การพัฒนาการตรวจจับการฉ้อโกงในระบบการเงินโดยการใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์ข้อมูลเครือข่าย

สุเชฏฐ์ เหร่บุตร

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

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

SUCHET REABOOT

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Big Data Engineering,
College of Innovative Technology and Engineering,
Dhurakij Pundit University
Academic Year 2022

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อสารนิพนธ์ การพัฒนาการตรวจจับการฉ้อโกงในระบบการเงิน โดยการใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์ข้อมูลเครือข่าย

เสนอโดย สุเชฏฐ์ เหร่บุตร

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

_____ ประธานกรรมการ
(รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด)

_____ กรรมการที่ปรึกษาสารนิพนธ์
(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

_____ กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

_____ คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
(ดร.ชัยพร เชมะภาตะพันธ์)
วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อสารนิพนธ์	การพัฒนาการตรวจจับการฉ้อโกงในระบบการเงิน โดยการใช้เทคนิคการเรียนรู้ของเครื่องและการวิเคราะห์ข้อมูลเครือข่าย
ชื่อผู้เขียน	สุเชฏฐ์ เหร่บุตร
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
หลักสูตร	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2565

บทคัดย่อ

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

คำสำคัญ: การตรวจจับการฉ้อโกง, การวิเคราะห์เครือข่าย, การเรียนรู้ของเครื่อง, การจัดการข้อมูลที่ไม่สมดุล

.....

อาจารย์ที่ปรึกษา

Thematic Paper Title	ENHANCING FRAUD DETECTION IN FINANCIAL SYSTEMS USING MACHINE LEARNING AND NETWORK ANALYSIS TECHNIQUES
Author	SUCHET REABOOT
Thematic Paper Advisor	Dr. Eakasit Pacharawongsakda
Program	Big Data Engineering
Academic Year	2022

ABSTRACT

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

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

Advisor

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้โดยการให้ความช่วยเหลือของ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้คำแนะนำ ตรวจสอบ และแก้ไขข้อบกพร่องต่างๆมาโดยตลอด เพื่อให้สารนิพนธ์ฉบับนี้สมบูรณ์ผู้เขียนจึงขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ขอขอบคุณ รศ.ดร.วฤษาย์ ร่มสายหยุด ที่กรุณาให้เกียรติเป็นประธานโดยมี ผศ.ดร.ดวงใจ จิตคงชื่น เป็นกรรมการในการสอบสารนิพนธ์ ซึ่งได้กรุณาตรวจ แก้ไขสารนิพนธ์ฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น และนางสาวกุลธิดา รอดบุญ ที่ให้ความสะดวกด้านอำนวยการ และประสานงาน ในการทำสารนิพนธ์ให้กับผู้เขียนมาโดยตลอด

สุดท้ายนี้ขอขอบคุณ บิดา มารดา ครอบครัวและเพื่อนๆ ที่คอยช่วยส่งเสริม สนับสนุนและให้กำลังใจ ทำให้การศึกษาวิจัยในครั้งนี้สำเร็จลุล่วงไปด้วยดี

สุเชฏฐ์ เหร่บุตร

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ง
บทคัดย่อภาษาอังกฤษ..... จ
กิตติกรรมประกาศ..... ฉ
สารบัญ..... ช
สารบัญตาราง..... ฌ
สารบัญภาพ..... ญ
บทที่
1. บทนำ..... 1
1.1 ที่มาและความสำคัญของปัญหา..... 1
1.2 วัตถุประสงค์ของการศึกษาหรือวิจัย..... 1
1.3 ขอบเขตของงานวิจัย..... 2
1.4 ประโยชน์ที่คาดว่าจะได้รับ..... 2
1.5 นิยามศัพท์..... 2
2. แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง..... 3
2.1 ธุรกรรมทางการเงินที่เกิดการฉ้อโกง (Fraud Transactions)..... 3
2.2 การเรียนรู้ของเครื่อง..... 4
2.3 Centrality..... 4
2.4 Decision Tree..... 6
2.5 Random Forest..... 7
2.6 Extreme Gradient Boosting Tree (XGBoost)..... 7
2.7 การคัดเลือกคุณลักษณะของข้อมูล..... 8
2.8 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data)..... 8
2.9 ตัววัดประสิทธิภาพของโมเดล..... 9
2.10 Web Application..... 11
2.11 งานวิจัยที่เกี่ยวข้อง..... 12

สารบัญ (ต่อ)

บทที่		หน้า
3.	ระเบียบวิธีวิจัย	14
	3.1 การรวบรวมข้อมูล (Data Gathering)	15
	3.2 การเตรียมข้อมูล (Data Preprocessing)	19
	3.3 การสร้างตัวแปร (Feature Engineering)	22
	3.4 การจัดการข้อมูลที่ไม่สมดุล (Handle Imbalance data)	22
	3.5 การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพ	23
	3.6 การคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)	24
	3.7 การสร้างโมเดล	25
	3.8 การประเมินผล	25
	3.9 การสร้าง Web-application	25
	3.10 เครื่องมือที่ใช้ในงานวิจัย	26
4.	ผลการวิจัย	28
	4.1 ผลการคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)	28
	4.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล	29
	4.3 การนำไปใช้บน Web-application	32
5.	สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	36
	5.1 สรุปผลการศึกษา	36
	5.2 ข้อเสนอแนะ	37
บรรณานุกรม		38
ภาคผนวก		42
	ก Feature importance	43
	ข Confusion Matrix	50
	ค Decision tree ในโมเดล XGBoost ในวิธีการที่ 1	54
ประวัติผู้เขียน		57

สารบัญตาราง

ตารางที่ **หน้า**

2.1 Confusion Matrix in 2 class (Legitimate and Fraud)..... 10

3.1 รายละเอียดของชุดข้อมูล..... 15

3.2 ตัวอย่างการคำนวณแบบ Cumulative sum of transactions..... 22

3.3 Feature Importance in Extra trees Classifier..... 24

4.1 Top 9 Feature Importance in Extra trees Classifier..... 28

4.2 ผลของโมเดลที่มีการใช้ feature centrality และมีการทำ feature selection..... 29

4.3 ผลของโมเดลที่ไม่มีการใช้ feature centrality และไม่ได้ทำ feature selection..... 30

4.4 ผลของโมเดลที่ใช้เพียง feature หลังจากการ prepare data..... 30

4.5 เปรียบเทียบผลโมเดล Extreme Gradient Boosting (XGBoost) ของ 3 วิธี เทียบความ overfitting..... 31

5.1 Confusion Matrix ของโมเดล XGBoost ในวิธีการที่ 1..... 36

สารบัญภาพ

ภาพที่ **หน้า**

2.1 ลักษณะของ Undirected Graph สำหรับตัวอย่างการคำนวณ Degree Centrality..... 5

2.2 ลักษณะของ Undirected Graph สำหรับตัวอย่างการคำนวณ Closeness Centrality... 6

2.3 โครงสร้างของ Decision Tree..... 6

2.4 โครงสร้างของ Random Forest..... 7

2.5 โครงสร้างของ XGBoost..... 8

2.6 ลักษณะของข้อมูลที่ไม่สมดุล..... 9

2.7 ประกอบการอธิบาย Confusion matrix เพิ่มเติม..... 10

3.1 ตัวอย่างข้อมูลที่นำมาใช้..... 16

3.2 ประเภทของ transactions ใน datasets..... 17

3.3 จำนวน transactions Legitimate เปรียบเทียบ จำนวน transactions fraud..... 18

3.4 ประเภทของ transactions ที่เกิด Fraud..... 19

3.5 transactions หลังจากเลือก type เฉพาะ CASH_OUT และ TRANSFER..... 20

3.6 Number of Fraudulent Transactions by Step..... 21

3.7 สัดส่วนของ class Fraud ในข้อมูล training data และ testing data..... 23

4.1 หน้า User interface บน Web-Application..... 33

4.2 ตัวอย่างผลลัพธ์ fraud ที่ได้จากการทำนายบน Web-Application..... 34

4.3 ตัวอย่างผลลัพธ์ Legitimate ที่ได้จากการทำนายบน Web-Application..... 35

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

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

การเข้าถึงการทำธุรกรรมที่ง่ายขึ้น จึงเป็นอีกสาเหตุหนึ่งของมิจฉาชีพที่ใช้ช่องทางนี้ในการฟอกเงิน ในรูปของการรับจ้างเปิดบัญชี (บัญชีม้า) คือมิจฉาชีพที่จ้างให้ตนไปเปิดบัญชีธนาคารเป็นชื่อของตน แล้วก็เอาสมุดคู่ฝากกับบัตรเอทีเอ็มของตนไปเป็นของตัวเอง โดยมีจุดประสงค์เพื่อนำบัญชีไปรับโอนเงินที่ได้มาแบบผิดกฎหมาย เช่น ค้ายา การพนัน ฉ้อโกง ทำให้ตำรวจตามตัวได้ยาก [1]

ถ้าหากได้นำเส้นทางการเงินของกลุ่มที่รับจ้างเปิดบัญชี มาวิเคราะห์ จะพบว่าเครือข่ายทางการเงินของบัญชีเหล่านี้ว่ามีการโอนต่อไปให้กลุ่ม/ตัวบุคคล หรือมีการรับเงินค่าจ้างจากกลุ่ม/ตัวบุคคลเหล่านั้น

ทั้งนี้ด้วยข้อจำกัดของข้อมูลจริงที่เกี่ยวกับธุรกรรมการฉ้อโกงนั้นหาได้ยาก งานวิจัยนี้จึงได้ใช้ข้อมูลที่จำลองขึ้นมาบนพื้นฐานข้อมูลจริงของการทำธุรกรรมบนมือถือของผู้ให้บริการในแถบประเทศแอฟริกา [2]

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

1.2 วัตถุประสงค์ของการศึกษาหรือวิจัย

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

1.3 ขอบเขตของงานวิจัย

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

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้โมเดลที่สามารถตรวจจับ transactions ที่ เกิดการฉ้อโกงได้
2. ได้เครื่องมือในการพร้อมทั้งแสดงผลการทำนายที่เข้าใจง่ายในรูปแบบ web-application

1.5 นิยามศัพท์

1. “ธุรกรรมฉ้อโกง” หมายถึง ธุรกรรมทางการเงินที่มีการใช้บัญชีเงินฝาก หรือบัญชี e-Money ที่โอนเงินได้ของบุคคลอื่นมาเป็นช่องทางในการรับเงินและถ่ายโอนเงินที่ได้มาจากการกระทำความผิด หรือ บัญชีม้า) ซึ่งธุรกรรมนั้นมีลักษณะจำนวนความถี่สูง , มีการโอนเงินไปยังบัญชีอื่นที่ต้องสงสัย และ บัญชีที่มีการโอนเงินเข้ามาจำนวนมากโดยโอนเข้ามาจากหลายบัญชีแต่โอนออกไปยังบัญชีปลายทางเพียงไม่กี่บัญชี เป็นต้น [3]

2. “Web Application” หมายถึง แอปที่ถูกเขียนขึ้นมาให้สามารถทำงานบนเบราว์เซอร์ โดยมีการพัฒนาด้วยเทคโนโลยีเว็บ เช่น HTML, CSS

บทที่ 2

ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.1 Fraud Transactions

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

2.3 Centrality

2.4 Decision Tree

2.5 Random Forest

2.6 XGBoost

2.7 การคัดเลือกคุณลักษณะของข้อมูล (Feature selection)

2.8 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data)

2.9 ตัววัดประสิทธิภาพของโมเดล

2.10 Web Application

2.11 งานวิจัยที่เกี่ยวข้อง

2.1 ธุรกรรมทางการเงินที่เกิดการฉ้อโกง (Fraud Transactions)

ธุรกรรมทางการเงินที่นับว่าเข้าข่ายทุจริต โดยมีการใช้บัญชีเงินฝาก หรือบัญชี e-Money ที่โอนเงินได้ของบุคคลอื่นมาเป็นช่องทางในการรับเงินและถ่ายโอนเงินที่ได้มาจากการกระทำความผิด หรือ บัญชีม้า) ซึ่งธุรกรรมนั้นมีลักษณะจำนวนความถี่สูง , มีการโอนเงินไปยังบัญชีอื่นที่ต้องสงสัย และ บัญชีที่มีการโอนเงินเข้ามาจำนวนมากโดยโอนเข้ามาจากหลายบัญชีแต่โอนเงินออกไปยังบัญชีปลายทางเพียงไม่กี่บัญชี เป็นต้น [3]

2.2 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning) คือ การสอนให้ระบบคอมพิวเตอร์ทำการเรียนรู้ด้วยตัวเองโดยการใช้ข้อมูล การเรียนรู้ของเครื่องนั้นเป็นได้สองรูปแบบใหญ่ ๆ คือ Supervised Learning คือการที่คอมพิวเตอร์เรียนรู้ด้วยการที่มีข้อมูลมาสอน และ Unsupervised Learning คือการที่คอมพิวเตอร์เรียนรู้โดยที่ไม่ต้องมีข้อมูลมาสอน [4]

