า Precision ลดลงอย่างมาก แม้จะเพิ่ม Recall ได้ก็ตาม ซึ่งอาจส่งผลให้มีการแจ้งเตือนผิดพลาด (False Positive) จำนวนมาก ขณะที่การใช้ Tomek Links สามารถรักษาสมดุลระหว่าง Precision และ Recall ได้ดีกว่า สะท้อนให้เห็นถึงความสำคัญในการเลือกเทคนิคที่เหมาะสมกับบริบทการใช้งาน

การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) แสดงให้เห็นผลลัพธ์ที่น่าสนใจ โดยส่งผลต่อประสิทธิภาพของโมเดลไม่มากนักในกรณีที่ใช้ SMOTE แต่มีผลชัดเจนในการปรับปรุงประสิทธิภาพเมื่อใช้ร่วมกับ Tomek Links หรือเทคนิคแบบ Hybrid ผลการทดลองนี้แตกต่างจากงานวิจัยของ Huang และคณะ (2024) ที่พบว่าการใช้ Hybrid Neural Network ร่วมกับ Clustering-based Undersampling ให้ผลลัพธ์ที่ดีกว่าการใช้เทคนิคเดี่ยว

2.2 การวิเคราะห์ความสำคัญของคุณลักษณะ

ผลการวิเคราะห์ SHAP แสดงให้เห็นว่า merchant_total_trans และ amount เป็นคุณลักษณะที่มีความสำคัญสูงสุดในการตรวจจับธุรกรรมฉ้อโกง ซึ่งสอดคล้องกับผลการวิจัยของ Sinap (2024) ที่พบว่ามูลค่าธุรกรรมและพฤติกรรมการใช้จ่ายเป็นปัจจัยสำคัญในการตรวจจับการฉ้อโกงบัตรเครดิต ความสัมพันธ์ระหว่างคุณลักษณะที่พบจากการวิเคราะห์ SHAP มีความน่าสนใจ โดยพบว่าร้านค้าที่มีประวัติธุรกรรมน้อยแต่มีธุรกรรมมูลค่าสูงมีความเสี่ยงต่อการฉ้อโกงมากกว่า ซึ่งเป็นข้อค้นพบที่มีคุณค่าในการพัฒนาระบบเตือนภัยล่วงหน้า (Early Warning System) สำหรับธุรกรรมที่มีความเสี่ยง

ผลการวิจัยนี้ยังสอดคล้องกับผลการสำรวจข้อมูลในบทที่ 3 ที่พบว่าธุรกรรมที่มีมูลค่าสูงมีสัดส่วนการฉ้อโกงสูงขึ้น โดยเฉพาะในหมวดหมู่ Leisure และ Travel ที่มีเปอร์เซ็นต์การฉ้อโกงสูงถึง 95.0% และ 79.4% ตามลำดับ

2.3 ผลกระทบของความไม่สมดุลของข้อมูลต่อประสิทธิภาพของโมเดล

การวิจัยนี้ยืนยันว่าความไม่สมดุลของข้อมูล (1.21% ธุรกรรมฉ้อโกง, 98.79% ธุรกรรมปกติ) เป็นปัญหาสำคัญในการพัฒนาโมเดลตรวจจับการฉ้อโกง ซึ่งสอดคล้องกับงานวิจัยของ Awoyemi และคณะ (2017) ที่พบว่าการสุ่มตัวอย่างแบบผสมช่วยปรับปรุงประสิทธิภาพของโมเดลอย่างมีนัยสำคัญ ผลการทดลองในงานวิจัยนี้แสดงให้เห็นว่า การใช้เทคนิคที่แตกต่างกันในการจัดการข้อมูลไม่สมดุลส่งผลต่อความสมดุลระหว่าง Precision และ Recall อย่างมีนัยสำคัญ โดย:

1. Tomek Links ช่วยรักษาสมดุลระหว่าง Precision และ Recall ได้ดีที่สุด
2. SMOTE และ Hybrid เน้นเพิ่ม Recall แต่ลด Precision
3. Class Weight ให้ค่า Recall สูงสุดแต่ Precision ต่ำที่สุด

ผลการทดลองนี้เป็นข้อมูลสำคัญสำหรับองค์กรในการเลือกเทคนิคที่เหมาะสมกับความต้องการ ขึ้นอยู่กับค่าใช้จ่ายและผลกระทบของ False Positive และ False Negative ในบริบทของธุรกิจนั้น ๆ