การเรียนรู้แบบมีผู้สอน (Supervised learning) จะมีกระบวนการโดยอัลกอริทึมจำเป็นต้องใช้ข้อมูลในส่วนสำหรับ train (training data) และส่วนที่รับกลับมาเพื่อปรับปรุง (feedback) จากมนุษย์เพื่อที่จะเรียนรู้ความสัมพันธ์ระหว่างข้อมูลที่ถูกป้อนเข้ามาสู่ข้อมูลที่ออกไป ซึ่งการเรียนรู้แบบมีผู้สอน มีอยู่ 2 ประเภทคือ การแบ่งแยกประเภท (Classification) จะใช้สำหรับข้อมูล class คำตอบที่เป็นค่าไม่ต่อเนื่อง และการถดถอย (Regression) จะใช้สำหรับข้อมูล class คำตอบที่เป็นค่าต่อเนื่อง [4]

การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) จะมีกระบวนการโดยอัลกอริทึมจะตรวจสอบเฉพาะข้อมูลที่ป้อนเข้ามาเท่านั้นโดยปราศจากการให้ผลลัพธ์ที่จะเกิดขึ้น [4]

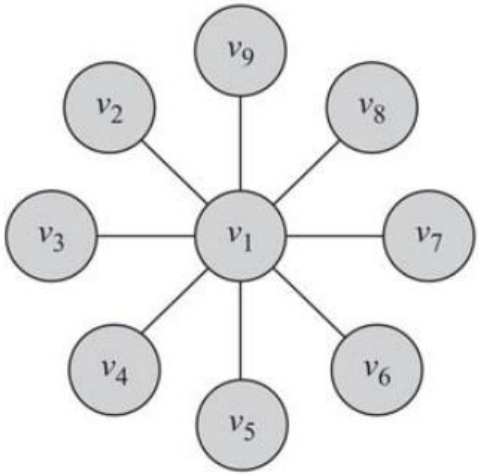
2.3 Centrality

งานวิจัยนี้ได้ใช้ค่า Degree Centrality และ Closeness Centrality เพื่อมาช่วยคำนวณหาความเป็นศูนย์กลางของบัญชี เพื่อใช้ระบุบัญชีของผู้ทำการฉ้อโกง เนื่องจากบัญชีม้ามีลักษณะคือ บัญชีที่เจ้าของบัญชีตัวจริงไม่ได้เปิดเพื่อใช้เอง แต่ขายบัญชีให้คนร้าย ยอมให้คนร้ายเอาไปใช้หรือถูกคนร้ายขโมยข้อมูล หรือถูกสวมตัวตนมาเปิดบัญชี ใช้เป็นช่องทางในการรับ / โอนเงินที่ได้มาจากการกระทำความผิดเพื่อปกปิดไม่ให้มีหลักฐานหรือถูกเชื่อมโยงมาถึงตัวได้ [5]

2.3.1 Degree Centrality เป็นแนวคิดเกี่ยวกับตัววัด ซึ่งจะดูเพียงจำนวนเส้นที่เชื่อมกับโหนด (Node) นั้น ๆ ใน undirected Graph ให้ Degree Centrality = C_d สำหรับ Node $\boldsymbol{\upsilon} i$ เขียนเป็นสมการได้ดังนี้ [6]

$$C_d(v_i) = d_i$$

จากภาพที่ 2.1 จะคำวณค่า Degree Centrality ของ Node $\boldsymbol{\upsilon_1}$ = 8 และ Node อื่นๆ = 1



ภาพที่ 2.1 ลักษณะของ Undirected Graph สำหรับตัวอย่างการคำนวณ Degree Centrality

ที่มา : หนังสือ Social Media mining

2.3.2 Closeness Centrality (Undirected Graph) เป็นแนวคิดเกี่ยวกับตัววัด โดยจะพิจารณาเส้นเชื่อมโยงของจุดที่เกิดขึ้นโดยตรงและทางอ้อมผ่านจุดอื่น ๆ ในเครือข่าย จุดใดมีค่าความเป็นศูนย์กลางสูง ย่อมหมายถึงความสามารถในการติดต่อกับจุดอื่นนั้นเป็นไปได้อย่างรวดเร็ว นอกจากนี้ค่าความใกล้ชิดที่มีค่าสูง ยังสื่อถึงความมีประสิทธิภาพในการสื่อสารข่าวสารข้อมูลหรือข้อความเห็นได้ทั่วถึงตลอดทั้งเครือข่าย และมีความจำเป็นน้อยที่ต้องพึ่งพาจุดอื่น ๆ ในการส่งผ่านข่าวสารข้อมูล เขียนเป็นสมการได้ดังนี้ [6]

$$C_c(\upsilon_i) = \frac{1}{\bar{I}_{\upsilon_i}} \quad \text{โดยที่ } \bar{I}_{\upsilon_i} = \frac{1}{n-1}\sum_{\upsilon_j \neq \upsilon_i} I_{i,j}$$

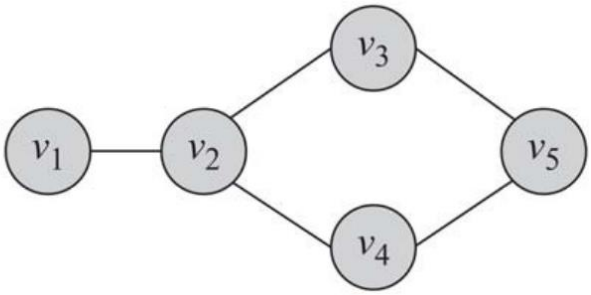
จากรูปที่ 2.2 จะคำวณค่า Closeness Centrality ได้ดังนี้

Node $\boldsymbol{\upsilon_1}$ = 1 / ((1+2+2+3)/4) = 0.5

Node $\boldsymbol{\upsilon_2}$ = 1 / ((1+1+1+2)/4) = 0.8

Node $\boldsymbol{\upsilon_3}$ = Node $\boldsymbol{\upsilon_4}$ = 1 / ((1+1+2+2)/4) = 0.66

Node $\mathbf{5}$ = 1 / ((1+1+2+3)/4) = 0.57

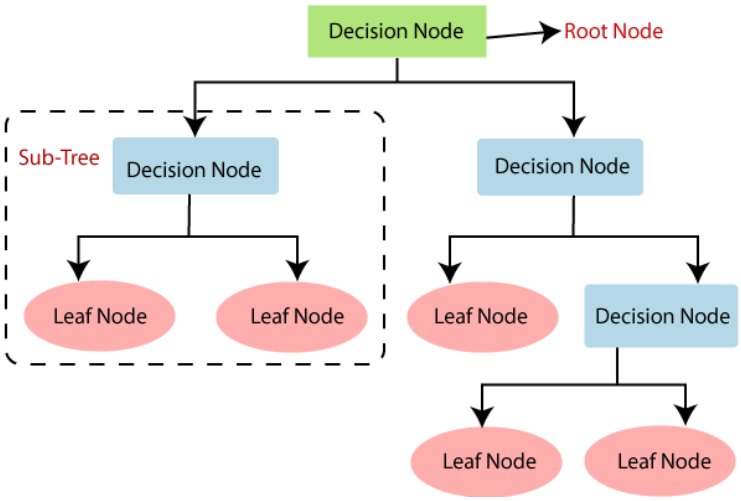


ภาพที่ 2.2 ลักษณะของ Undirected Graph สำหรับตัวอย่างการคำนวณ Closeness Centrality

ที่มา : หนังสือ Social Media mining

2.4 Decision Tree

งานวิจัยนี้ได้เลือกใช้ต้นไม้ตัดสินใจ (Decision Tree) เนื่องมาจากสามารถตีความได้ง่าย โดย Decision Tree มีกระบวนการทำงานในการแบ่งแยกข้อมูล โดยให้ตัวแปรที่สามารถจำแนกข้อมูลได้มากที่สุดเป็น node แรก (root node) และชั้นความลึกถัดไปจะเป็นตัวแปรที่แบ่งแยกข้อมูลได้รองลงมา ทำเช่นนี้จนกระทั่งครบความลึกที่ได้ระบุไว้หรือจนไม่สามารถแบ่งแยกข้อมูลออกมาได้ [7]

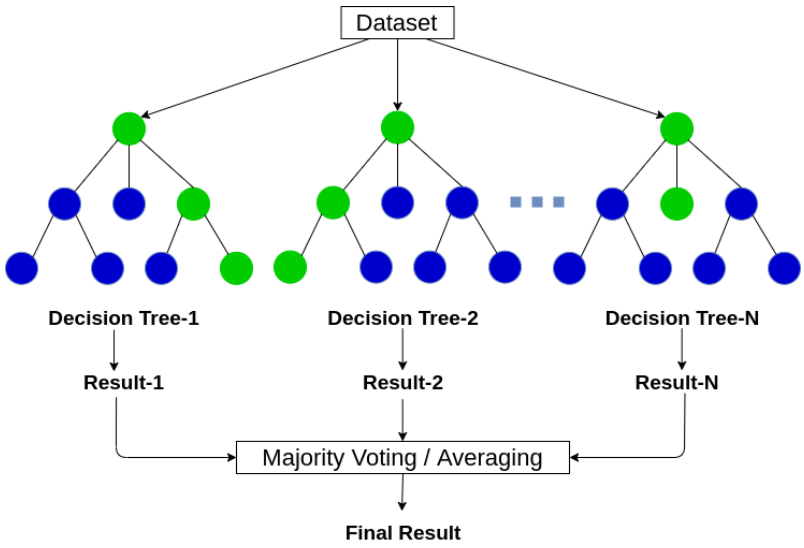


ภาพที่ 2.3 โครงสร้างของ Decision Tree

ที่มา: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

2.5 Random Forest

งานวิจัยนี้ได้เลือกใช้ป่าสุ่ม (Random Forest) เนื่องมาจากมีการใช้ Decision tree หลายต้น (Ensemble of Decision Trees) มาร่วมกันตัดสินเพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพเพิ่มขึ้น โดยในการ trained โมเดลนั้นจะใช้วิธี bagging method (random sampling with replacement) ซึ่งจะนำมาสร้างแบบจำลองต้นไม้โดยแต่ละต้นจะมีชุดข้อมูลสำหรับ trained ไม่ซ้ำกัน (subsets of the training set) โดยแบบจำลองจะมีการทำนายผลออกมาซึ่งจะนำผลการทำนายที่ได้มาโหวตหาผลการทำนายที่ได้รับการโหวตมากที่สุด [7]

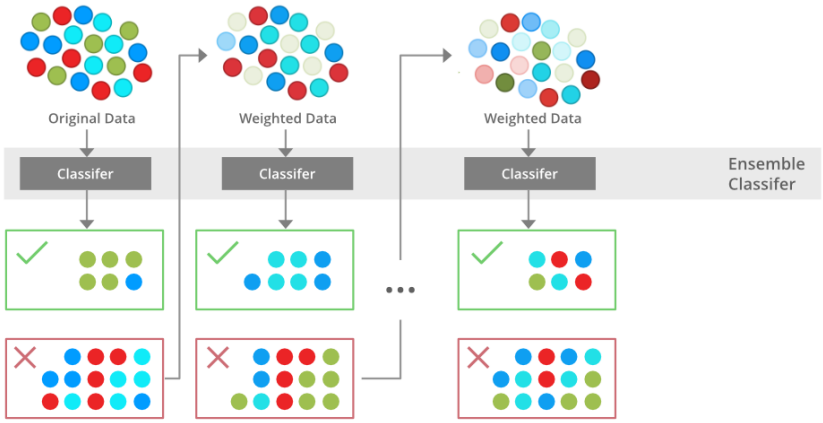


ภาพที่ 2.4 โครงสร้างของ Random Forest

ที่มา: https://www.tibco.com/reference-center/what-is-a-random-forest

2.6 Extreme Gradient Boosting Tree (XGBoost)

งานวิจัยนี้ได้เลือกใช้ XGBoost เนื่องมาจากอัลกอริทึมนี้ถูกพัฒนามาจาก Gradient Boosting ซึ่งออกแบบให้มีประสิทธิภาพสูง , ยืดหยุ่น และสามารถนำไปใช้ได้กับระบบต่างๆ โดยการทำงานของมันจะใช้เทคนิคการนำการเรียนรู้ต้นไม้ตัดสินใจจำนวนหลายๆโมเดลมาทำนายต่อกัน ซึ่งแต่ละต้นไม้ตัดสินใจจะได้เรียนรู้จากค่าความผิดพลาดของต้นไม้ตัดสินใจก่อนหน้า โดยนำค่าความผิดพลาดนั้นมาปรับปรุงในการสร้างโมเดลถัดไป [7], [8]



ภาพที่ 2.5 โครงสร้างของ XGBoost

ที่มา: https://www.geeksforgeeks.org/xgboost

2.7 การคัดเลือกคุณลักษณะของข้อมูล

ก่อนจะนำข้อมูลไปสร้างโมเดลจะต้องมีขั้นตอนการคัดเลือกคุณลักษณะหรือตัวแปรที่มีความสำคัญและเหมาะสมเพื่อให้ได้โมเดลที่มีประสิทธิภาพและยังช่วยให้การทำงานไวขึ้น โดยใน supervised models จะมีทั้งสิ้น 3 วิธีการในการคัดเลือก ดังนี้ 1.filter method, 2.Wrapper method และ 3.Intrinsic method [9]

การคัดเลือกคุณลักษณะของข้อมูล โดยใช้ Extra trees Classifier เป็นวิธีการคัดเลือกตัวแปรที่ด้วยวิธีการ filter metod โดยจะใช้โมเดลที่ได้ทำการเรียนรู้กับตัว training data หา feature ที่สำคัญและทำการเลือกจำนวนตัวแปรไปใช้ต่อ ซึ่ง Extra trees มีวิธีการคล้ายกับ random forest แต่ต้นไม้ในโมเดล จะมีความสัมพันธ์กันน้อยกับต้นไม้ต้นอื่นในโมเดล จึงทำให้ผลลัพธ์ที่ได้นั้น overfit กับตัว data น้อยลง [10]

2.8 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data)

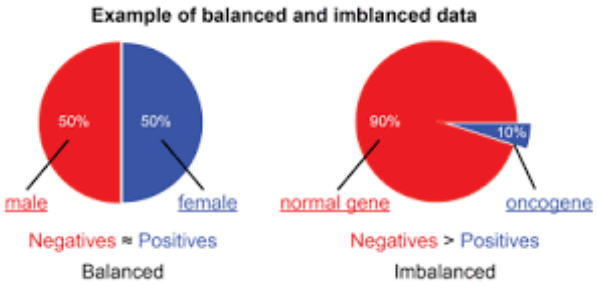
ปัญหา imbalanced classification เกิดขึ้นเมื่อจำนวนข้อมูลในแต่ละ Class คำตอบมีจำนวนข้อมูลแตกต่างกันมาก การมีข้อมูลที่ไม่สมดุลนี้ส่งผลกระทบต่อประสิทธิภาพการทำงานของโมเดล โดยเฉพาะอย่างยิ่ง Minority Class ที่มีความสำคัญมากกว่า เช่น Fraud detection, Churn prediction และจะทำให้ได้ผลลัพธ์ที่โน้มเอียงไปหา Majority Class (Biased Sampling) ซึ่งจะส่งผลต่อการวัดผลที่ไม่มีประสิทธิภาพ

ตามมา ทั้งนี้ในการจัดการปัญหา imbalanced data หลักๆ จะมีสองแนวทางคือ Data Sampling และ Class weighting

การให้น้ำหนักในคลาสเป็นวิธีที่นิยมในการจัดการกับปัญหาความไม่สมดุลของคลาสในโมเดลการเรียนรู้ของเครื่อง โดยการกำหนดน้ำหนักที่แตกต่างกันให้กับแต่ละคลาสในระหว่างการฝึกฝน เพื่อให้แน่ใจว่าทุกคลาสจะได้รับความสำคัญที่เท่ากัน [11]

ในงานวิจัยนี้เลือกใช้วิธีการจัดการปัญหา imbalanced data ด้วยการทำ Class weighting โดยการใช้ Cost-sensitive Algorithms ซึ่งเป็นหนึ่งในวิธีการของ Cost-sensitive learning ซึ่งจะทำการปรับค่า weight ใน class ต่างๆผ่านทาง cost matrix ในตัว Algorithms ทั้งนี้ library scikit-learn มี best practice heuristic สำหรับ class weighting นั้นคือ balanced โดยมีวิธีคำนวณ weight ดังนี้ [12]

$$weighting = \frac{n - samples}{(n\_classes \ * \ n\_samples\_with\_class)}$$



ภาพที่ 2.6 ลักษณะของข้อมูลที่ไม่สมดุล

ที่มา: https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

2.9 ตัววัดประสิทธิภาพของโมเดล

ในงานวิจัยนี้ได้ใช้ Confusion Matrix หรือ error matrix ในการเปรียบเทียบประสิทธิภาพของโมเดลภายใต้เงื่อนไขต่างๆ โดยทั่วไปจะใช้เปรียบเทียบค่าที่เกิดจากการทำนาย และค่าที่เกิดขึ้นจริง [13]

ตารางที่ 2.1 Confusion Matrix in 2 class (Legitmate and Fraud)

CONFUSION MATRIX		PREDICTED	
		LEGITIMATE	FRAUD
TRUE	LEGITIMATE	TN	FP
	FRAUD	FN	TP

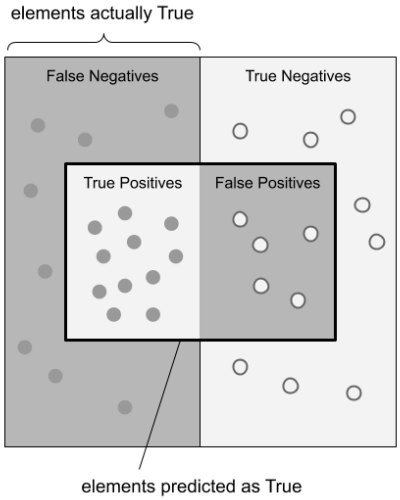
จากตารางที่ 2.1 อธิบายได้ดังนี้

True Positive (TP) คือการที่โมเดลทำนายว่าเป็น Class Fraud และตรงกับค่าจริง

True Negative (TN) คือการที่โมเดลทำนายว่าเป็น Class Legitimate และตรงกับค่าจริง

False Positive (FP) คือการที่โมเดลทำนายว่าเป็น Class Fraud แต่ค่าจริงไม่ใช่ (ค่าจริงเป็น Class Legitimate)

False Negative (FN) คือการที่โมเดลทำนายว่าเป็น Class Legitimate แต่ค่าจริงไม่ใช่ (ค่าจริงเป็น Class Fraud)



ภาพที่ 2.7 ประกอบการอธิบาย Confusion matrix เพิ่มเติม

ที่มา: https://www.ml-science.com/confusion-matrix

ซึ่งจากค่า Measurement สามารถนำมาคำนวณเพิ่มเติมเพื่อง่ายต่อการวัดผลได้ และทั้งนี้งานวิจัยนี้ได้เลือกใช้การคำนวณการวัดผลทั้งสิ้น 4 ค่าได้แก่ Accuracy, Precision, Recall และ F1 Score

Accuracy คืออัตราส่วนที่บอกถึงความถูกต้องของโมเดลว่าทำนายถูกต้องตรง เทียบกับค่าที่เกิดขึ้นจริงทั้งหมด สามารถคำนวณได้จากสมการ

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

Precision คืออัตราส่วนของข้อมูลที่โมเดลทำนาย class positive ถูกต้อง เทียบกับจำนวนค่า positive ที่ได้มาจากการทำนายทั้งหมด สามารถคำนวณได้จากสมการ

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

Recall คืออัตราส่วนของข้อมูลที่โมเดลทำนาย class positive ถูกต้อง เทียบกับจำนวนของ class positive ทั้งหมด (ค่า positive จากการทำนายและค่า positive จากค่าจริง) สามารถคำนวณได้จากสมการ

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

F1 Score คือค่าเฉลี่ยค่าความแม่นยำระหว่าง precision และ recall สามารถคำนวณได้จากสมการ

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

2.10 Web Application

Web Application คือแอปที่ถูกเขียนขึ้นมาให้สามารถเปิดใช้ในเว็บเบราว์เซอร์ได้โดยตรงทำให้โดยรวมแล้วกินทรัพยากรค่อนข้างน้อย สามารถเปิดใช้งานได้รวดเร็ว ตัวอย่างการใช้งาน เช่นเว็บแอปสำหรับคิดเลข เว็บแอปสำหรับจับเวลา เป็นต้น โดยส่วนมากแล้วจะมีความเรียบง่าย รวดเร็ว และสบายตากว่าเว็บไซต์ปกติ

เนื่องจากเน้นใช้งานในเรื่องใดเรื่องหนึ่งเป็นหลัก [14]

2.11 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการตรวจจับการฉ้อโกง (Fraud detection) หรือการศึกษาตัวแปรที่มีความสำคัญกับการการการตรวจจับการฉ้อโกง (Fraud detection) ที่ผู้วิจัยได้ศึกษาสรุปได้ดังนี้

Hajek et al. (2022) [15] ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนระบบจำลองการชำระเงินผ่านมือถือ โดยได้มีการเทียบประสิทธิภาพของโมเดล ทั้งแบบใช้วิธี Supervised Learning Method และ Outlier Detection Methods ในงานวิจัยนี้พบว่า semi-supervised ensemble model integrating multiple unsupervised outlier detection และใช้ algorithms XGBoost classifier (XGBOD) ให้ผลลัพธ์ที่ดีที่สุด ให้ค่า accuracy เท่ากับ 0.9994 , F1 เท่ากับ 0.8737 , Precision เท่ากับ 0.9942 และ Recall เท่ากับ 0.7793 ขณะที่โมเดลที่ดีที่สุดในแง่ของ cost saving of fraud detection system (ค่า recall มากที่สุด) ได้แก่ combining random under-sampling and XGBoost (RUS+XGBoost) ให้ค่า accuracy เท่ากับ 0.9963 , F1 เท่ากับ 0.2812 , Precision เท่ากับ 0.1637 และ Recall เท่ากับ 0.9976

Wen et al. (2022) [16] ในงานวิจัยนี้ได้นำเสนอวิธีการตรวจจับการฉ้อโกงในภาคการเงินโดยอาศัย kg (knowledge graph)ในการศึกษาความสัมพันธ์ระหว่างผู้จัดการกับสถาบันการเงินที่เกี่ยวข้อง โดยได้มีการเทียบประสิทธิภาพของโมเดลทั้งหมด 4 โมเดล และโมเดลที่มีการใช้ kg อีก 4โมเดล โดยได้ผลการทดลองว่าโมเดลที่มีประสิทธิภาพสูงสุดได้แก่ SVM ที่มีการใช้ centrality measure ให้ค่า average Accuracy เท่ากับ 0.9318 , average F1 เท่ากับ 0.5852 , average Precision เท่ากับ 0.6845 และให้ค่า average Recall เท่ากับ 0.5611

Sahu et al. (2020) [17] ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนเครดิตการ์ด ในงานวิจัยนี้ได้เปรียบเทียบวิธีในการจัดการปัญหาความไม่สมดุลของข้อมูล 2 วิธี 1. Oversampling minority class และ 2. Cost based โดยการปรับค่าน้ำหนักให้กับ minority class เพื่อให้มีผลกระทบสูงขึ้นต่อการสร้างโมเดล โดยได้ทำการทดลองทั้งสิ้น 5 โมเดล ดังนี้ ANN, LR, SVM, DT, RF ซึ่งโมเดลที่มีประสิทธิภาพสูงสุดในแง่ของค่า F1 ได้แก่โมเดล Random forest ทั้ง 2 วิธี โดยวิธี Oversampling minority class ให้ค่า accuracy เท่ากับ 96.57, F1 เท่ากับ 95.21, precision เท่ากับ 97.59 และ recall เท่ากับ 94.95 และวิธี Cost based ให้ค่า accuracy เท่ากับ 96.48 , F1 เท่ากับ 92.03 , precision เท่ากับ 94.1 และ recall เท่ากับ 86.88

Varmedja et al. (2019) [18] ในงานวิจัยนี้ได้นำเสนอการใช้ Machine Learning ในตรวจจับการฉ้อโกงบนเครดิตการ์ด ซึ่งได้ใช้เทคนิค SMOTE ในการจัดการปัญหาความไม่สมดุลของข้อมูล โดยได้ทำการทดลองทั้งสิ้น 4 โมเดล ดังนี้ Logistic Regression, Random Forest, Naïve Bayes และ Multilayer Perceptron ซึ่งโมเดลที่มีประสิทธิภาพสูงสุดได้แก่ Random forest ให้ค่า accuracy เท่ากับ 99.96 , precision เท่ากับ 96.38 และ recall เท่ากับ 81.63

ในงานวิจัยที่เกี่ยวกับการตรวจจับการฉ้อโกง (Fraud detection) วิธีการที่ได้รับผลดีที่สุดคือการใช้ XGBoost และ Random Forrest ซึ่งเป็นโมเดล tree based และการใช้ค่า centrality measure ทำให้การตรวจจับมีประสิทธิภาพมากยิ่งขึ้นเมื่อเทียบกับการไม่มีการประยุกต์ใช้ค่าดังกล่าว ด้วยเหตุนี้งานวิจัยนี้จึงได้ทำการประยุกต์ใช้ความสามารถและข้อดีของวิธีการต่าง ๆ ดังกล่าว

บทที่ 3

วิธีวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยมีขั้นตอนการดำเนินการดังต่อไปนี้

3.1 การรวบรวมข้อมูล (Data Gathering)

งานวิจัยนี้ได้ใช้ข้อมูล PaySim จาก Kaggle ซึ่งเป็นจำลองการทำธุรกรรมเงินสดผ่านมือถือในประเทศแอฟริกาใดประเทศหนึ่งโดยอิงจากตัวอย่างการทำธุรกรรมจริงที่สกัดมาจาก financial logs ในระยะเวลา 1 เดือน ทั้งนี้ชุดข้อมูลจำลองนี้ถูกลดขนาดลงเป็น 1/4 จากชุดข้อมูลดั้งเดิมและสร้างขึ้นเพื่อใช้งานใน Kaggle เท่านั้น [19]