3. ข้อเสนอแนะ

3.1 ข้อเสนอแนะในการนำผลการวิจัยไปประยุกต์ใช้

3.1.1 การเลือกโมเดลและเทคนิคตามบริบทธุรกิจ

1. กรณีต้องการความสมดุลระหว่างการตรวจจับและการลดการแจ้งเตือนผิดพลาด ควรใช้ XGBoost ร่วมกับ Tomek Links

2. กรณีต้องการตรวจจับธุรกรรมฉ้อโกงให้ได้มากที่สุดและยอมรับการแจ้งเตือนผิดพลาดได้ ควรใช้ LightGBM ร่วมกับ Class Weight

3. กรณีมีข้อจำกัดด้านทรัพยากรคอมพิวเตอร์ ควรใช้ Random Forest ร่วมกับ Class Weight ซึ่งให้ความสมดุลที่ดีและใช้ทรัพยากรน้อยกว่า

3.1.2 การพัฒนาระบบเตือนภัยแบบหลายระดับ (Tiered Alert System)

1. ระดับความเสี่ยงสูง: ใช้ XGBoost + Tomek Links เพื่อระบุธุรกรรมที่มีความเสี่ยงสูงและต้องการการตรวจสอบทันที

2. ระดับความเสี่ยงปานกลาง: ใช้ CatBoost + SMOTE เพื่อระบุธุรกรรมที่ควรตรวจสอบ

3. ระดับความเสี่ยงต่ำ: ใช้ LightGBM + Class Weight เพื่อการเฝ้าระวังทั่วไป

3.2 การใช้คุณลักษณะที่มีความสำคัญสูง

ให้ความสำคัญกับการตรวจสอบร้านค้าที่มีประวัติธุรกรรมน้อยและมีธุรกรรมมูลค่าสูง พัฒนาดัชนีความเสี่ยง (Risk Index) โดยใช้ความสัมพันธ์ระหว่างมูลค่าธุรกรรมและพฤติกรรมของร้านค้า ตรวจสอบธุรกรรมในหมวดหมู่ที่มีความเสี่ยงสูง เช่น Leisure และ Travel อย่างเข้มงวด

3.3 การปรับปรุงโมเดลอย่างต่อเนื่อง

พัฒนาระบบการเรียนรู้แบบต่อเนื่อง (Continuous Learning) เพื่อรองรับรูปแบบการฉ้อโกงใหม่ ใช้ผลตอบรับจากผู้เชี่ยวชาญ (Expert Feedback) เพื่อปรับปรุงการทำนายของโมเดล ปรับ Threshold ของโมเดลตามสถานการณ์ เช่น เพิ่ม Threshold ในช่วงที่มีความเสี่ยงสูง

3.4 ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

ผู้วิจัยมีข้อเสนอแนะสำหรับการทำวิจัยในอนาคต ดังนี้

1. ควรศึกษาประสิทธิภาพของโมเดลขั้นสูง เช่น Deep Learning หรือ Ensemble Models เพื่อเพิ่มประสิทธิภาพการตรวจจับธุรกรรมฉ้อโกง

2. ควรศึกษาเทคนิคการจัดการข้อมูลไม่สมดุลเพิ่มเติม เช่น ADASYN หรือ One-Class Classification

3. ควรพัฒนาโมเดลที่รองรับการเปลี่ยนแปลงของรูปแบบการฉ้อโกง (Concept Drift) โดยใช้เทคนิค Online Learning

4. ควรทดสอบโมเดลกับข้อมูลจากหลากหลายอุตสาหกรรมเพื่อศึกษาความสามารถในการทำงานข้ามบริบท

5. ควรพัฒนาระบบอธิบายผลการทำนายที่เข้าใจง่ายสำหรับผู้ใช้งานที่ไม่มีความรู้ด้านเทคนิค

3.5 ข้อจำกัดของการวิจัย

1. ข้อจำกัดด้านข้อมูล: ข้อมูลที่ใช้มาจากแหล่งเดียวและไม่ได้อัปเดตแบบเรียลไทม์ ทำให้อาจไม่ครอบคลุมรูปแบบการฉ้อโกงที่หลากหลายและทันสมัย นอกจากนี้ข้อมูลยังมีความไม่สมดุลสูงมาก (ธุรกรรมฉ้อโกงเพียง 1.21%) ซึ่งส่งผลต่อการฝึกโมเดล