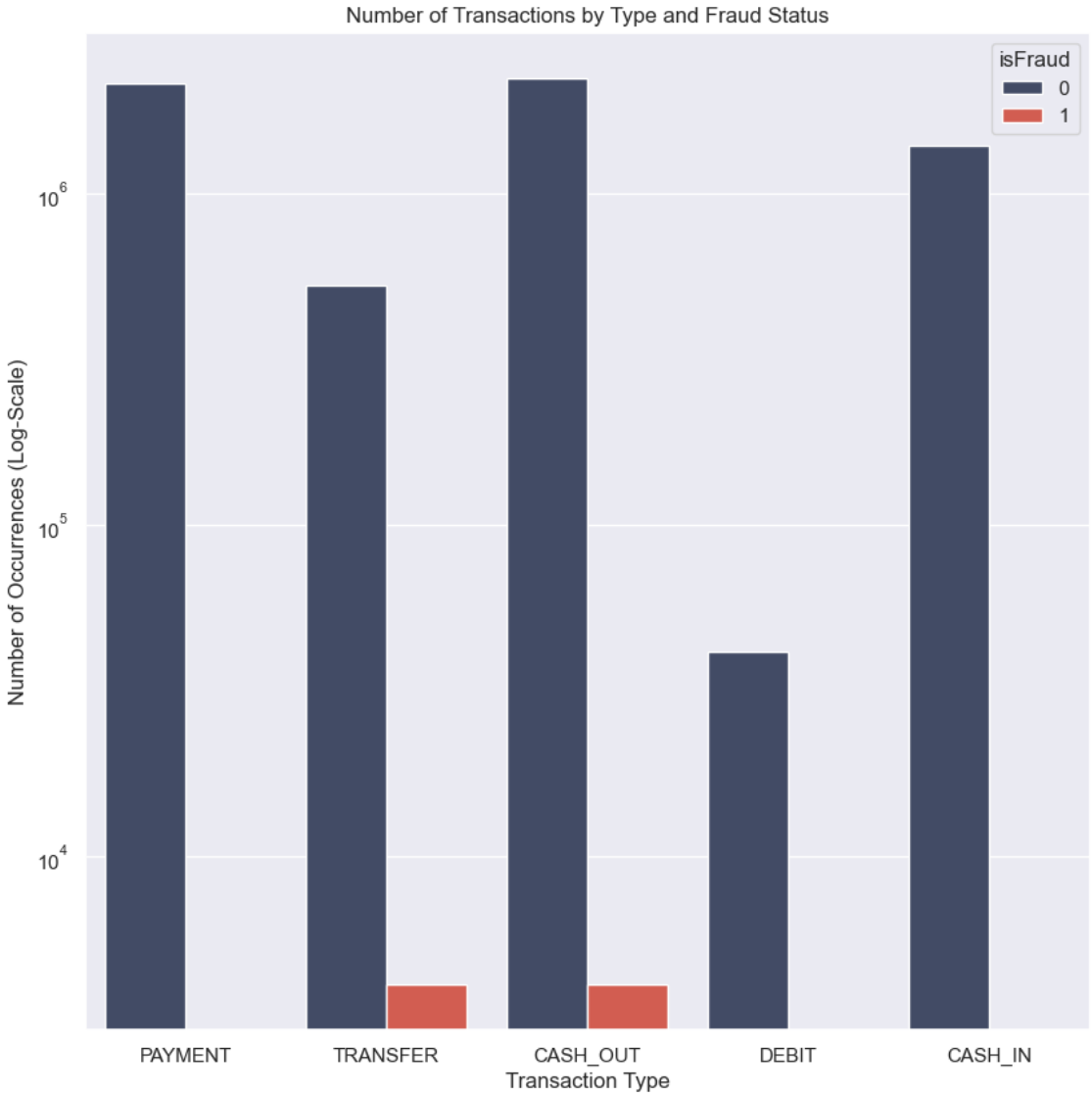
ตารางที่ 3.1 รายละเอียดของชุดข้อมูล

	Field	คำอธิบายข้อมูล
1	step	maps a unit of time in the real world.
2	type	ประเภทของ transaction
3	amount	จำนวนเงิน
4	nameOrig	ชื่อลูกค้าที่ทำ transaction
5	oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำ transaction
6	newbalanceOrig	Balance ของคนต้นทาง หลังทำ transaction
7	nameDest	ชื่อลูกค้าที่รับ transaction
8	oldbalanceDest	Balance ของคนปลายทาง ก่อนทำ transaction
9	newbalanceDest	Balance ของคนปลายทาง หลังทำ transaction
10	isFraud	Transaction ที่เกิด fraud หรือ Legitimate
11	isFlaggedFraud	Flag ความพยายามที่ผิดกฎหมายในการทำ transaction ด้วยเงินจำนวนมาก

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.00	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	0	0
...	...	...	...	...	...	...	...	...	...	...	...
6362615	743	CASH_OUT	339682.13	C786484425	339682.13	0.00	C776919290	0.00	339682.13	1	0
6362616	743	TRANSFER	6311409.28	C1529008245	6311409.28	0.00	C1881841831	0.00	0.00	1	0
6362617	743	CASH_OUT	6311409.28	C1162922333	6311409.28	0.00	C1365125890	68488.84	6379898.11	1	0
6362618	743	TRANSFER	850002.52	C1685995037	850002.52	0.00	C2080388513	0.00	0.00	1	0
6362619	743	CASH_OUT	850002.52	C1280323807	850002.52	0.00	C873221189	6510099.11	7360101.63	1	0

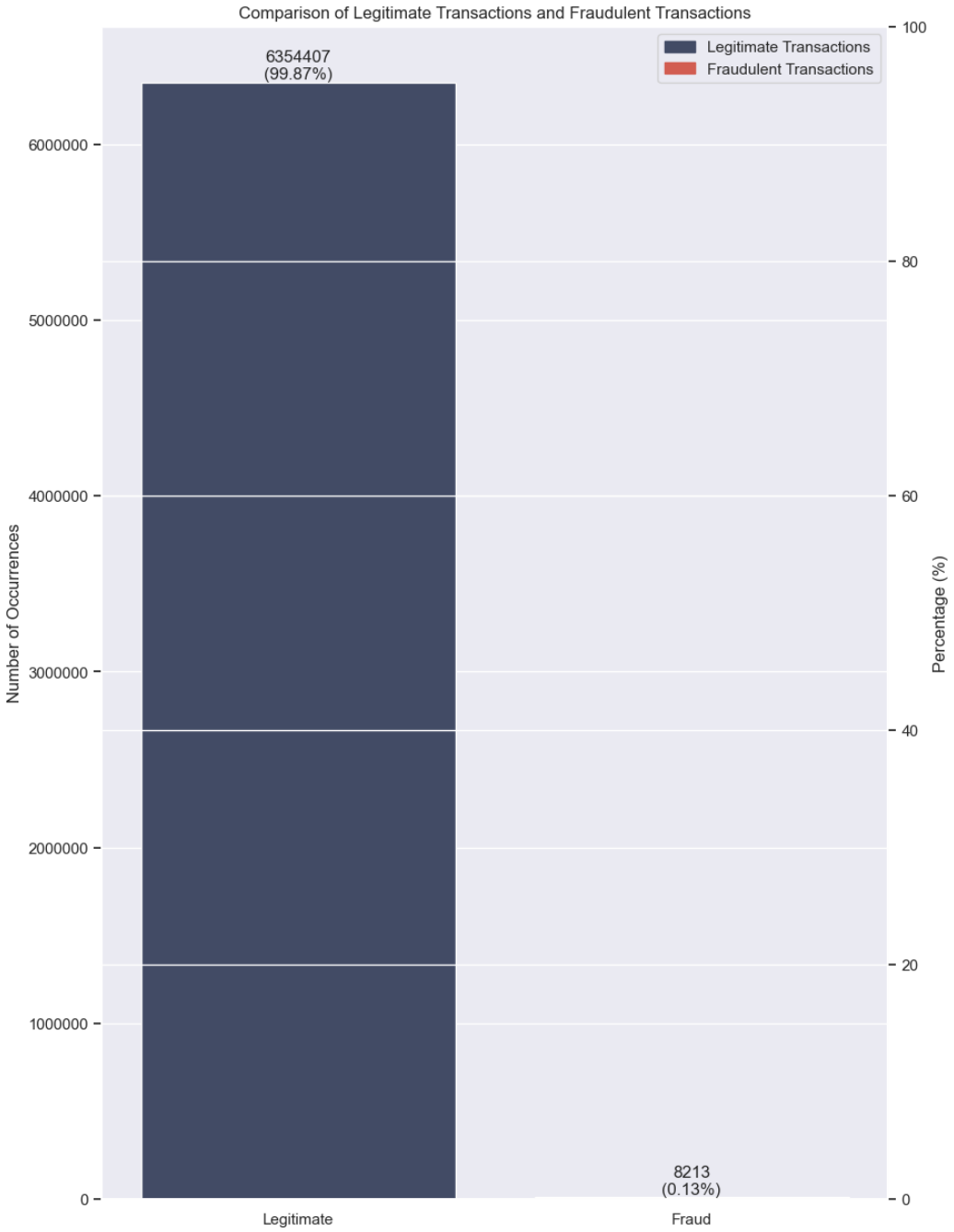
6362620 rows × 11 columns

ภาพที่ 3.1 ตัวอย่างข้อมูลที่นำมาใช้



ภาพที่ 3.2 ประเภทของ transactions ใน datasets

จากภาพที่ 3.2 จะเห็นได้ว่า type ของ transactions ที่เกิด fraud นั้นมีเพียง type : TRANSAFER และ CASH_OUT



ภาพที่ 3.3 จำนวน transactions Legitimate เปรียบเทียบ จำนวน transactions fraud

จากภาพที่ 3.3 จะเห็นได้ว่า fraud transactions มีค่าเท่ากับ 8,213 หรือ 0.13% ของจำนวน transactions ทั้งหมด

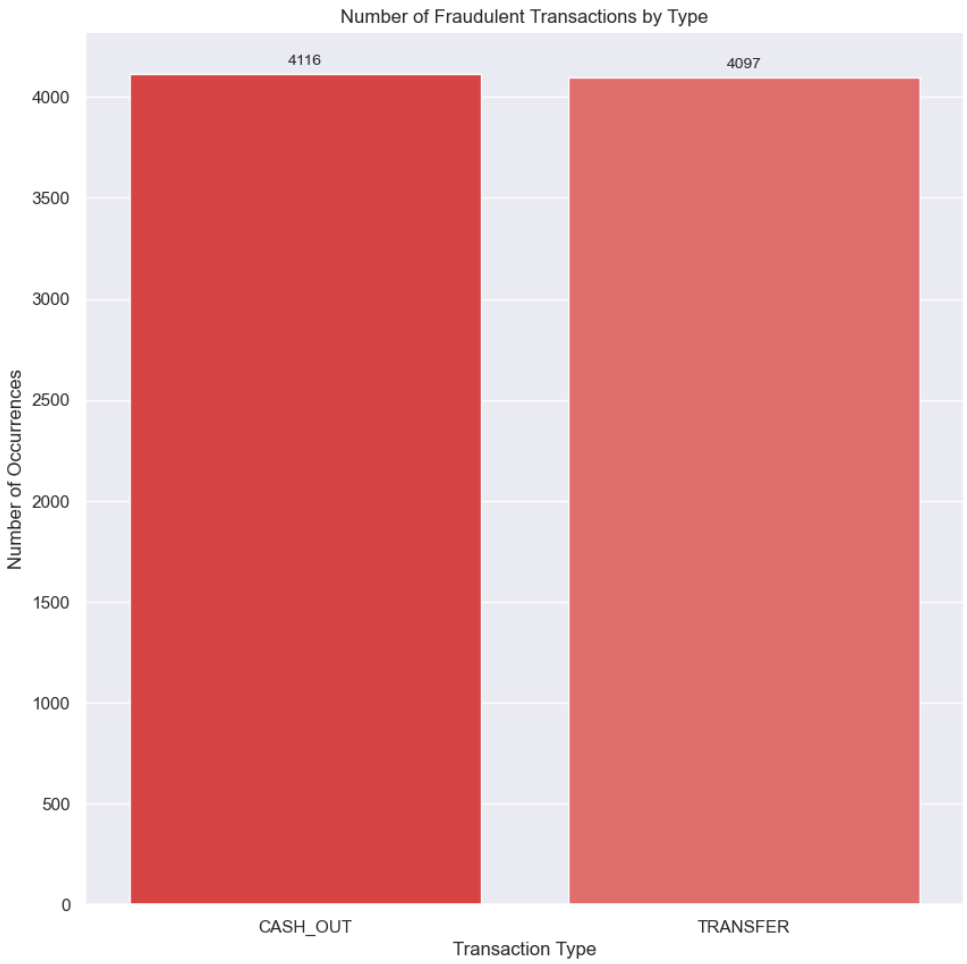
3.2 การเตรียมข้อมูล (Data Preprocessing)

3.2.1 ตัวแปรที่ใช้ในการศึกษา

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

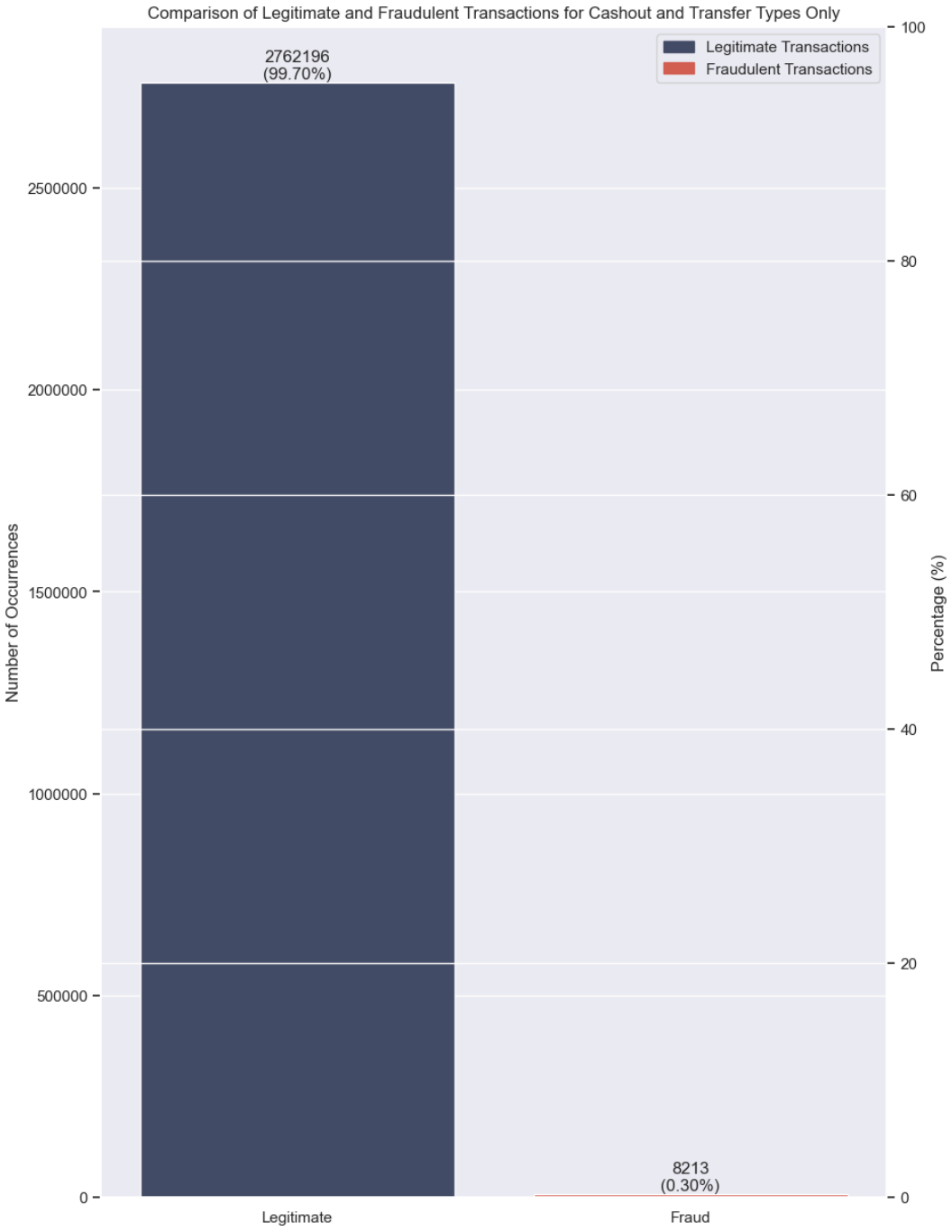
3.2.2 ข้อมูลที่ใช้ในการศึกษา

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



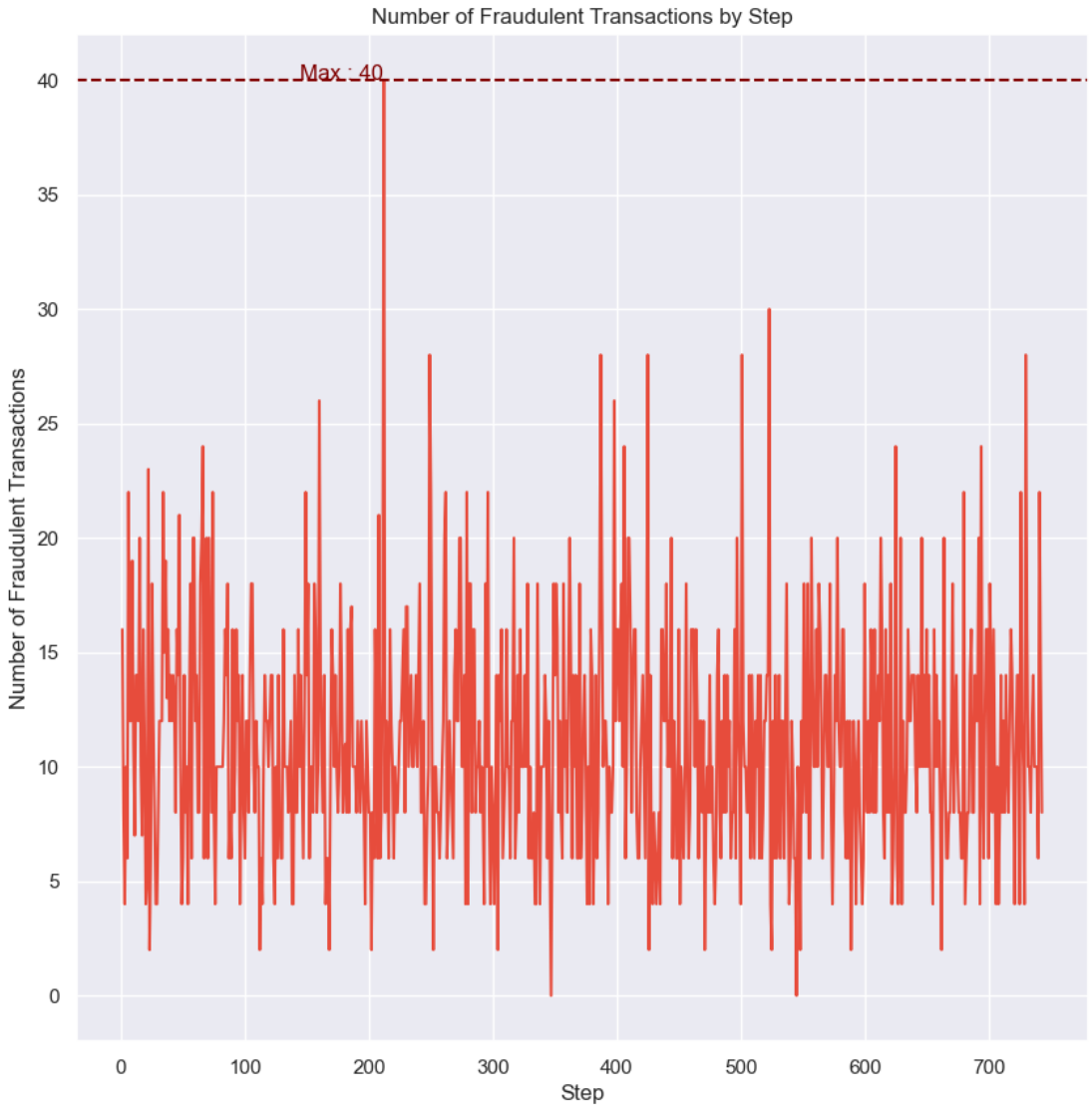
ภาพที่ 3.4 ประเภทของ transactions ที่เกิด Fraud

จากภาพที่ 3.4 จะเห็นได้ว่า type ของ transactions ที่เกิด fraud นั้น type CASH_OUT มากกว่า type TRANSFER เล็กน้อย



ภาพที่ 3.5 transactions หลังจากเลือก type เฉพาะ cashout และ transfer

จากภาพที่ 3.5 หลังจากเลือก type ของ transactions ที่เกิด fraud (type CASH_OUT และ TRANSFER) จะมีสัดส่วนของ fraud transactions เพิ่มขึ้นมาเป็น 0.30% เมื่อเทียบกับจำนวน transactions ทั้งหมด



ภาพที่ 3.6 Number of Fraudulent Transactions by Step

จากภาพที่ 3.6 เป็นการ Visualization สัดส่วนของ fraud transactions โดยดูตามช่วงเวลา(ตัวแปร Step)

3.3 การสร้างตัวแปร (Feature Engineering)

3.3.1 Centrality

3.3.3.1 สร้าง Network การทำ transactions ระหว่าง nameOrig กับ nameDest โดยให้ nameOrig เป็น source และ nameDest เป็น target

3.3.3.2 degree centrality : ทำการคำนวณค่า Degree Centrality ใน network ที่ได้สร้างไว้

3.3.3.3 closeness centrality : ทำการคำนวณค่า Closeness Centrality ของ name account ใน network ที่ได้สร้างไว้

3.3.2 Cumulative sum of transactions

3.3.2.1 ทำการคำนวณ transactions ของ type Transfer และ type Cashout ในลักษณะของ Cumulative sum เพื่อนับจำนวน transactions ที่เกิดขึ้น ของทั้ง nameOrig กับ nameDest

ตารางที่ 3.2 ตัวอย่างการคำนวณแบบ Cumulative sum of transactions

step	type	nameOrig	nameDest	nameOrig		nameDest	
				transfer_out_count	cashout_out_count	transfer_in_count	cashout_in_count
18	TRANSFER	C24957224	C1917728887	1	0	1	0
33	TRANSFER	C24957224	C1704862259	2	0	1	1

จากตาราง 3.2 จะเป็นตัวอย่างการคำนวณค่าของ nameOrig ที่มีค่าเท่ากับ C24957224 ที่มี transaction ใน type transfer โดยจะเห็นได้ว่า step 33 มีค่า transfer_out_count อัพเดตขึ้นมา 1 ค่า จาก step 18 ส่วนค่า cashout_out_count ไม่เปลี่ยนแปลง

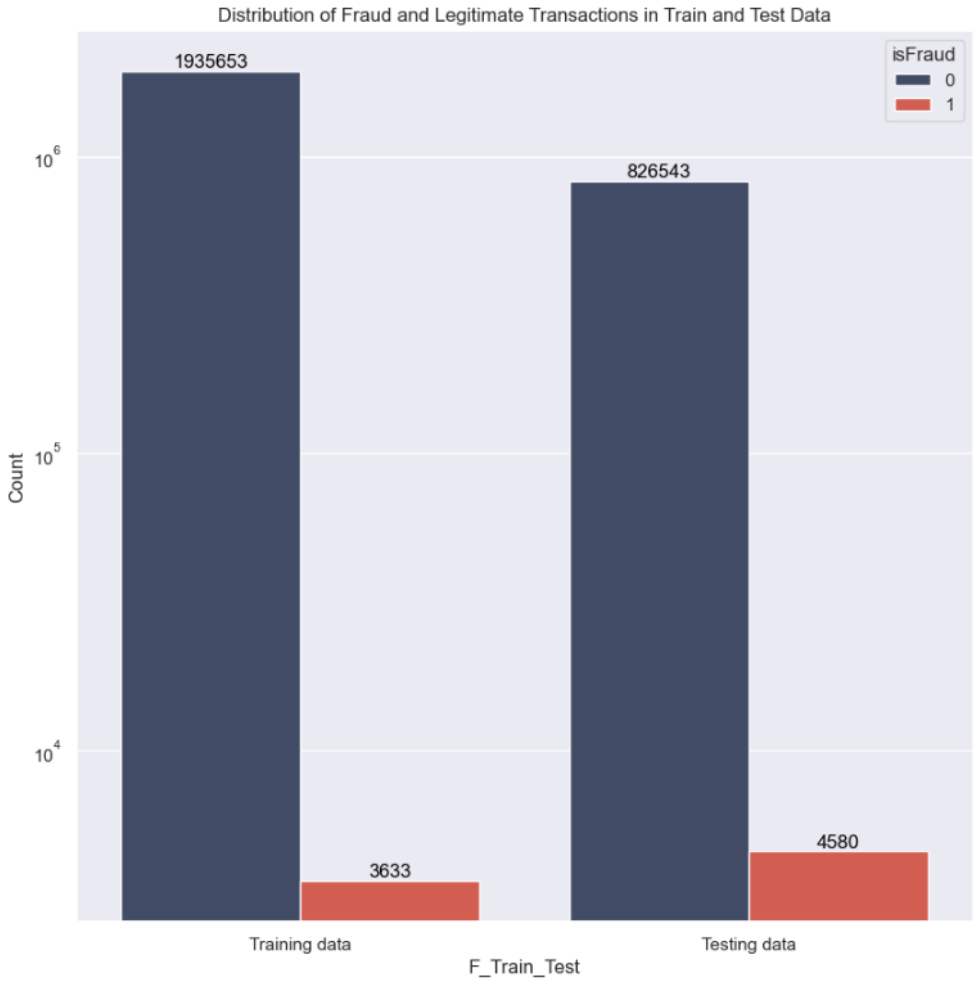
3.4 การจัดการข้อมูลที่ไม่สมดุล (Handle Imbalance data)