2. ข้อจำกัดด้านโมเดล: การวิจัยเน้นโมเดลการเรียนรู้ของเครื่องแบบดั้งเดิม ไม่ได้ครอบคลุมเทคนิคขั้นสูงอย่าง Deep Learning ซึ่งอาจมีความสามารถในการจับรูปแบบซับซ้อนได้ดีกว่า การปรับแต่งพารามิเตอร์ยังทำได้เพียงบางส่วนเนื่องจากข้อจำกัดด้านทรัพยากร

3. ข้อจำกัดด้านการทดสอบ: ยังไม่ได้ทดสอบโมเดลในสภาพแวดล้อมการใช้งานจริง ซึ่งมีความซับซ้อนมากกว่าและต้องการการประมวลผลแบบเรียลไทม์ รวมถึงการรับมือกับปัญหา Concept Drift

4. ข้อจำกัดด้านประสิทธิภาพการประมวลผล: เนื่องจากข้อมูลมีจำนวนมาก (594,643 รายการ) และต้องทดลองกับหลายโมเดลและเทคนิค ทำให้ใช้เวลาในการประมวลผลค่อนข้างนาน โดยเฉพาะในขั้นตอนการปรับแต่งพารามิเตอร์

5. ข้อจำกัดด้านการอธิบายผล: การอธิบายผลลัพธ์ของโมเดลประเภท Ensemble และ Boosting ยังมีความซับซ้อน ทำให้ยากต่อการสื่อสารเหตุผลในการตัดสินใจแก่ผู้ใช้งานทั่วไป

4. บทสรุป

งานวิจัยนี้นำเสนอการศึกษาเชิงเปรียบเทียบของโมเดลการเรียนรู้ของเครื่องและเทคนิคการจัดการข้อมูลไม่สมดุลในการตรวจจับธุรกรรมฉ้อโกงทางการเงิน ผลการวิจัยแสดงให้เห็นว่าโมเดลประเภท Gradient Boosting โดยเฉพาะ XGBoost ร่วมกับเทคนิค Tomek Links ให้ประสิทธิภาพโดยรวมดีที่สุด ในแง่ของความสมดุลระหว่าง Precision และ Recall การวิเคราะห์ SHAP แสดงให้เห็นว่าพฤติกรรมของร้านค้าและมูลค่าธุรกรรมเป็นปัจจัยสำคัญในการบ่งชี้การฉ้อโกง โดยรูปแบบที่น่าสงสัยคือร้านค้าที่มีประวัติธุรกรรมน้อยแต่มีธุรกรรมมูลค่าสูง หรือลูกค้าที่มีประวัติทำธุรกรรมเยอะในหมวดหมู่ที่มีธุรกรรมน้อย การเลือกเทคนิคการจัดการข้อมูลไม่สมดุลควรพิจารณาจากบริบทธุรกิจและความต้องการเฉพาะ หากต้องการความสมดุลระหว่างการแจ้งเตือนที่ถูกต้องและการตรวจจับครบถ้วน ควรใช้ Tomek Links หากต้องการตรวจจับให้ได้มากที่สุดโดยยอมรับการแจ้งเตือนผิดพลาดได้ ควรใช้ SMOTE หรือ Class Weight

การสร้างระบบเตือนภัยแบบหลายระดับที่ใช้โมเดลหลายแบบร่วมกันอาจเป็นแนวทางที่มีประสิทธิภาพที่สุดในการตรวจจับธุรกรรมฉ้อโกง โดยใช้ XGBoost + Tomek Links สำหรับความเสี่ยงสูง CatBoost + SMOTE สำหรับความเสี่ยงปานกลาง และ LightGBM + Class Weight สำหรับการเฝ้าระวังทั่วไป การวิจัยในอนาคตควรมุ่งเน้นการพัฒนาโมเดลขั้นสูงเช่น Deep Learning การจัดการกับ Concept Drift และการพัฒนาระบบอธิบายได้ เพื่อรองรับรูปแบบการฉ้อโกงที่มีความซับซ้อนและเปลี่ยนแปลงตลอดเวลา