หลังจากได้มีกระบวนการจัดเตรียมข้อมูลโดยได้เลือกเฉพาะ type transactions ที่เกิด fraud เท่านั้น ทำให้จำนวนข้อมูลเหลืออยู่ 2,770,409 transactions ซึ่ง class fraud มีจำนวนทั้งสิ้น 8,213 transactions คิดเป็นสัดส่วน 0.30 % ของ transactions ทั้งหมด

เนื่องจากข้อมูลคำตอบ (Label data) มีสัดส่วนที่ไม่สมดุลกันจึงได้เลือกใช้วิธีปรับค่า weight ใน class คำตอบ ซึ่งเป็นการปรับค่าน้ำหนัก (Weight) ให้แต่ละคลาสคำตอบไม่เท่ากัน โดยจะปรับน้ำหนักของ Minority class ให้มีค่ามากกว่า Majority class

3.5 การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพ

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



ภาพที่ 3.7 สัดส่วนของ class Fraud ในข้อมูล training data และ testing data

จากภาพที่ 3.7 เป็นการเทียบสัดส่วนของ class คำตอบ ในชุดข้อมูล training data และ testing data ซึ่งข้อมูล fraud transactions ใน testing data มีสัดส่วนอยู่ที่ 0.55% และ training data มีสัดส่วนอยู่ที่ 0.19%

3.6 การคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)

การสร้างโมเดลจะต้องมีการคัดเลือกตัวแปรที่มีความสำคัญ 9 ตัว (เท่ากับจำนวนตัวแปรก่อนมีการสร้าง feature centrality) เพื่อให้โมเดลมีประสิทธิภาพสูง ซึ่งในงานวิจัยนี้เลือกใช้เทคนิค ใช้ model Extra trees Classifier หาค่า Feature importance ของชุดข้อมูล เพื่อดูความสัมพันธ์ระหว่างตัวแปรต่างๆ และตัวแปรคำตอบ (Label) ซึ่งตัวแปรที่มีค่าน้ำหนักมากหมายถึงมีความสำคัญมากกว่าอีกตัวแปรที่มีค่าน้ำหนักน้อยกว่า

โดยค่าพารามิเตอร์ที่ใช้ในการสร้างโมเดล Extra trees Classifier ได้แก่ number of tees ที่ 100, criterion เป็น gini และ maximal depth ที่ 10

เมื่อนำโมเดล Extra trees Classifier ไป fitting ในชุด training data พบว่ามีตัวแปรที่สำคัญ ดังตารางที่ 3.3

ตารางที่ 3.3 Feature Importance in Extra trees Classifier

Feature	Feature Description	Importance
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.191760
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.166195
oldbalanceOrg	Balance ของคนต้นทางก่อนทำการโอน	0.152379
amount	จำนวนเงิน	0.151416
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.094640
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.067429
Degree_Score_Dest	ค่า Degree centrality ของคนปลายทาง	0.047205
Clo_Score_Orig	ค่า Closeness centrality ของคนต้นทาง	0.041692
Clo_Score_Dest	ค่า Closeness centrality ของคนปลายทาง	0.037046
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.035120
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.009176
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.005915
Degree_Score_Orig	ค่า Degree centrality ของคนต้นทาง	0.000027

3.7 การสร้างโมเดล

หลังจากทำการคัดเลือกคุณลักษณะ (Feature Selection) ที่ดีที่สุด 9 ลำดับแรก เสร็จเรียบร้อยแล้ว ก็นำตัวแปรที่ได้ไปสร้างโมเดล

โดยงานวิจัยนี้ได้ทดลองใช้ทั้งหมด 3 algorithm ได้แก่ Decision Tree , Random Forest และ Extreme Gradient Boosting (XGBoost) และได้ทำการปรับจูนค่าพารามิเตอร์ดังนี้

3.7.1 โมเดล Decision Tree

ในงานวิจัยนี้ได้ทำการสร้างโมเดล Decision Tree โดยค่าพารามิเตอร์ที่ใช้ในการสร้างโมเดล ได้แก่ class weight เป็น balanced ,criterion เป็น gini และmaximal depth ที่ 10

3.7.2 โมเดล Random Forest

ในงานวิจัยนี้ได้ทำการสร้างโมเดล Random Forest โดยค่าพารามิเตอร์ที่ใช้ในการสร้างโมเดล ได้แก่ number of tees ที่ 100, class weight เป็น balanced ,criterion เป็น gini และmaximal depth ที่ 10

3.7.3 โมเดล Extreme Gradient Boosting (XGBoost)

ในงานวิจัยนี้ได้ทำการสร้างโมเดล Extreme Gradient Boosting โดยค่าพารามิเตอร์ที่ใช้ในการสร้างโมเดล ได้แก่ number of tees ที่100, scale pos weight เป็นค่าของ sum (negative class)/sum(positive class), objective = binary: logistic และmaximal depth ที่ 10

3.8 การประเมินผล

ขั้นตอนนี้เป็นการวัดประสิทธิภาพของโมเดลเพื่อดูว่าผลลัพธ์ที่ได้มีความน่าเชื่อถือมากน้อยเพียงใด หลังจากที่ได้ทำการแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนที่ทำการฝึกฝนโมเดลและส่วนที่สำหรับประเมินโมเดล โดยได้ใช้เกณฑ์ประเมินซึ่งวัดจากค่า F-1 score, Precision, Recall และ Accuracy ทั้งนี้ยังได้เปรียบเทียบผลลัพธ์ระหว่างโมเดลที่ใช้และไม่ใช้ตัวแปร centrality โดยคำนึงถึง Overfitting ของตัวโมเดล

3.9 การสร้าง Web-application

หลังจากวัดผลและทำการเลือกโมเดลเสร็จเรียบร้อย จะนำโมเดลนั้นมาอยู่เบื้องหลัง Application ซึ่งพัฒนาโดยใช้ library flask ซึ่งเขียนโดย python โดยมีวิธีการดังนี้

1. ทำการ Save model ที่เลือกลงในไฟล์ .pkl
2. สร้างตัวแปรในการรองรับค่าสำหรับใช้ Mapping input
3. ออกแบบจำนวน input สำหรับส่งค่ามายัง application

4. นำค่า input ที่ได้รับมาทำการ Mapping กับตัวแปร Cumulative และ centrality ที่ได้จัดเตรียมไว้

5. ส่งค่าหลังจาก Mapping ข้อมูลแล้วเข้าสู่ model

6. แสดงผลลัพธ์ว่าเป็น fraud หรือ ไม่ใช่ fraud (Legitimate)

3.10 เครื่องมือที่ใช้ในงานวิจัย

3.10.1 Jupyter Notebook เป็น Web-Application ประกอบด้วย ช่อง ๆ cell เรียงต่อกันลงไป โดยแต่ละ cell สามารถเป็นเนื้อหา static content ต่าง ๆ เช่น ข้อความ รูปภาพ กราฟ วิดีโอ เสียง หรือ เป็นโค้ดโปรแกรมคอมพิวเตอร์ ภาษา Python ที่สามารถรันคำสั่งประมวลผล แสดงผลลัพธ์ออกมาได้ และเป็นเครื่องมือที่ใช้ในการเตรียมข้อมูล แปลงข้อมูล การทำ Machine learning การทำ Visualization

3.10.2 Visual Studio Code (VS Code) เป็นโปรแกรมแก้ไข code โดยถูกออกแบบเพื่อจัดการกับภาษาทางด้านการเขียนโปรแกรมและให้คำแนะนำ completion based on imported modules, function definitions, and variable types ซึ่งในงานวิจัยนี้ใช้ VS Code ในการพัฒนาโมเดลด้วย python และการทำ Web-Application ด้วย HTML และ CSS

3.10.3 Library ที่ใช้สำหรับงานวิจัยนี้

Networkx เป็น library ที่ใช้สำหรับการจัดการและวิเคราะห์ network รวมถึง provide Network Algorithms เช่น Breadth-First Search (BFS), Depth-First Search (DFS), Centrality เป็นต้น ทั้งนี้สามารถทำงานร่วมกับ library Matplotlib เพื่อทำการ visualization

Pandas เป็น library ที่ได้รับความนิยมสำหรับการจัดการและวิเคราะห์ข้อมูล ทำให้ผู้ใช้สามารถจัดโครงสร้างข้อมูลให้เหมาะสมสำหรับการวิเคราะห์ผ่านทางสองโครงสร้างข้อมูลหลักคือ Series และ Data Frame และสามารถโหลดข้อมูลจากรูปแบบต่างๆ เช่น CSV, Excel เป็นต้น

Numpy เป็น library ที่ใช้สำหรับการคำนวณตัวเลขใน Python ให้การสนับสนุนอาร์เรย์ขนาดใหญ่และเมทริกซ์หลายมิติ พร้อมกับสมัยฟังก์ชันทางคณิตศาสตร์ที่หลากหลาย

Matplotlib เป็น library ที่ใช้สำหรับการทำ Data Visualization เช่น สร้างกราฟวงกลม กราฟแท่ง กราฟเส้น เป็นต้น และใช้ในการปรับแต่ง layout ของกราฟ

Seaborn เป็น library ที่ใช้สำหรับการทำ Data Visualization ในรูปแบบของ High-level และมี built-in function เช่น heatmaps, pair plots ซึ่ง Seaborn มีการปรับแต่ง styles และ themes ได้ง่ายกว่า Matplotlib

Scikit-learn เป็น library ที่ใช้สำหรับการทำ machine-learning และงาน data science โดยมีการ provide algorithms ทั้ง supervised and unsupervised learning. มีเครื่องมือที่ใช้สำหรับการทำ Data Preprocessing รวมไปถึงการทำ Model evaluation และการ Tuning Model

XGBoost เป็น library ที่ใช้สำหรับการเรียกใช้ algorithm Extreme Gradient Boosting (XGBoost)

บทที่ 4

ผลการศึกษา

งานวิจัยนี้เป็นการศึกษาเพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยมีรายละเอียดของผลการศึกษาดังต่อไปนี้

4.1 ผลการคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)

งานวิจัยนี้เลือกคุณลักษณะที่สำคัญโดยใช้ model Extra trees Classifier หาค่า Feature importance ของชุดข้อมูล เพื่อดูความสัมพันธ์ระหว่างตัวแปรต่างๆ และตัวแปรคำตอบ (Label) ซึ่งตัวแปรที่มีค่าน้ำหนักมากหมายถึงมีความสำคัญมากกว่าอีกตัวแปรที่มีค่าน้ำหนักน้อยกว่า

โดยตัวแปร 9 ลำดับแรกที่มีค่า Feature importance แสดงดังตารางที่ 4.1

ตารางที่ 4.1 Top 9 Feature Importance in Extra trees Classifier

Feature	Feature Description	Importance
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.191760
cash_out_count	จำนวนครั้งของการโอนเงิน(type cash) ของคนต้นทาง	0.166195
oldbalanceOrg	Balance ของคนโอน ก่อนทำการโอน	0.152379
amount	จำนวนเงิน	0.151416
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.094640
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.067429
Degree_Score_Dest	ค่า Degree centrality ของคนปลายทาง	0.047205
Clo_Score_Orig	ค่า Closeness centrality ของคนโอน	0.041692
Clo_Score_Dest	ค่า Closeness centrality ของคนปลายทาง	0.037046

4.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล

งานวิจัยนี้ได้ทำการทดลองสร้างโมเดล Decision Tree, Random Forrest และ Extreme Gradient Boosting (XGBoost) ได้ทำการเปรียบเทียบกับโมเดลที่ไม่มี feature centrality และเปรียบเทียบกับโมเดลที่ไม่มีทั้ง feature centrality และ feature cumulative sum transactions

ตารางที่ 4.2 ผลของโมเดลที่มีการใช้ feature centrality และมีการทำ feature selection

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

จากตารางพบว่าโมเดล Extreme Gradient Boosting (XGBoost) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight , มีการใช้ feature centrality และใช้ตัวแปรที่มีค่า feature importance 9 ลำดับแรก มีประสิทธิภาพในตรวจจับ transactions ที่เกิดการฉ้อโกงได้ดีที่สุดโดยให้ค่า Precision อยู่ที่ 74.88 % ค่า Recall อยู่ที่ 88.47% ค่า F-1 Score อยู่ที่ 81.11% และ Accuracy อยู่ที่ 99.77% ซึ่งสาเหตุที่เลือกโมเดลนี้เนื่องจากพิจารณาจากค่า F-1 Score ที่มากที่สุด เพราะต้องการให้ความสำคัญของทั้งเหตุการณ์ False Negative และ False Positive

ตารางที่ 4.3 ผลของโมเดลที่ไม่มีการใช้ feature centrality และไม่ได้ทำ feature selection

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

จากตารางพบว่าโมเดล Extreme Gradient Boosting (XGBoost) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight ,ไม่มีการใช้ feature centrality และ ไม่มีการทำ feature selection (เนื่องจากมี feature ทั้งสิ้น 9 ตัว) มีประสิทธิภาพในตรวจจับ transactions ที่เกิดการฉ้อโกงได้ดีที่สุดโดยให้ค่า Precision อยู่ที่ 91.18% ค่า Recall อยู่ที่ 92.99% ค่า F-1 Score อยู่ที่ 92.08% และ Accuracy อยู่ที่ 99.91% ซึ่งสาเหตุที่เลือกโมเดลนี้เนื่องจากพิจารณาจากค่า F-1 Score ที่มากที่สุด เพราะต้องการให้ความสำคัญของทั้งเหตุการณ์ False Negative และ False Positive

ตารางที่ 4.4 ผลของโมเดลที่ใช้เพียง feature หลังจากการ prepare data

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

จากตารางพบว่าโมเดล Extreme Gradient Boosting (XGBoost) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight และไม่มีการใช้ feature ที่ได้จากการทำ feature engineering เพิ่มเติม (มีตัวแปรทั้งสิ้นดังนี้ amount ,oldbalanceOrg ,newbalanceOrig ,oldbalanceDest ,newbalanceDest, type_CASH_OUT และ type_TRANSFER ซึ่งทั้ง 2 ตัวแปรสุดท้ายได้มาจากการทำ one-hot encoding กับตัวแปร type) มีประสิทธิภาพในตรวจจับ transactions ที่เกิดการฉ้อโกงได้ดีที่สุดโดยให้ค่า Precision อยู่ที่ 91.79% ค่า Recall อยู่ที่ 93.03% ค่า F-1 Score อยู่ที่ 92.41% และ Accuracy อยู่ที่ 99.92% ซึ่งสาเหตุที่

เลือกโมเดลนี้เนื่องจากพิจารณาจากค่า F-1 Score ที่มากที่สุด เพราะต้องการให้ความสำคัญของทั้งเหตุการณ์ False Negative และ False Positive

จากตารางที่ 4.2 ,4.3 และ 4.4 พบว่า Extreme Gradient Boosting (XGBoost) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight ได้ผลดีที่สุดทั้ง 3 วิธี

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

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

วิธีที่ 3. ไม่มีการใช้ทั้ง feature centrality และไม่มี feature cumulative sum transactions

จากทั้ง 3 วิธีจึงได้ทำการเปรียบเทียบความ Overfitting ดังตารางที่ 4.5

ตารางที่ 4.5 เปรียบเทียบผลโมเดล Extreme Gradient Boosting (XGBoost) ของ 3 วิธี เทียบความ overfitting

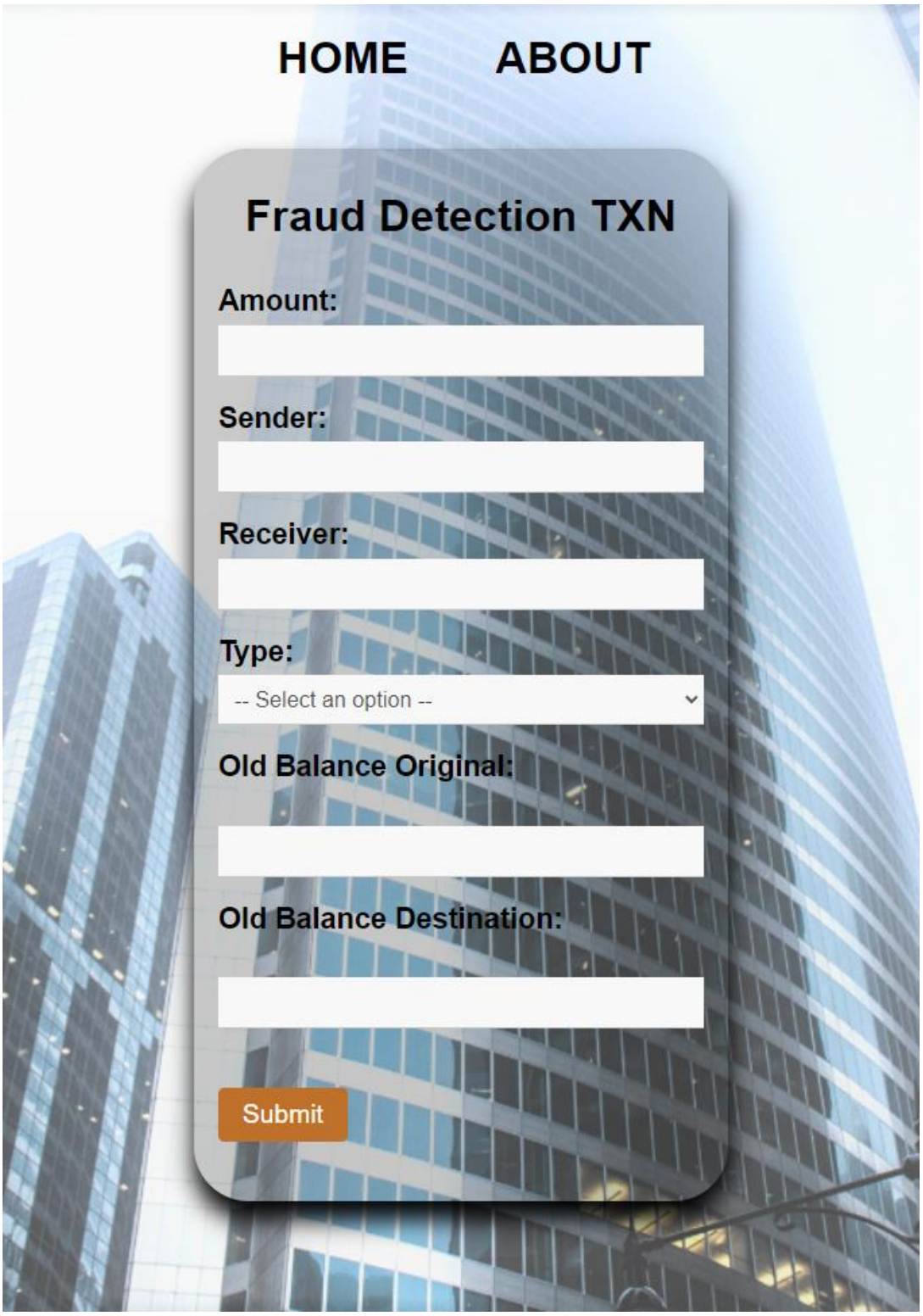
Approach	Dataset	Precision Class Fraud	Recall Class Fraud	F-1 Score Class Fraud	Accuracy
1	Training Data	55.50%	100%	71.38%	99.85%
1	Testing Data	74.88%	88.47%	81.11%	99.77%
2	Training Data	87.25%	100%	93.19%	99.97%
2	Testing Data	91.18%	92.99%	92.08%	99.91%
3	Training Data	87.19%	100%	93.15%	99.77%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

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

ดังนั้นจึงพิจารณาเลือกวิธีการที่ 1 ในการนำโมเดลไปใช้ในการจำลองการตรวจจับ transactions ที่เกิดการฉ้อโกง

4.3 การนำไปใช้บน Web-application

จากการทดลองหากนำโมเดลที่ได้ไปใช้จริง ผลการตรวจจับจะถูกนำไปแสดงเป็นโมเดลในการตรวจจับ transactions ใน Web-application โดยในหน้าเว็บจะมี User Interface สำหรับใส่ input data ในการส่งค่าและรับค่าผลลัพธ์ที่ได้จากโมเดล



ภาพที่ 4.1 หน้า User interface บน Web-Application

ภาพที่ 4.2 ตัวอย่างผลลัพธ์ fraud ที่ได้จากการทำนายบน Web-Application

ภาพที่ 4.3 ตัวอย่างผลลัพธ์ Legitimate ที่ได้จากการทำนายบน Web-Application

บทที่ 5

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเกี่ยวกับการพัฒนาโมเดลเพื่อช่วยตรวจจับธุรกรรมทางการเงินที่เกิดการฉ้อโกง ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) และนำความรู้ทางด้านกราฟมาใช้ โดยสามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการศึกษา

1. ได้หาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมสำหรับการตรวจจับธุรกรรมการฉ้อโกงโดยการประยุกต์ใช้ค่า Centrality นำมาสร้างเป็นตัวแปรในโมเดล ซึ่งพบว่าสามารถลดความ Overfitting กับตัวข้อมูลของโมเดลที่นำมาใช้ได้ เมื่อเปรียบเทียบกับวิธีการอื่นที่ไม่ได้มีการใช้ตัวแปรที่เกี่ยวข้องกับค่า Centrality

2. ได้พัฒนาโมเดล Extreme Gradient Boosting (XGBoost) (ตามวิธีการทดลองที่ 1) ที่มีการจัดการ Imbalance ด้วยวิธี class-weight, มีการใช้ feature degree centrality และ closeness centrality และใช้ตัวแปรที่มีค่า feature importance 9 ลำดับแรกซึ่งให้ค่า Precision อยู่ที่ 74.88% ค่า Recall อยู่ที่ 88.47% ค่า F-1 Score อยู่ที่ 81.11% และ Accuracy อยู่ที่ 99.77% ซึ่งผลการทดสอบดังตารางที่ 5.1

ตารางที่ 5.1 CONFUSION MATRIX ของโมเดล XGBoost ในวิธีการที่ 1

CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	825,184	1,359	99.84%
	FRAUD	528	4,052	88.47%
Class precision		99.94%	74.88%	

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

5.2 ข้อเสนอแนะ

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

2. พัฒนาประสิทธิภาพของโมเดล หรือทดลองทำโมเดลอื่นเพิ่มเติมเพื่อให้ได้ผลการตรวจจำที่แม่นยำมากขึ้น

3. พิจารณาหรือทดลองการจัดการข้อมูลที่ไม่สมดุลแนวทางอื่นเพิ่มเติมเพื่อให้ได้ผลการ ตรวจจำแม่นยำมากขึ้น

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

บรรณานุกรม

บรรณานุกรม

[1] ออมตังค์ ความรู้ทางการเงินออนไลน์. *“ใครมาชวน "รับจ้างเปิดบัญชี" ห้ามหลงเชื่อเด็ดขาด!,”* [ออนไลน์]. https://oomtang.gsb.or.th/kms/kms_view/52#:~:text=%22%E0%B8%82%E0%B8%9A%E0%B8%A7%E0%B8%99%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%88%E0%B9%89%E0%B8%B2%E0%B8%87%E0%B9%80%E0%B8%9B%E0%B8%B4%E0%B8%94%E0%B8%9A%E0%B8%B1%E0%B8%8D%E0%B8%8A%E0%B8%B5%22,%E0%B9%80%E0%B8%87%E0%B8%B4%E0%B8%99%E0%B8%84%E0%B9%88%E0%B8%B2%E0%B8%88%E0%B9%89%E0%B8%B2%E0%B8%87%E0%B9%83%E0%B8%AB%E0%B9%89%E0%B9%80%E0%B8%A3%E0%B8%B2. (เข้าถึงเมื่อ: 8 มิถุนายน 2566).

[2] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, “Paysim : A financial mobile money simulator for fraud detection”, in 28th European Modeling and Simulation Symposium, EMSS, 2016, pp. 249–255.

[3] ธนาคารแห่งประเทศไทย. *“นำส่งแนวนโยบายการบริหารจัดการภัยทุจริตจากการทำธุรกรรมทางการเงิน,”* [ออนไลน์]. https://www.bot.or.th/content/dam/bot/fipcs/documents/FOG/2566/ThaiPDF/25660066.pdf. (เข้าถึงเมื่อ: 8 มิถุนายน 2566).

[4] สมาคมโปรแกรมเมอร์ไทย. *“อะไรคือ การเรียนรู้ของเครื่อง (Machine Learning)? (ฉบับมือใหม่),”* [ออนไลน์]. https://www.thaiprogrammer.org/2018/12/%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3%E0%B8%84%E0%B8%B7%E0%B8%AD-%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B9%80%E0%B8%A3%E0%B8%B5%E0%B8%A2%E0%B8%99%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%82%E0%B8%AD%E0%B8%87%E0%B9%80. (เข้าถึงเมื่อ: 10 มิถุนายน 2566).

[5] ธนาคารกสิกรไทย. *“บัญชีม้า,”* [ออนไลน์]. https://www.kasikornbank.com/th/personal/digital-banking/kbankcyberrisk/pages/sellingbankaccount.html. (เข้าถึงเมื่อ: 16 กรกฎาคม 2566).

บรรณานุกรม (ต่อ)

[6] R. Zafarani, M. A. Abbasi and H. LIU, *Social Media Mining : An Introduction*. Cambridge University Press, 2014. Accessed on : Jun. 10, 2023 [Online]. Available : http://www.socialmediamining.info/SMM.pdf

[7] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TernsorFlow,* 2nd ed. California: O'Reilly Media, 2019.

[8] DMLC UW. *"XGBoost: A Scalable Tree Boosting System*," [Online]. https://dmlc.cs.washington.edu/xgboost.html. (Accessed: Jun 28, 2023).

[9] J. Brownlee, "*How to Choose a Feature Selection Method For Machine Learning*," [Online]. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data. (Accessed: Jul. 23, 2023).

[10] Geeksforgeeks. "*ML | Extra Tree Classifier for Feature Selection*," [Online]. https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection. (Accessed: Jun. 10, 2023).

[11] A. Kumar, "*Dealing with Class Imbalance in Python: Techniques*," [Online]. https://vitalflux.com/class-imbalance-class-weight-python-sklearn. (Accessed: Jul. 16, 2023).

[12] J. Brownlee, *Imbalanced Classification with Python : Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning,* 2020. Accessed on : Jun. 29, 2023. [Online]. Available: https://machinelearningmastery.com/imbalanced-classification-with-python.