ในภาพรวม งานวิจัยนี้ช่วยเพิ่มความเข้าใจเกี่ยวกับการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องในการตรวจจับธุรกรรมฉ้อโกง และนำเสนอแนวทางที่มีประสิทธิภาพในการแก้ไขปัญหาความไม่สมดุลของข้อมูล ซึ่งเป็นความท้าทายสำคัญในงานด้านนี้ ผลการวิจัยนี้มีคุณค่าทั้งในเชิงวิชาการและการประยุกต์ใช้ในอุตสาหกรรมการเงิน โดยสามารถนำไปพัฒนาต่อยอดเพื่อสร้างระบบตรวจจับการฉ้อโกงที่มีประสิทธิภาพมากยิ่งขึ้นในอนาคต

บรรณานุกรม

Aslam, A., และ Hussain, A. (2024). A Performance Analysis of Machine Learning Techniques for Credit Card Fraud Detection. *Journal on Artificial Intelligence*, *6*(1), 1-21. https://doi.org/10.32604/jai.2024.047226

Association of Certified Fraud Examiners. (2024). *Occupational fraud 2024: A report to the nations*. https://www.acfe.com/about-the-acfe/newsroom-for-media/press-releases

Awoyemi, J. O., Adetunmbi, A. O., และ Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI), Computing Networking and Informatics (ICCNI), 2017 International Conference on* (pp. 1-9): IEEE.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010950718922

Chang, V., Doan, L. M. T., Di Stefano, A., Sun, Z., และ Fortino, G. (2022). Digital payment fraud detection methods in digital ages and Industry 4.0. *Computers and Electrical Engineering*, *100*. https://doi.org/10.1016/j.compeleceng.2022.107734

Dhawan, R. a. G. (2024). A Comparative Performance Analysis of Machine Learning Techniques for Credit Card Fraud Detection.

Dornadula, V. N., และ Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631-641. https://doi.org/https://doi.org/10.1016/j.procs.2020.01.057

Elhassan, T., M, A., F, A.-M., และ Shoukri, M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Global Journal of Technology and Optimization*, *01*. https://doi.org/10.4172/2229-8711.S1111

Elreedy, D., Atiya, A. F., และ Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903-4923. https://doi.org/10.1007/s10994-022-06296-4

Federal Trade Commission. (2024). *Claims Journal: National News*. https://www.claimsjournal.com/news/national/2024/02/12/321928.htm

Gor, M., Dobriyal, A., Wankhede, V. A., Sahlot, P., Grzelak, K., Kluczyński, J., และ Łuszczek, J. (2022). Density Prediction in Powder Bed Fusion Additive Manufacturing: Machine Learning-Based Techniques. *Applied Sciences*, *12*, 1-19. https://doi.org/10.3390/app12147271

Hajek, P., Abedin, M. Z., และ Sivarajah, U. (2022). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. *Inf Syst Front*, 1-19. https://doi.org/10.1007/s10796-022-10346-6

Huang, H., Liu, B., Xue, X., Cao, J., และ Chen, X. (2024). Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique. *Applied Soft Computing*, *154*. https://doi.org/10.1016/j.asoc.2024.111368

Khan, M. Y., Qayoom, A., Nizami, M., Siddiqui, M. S., Wasi, S., และ Syed, K.-U.-R. R. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*. https://doi.org/10.1155/2021/2553199

Pereira, R. M., Costa, Y. M. G., และ Silla Jr, C. N. (2020). MLTL: A multi-label approach for the Tomek Link undersampling algorithm. *Neurocomputing*, *383*, 95-105. https://doi.org/10.1016/j.neucom.2019.11.076

Siamrath. (2023). การหลอกลวงทางออนไลน์ในประเทศไทย ปี *2566*. Retrieved February 9, 2025 from https://siamrath.co.th/c/578102

Sinap, V. (2024). Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets. *Turkish Journal of Engineering*, *8*(2), 196-208. https://doi.org/10.31127/tuje.1386127

Sorour, S. E., AlBarrak, K. M., Abohany, A. A., และ El-Mageed, A. A. A. (2024). Credit card fraud detection using the brown bear optimization algorithm. *Alexandria Engineering Journal*, *104*, 171-192. https://doi.org/10.1016/j.aej.2024.06.040

Wang, J.-H., Liu, C.-Y., Min, Y.-R., Wu, Z.-H., และ Hou, P.-L. (2024). Cancer Diagnosis by Gene-Environment Interactions via Combination of SMOTE-Tomek and Overlapped Group Screening Approaches with Application to Imbalanced TCGA Clinical and Genomic Data. *Mathematics*, *12*(14). https://doi.org/10.3390/math12142209

ประวัติผู้เขียน