[13] The Science of Machine Learning. "*Confusion Matrix*," [Online]. https://www.ml-science.com/confusion-matrix. (Accessed: Jun. 10, 2023).

[14] WEBSITE CRAFTER (ม.ป.ป.). *"Web application คืออะไร? ต่างจากเว็บไซต์ทั่วไปอย่างไร?,"* [ออนไลน์]. https://1stcraft.com/website-application-vs-general-website. (เข้าถึงเมื่อ: 13 มิถุนายน 2566).

บรรณานุกรม (ต่อ)

[15] P. Hajek, M. Z. Abedin and U. Sivarajah, “Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework” Inf Syst Front (2022). doi : 10.1007/s10796-022-10346-6.

[16] S. Wen, J. Li, X. Zhu, and M. Liu, “Analysis of financial fraud based on manager knowledge graph,” Procedia Computer Science, 2022, vol. 199, pp. 773-779, doi: 10.1016/j.procs.2022.01.096.

[17] A. Sahu, H. GM and M. K. Gourisaria, “A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques,” 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-7, doi: 10.1109/INDICON49873.2020.9342462.

[18] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[19] E. A. Lopez-Rojas, “*Synthetic Financial Datasets For Fraud Detection*,” [Online]. https://www.kaggle.com/datasets/ealaxi/paysim1. (Accessed: Jul. 16, 2023).

ภาคผนวก

ภาคผนวก ก

Feature importance

คุณลักษณะที่สำคัญของโมเดล ที่มีมีการใช้ feature degree centrality และ closeness centrality และใช้ตัวแปรที่มีค่า feature importance 9 ลำดับแรก จากการทำ Feature selection ด้วยโมเดล Extra tree

โมเดล Decision Tree รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.626928
amount	จำนวนเงิน	0.317575
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.043295
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.007849
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.001917
Degree_Score_Dest	ค่า Degree centrality ของคนปลายทาง	0.001700
Clo_Score_Dest	ค่า Closeness centrality ของคนปลายทาง	0.000568
Clo_Score_Orig	ค่า Closeness centrality ของคนต้นทาง	0.000144
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.000024

โมเดล Random Forest รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.451692
amount	จำนวนเงิน	0.239270
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.190836
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.034949
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.030475
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.025116
Degree_Score_Dest	ค่า Degree centrality ของคนปลายทาง	0.019546
Clo_Score_Dest	ค่า Closeness centrality ของคนปลายทาง	0.004881
Clo_Score_Orig	ค่า Closeness centrality ของคนต้นทาง	0.003236

โมเดล Extreme Gradient Boosting (XGBoost)

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.321038
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.287387
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.165961
amount	จำนวนเงิน	0.140330
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.034144
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.017045
Clo_Score_Dest	ค่า Closeness centrality ของคนปลายทาง	0.013691
Clo_Score_Orig	ค่า Closeness centrality ของคนต้นทาง	0.011406
Degree_Score_Dest	ค่า Degree centrality ของคนปลายทาง	0.008998

คุณลักษณะที่สำคัญของโมเดลที่ไม่มีมีการใช้ feature degree centrality และ closeness centrality และไม่มีการทำ feature selection

โมเดล Decision Tree รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.549392
amount	จำนวนเงิน	0.190014
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.129413
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.115854
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.010749
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.002080
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.001317
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.001013
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.000167

โมเดล Random Forest รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.425374
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.160570
amount	จำนวนเงิน	0.155006
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.081304
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.078973
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.051349
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.023040
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.020409
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.003975

โมเดล Extreme Gradient Boosting (XGBoost)

Feature	Feature Description	Importance
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.575921
oldbalanceOrg	Balance ของคนโอน ก่อนทำการโอน	0.196452
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.089050
amount	จำนวนเงิน	0.063704
transfer_out_count	จำนวนครั้งของการโอนเงิน (type transfer) ของคนต้นทาง	0.028755
cash_in_count	จำนวนครั้งของการรับเงิน (type cash) ของคนปลายทาง	0.025101
cash_out_count	จำนวนครั้งของการโอนเงิน (type cash) ของคนต้นทาง	0.014789
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.004523
transfer_in_count	จำนวนครั้งของการรับเงิน (type transfer) ของคนปลายทาง	0.001705

คุณลักษณะที่สำคัญของโมเดลที่ไม่มีมีการใช้ทั้ง feature centrality และ feature cumulative sum transactions

โมเดล Decision Tree รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.548983
amount	จำนวนเงิน	0.190515
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.129828
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.115759
type_TRANSFER	Flag transaction ว่าเป็น type transfer	0.011187
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.003498
type_CASH_OUT	Flag transaction ว่าเป็น type cash_out	0.000230

โมเดล Random Forest รายละเอียดดังนี้

Feature	Feature Description	Importance
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.397940
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.186452
amount	จำนวนเงิน	0.158600
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.121398
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.071950
type_CASH_OUT	Flag transaction ว่าเป็น type cash_out	0.032210
type_TRANSFER	Flag transaction ว่าเป็น type transfer	0.031450

โมเดล Extreme Gradient Boosting (XGBoost)

Feature	Feature Description	Importance
newbalanceOrig	Balance ของคนต้นทาง หลังทำการโอน	0.588354
oldbalanceOrg	Balance ของคนต้นทาง ก่อนทำการโอน	0.196790
newbalanceDest	Balance ของคนปลายทาง หลังทำการโอน	0.089720
type_CASH_OUT	Flag transaction ว่าเป็น type cash_out	0.062693
amount	จำนวนเงิน	0.057851
oldbalanceDest	Balance ของคนปลายทาง ก่อนทำการโอน	0.004593
type_TRANSFER	Flag transaction ว่าเป็น type transfer	0.000000

ภาคผนวก ข

Confusion Matrix

Confusion Matrix ของโมเดลที่มีการใช้ feature degree centrality และ closeness centrality และใช้ตัวแปรที่มีค่า feature importance 9 ลำดับแรก จากการทำ Feature selection ด้วยโมเดล Extra tree Classifier

DECISION TREE CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	778,588	47,955	94.20%
	FRAUD	59	4,521	98.71%
Class precision		99.99%	8.62%	

RANDOM FOREST CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	795,920	30,623	96.30%
	FRAUD	307	4,273	93.30%
Class precision		99.96%	12.24%	

Extreme Gradient Boosting CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	825,184	1,359	99.84%
	FRAUD	528	4,052	88.47%
Class precision		99.94%	74.88%	

Confusion Matrix ของโมเดลที่ไม่มีการใช้ feature degree centrality และ closeness centrality และไม่มีการทำ feature selection

DECISION TREE CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	816,936	9,607	98.84%
	FRAUD	138	4,442	96.99%
Class precision		99.98%	31.62%	

RANDOM FOREST CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	811,375	15,168	98.16%
	FRAUD	89	4,491	98.06%
Class precision		99.99%	22.84%	

Extreme Gradient Boosting CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	826,131	412	99.95%
	FRAUD	321	4,259	92.99%
Class precision		99.96%	91.18%	

Confusion Matrix ของโมเดลที่ไม่มีมีการใช้ทั้ง feature centrality และ feature cumulative sum transactions

DECISION TREE CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	817,325	9,218	98.88%
	FRAUD	123	4,457	97.31%
Class precision		99.98%	32.59%	

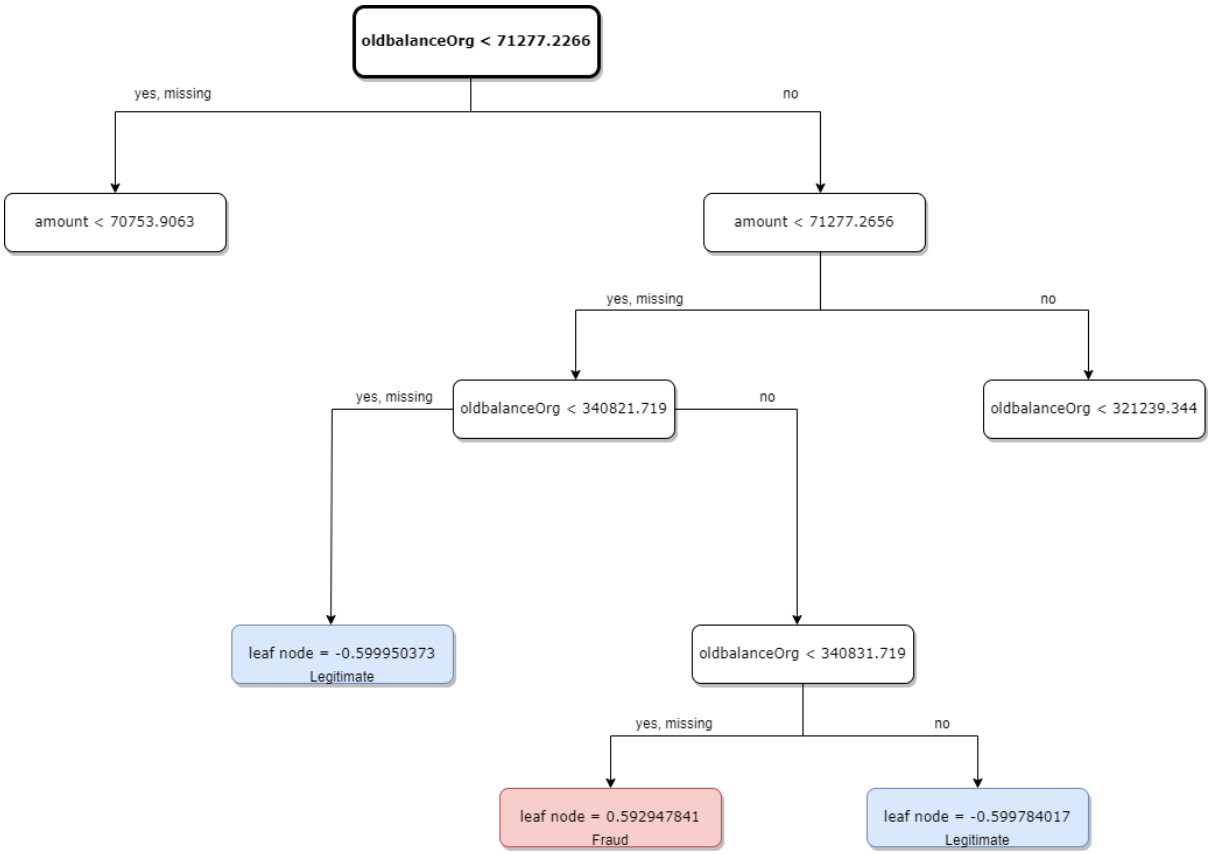
RANDOM FOREST CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	811,056	15,487	98.13%
	FRAUD	117	4,463	97.45%
Class precision		99.99%	22.37%	

Extreme Gradient Boosting CONFUSION MATRIX		PREDICTED		Class recall
		LEGITIMATE	FRAUD	
TRUE	LEGITIMATE	826,162	381	99.95%
	FRAUD	319	4,261	93.03%
Class precision		99.96%	91.79%	

ภาคผนวก ค

Decision tree ในโมเดล XGBoost ในวิธีการที่ 1

XGBoost ในวิธีการที่ 1 นั้นประกอบไปด้วย decision tree 100 ต้น ทั้งนี้จะหยิบ decision tree ต้นแรกในโมเดลมาอธิบายบาง Internal nodes และบาง Leaf nodes ดังนี้



จากรูปจะอธิบายได้ว่า

ในกรณีที่ 1 ตัว decision tree ในโมเดลนี้จะให้ค่าข้อมูลนี้ว่าเป็น Legitimate transaction ถ้าข้อมูลมีค่า oldbalanceOrg >= 71,277.2266 และมีค่า amount < 71,277.2656 หรือมีค่า amount = missing value และมีค่า oldbalanceOrg < 34,0821.719 หรือมีค่า oldbalanceOrg = missing value

ในกรณีที่ 2 ตัว decision tree ในโมเดลนี้จะให้ค่าข้อมูลนี้ว่าเป็น Legitimate transaction ถ้าข้อมูลมีค่า oldbalanceOrg >= 71,277.2266 และมีค่า amount < 71,277.2656 หรือมีค่า amount = missing value และมีค่า oldbalanceOrg >=34,0831.719

ในกรณีที่ 3 ตัว decision tree ในโมเดลนี้จะให้ค่าข้อมูลนี้ว่าเป็น Fraud transaction ถ้าข้อมูลมีค่า oldbalanceOrg >= 71,277.2266 และมีค่า amount < 71,277.2656 หรือมีค่า amount = missing value และมีค่า oldbalanceOrg>= 34,0821.719 แต่ไม่เกิน 34,0831.719

ประวัติผู้เขียน

ชื่อ – นามสกุล สุเชฏฐ์ เหร่บุตร

ประวัติการศึกษา

พ.ศ. 2559 - ปริญญาตรี สาขาเศรษฐศาสตร์ มหาวิทยาลัยเชียงใหม่

ประสบการณ์ทำงาน

พ.ศ. 2566 - Business Data Analytics ธนาคาร CIMB Thai

พ.ศ. 2565 - SrOfficer Fraud Management of Fraud Analysis Unit,
Fraud Management Department บริษัท เงินติดล้อ จำกัด (มหาชน)